Tea plant genomics: achievements, challenges and perspectives

En-Hua Xia1, Wei Tong1, Qiong Wu1, Shu Wei1, Jian Zhao1, Zheng-Zhu Zhang1, Chao-Ling Wei1 and Xiao-Chun Wan1

Abstract
Tea is among the world’s most widely consumed non-alcoholic beverages and possesses enormous economic, health, and cultural values. It is produced from the cured leaves of tea plants, which are important evergreen crops globally cultivated in over 50 countries. Along with recent innovations and advances in biotechnologies, great progress in tea plant genomics and genetics has been achieved, which has facilitated our understanding of the molecular mechanisms of tea quality and the evolution of the tea plant genome. In this review, we briefly summarize the achievements of the past two decades, which primarily include diverse genome and transcriptome sequencing projects, gene discovery and regulation studies, investigation of the epigenetics and noncoding RNAs, origin and domestication, phylogenetics and germplasm utilization of tea plant as well as newly developed tools/platforms. We also present perspectives and possible challenges for future functional genomic studies that will contribute to the acceleration of breeding programs in tea plants.

Introduction
The tea plant (Camellia sinensis) is an economic crop with significant importance1. Its leaves are generally used to produce tea—the world’s most popular non-alcoholic beverage. Tea comprises abundant characteristic compounds, such as tea polyphenols, theanine, and caffeine, which not only determine the quality of tea but also confer it with tremendous health benefits2. However, compared to other crops, such as rice3, research on the functional genomics of tea plants is lagging behind, which has hampered the utilization of genetic resources for modern molecular breeding, which is urgently needed. The rapid development and innovation of biotechnologies, particularly next-generation sequencing (NGS), has resulted in significant progress in understanding the genomics and genetics of tea plants in the past two decades, which can be briefly summarized according to five aspects: (1) the deciphering of the genome and transcriptome of tea plants using diverse NGS technologies; (2) the identification and functional analysis of tea quality-related genes and their regulatory networks; (3) the investigation of the epigenetics and noncoding RNAs (ncRNAs) involved in stress responses and other biological processes; (4) the exploration of the diversification and domestication of tea plants; and (5) the development of innovative tools and techniques for the characterization of novel genes and alleles. In this review, we briefly summarize the main achievements of the last two decades and present potential challenges and perspectives for future functional genomic studies that would contribute to the acceleration of breeding programs in tea plants.
tea plants and that 30 chromosomes (2n = 30) occurred in the zygotes of *C. japonica*, a closely related species of tea plant. These findings suggested that the diploid tea plant has 30 chromosomes and that the chromosome number might be conserved in genus *Camellia*. New innovations in biotechnologies gradually confirmed these conclusions among globally collected tea clones. In particular, the investigation of the chromosome numbers and ploidy levels of 139 tea individuals representing almost all sections of genus *Camellia* revealed the diploid nature of the majority (90.65%) of *Camellia* species, with a chromosome number of 2n = 30. Only a few tea plant species are polyploid, which exhibit 45–120 chromosomes. These species predominantly come from section *Camellia* (*C. chekiangoleosa*, *C. mairei*, and *C. reticulata*), section *Oleifera* (*C. oleifera* and *C. sasanqua*), and section *Arbica* (*C. granthamiana*).

Genome size refers to the DNA content within the haploid genome of an organism. The estimation of genome size is not only of great significance for cytogenetic studies but also provides indispensable basic data for genome sequencing, comparative genomics, and evolutionary analyses. The genome size of tea plants was initially estimated to be 3.8–4.0 Gb, which was further demonstrated by the investigation of 36 Indian tea accessions, where all the investigated accessions exhibited diploid genomes with sizes ranging from 3.5 to 3.8 Gb. The recent genome sequencing of two representative elite tea plant cultivars, *C. sinensis* var. *sinensis* cv. shuchazao and *C. sinensis* var. *assamica* cv. yunkang#10, showed an ~3.0 Gb genome size, which is similar to that of maize but much larger than those of coffee and cocoa. While genome size is likely to be conserved in cultivated tea plants, intraspecific and interspecific variations are also observed in genus *Camellia*, possibly due to the frequent hybridization and polyploidization events that occur after speciation.

**Genetic linkage maps of tea plants**

The construction of genetic linkage maps is the basis of molecular biology and is essential for a wide range of genetics and genomic studies, such as quantitative trait mapping, molecular marker-assisted breeding and comparative genomic studies. Unlike other crops, tea plants are self-incompatible species with relatively low cross-fertility rates, resulting in a lack of a high generation segregation populations and sufficient offspring for genetic map construction. More than 10 genetic linkage maps have been generated for tea plants, most of which are produced based on an F1 population consisting of <150 offspring. The dominant molecular markers used for genetic map construction include random amplification of polymorphic DNA (RAPD) markers, amplified fragment length polymorphism (AFLPs), inter-simple sequence repeats (ISSRs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs). These reported genetic maps vary in length and density, with the total map length ranging from 1180.9 to 3314.3 cM and the average distance between adjacent markers from 0.4 to 20.1 cM. Such differences are principally affected by the number of individuals and types of molecular markers applied for map construction. The highest-density genetic map of tea plants obtained to date was constructed using a total of 4217 polymorphic SNP markers developed from 327 individuals of the F1 segregating population (LJ43 × BHZ) via 2b-RAD sequencing. The resultant genetic maps consisted of 15 linkage groups ranging from 87.46 to 146.93 cM. The total map length was 1679 cM, with an average interval of 0.4 cM between 4217 adjacent markers. LG01 and LG03 were the two largest linkage groups, with genetic distances of 146.93 cM (374 markers) and 146.61 cM (374 markers), respectively. Despite the numerous genetic maps currently available, the density and quality of these maps need to be further improved by investigating larger populations and applying more advanced biotechnologies. The resultant data will be important for future investigations of tea plant biology that will benefit the whole tea industry.

**Current progress in tea plant genome sequencing**

With the revolution of sequencing technology, an increasing number of labs around the world have successfully released more than 236 plant genomes, and the tea plant genome is probably among the most complex of these. The assemblies of ~3 Gb of genomic sequences with an average scaffold N50 size of 920 kb, which is much larger than those from other plants with complex genomes, such as the orchid (scaffold N50 = 359 Kb) and moso bamboo (scaffold N50 = 329 Kb). Nevertheless, compared to other model species, such as rice, the scaffold N50 sizes of the current assemblies are still somewhat disappointing, due largely to the limitation imposed by the short read length generated from NGS technology (Fig. 1). Unassembled regions account for ~5% of the current assemblies. In the future, additional efforts will be needed to further improve the quality and completeness of the genome assemblies of tea plants, particularly with the upcoming advances in sequencing.
technologies (e.g., 3rd-generation sequencing) and new algorithm-derived assemblers.

There are 33,932 and 36,951 protein-coding genes in the CSS and CSA tea plant genomes, respectively. Transposable elements (TEs) account for 64% and 80% of the CSS and CSA genomes, respectively. LTR retrotransposons are among the most dominant TEs, independently representing 58% and 67% of the CSS and CSA genomes, respectively. The rapid propagation of TEs, particularly the proliferation and persistence of a single Ty3/gypsy retrotransposon family (TL001) for over 50 million years, has been proven to have led to a considerable increase in genome size in tea plants\(^1\). CSA and CSS diverged from their common ancestor ~0.38–1.54 million years ago (mya). However, this divergence time might have been underestimated because of the draft nature of the two current genome assemblies, which generated only a small proportion of collinear genes for dating. Similar to other plants, tea plants underwent two whole-genome duplication (WGD) events, including recent and ancient events occurring 30–40 and 90–100 mya, respectively\(^2\).

Tea plants contain rich and diverse secondary metabolites that not only confer tea quality characteristics, such as color, aroma and taste, but also play critical roles in the responses to biotic and abiotic stresses\(^3\). Several genes encoding enzymes associated with the biosynthesis of secondary metabolites were significantly amplified in the tea plant genome. In particular, the striking expansion of serine carboxypeptidase-like (SCPL) acyltransferase-encoding genes might directly contribute to the high accumulation of galloylated catechins that determine tea palatability\(^2\). Compared to kiwifruit, tomato, potato, cacao, and Arabidopsis thaliana, tea plants show expansion of disease resistance genes, including nucleotide-binding sites with leucine-rich repeats (NBS-LRRs) and pattern-recognition receptors (RLK-LRRs)\(^1\). This contributes to the enhancement of the immune system of tea plants and, thus, their adaptation to diverse global environments. WGD events and subsequent tandem

| Mapping population                  | Size\(^a\) | Marker type | No. of markers | Total length (cM) | Marker density (cM) | No. of linkage groups | Reference         |
|-------------------------------------|------------|-------------|----------------|-------------------|---------------------|----------------------|-------------------|
| Sayamakaori × Kana-CK17            | 54         | RAPD        | 140            | 1640              | 11.7                | 17                   | Ota et al.\(^{23}\) |
| SFS150 × TN14/3                    | 90         | RAPD, AFLP  | 126            | 1350              | 11.7                | 15                   | Hackett et al.\(^{18}\) |
| Keemun #4 × Chaoandawuye           | 69         | AFLP        | 208            | 2458              | 11.9                | 17                   | Huang et al.\(^{19}\) |
| Fudingdabaicha × Zhenong #12       | 94         | ISSR, RAPD  | 62             | 1181              | 20.1                | 7                    | Huang et al.\(^{24}\) |
| SFS150 × S15/10                    | 42         | SSR, RAPD, AFLP | 100        | 1412              | 14.1                | 30                   | Kamunya et al.\(^{20}\) |
| Sayamakaori × Kana-CK17            | 54         | SSR, STS, CAPS, RAPD | 279        | 1218              | 4.35                | 15                   | Taniguchi et al.\(^{25}\) |
| Yingshuang × Beiyuedanzhu           | 183        | SSR         | 406            | 1144              | 2.9                 | 15                   | Ma et al.\(^{27}\)  |
| Yingshuang × Beiyuedanzhu           | 148        | SSR, SNP    | 6448           | 3965              | 1.0                 | 15                   | Ma et al.\(^{21}\)  |
| Long #43 × Baihaozao                | 170        | SSR         | 483            | 1226              | 2.5                 | 15                   | Tan et al.\(^{28}\) |
| Fushun × Kemsull                   | 79         | SSR, RAPD, AFLP | 678          | 1442              | 4.7                 | 15                   | Chang et al.\(^{26}\) |
| Long #43 × Baihaozao                | 327        | SNP         | 4217           | 1679              | 0.4                 | 15                   | Xu et al.\(^{29}\)  |

\(^{a}\)Population size
duplications are the two major evolutionary forces that drive such gene expansions.

Tea-processing suitability refers to the characteristics of tea varieties that make them suitable for the manufacture of certain types of tea to achieve the best quality. It is accepted that different varieties of tea plants differ in terms of flavors and tastes. The sequencing of the tea plant genome together with the transcriptomes and metabolomes of 25 *Camellia* species representing almost all the sections from genus *Camellia* showed that, despite the conservation of three metabolic pathways (catechins, theanine, and caffeine) among *Camellia* species, most of the flavonoid and caffeine, but not the theanine-related genes, exhibited higher expression levels in species from section *Thea* than in non-*Thea* species. This elucidates why the leaves of tea plants from non-*Thea* sections, such as well-known ornamental camellias (e.g., *C. japonica*) and the traditional oil tea plant (*C. oleifera*), accumulate such low contents of catechins and caffeine but not theanine, making them unsuitable for tea production. Further investigation of the roles of natural selection and artificial selection in shaping the genes involved in the three metabolic pathways will contribute to a deep understanding of the processing suitability and quality of tea. In addition, caffeine (1,3,7-trimethylxanthine) is among the most well-known purine alkaloids in plants, and its biosynthesis is mainly catalyzed by the N-methyltransferases (NMTs). The NMT genes have undergone recent rapid and independent evolution in tea plants relative to cocoa and coffee.

**Organelle genome sequencing**

Plants generally harbor two independent organelle genomes (chloroplast and mitochondria), which provide invaluable resources for a range of functional, evolutionary and comparative genomic studies. The cultivated tea plant harbors a circular chloroplast (cp) genome of 157,096 bp in length with an overall GC content of 37.3%. It exhibits the typical plant cp genome structure, including a pair of inverted repeat regions (IRs, 26,080 bp) separated by a large single-copy region (LSC, 86,653 bp) and a small single-copy region (SSC, 18,283 bp). The whole-genome sequence encodes a total of 133 cp genes, 86 of which are protein-coding genes, while 8 are rRNA genes, and 39 are tRNA genes. The cp genomes have been assembled for diverse well-known tea cultivars, such as "Longjing #43", "Yunkang #10", and "Shuchazao". In addition to cultivated tea plants, an increasing number of cp genomes from species of genus *Camellia* have also been released, including those of *C. japonica* (NC_036830), *C. mairei* (NC_035688), *C. szechuanensis* (NC_035651), *C. elongata* (NC_035652), *C. azalea* (NC_035574), *C. pubicosta* (NC_024662), *C. petelottii* (NC_024661), *C. reticulata* (NC_024663), *C. grandibracteata*...
that determine tea quality. Subsequently, growing biosynthesis of characteristic compounds of tea plants and hormone responses and the mechanisms underlying mass, total numbers of genes, and sequence homology, suggesting extreme genome conservation during evolution.

In sharp contrast to the chloroplast genome, only a partially assembled mitochondrial (mt) genome is available for cultivated tea plants. The current release of the tea plant mt genome consists of two circular scaffolds with total lengths of 702,253 bp (45.63% GC) and 178,082 bp (45.81% GC). It encodes 71 mt genes, including 44 protein-coding genes, 24 tRNA genes, and 3 rRNA genes. Compared to the cp genome, the tea plant mt genome shows a repeat-rich nature, including a total of 38,027 bp of long repeat sequences.

Transcriptome sequencing and gene discovery in tea plants

Transcriptome sequencing of tea plants

Transcriptome sequencing has revolutionized genetic and functional genomic studies of organisms, particularly for nonmodel species without available sequenced genomes. In the last decade, innovations overcoming challenges in tea plant genome sequencing have greatly accelerated the transcriptomic investigation of this economically important crop. Shi and colleagues examined the first transcriptome of tea plants (cultivar shuchazao) using Illumina sequencing technology and obtained a total of 127,094 unigenes, which were applied for the in-depth exploration of candidate genes involved in the biosynthesis of characteristic compounds of tea plants that determine tea quality. Subsequently, growing transcriptome-sequencing projects were launched to further investigate the gene expression dynamics of tea plants under cold acclimation, drought stress, and hormone responses, and the mechanisms underlying self-incompatibility, nitrogen utilization, trichome formation, and tea quality. These results greatly broadened our understanding of tea plant biology. With the advent of single-molecule sequencing technology, a recent study produced a more accurate full-length transcriptome of tea plants. In addition to the cultivated tea plants, several transcriptomes from closely related species from genus Camellia have also been reported, which have provided indispensable resources for comparative transcriptomic studies. (Supplementary Table 1). The sequencing of the transcriptome of oil tea plants (C. oleifera) identified 3022 orthologous gene pairs between cultivated tea plants and oil tea plants, among which 211 exhibited evidence of positive selection. Compared to the cultivated tea plants, C. taliensis showed extraordinary amplification of cold tolerance-related genes. Most of the genes associated with triacylglycerol biosynthesis in C. reticulata and C. sinensis are multiple-copy genes, suggesting the potential occurrence of WGD events in the common ancestor of genus Camellia. An increasing number of transcriptome-sequencing projects have also been carried out in other Camellia species, such as C. sasanqua, C. chekiangoleosa, and C. nitidissima. These datasets expanded the gene pool of tea plants and will be of particular importance for future tea plant breeding, as well as investigations of functional genomics, phylogenomics, and comparative transcriptomics. It should be noted that the current reported transcriptomes from both tea plants and other Camellia species vary in the total number of assembled transcripts and N50 sizes, primarily due to the different sequencing depths and assemblers adopted. In the future, more efforts should be made to evaluate the optimized-sequencing depth and assembler to better construct the transcriptomes of tea plants.

Functional gene sets and their regulatory networks

One fundamental mission of the molecular biology research on tea plants is to understand the functions and regulatory mechanisms of genes encoded by the genome. During recent decades, rapid advances in biotechnologies have facilitated the cloning and functional characterization of an increasing number of genes in tea plants. Most of these genes have been identified from CSS and CSA, in which the numbers of cloned genes are much greater than those from other closely related tea plant species. Their functions can generally be classified into three major categories, including secondary metabolite biosynthesis, abiotic and biotic stress responses, and aroma formation. Among these genes, those associated with secondary metabolite biosynthesis and aroma formation are the most studied because they directly determine the quality of tea. Supplementary Table 2 summarizes the details of the currently available cloned genes in tea plants. We here selectively highlight some representatives that would be preferred candidates for the future genetic improvement of tea plants.

Caffeine is the most abundant purine alkaloid in the majority of tea plants. It is a crucial component of tea quality and is significantly related to the bitterness of tea. Nevertheless, excessive intake of caffeine has been reported to have some side effects for human health, such as increasing the risk of cardiovascular disease, palpitations, and insomnia. The documentation of the genes controlling caffeine biosynthesis is therefore essential for the future of breeding new varieties with a low caffeine content. Kato et al. cloned the gene coding...
caffeine synthase (TCS) from the young leaves of tea plants using the rapid amplification of complementary DNA ends (RACE) technique. This gene is 1438 bp in size and encodes 369 amino acids. The expression of the TCS gene in *Escherichia coli* enables high production of caffeine after feeding on xanthine and S-adenosylmethionine as substrates, confirming the caffeine synthase activity of TCS.

Theanine is another characteristic amino acid of tea plants and is related to the fresh tastes of tea. The gene encoding enzyme associated with theanine biosynthesis (*CsTSI*) was identified through a combination of genomic and transcriptomic analyses. This gene shares high homology with the glutamine synthetase (PtGS) gene of *Pseudomonas taetrolens*, indicating its potential bacterial origin. The overexpression of *CsTSI* in *A. thaliana* significantly increases the accumulation of theanine after ethylamine feeding. Tea polyphenols are the major secondary metabolites of tea plants and are closely related to the astringent and bitter taste of tea. Two enzymes, UDPglucose:galloyl-1-O-β-D-glucosyltransferase (UGGT) and epicatechin:1-O-galloyl-β-D-glucose O-galloyltransferase (ECGT), were purified and shown to be involved in the biosynthesis of galloylated polyphenols.

### Table 2  List of representative functional genes in tea plants.

| Gene symbol | NCBI accession | Length (bp) | Function description | Ref. |
|-------------|----------------|------------|----------------------|-----|
| **Polyphenol biosynthesis** | | | | |
| CsANR1 | GU992402 | 1044 | Encoding anthocyanidin reductase | 77 |
| CsANR2 | GU992400 | 1014 | Encoding anthocyanidin reductase | 77 |
| CsLAR | GU992401 | 1282 | Involved in flavan-3-ol biosynthesis | 77 |
| CsF3′H | KT180309 | 1706 | Encoding flavonoid 3′-hydroxylase | 92 |
| CsC4H | KY615675 | 1518 | Encoding cinnamate 4-hydroxylase in synthesis of flavonoids and lignins | 93 |
| CsF3H | KY615688 | 1107 | Encoding flavanone-3-hydroxylase | 94 |
| CsMYB4a | KY774676 | 735 | Negatively regulates phenylpropanoid and shikimate pathway | 78 |
| CsMYB5 | KY827396 | 930 | Regulate anthocyanin and proanthocyanidin biosynthesis | 79 |
| CsMYB75 | GGTM01017033 | 765 | Regulate anthocyanin hyperaccumulation | 76 |
| CsGST | MK431867 | 642 | Encoding glutathione-S-transferase | 95 |
| **Caffeine biosynthesis** | | | | |
| CsTCS | AB031280 | 1438 | Caffeine synthase gene | 73 |
| **Theanine biosynthesis** | | | | |
| CsTSI | TEA015198* | 2544 | Theanine synthetase gene | 2 |
| CsAAPs | TEA031577* | 2339 | Encoding amino acid permease involved in theanine transportation | 96 |
| **Aroma formation** | | | | |
| Csβ-PD | AB088027 | 1729 | Encoding beta-primeverosidase in floral aroma formation | 88 |
| CsGT1 | AB847092 | 1458 | Converting volatile organic compounds into β-primeverosides for aroma biosynthesis | 90 |
| CsSAMT | MG459470 | 1104 | Encoding the salicylic acid carboxyl methyltransferase for methyl salicylate formation | 97 |
| CsUS/UES | KFO06849 | 2050 | Producing (E)-nerolidol and linalool in vitro | 87 |
| CsUGT85A53 | NA | 1455 | Formation of (Z)-3-hexenyl glucoside, novel insect pest control | 81 |
| **Abiotic and biotic stress** | | | | |
| CsCOR1 | EU563236 | 755 | Involved in cold, salinity and dehydration tolerance | 82 |
| CsHSP17.2 | KU244518 | 734 | Mediates tea plant thermo tolerance | 98 |
| CsICE1 | GQ229032 | 1964 | Inducer of CBF expression | 84 |
| CsCBF1 | EU563238 | 1210 | C-repeat-binding factor | 84 |
| CsTLP | DQ444296 | 681 | Increases tolerance to fungal pathogens overexpressed in potato | 83 |
| CsGolS2 | KP757767 | 1574 | Response to *Ectropic oblique* attack | 86 |

*CSS locus ID (from TPIA: http://tpla.teaplant.org)"
catechins. Three UDP-glycosyltransferase (UGT) genes, CsUGT84A22, CsUGT78A14, and CsUGT78A15, were found to be involved in the biosynthesis of β-glucogallin and glycosylated flavonols. Anthocyanin is another type of polyphenol that typically accumulates in purple tea varieties. It has considerable health benefits and has been used to develop anthocyanin-enriched beverages. The expression of a gene encoding leucoanthocyanidin reductase (CsLAR) in E. coli results in the production of 2R,3S-trans-flavan-ol (+)-catechin after the feeding of leucocyanidin as a substrate. Similarly, the expression of two anthocyanidin reductase (CsANR1 and CsANR2)-encoding genes in E. coli enables the conversion of cyanidin to a mixture of (+)-epicatechin and (−)-catechin. More recently, CsMYB75 and CsGSTF1 were found to be involved in anthocyanin accumulation in purple tea. CsMYB75 can promote the expression of CsGSTF1 in transgenic tobacco plants. CsGSTF1 enables the transfer of anthocyanin from the endoplasmic reticulum (ER) to vacuoles. In tea plants, the regulation of the catechin metabolic pathway is complex, and several transcription factors (e.g., MBW complexes) are involved.

Aroma is vital for tea quality and for attracting global interest. Since the advent of new chemical analytical techniques, such as mass spectrometry, considerable efforts have been made to identify volatile constituents of different types of teas and to assess volatile odor activities and their contribution to tea aroma. Significant studies on tea aroma biology have revealed that hundreds of volatile terpenoids, in addition to some other volatiles, such as cis-3-Hexen-1-ol, are present in tea as glycosides, which can be released during the tea manufacturing process. A tea beta-primeverosidase gene was isolated and expressed in E. coli. The prokaryotically expressed mature protein was found to be able to hydrolyze beta-primeverosides, liberating a primerose unit and aglycons and playing roles in tea plant defense and floral aroma formation during the tea manufacturing process. Two UGTs from C. sinensis, UGT85K11 (CsGT1) and UGT94P1 (CsGT2), were found to convert volatiles into β-primeverosides via sequential glucosylation and xylosylation, respectively. Recently, a UGT gene (CsUGT85A53) that was functionally validated was shown to enable the conversion of exogenous (Z)-3-hexenol from damaged adjacent tea leaves to its glucoside form, enhancing the ability of tea plants to defend against pests. Studies on tea volatile biosynthesis have also revealed an interesting bifunctional linalool/nerolidol synthase gene (CsLIS/NES) in tea plants. This gene transcribes two transcript isoforms, CsLIS/NES-1 and CsLIS/NES-2, which possess distinct subcellular localizations and molecular functions, with CsLIS/NES-1 localizing to chloroplasts and functioning as linalool synthase, while CsLIS/NES-2 localizes to the cytosol and potentially acts as a nerolidol synthase. In addition, the biosynthesis of the tea volatiles indole, linalool, and nerolidol was found to be controlled by the corresponding genes in tea. Collectively, the aforementioned genes that were cloned and functionally validated in tea plants have broadened our understanding of the genetic basis of tea quality and will be particularly useful for the future genetic improvement of tea plants.

DNA methylation of tea plants

DNA methylation is among the most essential and ubiquitous epigenetic modifications in plants. It plays crucial roles in gene expression regulation, cell differentiation, and transposon element (TE) silencing. DNA methylation is mostly investigated in model plants, such as Arabidopsis and rice, due to their small genome sizes and low genome complexity. Tea plants are non-model woody plants with complex nuclear genomes. The sequencing of the tea plant genome has revealed an extraordinary amplification of genome size driven primarily by a burst of TEs. DNA methylation was confirmed to be closely linked with TE burst in tea plants. Compared to asterids such as potato, tea plants exhibit higher CHG methylation levels, similar to those found in maize and Norway spruce, two plant species with a comparable genome size and repeat content to tea plants. Remarkably, the DNA methylation levels of TEs vary with their evolutionary distance in tea plants, with the methylation level increasing in recent, active TEs and decreasing in ancient TEs. It is widely accepted that DNA methylation can spread from TE boundaries to close genomic regions, resulting in increased methylation levels of adjacent genes. High TE contents in tea plants might therefore contribute to the high level of DNA methylation. Several genes involved in the biosynthesis of secondary metabolites (catechins, theanine, and caffeine) were found to be TE-related in tea plants. This suggests that TE-mediated DNA methylation may have some effects on the formation of tea quality.

Chilling damage caused by low temperature occurs commonly in tea plants and has severely affected the sustainable development of the global tea industry. Approximately 49–51% of cytosine residues are methylated during the cold acclimation of tea plants. Compared to preacclimation, the DNA methylation level increases significantly during cold acclimation and is maintained at a higher level after deacclimation. The gene encoding DNA methyltransferase, CsDRM2, was cloned from an elite cultivar of tea plant (Longjing #43) and found to exhibit high expression levels under cold acclimation. In the future, additional DNA methylation projects need to be performed in relation to other aspects of tea plants to elucidate the roles of these modifications in plant development, environmental adaptation, stress...
responses, and even genome structure variation and stabilization.

ncRNAs of tea plants

MicroRNAs (miRNAs), a class of endogenous small ncRNAs, have been recognized as a critical genetic regulator primarily engaged in posttranscriptional regulation. In tea plants, a series of studies have been undertaken to examine the posttranscriptional regulation of miRNAs in development, metabolism, and responses to biotic/abiotic stresses. Tea plant miRNAs can negatively regulate catechins and terpenoid biosynthesis by down-regulating target genes related to their biosynthesis at both the transcriptional and posttranscriptional levels. For instance, Cs-miR156 regulates catechin accumulation in tea plants by suppressing the expression of the target gene SPL in the presence of different nitrogen forms. The characterization of tea plant miRNA regulatory networks in response to Colletotrichum gloeosporioides, which is considered one of the dominant endophytic taxa in tea plants, indicated that miRNAs may be involved in the response to C. gloeosporioides attack. miRNA characterization in tea plants under cold and drought stresses suggested the potential existence of an miRNA-mediated regulatory mechanism under abiotic stresses showing coherent or incoherent relationships with target genes to prevent tea plants from being injured. Csn-miR398a-3p-1 of tea plants was experimentally proven to directly cleave CsCSD4, a superoxide dismutase (SOD) gene related to the removal of reactive oxygen species (ROS), and the expression pattern of csn-miR398-3p-1 is opposite that of CsCSD4 under cold treatment. Roles of miRNAs have also been found in the bud dormancy of tea plants. Cs-miR139c regulates the release of tea plants from bud dormancy by suppressing the expression of its target gene, CsnTCP2 (Teosinte branched/Cycloidea/Proliferating cell factor 2). There is no doubt that the investigation of miRNAs provides a broad understanding of the posttranscriptional regulatory networks between protein-coding and ncRNAs; however, most current progress is limited to the identification, but not the functional verification, of miRNAs, which will require further investigation in tea plants.

In addition to miRNAs, long noncoding RNAs (lncRNAs) and circRNAs have been identified in tea plants using 170 RNA-seq datasets (~7157 million reads) from 11 tissues, generating more than 33,000 putative novel lncRNAs. The tea plant lncRNAs showed tissue-specific expression patterns and were suggested to be involved in major aroma formation pathways of black tea. circRNAs (circular RNAs) were recently discovered as a new class of ncRNAs that are covalently closed, single-stranded, and generated via back-splicing events. Since their first documentation in Arabidopsis, circRNAs have been detected in several other plants. A total of 3174 circRNAs have been characterized in tea plants. As similarly observed in most other plants, tea plant circRNAs exhibit tissue-specific expression patterns, and their expression shows a positive relationship with that of their parental genes. Tea plant circRNAs were found to exhibit potential miRNA-binding sites and to display crucial functions in the photosynthetic machinery and metabolite biosynthesis during leaf development. In other plants, circRNAs can act as miRNA sponges under tomato chilling injury, dehydration stress in wheat, and infection with the bacterial canker pathogen in kiwifruit. These studies have increasingly complicated our understanding of ncRNAs and their regulatory mechanisms in tea plants. Additionally, none of these identified ncRNAs (miRNAs, lncRNAs, and circRNAs) have been functionally well-verified regarding their roles in regulating posttranscriptional processes in tea plants. Further efforts aimed at accurate lncRNA characterization and the functional examination of their interaction mechanisms in competing endogenous RNA (ceRNA) networks should receive more attention to elucidate their roles in response to stress and the regulation of tea plant secondary metabolite formation.

Genetic diversity and population structure of tea plants

In the thousands of years since the tea plant was first discovered, recorded, and cultivated to produce tea in China, it has spread to more than 50 countries around the world in tropical and subtropical regions. The wide range of growing areas, long history of cultivation, self-incompatibility and allogamy of tea plants make these cultivars highly heterogeneous, resulting in high diversity in both their genetics and morphology. As the key component of crop genetic improvement, the available germplasms and the genetic diversity within a gene pool determine the success of breeding programs. The development of molecular markers, such as RAPD, AFLP, SSR, and SNP markers, has allowed the successful utilization of these markers to estimate the genetic diversity and determine the phylogenetic relationships of different tea germplasms. Early in 1995, the systematic assessment of the genetic variability of three major types (CSA, CSS, and Cambod type) of tea plants from Kenya and the UK suggested that the Cambod population possessed the lowest diversity. The further investigation of genetic and morphological traits revealed consistent findings in populations from other major tea plantation countries. Germplasm from the Korean population shows higher diversity than that from Japan due to more extensive plantations in Korea and longer, intensive tea selection programs in Japan. The genetic diversity
of the Chinese, Indian, and Sri Lankan populations is higher than those from other countries\textsuperscript{151}. Most Italian tea plants exhibit a close relationship with cultivars from Zhejiang Province\textsuperscript{159}. Kenyan tea plants show the highest diversity among all African germplasms, while the lowest is found in South Africa\textsuperscript{160,161}. CSA was indicated to have donated the largest amount of genetic material in African tea breeding\textsuperscript{161}. Narrow genetic diversity is detected in the Sri Lankan population, which is dominated by Assam-type tea\textsuperscript{162}. The estimation of worldwide tea plant genetic diversity has increased our understanding of the population structures and origins of tea plants, which will benefit the global tea industry.

China hosts abundant tea plant germplasms. A wide range of genetic diversity and population studies have been undertaken, particularly for Chinese germplasms\textsuperscript{148,149,152,163–167}. Most of the Chinese tea germplasms can be clustered based on their geographical origin and genetic background\textsuperscript{149}. Higher levels of genetic diversity are observed in Guangxi, Yunnan, and Guizhou Provinces\textsuperscript{148,167}. Compared to the wild tea plants, such as \textit{C. taliensis}, the cultivated tea plants display higher heterozygosity\textsuperscript{152}.

Molecular markers are useful and are commonly employed in crop breeding. Traditional experimental screening of marker polymorphism is laborious and time-consuming, especially for the broadly used SSR markers. The development of the CandiSSR pipeline has facilitated the efficient identification of polymorphic SSRs (PolySSRs) based on high-throughput sequencing data, and this pipeline has successfully identified 450 PolySSRs in the transcriptomes of four \textit{Camellia} species\textsuperscript{168}. Despite the diverse molecular and morphological markers developed, more efforts are still needed to develop extra-robust markers (e.g., SNPs) to better facilitate the conservation and utilization of global tea resources.
Platforms and tools for tea plant functional genomics

Databases

Biological databases provide scientists with an opportunity to centrally access a variety of biological data. Various tea plant databases have been established to help the tea-related research community to better understand the metabolome, health benefits, and genomics of tea169–171 (Fig. 2b). Yue and colleagues constructed the first comprehensive manually curated Tea Metabolome Database (TMDB)171, which contains over 1400 metabolites and 600 NMR datasets collected from 364 publications. Retrieval from the TMDB can thoroughly access a total of 24 basic attributes of tea compounds of interest, such as compound structure, molecular weight, and compound uses. The beneficial effects of tea are principally derived from diverse bioactive compounds. The establishment of the searchable TBC2health database has connected the relationships between tea compounds and their health beneficial effects well169. The current release of TBC2health compromises 240 bioactive tea compounds and their associated health effects on 206 diseases. The user-friendly response interface contains multiple easily visualized results, aiding in the efficient discovery of potential mechanisms of the health benefits of tea. Increasingly, studies have confirmed that the health benefits of tea are primarily achieved through the regulation of target gene expression or protein activities172,173. More recently, the TBC2target database was developed to provide candidate target genes for a total of 240 bioactive tea compounds170. The target genes were predicted using PharmMapper via a pharmacophore-mapping strategy174. With TBC2target in hand, scientists can explore the potential genes targeted by tea compounds and then reveal the possible health-promoting mechanisms of bioactive tea compounds.

Tea plants are rich in secondary metabolites that essentially determine tea quality2. Hundreds of studies on tea quality and physiology have been reported, which have generated a wide variety of specific biological data1,2. The recently developed Tea Plant Information Archive (TPIA) database employs the published tea plant genome as a basic framework and incorporates a wide range of publicly available genomic, transcriptomic, and metabolomic data of tea plants together with globally collected tea germplasms175. It also comprises over 70 transcriptomes widely collected from 21 Camellia species, plentiful gene expression data from various conditions (across species, stresses, and hormone treatments), orthologs, transcription factors, SSRs, metabolites (mainly catechins, theanine, and caffeine), correlations, and manually curated functional genes. Through long-term maintenance and timely updating with novel datasets, the TPIA is gradually becoming the central gateway for tea communities to investigate the biology of tea plants in more detail, thus benefitting the sustainable development of the global tea industry.

Quantitative trait loci (QTL) mapping and bulked-segregant analysis (BSA)

QTL mapping is an efficient tool for revealing the molecular basis of complex traits of organisms175. Despite the fact that QTL mapping is widely used in several crops177,178, the application of this method to tea plants is still challenged, mainly due to the self-incompatible nature of tea plants, which exhibit a low seed yield, leading to the propagation of insufficient populations for QTL mapping. Nevertheless, considering the importance of tea plants in global tea consumption, increasing efforts are being focused on the QTL mapping of several agronomic traits of tea plants, particularly tea yield20 and secondary metabolites22,27–29 (Fig. 2b). A total of 23 putative QTLs associated with tea yield were detected using 5-year yield data from 42 F1 individuals collected across two sites in Kenya20. These efforts led to the first QTL-mapping investigation in tea plants, but there was no shared QTLs discovered at the two sites, indicating an overestimated QTL effect due to insufficient mapping populations. In contrast to yield traits, QTLs related to tea quality have been widely explored. Using 2-year catechins data from 183 individuals crossbred from two tea plant varieties with distinct catechins composition, a total of 25 catechin-related QTLs were identified27. Nine of these QTLs were validated across years and found to cluster together in the chromosomal regions of LG03 and LG11, suggesting a potential tandem duplication origin of tea plant catechin genes. Similarly, 10 QTLs were found to be involved in the determination of theobromine and caffeine contents using 148 individuals from a pseudo-testcross population22. With the exception of one QTL related to caffeine content, which was validated across years and mapped on LG01, the other QTLs were validated from only 1 or 2 years of data, possibly due to the relatively small size of the mapping populations. With the expansion of mapping populations, a recent QTL mapping study detected several novel QTLs related to flavonoids29. Among the 27 identified QTLs associated with flavonoids, 7 were newly revealed to be involved in the determination of anthocyanin content and young shoot color. Interestingly, quite different from other recent reports22, QTLs controlling caffeine content were identified on LG11 in this study29. This implies that the caffeine content of tea plants is probably controlled by multiple genes located on different chromosomes or that the QTL effects have been overestimated because of the limitations of the mapping populations. In addition to tea quality and yield traits, QTL mapping has been applied to agronomic traits such as spring bud flush and leaf size (e.g., mature leaf length, width, and shape indexes)28. In the future,
efforts will be needed to further increase the mapping populations and improve the density of genetic linkage maps to precisely identify QTLs in tea plants.

Although over 80 QTLs have been detected for diverse agronomic traits of tea plants, it is still difficult to explore the candidate genes and/or alleles related to QTLs further owing to the insufficient genetic markers and genomic information available. The recently developed BSA method selects and pools samples from individuals with extreme phenotypes from biparental segregation populations and enables the efficient identification of genes and alleles governing complex traits through statistical genomic and phenotypic analyses. In tea plants, the utilization of the BSA method coupled with RNA sequencing has facilitated the discovery of a flavonoid 3',5'-hydroxylase (F3’5’H) gene that is significantly associated with catechin content based on individuals from the two tails of an F1 population that segregates with catechin content.

**Future challenges and perspectives**

**Completion of the tea plant genome with advanced technologies**

Tea originated from China and has spread to over 160 countries around the world. According to statistics from 2017 from the Food and Agriculture Organization of the United Nations (FAO; http://www.fao.org/), the global tea plantation area has exceeded 4.08 million hectares, and global total tea production has reached 6.10 million tons, with an average increase of 0.33 million hectares and 0.58 million tons over the past 5 years (Fig. 2a). In stark contrast to the flourishing development of the global tea industry, research on the major fundamental biological problems encountered in tea plants is still lagging behind, resulting in a low breeding efficiency and few excellent varieties. The rapid development and application of genomics approaches have effectively promoted the breeding programs of crops; however, genomic investigations of tea plants are still limited and challenging. Although the genome sequences of two major varieties (CSS and CSA) of tea plants were generated recently, their quality and completeness still need to be improved. Given the limitations of current NGS-based approaches related to highly repetitive and heterozygous genomes, advanced single-molecule-sequencing technologies (e.g., PacBio SMRT and Oxford Nanopore) will help resolve the highly repetitive regions of the tea plant genome. With the further inclusion of linkage map and Hi-C technology to interactively address heterozygous genomic regions, the eventually produced chromosome-level reference genome of tea plants will facilitate comparative genomics, evolutionary and population genetics studies and, thus, promote breeding programs in the future (Fig. 3).

**Towards tea plant pan-genomics and the phylogeny of genus Camellia**

Tea plants are widely distributed in tropical and subtropical regions around the world. They contain rich and unique characteristic secondary metabolites, such as catechins, theanine, and caffeine, which are essential to the formation of tea quality. However, the contents of these secondary metabolites vary greatly among different varieties and *Camellia* species, leading to large differences in tea processing suitability. They also differ significantly in several morphological traits (e.g., leaf size) and stress resistance characteristics (e.g., cold tolerance), showing a divergent genetic makeup. Therefore, the genome sequence of a single individual of a tea plant variety (e.g., shuchazao or yunkang #10) cannot represent the entire gene pool of tea plants. Pan-genome sequencing
is a newly developed strategy for investigating the genetic variations among different subspecies and/or individuals that has been widely used in several crops and model species, such as rice, soybean, maize, and Arabidopsis. With the completion of the tea plant genome, more efforts are still needed to further investigate the pan-genomics of tea plants from different varieties and/or ecotypes to expand the gene pool. However, the phylogeny of genus Camellia and the relationships among different tea populations remain largely unclear, which has seriously hindered the in-depth exploration of pan-genomics in tea plants.

Although several studies have attempted to resolve the phylogeny of genus Camellia using chloroplast and/or mitochondrial DNA sequences, the generated phylogenetic trees are somewhat poorly supported, primarily due to the high frequency of hybridization and polyploidyization, as well as the relative conservation of organelle DNA sequences among Camellia species. Compared to organelle fragments, massive nuclear gene sequences rapidly obtained from genome and transcriptome sequencing provide abundant candidate low-copy-number nuclear genes and informative sites for phylogenetic investigation and have been increasingly applied to clarify the phylogenies of some complex plant taxa. Therefore, the sequencing of additional genomes and/or transcriptomes of Camellia species, as well as elite tea cultivars in the foreseeable future will not only help to reveal their complex phylogenetic relationships but will in turn promote pan-genomic studies of tea plants to expand the gene pool.

Genome-wide association study (GWAS) of tea plants

GWAS is a widely accepted strategy to identify genes/alleles associated with complex traits in crops with the rapid innovation of NGS technologies. Genotypes (typically based on SNPs) and phenotypes (usually those of different traits) are the two fundamental requirements in GWAS, in which a higher SNP density and more diverse phenotypes of a population will theoretically produce more reliable and significant associations. However, conventional GWAS design and methods have resulted in challenges in the identification of multiple functional alleles within a single gene as well as rare alleles in the population. GWAS of tea plants has also been confronted with major difficulties in population construction and phenotyping related to germplasm collection and the ultraslow growth of tea plants when transplanted to nursery gardens. Nevertheless, considering the high level of genetic and morphological diversity in tea plants, GWAS is still the most efficient way to detect candidate loci of complex traits undergoing improvement. To apply GWAS to study the genetics underlying tea traits of interest, the following goals should be considered: (1) collection and construction of core germplasms with adequate genetic diversity; (2) transplantation of the germplasms to one or more local gardens for accurate multiple-year phenotyping; (3) further improvement of the quality and annotation accuracy of the current genome assembly; (4) integration of GWAS data with other omics datasets (e.g., transcriptomic and metabolomics data); and (5) construction of an efficient transgenic system to functionally validate the candidates. In this way, GWAS can be effectively applied to...
understand the genetic basis of agronomic traits associated with tea quality, thus enabling the molecular design of new cultivars in the near future.

**Origin and domestication of tea plants**

China is documented as the first country to cultivate and utilize tea plants in the world. It harbors abundant tea germplasms and has long been considered the origin of tea plants. However, since no wild ancestor of tea plants has been discovered in China, the origin and domestication of tea plants are still a mystery and attract worldwide attention. Although recent progress has been made in understanding the origin and domestication of tea plants based on molecular markers (primarily SSRs), the resultant conclusions are unilateral and even controversial due to the poor representativeness of sample collections and the lack of sufficient molecular markers for diversity evaluation. The recent release of tea plant genome sequences has laid a solid foundation for solving this problem; however, further efforts focusing on the following issues are still needed: (1) collection of globally representative tea plant samples; (2) investigation of the population structure of tea plants and identification of their putative wild ancestor; (3) estimation of population diversity using genome-wide SNP markers; (4) scanning of genomic regions with significantly lower diversity in cultivated but not wild tea plants and identification of the candidate genomic regions selected during domestication; and (5) functional investigation of the characteristics of genes within regions of domestication. It should be noted that most of the existing tea varieties are bred and propagated directly from natural populations. Artificial domestication may have had little impact on the variation in genome sequences. Therefore, tea plants may actually undergo adaptive evolution compared to other crops (e.g., rice) subjected to strong domestication pressure.

**Establishment of high-efficient transgenic system**

*Agrobacterium*-mediated genetic transformation has been extensively applied to verify the function of genes with the aim of improving the quality and yield of crops. In tea plants, the first transgenic plant was generated using *Agrobacterium*-mediated transformation in somatic embryos. However, at least 45 weeks were possibly needed for the transformation procedure alone, and additional years were required for further transplantation. It also required several years of experimental work to establish stability and germline transmission in these plants. Although several efforts have been made to optimize the transformation system of tea plants using different strategies, transformation efficiency still represents a challenge under the available methods. In the face of these challenges, some key points related to transgenic technology need to be further detailed: (1) optimization of the transformation and regeneration system; (2) clarification of the transformation mechanism of tea plants, particularly to elucidate the effects of tea polyphenols on *Agrobacterium* infection; and (3) identification of more suitable tea varieties for better transformation and regeneration. The most recent CRISPR system has enabled efficient genome editing in crops and shows great potential for accelerating precise crop improvement. Considering the difficulties encountered transgenic studies of tea plant, the establishment of the CRISPR system will promote the transgenic progress in tea plants. More recently, scientists demonstrated an efficient approach to overcome self-incompatibility in diploid potato using the CRISPR/Cas9 system through S-RNase locus knockout, which provided a useful example for the further application of transgenic technology and breeding in tea plants, considering their similar self-incompatibility features.

**Collection and conservation of tea plant germplasms**

Resources are clearly the key to crop genetic improvement. The success of plant breeding and conservation is largely dependent on the amount and distribution of genetic variations present in collections. Currently, tea germplasms are becoming the most valuable fundamental material for tea breeding and biotechnology studies, presenting great potential in the whole tea industry. The remarkable achievements made in worldwide tea plant germplasm surveys, collection, and conservation have preserved more than 15,000 tea plant accessions according to incomplete statistics from the China Tea Science Society (http://www.chinatss.cn/). These invaluable germplasms have driven the extraordinarily rapid development of tea plant genomics, genetics, and breeding. However, most of the current accessions that have been collected and used are varieties, and few “wild” or close relatives have been integrated. Another major problem that is currently faced by the field is the imbalanced utilization and protection of cultivars versus “wild” germplasms. The survival environment of “wild” tea resources has been constantly destroyed with rapid economic development and urbanization. In addition, local germplasms (landraces) with specific characteristics are on the edge of being lost due to the popularization of elite varieties in the past few years. Above all, the surveying, collection, and conservation of tea plant germplasms should be urgently emphasized. Otherwise, no diverse tea plant resources or only a few mono varieties will survive and be available for utilization for the genetic improvement of tea in the future. We therefore propose that a tea plant conservation alliance uniting the global tea-planting countries needs to be established to accelerate the preservation and utilization of global tea plant resources.
In-depth understanding of the complex secondary metabolism network

The overall importance of tea plants as an economic crop is due to the health-promoting functions of teas as non-alcoholic drinks that are popular worldwide. Numerous studies have documented the health benefits of various tea drinks in humans, and studies continue to produce more fundamental discoveries about the effects of functional tea components on the improvement of human health. It is not only the three major types of tea plant secondary metabolites (catechins, caffeine, and theanine) but also the volatile terpenoids, saponins, polysaccharides, and other phenolic conjugates found in tea that contribute to the beneficial health effects and the enjoyable flavors of various teas. As mentioned above, our understanding of these complicated secondary metabolites in tea plants in terms of their biosynthesis, transport, and storage as well as their regulation is very limited. This is one of the major obstacles to the breeding of high-quality tea plant varieties with biofortified nutrients or functional components or other special properties. Future work could focus on basic questions about tea secondary metabolism, which will facilitate the breeding of tea plants with desirable qualities and benefit human health improvement projects.

In conclusion, tea plants are perennial woody plants in nature with a long growth cycle, which severely limits their genetic breeding. Traditional cross-breeding is extremely difficult and time-consuming for tea plants because most of the existing tea varieties are bred and propagated directly from natural populations. Modern transgenic breeding technology has provided us a new solution for the molecular design of breeding strategies, which basically relies on the development of functional genomics and molecular biology. Although great progress has been made in the last two decades, the genomics and molecular biology of tea plants are still not fully understood, mainly due to difficulties in resource collection and identification, the generation and identification of mutant plants, and population construction. In particular, the limited knowledge of the functional genomics and developmental biology of tea plants has narrowed our understanding of their basic biological characteristics. Therefore, compared to other crops such as rice, there is still a long way to go in tea plant genomics and molecular biology research. In the near future, scientists should focus more on industry-oriented major basic biological research based on germplasm collection and utilization, particularly by deciphering the tea plant genome, as an opportunity for enhancing studies on the mechanisms of the biosynthesis and regulation of secondary metabolites, the genetic basis of important agronomic traits, the molecular mechanisms of stress and disease resistance, and the developmental biology of tea plants, to accelerate the molecular design breeding program and, thus, promote industrial development.

Acknowledgements

We would like to thank Dr. Penghui Li, Dr. Zhaolang Zhang, and Dr. Chuankui Song for his discussion and comments on the manuscript. We apologize for not being able to cite many of the excellent publications on tea plants due to the limit of paper length. The authors acknowledge support from the National Natural Science Foundation of China (31800180), the Natural Science Foundation of Anhui Province (1908085MC75), the National Key Research and Development Program of China (2018YFD0100601), the Science and Technology Project of Anhui Province (13Z03012), the Special Innovative Province Construction in Anhui Province (15zs08032), the Changjiang Scholars and Innovative Research Team in University (IRT1101), the China Postdoctoral Science Foundation (No. 2017M621992), and the Postdoctoral Science Foundation of Anhui Province, China (No. 2017B189).

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information

accompanies this paper at (https://doi.org/10.1038/s41438-019-0225-4).

Received: 10 July 2019 Revised: 17 October 2019 Accepted: 3 November 2019

Published online: 01 January 2020

References

1. Xia, E.-H. et al. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol. Plant 10, 866–877 (2017).
2. Wei, C. et al. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proc. Natl. Acad. Sci. USA 115, E4151–E4158 (2018).
3. Li, Y. et al. Rice functional genomics research: past decade and future. Mol. Plant 11, 359–380 (2018).
4. Li, B., Chen, X., Chen, G. & Wang, J. The analysis of karyotype in tea plant. J. Tea Sci. 6, 7–14 (1986).
5. Sheidai, M., Jahanbakht, H. & Sofi-Siyavash, P. Cytogenetic study of various types of tea (Camellia sinensis) cultivars in Iran. Iran. J. Sci. Technol. 28, 33–42 (2004).
6. Huang, H., Tong, Y., Zhang, Q. & Gao, L. Z. Genome size variation among and within Camellia species by using flow cytometric analysis. PLoS ONE 8, e64981 (2013).
7. Furukawa, K., Sugiyama, S., Ohta, T. & Ohmido, N. Chromosome analysis of tea plant (Camellia sinensis) and ornamental camellia (Camellia japonica). Chromosome Sci. 20, 9–15 (2017).
8. Morinaga, T., Fukushima, E., Kano, T., Maruyama, Y. & Yamasaki, Y. Chromosome numbers of cultivated plants I. Bot. Mag. 43, 589–594 (1929).
9. Morinaga, T. & Fukushima, E. Chromosome numbers of cultivated plants II. Bot. Mag. 45, 140–145 (1931).
10. Kondo, K. Chromosome numbers in the genus Camellia. Biotropica 9, 86–94 (1977).
11. Sharma, S. & Raina, N. Chromosome constitution of some Indian tea clones. Int. J. Tea Sci. 5, 21–28 (2006).
12. Rahman, H. et al. Cytomorphological characterization of tea cultivars. Pak. J. Bot. 42, 485–495 (2010).
13. Hanson, L., Mahon, K., Johnson, M. & Bennett, M. First nuclear DNA C-values for another 25 angiosperm families. Ann. Bot. 88, 851–858 (2001).
14. Tanaka, J., Taniguchi, F., Hirai, N. & Yamaguchi, S. Estimation of the genome size of tea (Camellia sinensis). Plant Mol. Biol. 238 (2017).
15. Sharma, S., Kaushik, S. & Raina, N. Estimation of nuclear DNA content and its variation among Indian Tea accessions by flow cytometry. Physiol. Mol. Biol. Plants 1, 1–8 (2018).
16. Ott, J., Wang, J. & Lesl, S. M. Genetic linkage analysis in the age of whole-genome sequencing. Nat. Rev. Genet. 16, 275–284 (2015).
17. Zhang, C. C. et al. Transcriptome analysis reveals self-incompatibility in the tea plant (Camellia sinensis) might be under gametophytic control. BMC Genom. 17, 359 (2016).

18. Hackett, C., Wachira, F., Paul, S., Powell, W. & Waugh, R. Construction of a genetic linkage map for Camellia sinensis (tea). Heredity 85, 346–355 (2000).

19. Huang, J. et al. Construction of AFLP molecular markers linkage map in tea plant. J. Tea Sci. 25, 7–15 (2005).

20. Kamunya, S. M. et al. Genomic mapping and testing for quantitative trait loci in tea (Camellia sinensis (L.) O. Kuntze). Tree Genet. Genomes 6, 915–929 (2010).

21. Ma, J. Q. et al. Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). PLoS ONE 10, e0126796 (2015).

22. Ma, J. Q. et al. Quantitative trait loci mapping for theobromine and caffeine contents in tea plant (Camellia sinensis). J. Agric. Food Chem. 66, 13321–13327 (2018).

23. Ota, S. RAPD-based linkage mapping using F1 segregating populations derived from crossings between tea cultivar ‘Sayamakori’ and strain Yana-Oki7. Breed. Res. 1, 16 (1999).

24. Huang, F., Liang, Y., Lu, J. & Chen, R. Genetic mapping of Ck17 (mitochondrial DNA) in tea plant (Camellia sinensis). Euphytica 213, https://doi.org/10.1007/s10681-017-1797-0 (2017).

25. Ma, J. Q. et al. Construction of a SSR-based genetic map and identification of QTLs for catechins content in tea plant (Camellia sinensis). PLoS ONE 9, e93131 (2014).

26. Tan, L.-Q. et al. SSR-based genetic mapping and QTL analysis for timing of spring bud flushing, young shoot color, and mature leaf size in tea plant (Camellia sinensis). Tree Genet. Genomes 11, https://doi.org/10.1007/s11295-016-1088-9 (2016).

27. Xu, L. Y. et al. High-density SNP linkage map construction and QTL mapping for flavonoid-related traits in a tea plant (Camellia sinensis) using 2b-RAD sequencing. BMC Genom. 19, 955 (2018).

28. Chen, F. et al. The sequenced angiosperm genomes and genome databases. Front. Plant Sci. 9, 418 (2018).

29. Cai, J. J. et al. The genome sequence of the orchid Phalaenopsis equestris. Nat. Genet. 47, 65–72 (2015).

30. Peng, Z. et al. The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). Nat. Genet. 45, 456–461 (2013).

31. Goff, S. A. et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296, 92–100 (2002).

32. Dodds, P. N. & Rathjen, J. P. Plant immunity: towards an integrated view of plant-pathogen interactions. Nat. Rev. Genet. 11, 539–548 (2010).

33. Ashihara, H. & Suzuki, T. Distribution and biosynthesis of caffeine in plants. J. Agric. Food Chem. 41, 229–234 (2011).

34. Rees, S. M. & Tenaillon, I. M. The comparative transcriptome of domesticated species moso bamboo (Phyllostachys bambusoides). Euphytica 213, https://doi.org/10.1007/s10681-017-1797-0 (2017).

35. Xia, E. H. et al. Transcriptome analysis of the oil-rich tea plant, Camellia oleifera. BMC Genom. 12, 757 (2011).

36. Zhang, H. B. et al. De novo transcriptome assembly of the wild relative of tea plant (Camellia sinensis). BMC Genom. 16, 660 (2015).

37. Wang, W. et al. Insight into catechins metabolic pathways of Camellia sinensis based on genome and transcriptome analysis. J. Agric. Food Chem. 66, 4281–4293 (2018).

38. Liu, J. et al. Transcriptome analysis reveals the accumulation mechanism of anthocyanins in Zulian tea (Camellia sinensis var. assamica (Watt) krahmann) leaves. Plant Growth Regul. 81, 51–61 (2016).

39. Xu, Q. et al. Transcriptome profiling using single-molecule direct RNA sequencing approach for in-depth understanding of genes in secondary metabolism pathways of Camellia sinensis. Front. Plant Sci. 8, 1205 (2017).

40. Xia, E. H. et al. Transcriptome analysis of the oil-rich tea plant, Camellia oleifera, reveals candidate genes related to lipid metabolism. PLoS ONE 9, e104150 (2014).

41. Zhang, H. B. et al. De novo transcriptome assembly of the wild relative of tea (Camellia taliensis) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. BMC Genom. 16, 298 (2015).

42. Yao, Q. Y., Huang, H., Tong, Y., Xia, E. H. & Gao, L. Z. Transcriptome analysis identifies candidate genes related to tricglycerol and pigment biosynthesis and photosynthetic flowering in the ornamental and oil-producing plant, Camellia reticulata (Theaceae). Front. Plant Sci. 7, 163 (2016).

43. Huang, H., Xia, E. H., Zhang, H. B., Yao, Q. Y. & Gao, L. Z. De novo transcriptome sequencing of Camellia sasanqua and the analysis of major candidate genes related to floral traits. Plant Physiol. Biochem. 120, 103–111 (2017).

44. Wang, Z. W. et al. Deep sequencing of the Camellia chalepogandica transcriptome revealed candidate genes for anthocyanin biosynthesis. Gene 538, 1–7 (2014).
90. Ohgami, S. et al. Volatile glycosylation in tea plants: sequential glycosylations
85. Deng, W. W. et al. Molecular cloning, functional analysis of three cinnamyl
83. Acharya, K. et al. Overexpression of
93. Xia, J. et al. Characterization and expression profiling of Camellia sinensis
82. Li, X. W. et al. A novel cold-regulated gene from
74. Liu, Y. et al. Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (Camellia sinensis). J. Biol. Chem. 287, 44406–44417 (2012).
91. Wang, D., Kubota, K., Kobayashi, A. & Juan, I. Analysis of glycosidically bound
78. gene from tea plant (Camellia sinensis) - CS CO R 1 , an enzyme involved in catechin galloylation in the tea plant (Camellia sinensis). J. Agric. Food Chem. 57, 11036–11045 (2009).
70. Zeng, L., Watanabe, N. & Yang, Z. Understanding the biosyntheses and
193 (2017).
76. Wei, K. et al. A coupled role for CsMYB75 and CsGSTF1 in anthocyanin
71. Song, C., Hartl, K., McGrath, K., Hoffman, T. & Schwab, W. Attractive but
toxic: emerging roles of glycosidically bound volatiles and glycosyltransf
72. Zhang, Z. et al. Advances in research on functional genes of tea plant.
74. L i u , Y. e t a l . P u r i f i c a t i o n and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (Camellia sinensis). J. Biol. Chem. 287, 44406–44417 (2012).
94. Han, Y. et al. Functional analysis of two flavanone-3-hydroxylase genes from Camellia sinensis: a critical role in flavonoid accumulation. Genes 8, 300 (2017).
95. Liu, Y. et al. Three Camellia sinensis glutathione S-transferases are involved in the storage of anthocyanins, flavonols, and proanthocyanidins. Planta 250, 1163–1175 (2019).
96. Dong, C. et al. Theanine transporters identified in tea plants (Camellia sinensis L.). Plant J. https://doi.org/10.1111/tpj.14517 (2019).
97. Deng, W.-W., Wang, R., Yang, T., Jiang, L. N. & Zhang, Z.-Z. Functional char
98. Wang, M. et al. Molecular cloning and expression analysis of low molecular weight heat shock protein gene CsHSP17.2 from Camellia sinensis. J. Nanjing Agric. Univ. 38, 389–394 (2015).
99. Mohanpuria, P., Kumar, V. & Yadav, S. K. Tea caffeine: metabolism, functions, and reduction strategies. Food Sci. Biotechnol. 19, 275–287 (2010).
100. Shirlow, M. & Mathers, C. A. Study of caffeine consumption and symptoms: indigestion, palpitations, tremor, headache and insomnia. Int. J. Epidemiol. 14, 289–298 (1985).
101. Schaffer, S. W. et al. Effect of taurine and potential interactions with caffeine on cardiovascular function. Amino Acids 46, 1147–1157 (2014).
102. Yamamoto, S., Wakayama, M. & Tachiki, T. Cloning and expression of Pseu
domonas 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 (2006).
103. Higdon, J. V. & Frei, B. Tea catechins and polyphenols: health effects, meta
bolism, and antioxidant functions. Crit. Rev. Food Sci. Nutr. 43, 89–143 (2003).
104. He, X. et al. Isolation and characterization of key genes that promote fla
vonoid accumulation in purple-leaf tea (Camellia sinensis L.). Sci. Rep. 8, 130 (2018).
105. Gohain, B. et al. Understanding Darjeeling tea flavour on a molecular basis. Plant Mol. Biol. 78, 527–592 (2012).
106. Han, Z. X. et al. Green tea flavour determinants and their changes over manufacturing processes. Food Chem. 212, 739–748 (2016).
107. Lee, J., Chambers, D. H. & Chambers, E. A. Comparison of the flavor of green teas from around the world. J. Sci. Food Agric. 94, 1315–1324 (2014).
108. Hu, C. J. et al. Formation mechanism of the oolong tea characteristic aroma during bruising and withering treatment. Food Chem. 269, 202–211 (2018).
109. Schuh, C. & Scheibe, P. Characterization of the key aroma compounds in the beverage prepared from Darjeeling black tea: quantitative differences between tea leaves and infusion. J. Agric. Food Chem. 54, 916–924 (2006).
110. Guo, W. et al. (S)-linalyl, 2-phenylethyl, and benzyl disaccharide glycosides isolated as aroma precursors from oolong tea leaves. Biosci. Biotechnol. Biochem. 70, 500–507 (2006).
111. Deng, W. W. et al. Molecular cloning, functional analysis of three cinnamyl alcohol dehydrogenase (CAD) genes in the leaves of tea plant, Camellia sinensis. J. Agric. Food Chem. 57, 11036–11045 (2009).
112. Hu, C. J. et al. Formation mechanism of the oolong tea characteristic aroma during bruising and withering treatment. Food Chem. 269, 202–211 (2018).
113. Schuh, C. & Scheibe, P. Characterization of the key aroma compounds in the beverage prepared from Darjeeling black tea: quantitative differences between tea leaves and infusion. J. Agric. Food Chem. 54, 916–924 (2006).
114. Guo, W. et al. (S)-linalyl, 2-phenylethyl, and benzyl disaccharide glycosides isolated as aroma precursors from oolong tea leaves. Biosci. Biotechnol. Biochem. 70, 500–507 (2006).
115. Wang, D., Kubota, K., Kobayashi, A. & Juan, I. Analysis of glycosidically bound aroma precursors in tea leaves. 3. Change in the glycoside content of tea leaves during the oolong tea manufacturing process. J. Agric. Food Chem. 49, 5391–5396 (2001).
116. Wang, D., Yoshimura, T., Kubota, K. & Kobayashi, A. Analysis of glycosidically bound aroma precursors in tea leaves. 1. Qualitative and quantitative analysis of glycosides with aglycons as aroma compounds. J. Agric. Food Chem. 48, 5411–5418 (2000).
117. Ho, C., Zheng, X. & Li, S. Tea aroma formation. Food Sci. Hum. Wellness 4, 9–27 (2015).
118. Mei, X. et al. Formation and emission of linalool in tea (Camellia sinensis) leaves infested by tea green leafhopper (Empoasca (Matsumurae) anulata Matsuda). Food Chem. 237, 356–363 (2017).
119. Tariq, M. & Paszkowski, J. DNA and histone methylation in plants. Trends Genet. 20, 244–251 (2004).
120. Zhang, H., Lang, Z. & Zhu, J. K. Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. 19, 489–506 (2018).
121. Cokus, S. J. et al. Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219 (2008).
122. Li, X. et al. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell 20, 259–276 (2008).
123. Wang, L. et al. DNA methylome analysis provides evidence that the expression of the tea genome is linked to TE bursts. Plant Biotechnol. J. https://doi.org/10.1111/pbi.13018 (2018).
124. Nystedt, B. et al. The Norway spruce genome sequence and confier genome evolution. Nature 497, 579–584 (2013).
170. Zhang, S. et al. TB2Citgear: a resource of predicted target genes of bioactive compounds. Front. Plant Sci. 9, 211 (2018).

171. Yue, Y. et al. TMDIB: a literature-curated database for small molecular compounds found from tea. BMC Plant Biol. 14, 243 (2014).

172. Kien, R. D. & Fischer, S. M. Black tea polyphenols inhibit IGF-induced signalling through Akt in normal prostate epithelial cells and DuI45 prostate carcinoma cells. Carcinogenesis 23, 217–221 (2002).

173. Li, M. et al. EGCG induces lung cancer AS49 cell apoptosis by regulating Ku70 acetylation. Onco. Rep. 35, 2339–2347 (2016).

174. Wang, X. et al. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. Nucleic Acids Res. 45, W556–W569 (2016).

175. Xia, E. H. et al. Tea Plant Information Archive: a comprehensive genomics and bioinformatics platform for tea plant. Plant Biotechnol. J. 17, 1958–1953 (2019).

176. Wurschum, T. & Mapping, Q. L. T. for agronomic traits in breeding populations. Theor. Appl. Genet. 125, 201–210 (2012).

177. McCough, S. R. & Doerge, R. W. QTL mapping in rice. Trends Genet. 11, 462–467 (1995).

178. Salama, S. J., Hostet, B. M., Leduea, J. R., Stubbe, C. W. & Holland, J. B. QTL mapping with near-isogenic lines in maize. Theor. Appl. Genet. 114, 1211–1228 (2007).

179. Jin, Q. J., Ma, J. Q., Yao, M. Z., Ma, C. L. & Chen, L. Functional natural allelic variants of flavonoid 3,5'-hydroxylase gene governing catechin traits in tea plant and its relatives. Planta 245, 523–538 (2017).

180. Rothenberg, D., Wang, H., Chen, M., Zheng, W. & Zhang, L. Metabolome and transcriptome sequencing analysis reveals anthocyanin metabolism in pink flowers of anthocyanin-rich tea (Camellia sinensis). Molecules 24, 1064 (2019).

181. Xu, Q. et al. Unraveling a crawlish regulatory network of temporal aroma accumulation in tea plant (Camellia sinensis) leaves by integration of metabonomics and transcriptomics. Environ. Exp. Bot. 149, 81–94 (2018).

182. Tai, Y. et al. Gene co-expression network analysis reveals coordinated regulatory sets. 176. Wuerschum, T. & Mapping, Q. L. T. for agronomic traits in breeding populations. Theor. Appl. Genet. 125, 201–210 (2012).

183. Huang, X. H. et al. Genome-wide association studies of 14 agronomic traits in Camellia sinensis. Nucleic Acids Res. 201, E4868 (2015).

184. Chand, Y. et al. TMDB: a literature-curated database for small molecular compounds. Environ. Exp. Bot. 1045, 245–243 (2011).

185. Li, Y. H. et al. De novo assembly of soybean wild relatives for pan-genome characterization using nuclear microsatellites and cpDNA sequence data. Plant Sci. 45, 500–505 (2017).

186. Wei, K. et al. Comparison of catechins and purine alkaloids in albino and normal green tea cultivars (Camellia sinensis L.) by HPLC. Food Chem. 130, 720–724 (2012).

187. Eniu, P. I. et al. Leaf morphology and anatomy of Camellia section Camellia (Theaceae). Bot. J. Linn. Soc. 159, 456–476 (2009).

188. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 50, 278–284 (2018).

189. Li, Y. H. et al. Development of transgenic tea plants from leaf explants by the biolistic gun method and their evaluation. Plant Cell Tiss. Org. 254 (2004).

190. Xu, X. et al. Resequencing 50 accessions of cultivated and wild rice identifies markers for identifying agronomically important genes. Nat. Biotechnol. 30, 105–111, https://doi.org/10.1038/nbt.2050 (2011).

191. Mondal, T. K., Bhattacharya, A., Ahuja, P. S. & Chang, P. K. Transgenic tea (Camellia sinensis (L.) O. Kuntze cv. Kangra Jat) plants obtained by mediated transformation of somatic embryos. Plant Cell Rep. 20, 270–272 (2001).

192. Lopez, S. J., Kumar, R. R., Pius, P. K. & Muraleedharan, N. Tumefaciens-Mediated genetic transformation in tea (Camellia sinensis (L.) O. Kuntze). Plant Mol. Biol. Rep. 20, 201–202 (2004).

193. Mondal, T. K., Bhattacharya, A., Lacmikumaran, M. & Ahuja, P. S. Recent advances of tea (Camellia sinensis) biotechnology. Plant Cell Tiss. Org. 76, 867–876 (2012).

194. Sandal, I. et al. Mediated genetic transformation of tea leaf explants: effects of counteracting bactericidity of leaf polyphenols without loss of bacterial virulence. Plant Cell Rep. 26, 169–176 (2007).

195. Mohanpuria, P., Kumar, V., Ahuja, P. S. & Yadav, S. K. Mediated silencing of caffeine synthesis through root transformation in Camellia sinensis L. Mol. Biotechnol. 48, 235–243 (2011).

196. Saini, U. et al. Optimising parameters for biolistic gun-mediated genetic transformation of tea [Camellia sinensis (L.) O. Kuntze] using면 유전자 변이를 이용한 진화의 특성. Int. J. Mol. Sci. 17, https://doi.org/10.3390/ijms17071132 (2016).
221. Tang, Y. et al. Development of a CRISPR/Cas9 constructed for genome editing of caffeine synthase in Camellia sinensis. J. Tea Sci. 36, 414–426 (2016).
222. Lv, Q. et al. Optimization of tumefaciens-mediated transformation systems in tea plant (Camellia sinensis). Hortic. Plant J. 3, 105–109 (2017).
223. Alagarsamy, K., Shamala, L. F. & Wei, S. Protocol: high-efficiency in planta-mediated transgenic hairy root induction of Camellia sinensis var. sinensis. Plant Methods 14, 17 (2018).
224. Kumar, N., Pandey, S., Bhattacharya, A. & Ahuja, P. S. Do leaf surface characteristics affect infection in tea [Camellia sinensis (L.) O Kuntze]? J. Biosci 29, 309–317 (2004).
225. Chen, K., Wang, Y., Zhang, R., Zhang, H. & Gao, C. CRISPR/Cas genome editing and precision plant breeding in agriculture. Annu. Rev. Plant Biol. https://doi.org/10.1146/annurev-arplant-050718-100049 (2019).
226. Ye, M. et al. Generation of self-compatible diploid potato by knockout of S-RNase. Nat. Plants 4, 651–654 (2018).
227. Kottawa-Arachchi, J. D., Gunasekare, M. K. & Ranatunga, M. A. Biochemical diversity of global tea [Camellia sinensis (L.) O. Kuntze] germplasm and its exploitation: a review. Genet. Resour. Crop Evol. 66, 259–273 (2019).
228. Koech, R. K. et al. Functional annotation of putative QTL associated with black tea quality and drought tolerance traits. Sci. Rep. 9, 1465 (2019).
229. Koech, R. K. et al. Identification of novel QTL for black tea quality traits and drought tolerance in tea plants (Camellia sinensis). Tree Genet. Genomes 14, 9 (2018).
230. da Silva Pinto, M. Tea: a new perspective on health benefits. Food Res. Int. 53, 558–567 (2013).
231. Trevisanato, S. I. & Kim, Y. I. Tea and health. Nutr. Rev. 58, 1–10 (2000).
232. Suzuki, Y., Miyoshi, N. & Isemura, M. Health-promoting effects of green tea. Proc. Jpn. Acad. Ser. B 88, 88–101 (2012).
233. Chatuvedula, V. S. P. & Prakash, I. The aroma, taste, color and bioactive constituents of tea. J. Med. Plants Res. 5, 2110–2124 (2011).