Characterization of genome-wide genetic variation between two varieties of tea plant (Camellia sinensis) and development of InDel markers for genetics research

Shengrui Liu
Anhui Agricultural University
Yanlin An
Anhui Agricultural University
Wei Tong
Anhui Agricultural University
Xiuju Qin
Guangxi LuYI institute of tea tree species
Lidia Samarina
Russian Research Institute of Floriculture and Subtropical Crops
Rui Guo
Anhui Agricultural University
Xiaobo Xia
Anhui Agricultural University
Chaoling Wei (weichl@ahau.edu.cn)

Research article

Keywords: Molecular markers, Genetic diversity, SNP, InDel, Catechin/caffeine biosynthesis, Camellia sinensis

Posted Date: September 16th, 2019

DOI: https://doi.org/10.21203/rs.2.14507/v1

Version of Record: A version of this preprint was published on December 5th, 2019. See the published version at https://doi.org/10.1186/s12864-019-6347-0.

Abstract

Single nucleotide polymorphisms (SNPs) and Insertions/Deletions (InDels) are the major genetic variations and distributed extensively across the plant whole genome. Few investigations of these variations, however, have been performed in the long-lived perennial tea plant. Here, we have investigated the genome-wide genetic variation between Camellia sinensis var. sinensis 'Shuchazao' and Camellia sinensis var. assamica 'Yunkang 10', generating 7,511,731 SNPs and 255,218 InDels based on their whole genome sequences, and subsequently analyzed their distinct types and distribution patterns. A total of 48 InDel markers that yielded polymorphic and unambiguous fragments were developed when screening six tea cultivars. These markers were further deployed on forty-six tea cultivars for transferability and genetic diversity analysis, exhibiting informative with an average 4.02 of the number of alleles (Na) and 0.457 of polymorphism information content (PIC). The dendrogram showed that the phylogenetic relationships among these tea cultivars are highly consistent with their genetic backgrounds or original places. Interestingly, we observed that the content of catechin/caffeine between 'Shuchazao' and 'Yunkang 10' were significantly different, and a large number of SNPs/InDels were identified within catechin/caffeine biosynthesis-related genes. The identified genome-wide genetic variation and newly-developed InDel markers will provide a valuable resource for tea plant genetics and genomics studies, especially those SNPs/InDels within catechin/caffeine biosynthesis-related genes can be served as pivotal candidates for elucidating the molecular mechanism of catechin/caffeine biosynthesis.

Background

Tea is the most popular non-alcoholic beverage, which possesses numerous crucial properties including attractive aroma, pleasant taste, and helpful and medicinal benefits [1–3]. The tea plant (*Camellia sinensis* (L.) O. Kuntze) is a perennial evergreen woody plant (2n = 2x = 30) belonging to the section *Thea* of the genus *Camellia* in the family Theaceae [4, 5]. Evidence is accumulating that the tea plant was originated from Yunnan province in southwest China [4–7]. Currently, cultivated tea plant varieties mainly belong to two groups, *Camellia sinensis* var. *sinensis (*CSS)and *Camellia sinensis* var. *assamica (*CSA), are extensively cultivated in tropical and subtropical regions around the world [6, 8]. Generally, CSS is a slower-growing shrub with a relative higher cold-resistance capacity, while CSA is quick-growing with larger leaves and high sensitivity to cold climate [9]. With the successively release of two draft genome sequences, CSA 'Yunkang 10' [10] and CSS 'Shuchazao' [9], it is rapidly becoming another tractable experimental model for the genetics and functional genomics research on tea trees. It is known that self-incompatibility and long-term allogamy contributed tremendously to the highly heterogeneous and plentiful genetic variation of tea plant [11, 12]. Therefore, it is of great importance to characterize genome-wide genetic variation between the two subspecies.

Molecular markers, based on DNA polymorphisms, are useful and powerful tools for genetics and breeding researches. Numerous molecular markers have been successfully developed and applied in genetic and genomics research in tea plant, such as restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), random amplification of polymorphic DNAs (RAPDs), cleaved amplified polymorphic sequence (CAPS), inter-simple sequence repeat (ISSRs), and simple sequence repeat (SSRs) [12, 13]. With the rapid development of the high-throughput sequencing approaches, the third-generation single nucleotide polymorphism (SNP) and Insertion/Deletion (InDel) markers are gradually becoming the most widely used molecular markers, demonstrating a promising future in plant breeding and genetics researches.

SNPs are the most abundant genetic variations in most plant species, and the exploitation of SNP markers in single-copy regions is much easier than use of the other DNA markers [14–16]. InDel markers have practical value for those laboratories with limited resources, which also showed reliable transferability between distinct populations [14, 17, 18]. Both SNPs and InDels have been extensively applied for breeding programs and genetic researches including pedigree analysis, origin and evolutionary analysis, population structure and diversity analysis, linkage and QTL mapping, and marker-assisted selection [14, 19–22]. Several studies have also reported the development and application of SNP/InDel markers in tea plant genetics and genomics studies. For instance, 16 expressed sequence tag (EST)-SNP based CAPS markers were developed and applied for tea plant cultivar identification [23]; A set of SNPs from EST databases were identified and verified [24]; Fang et al. (2014) validated 60 EST-SNPs, and constructed genetic relationships among tea cultivars and their specific DNA fingerprinting [25]; Based on specific locus amplified fragment sequencing (SLAF-seq), a total of 6,042 SNP markers were validated and a final genetic map containing 6,448 markers were constructed [26]; Through restriction site-associated DNA sequencing (RAD-Seq) approach, Yang et al. (2016) identified a vast number of SNPs from 18 cultivated and wild tea accessions, and found that 13 genes containing non-synonymous SNPs exhibited strong selective signals suggesting artificial selective footprints during domestication of these tea accessions [27]. By harnessing the two reference genomes, it is now suitable for identifying genome-wide SNPs/InDels between them, further to guide rapid and efficient development of markers for high-resolution genetic analysis.

The whole genome sequences of tea trees can provide an elegant platform for identifying abundant genetic variation and developing a plenty of genetic markers. The completion of the two reference genome sequences is a great advance for genetics and genomics studies and a basis for this study. The tea plant whole genome CSA ‘Yunkang 10’ was firstly reported based on the Illumina next-generation sequencing platform, producing a 3.02 Gb genome assembly containing 37,618 scaffolds with an N50 length of 449 Kb [10]. Subsequently, the genome assembly of CSS ‘Shuchazao’ was released by combined Illumina and PacBio sequencing platforms, yielding a 3.14 Gb genome assembly consisting of 36,676 scaffolds with an N50 length of 1.39 Mb [9]. In this study, several principal objectives were completed: The genome-wide genetic variation and their distribution patterns were investigated; A number of polymorphic and stable InDel markers were developed, providing informative genetic markers for genetic and genomics studies; The contents of catechin and caffeine of the two tea cultivars were detected, and SNPs/InDels within catechin/caffeine biosynthesis-related genes were characterized. The identified genome-wide genetic variations and newly-developed InDel markers provided valuable resources for tea plant genetics and genomics studies, and the identification of SNPs/InDels within catechin/caffeine biosynthesis-related genes can serve as important candidate loci for functional analysis.

Results

Mapping of clean reads to the reference genome ‘Shuchazao’

An observation was that *Camellia sinensis* var. *sinensis* ‘Shuchazao’ has significant differences in size of bud, leaf and budding flower compared with in *Camellia sinensis* var. *assamica* ‘Yunkang 10’ (Fig. 1). The completion of the two reference genome sequences (‘Shuchazao’ and ‘Yunkang 10’) is a great advance for the comparative genomics studies on tea plant in *Thea* section. Therefore, the two genome assemblies were used for identification of genome-wide genetic variation between them. The CSS genome as the reference genome for its relative higher quality assemblies, and several steps were conducted for identifying genetic variation: aligning the clean reads of CSA to the reference using Bowtie2 software, removing PCR repeats with Picard program, calling SNPs and InDels using

Samtools and filtering the low-quality variations through GATK program, annotation for the remaining variations using snpEFF, and statistics of variants with Vcftools (Fig. 2A).

After filtering of raw data, a total of 324,154,064 clean reads from CSA whole genome sequencing data were generated; these reads had a coverage depth of 10.4X the 'Yunkang 10' genome, with 100 bp in length and 43% GC content (Fig. 2B). Through alignment, a total of 317,878,025 clean reads were mapped to the reference genome, accounting for 98.1% of total reads. The mapped clean reads contained two types sequencing reads: pair-end and single-end reads, the former was predominantly type (317,063,284, 99.7%) while single-end reads accounted for only 0.3% (814,741 clean reads) (Fig. 2B).

Identification and distribution of SNP and InDel loci

After a series of filtering, a total of 7,071,433 SNP loci were generated, with an average of SNP density in the tea genome was estimated to be 2,341 SNPs/Mb. Based on nucleotide substitutions, the detected SNPs were classified as transitions (Ts: G/A and C/T) and transversions (Tv: A/C, A/T, C/G, and G/T), which accounted for 77.46% (5,818,773) and 22.54% (1,692,958), respectively (Fig. 3A), with a Ts/Tv ratio of 3.44. In transitions, the number of A/G is equivalent to the C/T type, which included 2,905,203 and 2,913,570, respectively. As for transversions, the number of four types (A/C, A/T, C/G and G/T) are almost evenly distributed, with an insignificant difference among them, which accounted for 27.23% (460,988), 24.72% (418,536), 20.84% (352,802) and 27.21% (460,632), respectively (Fig. 3A).

A total of 255,218 InDels were identified, with an average density of 84.5 InDels/Mb. The length distribution of InDels was analyzed by dividing the lengths into different groups and calculating the ratios for the corresponding length groups (Fig. 3B). It is obviously that mononucleotide InDels are the most abundant types, accounting for 44.27% (112,976) of the total number. The length of InDels ranging from 1 to 20 bp was predominantly abundant, accounting for more than 95.5% (243,749) of the total InDels. An obvious tendency was that the number of InDels was gradually decreased with the increasing of InDel length.

Location and functional annotation of SNPs and InDels

The annotation of the 'Shuchazao' reference genome was used to uncover the distribution of SNPs and InDels within distinct genomic regions. According to the gene structure of the reference genome, the overwhelming number of SNPs (94%) was identified in intergenic regions, while only 6% (440,298) of SNPs were located in genic regions (Fig. 4A). Among the SNPs located in genic regions, 89,511 SNPs were detected in the CDs region, which contained 38,670 synonymous and 50,841 non-synonymous SNPs, respectively. Similarly, a small proportion of InDels were located in the genic regions, which accounted for only 12% (31,130) of the total number (Fig 4B). Remarkably, 3,406 InDels were located in the CDs region, which can be regarded as the preference for developing InDel markers.

To better understand the potential functions of these genetic variations within genes, GO term enrichment analysis of genes containing SNPs/InDels within CDs region was performed. These genes were classified into biological process, cellular component and molecular function categories (Fig. 5). As for the genes containing SNPs, the GO terms of cellular process, metabolic process and single-organism process were dominantly abundant in the biological process (Fig. 5A). In the cellular component category, the top three enriched GO terms were membrane,

cell and cell part. Based on the molecular function category, catalytic activity and binding are predominantly enriched, while others accounted for a small proportion (Fig. 5A). Interestingly, almost a consensus result was obtained for GO terms analysis of genes containing InDels, nothing but the number of genes is less compared with the number of genes containing SNPs.

Validation and polymorphism of newly-developed InDel markers

Initially, all InDels were used for designing primer pairs using Primer3.0. To validate the InDels and develop polymorphic InDel markers, we selected 100 InDel markers that were distributed on different scaffolds. To facilitate screening and developing of more practical markers, the length of all selected Indels were ranged from 5 to 20 bp in length. To identify the reliability and polymorphisms of the primers, six tea cultivars were selected for testing their amplified fragments using Fragment Analyzer^{TM} 96. Of the total primer sets tested, 48 primer pairs were successfully amplified with unambiguous bands and length polymorphisms among the six tea cultivars, 19 primer sets generated non-polymorphic or empty amplifications, and 33 primer pairs yielded non-specific amplification or ambiguous bands. Consequently, the 48 primer sets were regarded as elegant InDel markers and used for further analysis.

To test cross-cultivars/subspecies transferability, the 48 InDel markers were conducted on a panel of 46 tea cultivars belonging to section *Thea* of genus *Camellia.* The detailed information of these selected tea cultivars is listed (Table S1). The results of 18 InDel markers testing on various tea cultivars were given (Fig. 6), demonstrating that unambiguous and polymorphic bands were obtained based on these markers. The majority of InDel markers generated high polymorphism in the 46 tea cultivars. The amplified allele sizes across them were within the ranges detected in the donor tea cultivar, implying that the amplified fragments were derived from the same loci and the primer binding sites of the alleles were highly conserved among distinct tea cultivars/subspecies. Several crucial parameters for evaluating polymorphism of markers were subsequently conducted, such as the *Na* per locus ranged from 2 (CsInDel15, CsInDel16, CsInDel21, CsInDel24, CsInDel25, CsInDel33, CsInDel35, CsInDel39, CsInDel41, CsInDel46, and CsInDel47) to 14 (CsInDel38) with an average of 4.02 alleles, the MAF ranged from the lowest 0.266 (CsInDel20) to the highest at 0.957 (CsInDel41 and CsInDel47) with an average of 0.585, the *Ho* ranged from 0.021 (CsInDel24) to 1.000 (CsInDel15, CsInDel19, and CsInDel29) with an average of 0.524 and the *He* ranged from 0.082 (CsInDel41 and CsInDel47) to 0.869 with an average of 0.528, the PIC values were from the lowest value 0.078 (CsInDel41 and CsInDel47) to the highest 0.849 (CsInDel38) with an average of 0.457 (Table 1). It is noteworthy that the value of *He* has the similar variation trend with PIC value, while has a distinct variation trend with *Ho* values. The primer sequences and genomic location of these newly-developed markers are listed (Table S2). These results showed that these newly-developed InDel markers are informative and possess good transferability among various tea subspecies/cultivars.

Population structure and genetic relationship analysis

Population structure analysis was performed on the selected tea cultivars using Structure2.3.3 software based on 48 newly-developed InDel markers. The Q-plot output presented our grouping results, indicating that the two groups were the optimal classification at K = 2 (Fig. 7A). Apparently, tea cultivars from south and southwest of China (Guangxi, Guangdong, Yunnan and Sichuan provinces) belonging to *Camellia sinensis* var. *assamica* were

clustered tightly together. In comparison, the tea cultivars possessing smaller leaf size and shorter height that cultivated in several other provinces were classified into another group (Fig. 7B).

To further confirm the applicability of developed InDel markers for classification, we constructed a phylogenetic tree based on their genetic distances (Fig. 7C). It was demonstrated that two major branches were generated (designated as α and β groups), which contained 17 and 29 tea cultivars, respectively. Group α can be further divided into two subgroups, which were designated as α–1 and α–2 subgroups and consisted of 13 and 4 members, respectively. The dendrogram reflects that the phylogenetic relationships among them are highly consistent with their backgrounds or places of origin, as well as displays consistency with the results from population structure analysis although a little discrepancy was appeared (Fig. 7C).

Identification of genetic variation in catechin/caffeine biosynthesis-related genes

Tea cultivars belonging to *Camellia sinensis* var. *assamica* possess significant differences on phenotypes (plant height, size of leaf and flower) and major characteristic secondary metabolites (such as catechin and caffeine, which contributed tremendously to tea quality)compared with *Camellia sinensis* var. *sinensis*. Therefore, we detected the contents of catechin (flavan–3-ols) and caffeine in both 'Shuchazao' and 'Yunkang 10' based on HPLC analysis. It was found that the total content of catechin in both bud and the second leaf from 'Yunkang 10' is higher than from 'Shuchazao' (Fig. 8A). To understand the potential molecular mechanism of difference, we performed the catechin biosynthesis pathway based on several previous studies (Fig. 8B). By searching, we identified a number of SNPs and InDels in some crucial genes that involved in catechin biosynthesis pathway, including phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4-coumarate-CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H), flavonoid 3',5'-hydroxylase (F3'5'H), dihydroflavonol 4-reductase (DFR), leucoanthocyanidin reductase (LAR), anthocyanidin synthase (ANS), anthocyanidin reductase (ANR), 1-O-galloyl-β-D-glucose O-galloyltransferase (ECGT, which belongs to subclade 1A of serine carboxypeptidase-like (SCPL) acyltransferases) (Table 2).

Detection of caffeine content in the two tea varieties demonstrated that the caffeine in both bud and the second leaf from 'Yunkang 10' is lower than from 'Shuchazao' (Fig. 9a). In figure 9b, the well-studied caffeine biosynthesis pathway was also performed based on previous studies [10, 28–31]. Likewise, a number of genetic variations within some critical regulatory genes were also detected, such as in IMP dehydrogenase (IMPDH), Guanosine synthase (GMPS), 5'-nucleotidase (5'-Nase) and Tea caffeine synthase (TCS) genes (Fig. 9c and Table 2). Collectively, these results indicate that certain genetic variations within these genes may be the potential reason for the significant difference of catechin/caffeine synthesis between 'Shuchazao' and 'Yunkang 10'.

Discussion

Identification of genetic variations in tea plant whole genome

The recent release of the 'Shuchazao' and 'Yunkang 10' genome sequences will tremendous facilitate the efficiency of comparative genomics and functional research in tea plant. This enable researchers to study numerous agronomic traits associated with the perennial tea trees with a completely set of tools, including identification and development of SNP/InDel markers. Nevertheless, genome-wide identification and development of SNP/InDel

markers are still in infancy, especially those genetic variations related to important agronomical traits. By mapping the clean reads of 'Yunkang 10' to the reference genome assembly 'Shuchazao', we comprehensively surveyed DNA polymorphisms at the genome-wide scale and revealed the high level of genetic diversity between them. The vast number of SNPs and InDels identified in this study will provide valuable resources for tea plant genetics and breeding studies.

After filtering, a total of 7,071,433 SNPs and 255,218 InDels were identified, their densities distributed in the tea plant genome were estimated to be 2,341 SNPs/Mb and 84.5 InDels/Mb, respectively. The density of SNPs and InDels were less than in maize [32] and quinoa [19]. The density of identified InDels in tea plant genome is also less than in *Arabidopsis* [33], *Brassica rapa* [17] and rice [34]. Interestingly, the density of SNPs and InDels were higher than in foxtail millet [35] and soybean [36]. These significant differences of SNP/InDel density among different plant species may be due to the distinct filtering protocols and/or the different genomic composition. Actually, tea cultivars belonging to distinct subspecies are highly heterogeneous with broad genetic variation due to their self-incompatibility and long-term allogamy [11]. In terms of SNPs, our results showed that A/G and C/T transitions are the most common pattern of nucleotide substitution, which is consistency with the results obtained in other plant species, such in foxtail millet [35], citrus [37], and soybean [36]. As for InDels, the most prevalent types in the tea plant genome are short InDels. The number of 1–5 bp InDels are the predominant types, accounting for 76% of all the InDels, and similar results were displayed in several other plant species [14, 35–37].

Knowing genomic positions of genetic variations in genetic markers or functional genes is tremendously important. It was showed that only minimal SNPs and InDels were distributed in CDs region, which can be explained by the fact that CDs region only accounted for a small proportion of the whole genome sequences and have relative higher conservation compared with in other regions. Among the 89, 511 SNPs that located in CDs region, a total of 50,841 SNPs were non-synonymous variations. Non-synonymous variations can usually have several functional impacts due to an altered amino acid sequence, such as hampering of interaction between proteins and affecting the gene expression due to the functional consequences of distinct motif binding at variation sites [36, 38]. It was worth noting that a total of 3,406 identified InDels were located in CDs region. InDels are tend to have more impact on protein structure and function than single base changes, especially those in CDs region [36]. Nevertheless, genetic variations at UTRs may also play important roles, such as modify regulatory elements affecting the interaction of the UTRs with proteins and miRNAs [39]. Overall, these SNPs and InDels can be served as important candidates for functional research; especially those InDels in CDs, which can be considered as a valuable resource for developing phylogenetic and/or functional markers.

Development and application of InDel markers

Molecular markers are becoming indispensable tools for evolutionary analysis, germplasm identification and conservation, and marker-assisted selection (MAS). SSR is an extensively used marker type among genetic markers, and a large number of highly polymorphic SSR markers have been developed and applied in various genetics studies in tea plant [8, 13]. These SSR markers, however, could easily result in non-specific amplifications and cause confusion in genotyping scoring [19], especially for plant species with large genome and high repetitive sequences. Therefore, more stable markers should be developed and used for genetics and breeding researches, such as InDel markers. It is noteworthy that InDel markers can be assayed using the same separation and detection approaches as SSR markers (such as capillary DNA fragment analyzer and polyacrylamide gel electrophoresis), which will facilitate the widely application of InDel markers. Through a series of screening, we developed a final of

48 polymorphic and stable InDel markers with 5–20 bp in length based on the genomic assembled sequences (Table 1). The length of fragments of the alleles amplified across tea cultivars were consistent with the expected sizes of the products, implying that the primer binding sites of the alleles were highly conserved. The large proportion of InDel markers displayed a moderately PIC value (0.25<PIC<0.5), and the average of PIC was 0.4 of all markers. It is obviously that the PIC values of most InDel markers was lower than the PIC of the majority SSR markers [2, 8, 40, 41], supporting that the InDels markers are stable and bi-allelic throughout the genome. Therefore, these newly-developed InDel markers are suitable for germplasm identification and conservation, genetic diversity analysis, population structure and phylogenetic relationship analysis. Besides, InDels can affect gene functions by causing the gain or loss of a frameshift and/or a stop codon, it is therefore suitable for developing functional markers that might be particularly valuable for MAS [19, 42].

Population structure analysis and phylogenetic trees can reflect the genetic diversity, pedigree relationships, and the geographic distances among plant species and/or varieties [2, 16, 22]. They can also be used to evaluate the reliability of molecular markers. To test the reliability and practicability of the newly-developed InDel markers, population structure and phylogenetic relationship analysis were employed, and a consistent result was established (Fig. 7). Apparently, the tea cultivars from south and southwest of China were clustered together, which originated from *C. sinensis* var. *assamica* populations. In comparison, most of tea cultivars from central China had relatively close relationships with each other, which have distinct phenotypes including small leaf size and short height of tea trees. These results indicate that the population structure analysis and phylogenetic tree reflect the relationships of the 46 tea cultivars, demonstrating a high reliability of these InDel markers for genetic analysis.

Genetic variations within catechin/caffeine biosynthesis-related genes

Catechin and caffeine are one of the most important components in tea plant leaves, which enormously affect the quality of tea products and pharmacy [9, 43]. It is well-known that the contents of catechin and caffeine were influenced by genotypic factors, and significant difference can be observed among distinct tea varieties/cultivars [31, 44, 45].

Based on HPLC detection, we found that the total catechin content from ‘Yunkang 10’ was significantly higher than from ‘Shuchazao’ in both bud and the second leaf (Fig. 8a). Evidence has shown that the total catechin content of tea varieties tended to decline from the southern to the northern regions [44, 45], and our result is consistent with this tendency. Because catechins are important factors for the oxidation degree and dark tea was produced with severely fermentation during the processing [43, 45], our results supported the fact that most tea cultivars belonging to *Camellia sinensis* var. *assamica* are more suitable for producing dark tea. To understand the potential molecular mechanisms, genetic variations within key genes associated with catechin biosynthesis pathway were investigated between the two varieties. Unsurprisingly, a large number of SNPs and InDels were identified and some of them were located in CDs (Table 2). Combining the results of detection of catechin constitutes, it is probably to successfully select certain candidate genetic variations associated with the genotypic factors. For instance, a study reported that a number of candidate allelic variants relating to catechin traits at the F3’5’H locus were identified, the genetic effects of SNP840/848 were the most robust among them [43].

The result of HPLC detection showed that the caffeine content from ‘Yunkang 10’ was significantly lower than from ‘Shuchazao’ (Fig. 9a). Remarkably, a number of SNPs and InDels were found within some genes that associated

with caffeine biosynthesis pathway (Fig. 9c). Previously, a study reported that a 252 bp InDel mutation in the 5'-UTR of TCS1 plays a crucial role in caffeine biosynthesis [46]. Thus, our results can provide valuable candidates for identifying variations within genes relating to caffeine biosynthesis. Overall, these valuable resources can be used for further validation, such as functional characterization, association analysis, or development of functional markers for marker-assisted selection.

Conclusion

Comparison of the whole genome sequences between 'Yunkang 10' and 'Shuchazao' revealed that a large amount of genetic variations including SNPs and InDels, demonstrating that the tea plant genome is highly variable. The types of SNPs and InDels were subsequently investigated, and their distributions and annotation were also analyzed. Based on these InDel loci, a total of 48 novel InDel markers with moderate polymorphism and high stability were developed. Population structure and phylogenetic relationship analysis were conducted based on these markers, revealing that tea cultivars from *Camellia sinensis* var. *assamica* were apparently clustered together, while the other tea cultivars from *Camellia sinensis* var. *sinensis* were clustered into another group. Remarkably, significant differences were observed of catechin and caffeine content between 'Yunkang 10' and 'Shuchazao', and a number of SNPs and InDels were identified within genes relating to catechin/caffeine biosynthesis pathways.

Methods

Plant materials and DNA extraction

A total of forty-six clonal tea cultivars were collected from the main tea-growing regions in China, and we have acquired permission to collect all the tea samples. The details of these samples including cultivar name, subspecies, germplasm type, cultivation region are listed (Table S1). Two individuals ('Keke 1' and 'Keke 2') were collected from local natural population in Guangdong province with the local government's permission; three clonal tea cultivars ('Liubaoxiye', 'Lingyun 2' and 'Zihong') were collected from the Tea Germplasm Repository of the Tea Research Institute of Guangxi province with permission; the rest of forty-one clonal tea cultivars were commercial cultivars and cultivated widespread in China, which were deposited in the National Tea Germplasm Repository (N31°49', E117°13', Hefei, China) of our Institute (State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University). Until now, a total of 107 national tea cultivars (NTCs) and 139 provincial tea cultivars (PTCs) were registered in China [47]. In this study, 20 NTCs and 13 PTCs were used (the deposition numbers of NTCs are included in Table S1), and the rest of 13 local tea cultivars (LTCs) were registered by the corresponding provincial government, while subspecies type of four tea cultivars ('Keke 1', 'Keke 2', 'Ziyan' and 'Zixian') was still undetermined.

Young leaves of these tea cultivars were collected and immediately frozen in liquid nitrogen, and subsequently stored at −80 °C until further utilization. Total genomic DNA was extracted using the EZgene™ CP Plant Miniprep Kit (Biomiga, USA) following the manufacturer's protocol. The quality and quantity of DNA samples were determined by 1% agarose gels electrophoresis and the NanoDrop 2000 UV-Vis spectrophotometer, respectively. The concentration of each sample was adjusted to approximately 30 ng/ul for further use in the subsequently PCR amplifications.

Identification of SNPs and InDels by genome-wide comparison

Considering the quality of genome assemblies of 'Shuchazao' is better than the assemblies of 'Yunkang 10', it is reasonable to choose the assemblies of 'Shuchazao' as the reference genome. Thus, several steps were applied to identify genetic variation between the two assemblies: aligning the clean reads of 'Yunkang 10' to the reference using Bowtie2 software, removing PCR repeats with Picard program, calling SNPs and InDels using Samtools and filtering the low-quality variations through GATK program, annotation for the remaining variations using snpEFF, and statistics of variations with Vcftools (Fig. 2A). These several softwares have been accurately and expediently applied in SNP calling from next-generation sequencing data [48, 49].

Validation and development of InDels markers

To develop suitable InDel markers for genetic researches, the InDel length ≥5 and ≤20 bp were used as the candidate loci. Specific primers were designed based on the sequences flanking the InDel loci through the Primer 3.0 program with the following parameters: amplicons length (bp) 150–350; primer length 20–22, with optimum length being 20 bp; *Tm* (°C) 50–60, with 55 °C as the optimum; GC content (%) 40–60, with 50% as the optimum.

A total of 100 primer pairs were randomly selected and preliminary screened on six tea cultivars ('Guyuxiang', 'Longjing 43', 'Echa 5', 'Guilv', Yungui', and 'Fudingdabaicha') using the Fragment Analyzer^{TM} 96 (Advanced Analytical Technologies, Inc., Ames, IA). Primers that gave polymorphic and unambiguous bands were further screened for identification against the 46 tea cultivars. Details refer to PCR reagents and amplification conditions were performed according to our previous study [2].

Genetic diversity analysis

The PROSizeTM 2.0 included in the Fragment Analyzer^{TM} 96 system was applied to visually select strong and clearly polymorphic DNA fragments for scoring, with the same strategy as described previously [8]. The values of expected heterozygosity (*He)* and observed heterozygosity (*Ho)* were determined by Popgene 32 version software. The number of alleles (*Na),,* major allele frequency (MAF), and polymorphism information content (PIC) were calculated using PowerMarker 3.25 [50]. Based on the PIC value, markers were divided into three types: highly informative (PIC>0.5), moderately informative (0.25<PIC<0.5) and slightly informative (PIC<0.25) [19].

Population structure analysis

Genetic structure analysis of distinct tea accessions was performed using the Structure 2.3.4 program [51]. To minimize Hardy-Weinberg and linkage disequilibrium within each group, the model-based Bayesian clustering algorithm was employed to assign individuals to groups with a predetermined number (*K,* it represents the number of inferred populations). Ten independent runs for each *K* ranging from 2 to 9 were employed and 10,000 iterations were conducted for estimation after a 10,000 iterations burn-in period [19]. Estimation of the subgroups and the best *K* value was according to a previous study [52].

Phylogenetic analysis

Nei's genetic distances of the 46 tea cultivars based on 48 InDel markers was calculated using the PowerMarker 3.25. The dendrogram was constructed using the neighbor-joining (NJ) algorithm as implemented in MEGA 7.0

[53], with bootstrap values at the default setting of 1000 replicates. Pairwise gap deletion mode was employed to guarantee the divergent domains could contribute to the topology of the tree [54].

Detection of catechin content using HPLC

The content of catechin and caffeine were extracted and examined according to the previous studies [55]. All examples were detected with three independent biological replicates and each independent sample was examined with two technical replicates. The content of (+)-Gallocatechin (GC), (+)-Gallocatechin gallate (GCG), (-)-Epicatechin (EC), (-)-Epicatechin gallate (ECG), (-)-Epigallocatechin gallate (EGCG), and caffeine were detected. The catechin biosynthesis pathways were established according to previous studies [43, 56–59]. The number of SNP/InDel within the catechin/caffeine biosynthesis-related genes was also identified based on the result of alignment and functional annotation.

Abbreviations

SNPs, single nucleotide polymorphisms; InDels, Insertions/Deletions; RFLPs, restriction fragment length polymorphisms; AFLPs, amplified fragment length polymorphisms; RAPDs, random amplification of polymorphic DNAs; CAPS, cleaved amplified polymorphic sequence, ISSRs, inter-simple sequence repeats; SSRs, simple sequence repeat; EST, expressed sequence tag; SLAF-seq, specific locus amplified fragment sequencing; RAD-seq, restriction site-associated DNA sequencing; *He,* expected heterozygosity; *Ho,* observed heterozygosity; *Na,* number of alleles; MAF, major allele frequency; PIC, polymorphism information content.

Declarations

Acknowledgments

The authors thank the other members of our groups for technical assistance and appreciate the anonymous reviewers for constructive comments on this manuscript.

Funding

This work was financially supported by the Key R&D Program of China (2018YFD1000601), the Anhui Provincial Natural Science Foundation (1808085QC92), the China Postdoctoral Science Foundation (2017M621991), the Natural Science Foundation of Anhui Provincial Department of Education (KJ2018A0131), and the National Natural Science Foundation of China (31800585). The funding bodies had no role in the design of the study, collection, analysis, and interpretation of data, and in writing the manuscript.

Availability of data and materials

All the data and materials associated with the current study are available from the corresponding author on reasonable request.

Authors’ contributions

SRL performed data analysis and manuscript drafting. YLA conducted DNA extraction, primers design, PCR amplification, and InDel markers validation. WT were involved in identification and analysis of variation loci. XJQ and LS were involved in sample collection and data analysis. XBX and RG are involved in DNA extraction and PCR amplification. CLW conceived and designed research. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1.Yang CS, Wang X, Lu G, Picinich SC. Cancer prevention by tea: animal studies, molecular mechanisms and human relevance. Nat Rev Cancer.2009; 9(6):429–439.

2.Liu SR, Liu HW, Wu AL, Hou Y, An YL, Wei CL. Construction of fingerprinting for tea plant (Camellia sinensis) accessions using new genomic SSR markers. Mol Breed.2017; 37:93.

3.Zhang XC, Wu HH, Chen LM, Liu LL, Wan XC. Maintenance of mesophyll potassium and regulation of plasma membrane H+-ATPase are associated with physiological responses of tea plants to drought and subsequent rehydration. Crop J.2018; 6:611–620.

4.Hashimoto M, Takasi S. Morphological studies on the origin of the tea plant V, a proposal of one place of origin by cluster analysis. Jpn J Trop Agric.1978; 21:93–101.

5.Chen L, Yu FL, Tong QQ. Discussions on phylogenetic classification and evolution of section Thea. J Tea Sci.2000; 20:89–94.

6.Chang HT. Thea—a section of beveragial tea trees of the genus Camellia. Acta Sci Nat Univ Sunyats.1981; 1:87–99.

7.Yu FL. Discussion on the originating place and the originating center of tea plants. J Tea Sci.1986; 6(1).

8.Liu SR, An YL, Li FD, Li SJ, Liu LL, Zhou QY, Zhao SQ, Wei CL. Genome-wide identification of simple sequence repeats and development of polymorphic SSR markers for genetic studies in tea plant (Camellia sinensis). Mol Breed.2018; 38:59.

9.Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, Xia E, Lu Y, Tai Y, She G *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 U S A.2018; 115(18):E4151-E4158.

10.Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, Zhang Y, Liu Y, Zhu T, Li W *et al.* The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. Mol Plant.2017; 10(6):866–

877.

11.Chen L, Gao QK, Chen DM, Xu CJ. The use of RAPD markers for detecting genetic diversity, relationship and molecular identification of Chinese elite tea genetic resources [Camellia sinensis (L.)O. Kuntze] preserved in tea germplasm repository. Biodiv Conserv.2005; 14(6):1433–1444.

12.Ni S, Yao MZ, Chen L, Zhao LP, Wang XC. Germplasm and breeding research of tea plant based on DNA marker approaches. Front Agric China.2008; 2(2):200–207.

13.Mukhopadhyay M, Mondal TK, Chand PK. Biotechnological advances in tea (Camellia sinensis [L.] O. Kuntze): a review. Plant Cell Rep.2016; 35:255–287.

14.Hu YY, Mao BG, Peng Y, Sun YD, Pan YL, Xia YM, Sheng XB, Li YK, Tang L, Yuan LP *et al.* Deep re-sequencing of a widely used maintainer line of hybrid rice for discovery of DNA polymorphisms and evaluation of genetic diversity. Mol Genet Genomics.2014; 289(3):303–315.

15.Garrido-Cardenas JA, Mesa-Valle C, Manzano-Agugliaro F. Trends in plant research using molecular markers. Planta.2018; 247(3):543–557.

16.Villano C, Esposito S, Carucci F, Iorizzo M, Frusciante L, Carputo D, Aversano R. High-throughput genotyping in onion reveals structure of genetic diversity and informative SNPs useful for molecular breeding. Mol Breed.2019; 39:5.

17.Liu B, Wang Y, Zhai W, Deng J, Wang H, Cui Y, Cheng F, Wang XW, Wu J. Development of InDel markers for Brassica rapa based on whole-genome re-sequencing. Theor Appl Genet.2012; 126:231–239.

18.Thakur O, Randhawa GS. Identification and characterization of SSR, SNP and InDel molecular markers from RNA-Seq data of guar (Cyamopsis tetragonoloba, L. Taub.) roots. BMC Genomics.2018; 19(1):951.

19.Zhang TF, Gu MF, Liu YH, Lv YD, Zhou L, Lu HY, Liang SQ, Bao HB, Zhao H. Development of novel InDel markers and genetic diversity in Chenopodium quinoa through whole-genome re-sequencing. BMC Genomics.2017; 18:685.

20.Belaj A, de la Rosa R, Lorite IJ, Mariotti R, Cultrera NGM, Beuzón CR, González-Plaza JJ, Muñoz-Mérida A, Trelles O, Baldoni L. Usefulness of a New Large Set of High Throughput EST-SNP Markers as a Tool for Olive Germplasm Collection Management. Front Plant Sci.2018; 9:1320.

21.Zhang N, Zhang H, Ren Y, Chen L, Zhang J, Zhang L. Genetic analysis and gene mapping of the orange flower trait in Chinese cabbage (Brassica rapa L.). Mol Breed.2019; 39:76.

22.Sarkar D, Kundu A, Das D, Chakraborty A, Mandal NA, Satya P, Karmakar PG, Kar CS, Mitra J, Singh NK. Resolving population structure and genetic differentiation associated with RAD-SNP loci under selection in tossa jute (Corchorus olitorius L.). Mol Genet Genomics.2019; 294:479–492.

23.Ujihara T, Taniguchi F, Tanaka J, Hayashi N. Development of Expressed Sequence Tag (EST)-based Cleaved Amplified Polymorphic Sequence (CAPS) markers of tea plant and their application to cultivar identification. J Agric Food Chem.2011; 59:1557–1564.

24.Zhang CC, Wang LY, Wei K, Cheng H. Development and characterization of single nucleotide polymorphism markers in Camellia sinensis (Theaceae). Genet Mol Res.2014; 13(3):5822–5831.

25.Fang WP, Meinhardt LW, Tan HW, Zhou L, Mischke S, Zhang D. Varietal identification of tea (Camellia sinensis) using nanofluidic array of single nucleotide polymorphism (SNP) markers. Hortic Res.2014; 1:14035.

26.Ma JQ, Huang L, Ma CL, Jin JQ, Li CF, Wang RK, Zheng HK, Yao MZ, Chen L. 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.2015; 10(6):e0128798.

27.Yang H, Wei C-L, Liu H-W, Wu J-L, Li Z-G, Zhang L, Jian J-B, Li Y-Y, Tai Y-L, Zhang J *et al.* Genetic Divergence between Camellia sinensis and Its Wild Relatives Revealed via Genome-Wide SNPs from RAD Sequencing. Plos One.2016; 11(3):e0151424.

28.Deng W-W, Han J, Fan Y, Tai Y, Zhu B, Lu M, Wang R, Wan X, Zhang Z-Z. Uncovering tea-specific secondary metabolism using transcriptomic and metabolomic analyses in grafts of Camellia sinensis and C. oleifera. Tree Genet Genomes.2018; 14:23.

29.Guo Y, Zhu C, Zhao S, Zhang S, Wang W, Fu H, Li X, Zhou C, Chen L, Lin Y *et al.* De novo transcriptome and phytochemical analyses reveal differentially expressed genes and characteristic secondary metabolites in the original oolong tea (Camellia sinensis) cultivar 'Tieguanyin' compared with cultivar 'Benshan'. BMC Genomics.2019; 20(1):265.

30.Han J, Lu M, Zhu B, Wang R, Wan X, Deng W-W, Zhang Z-Z. Integrated transcriptomic and phytochemical analyses provide insights into characteristic metabolites variation in leaves of 1-year-old grafted tea (Camellia sinensis). Tree Genet Genomes.2019; 15:58.

31.Zhu B, Chen LB, Lu M, Zhang J, Han J, Deng WW, Zhang ZZ. Caffeine Content and Related Gene Expression: Novel Insight into Caffeine Metabolism in Camellia Plants Containing Low, Normal, and High Caffeine Concentrations. J Agric Food Chem.2019; 67(12):3400–3411.

32.Vroh Bi I, McMullen MD, Sanchez-Villeda H, Schroeder S, Gardiner J, Polacco M, Soderlund C, Wing R, Fang Z, Coe EH. Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci.2006; 46(1):12–21.

33.Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. Arabidopsis map-based cloning in the post-genome era. Plant Physiol.2002; 129(2):440–450.

34.Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X *et al.* Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol.2004; 135(3):1198–1205.

35.Bai H, Cao Y, Quan J, Dong L, Li Z, Zhu Y, Zhu L, Dong Z, Li D. Identifying the genome-wide sequence variations and developing new molecular markers for genetics research by re-sequencing a Landrace cultivar of foxtail millet. PLoS One.2013; 8(9):e73514.

36.Ramakrishna G, Kaur P, Nigam D, Chaduvula PK, Yadav S, Talukdar A, Singh NK, Gaikwad K. Genome-wide identification and characterization of InDels and SNPs in Glycine max and Glycine soja for contrasting seed permeability traits. BMC Plant Biol.2018; 18(1):141.

37.Zhang JZ, Liu SR, Hu CG. Identifying the genome-wide genetic variation between precocious trifoliate orange and its wild type and developing new markers for genetics research. DNA Res.2016; 23(4):403–414.

38.García-Lor A, Luro F, Navarro L, Ollitrault P. Comparative use of InDel and SSR markers in deciphering the interspecific structure of cultivated citrus genetic diversity: a perspective for genetic association studies. Mol Genet Genomics.2012; 287(1):77–94.

39.Steri M, Idda ML, Whalen MB, Orru V. Genetic variants in mRNA untranslated regions. Wiley Interdiscip Rev RNA.2018; 9(4):e1474.

40.Yao MZ, Ma CL, Qiao TT, Jin JQ, Chen L. Diversity distribution and population structure of tea germplasms in China revealed by EST-SSR markers. Tree Genet Genomes.2012; 8:205–220.

41.Tan LQ, Peng M, Xu LY, Wang LY, Chen SX, Zou Y, Qi GN, Cheng H. Fingerprinting 128 Chinese clonal tea cultivars using SSR markers provides new insights into their pedigree relationships. Tree Genet Genomes.2015; 11:90.

42.Liu TJ, Li YP, Zhou JJ, Hu CG, Zhang JZ. Genome-wide genetic variation and comparison of fruit-associated traits between kumquat (Citrus japonica) and Clementine mandarin (Citrus clementina). Plant Mol Biol.2018; 96(4–5):493–507.

43.Jin JQ, Ma JQ, Yao MZ, Ma CL, Chen L. Functional natural allelic variants of flavonoid 3',5'-hydroxylase gene governing catechin traits in tea plant and its relatives. Planta.2017; 245(3):523–538.

44.Chen L, Zhou ZX. Variations of main quality components of tea genetic resources [Camellia sinensis (L.) O. Kuntze] preserved in the China national germplasm tea repository. Plant Foods Hum Nutr 2005; 60:31–35.

45.Jin JQ, Ma JQ, Ma CL, Yao MZ, Chen L. Determination of catechin content in representative Chinese tea germplasms. J Agric Food Chem.2014; 62:9436–9441.

46.Jin JQ, Yao MZ, Ma CL, Ma JQ, Chen L. Natural allelic variations of TCS1 play a crucial role in caffeine biosynthesis of tea plant and its related species. Plant Physiol Biochem.2016; 100:18–26.

47.Yang YJ, Liang YR. Clonal tea cultivars in China. Shanghai Scientific and Technical Publishers.2014; Shanghai.

48.Wright B, Farquharson KA, McLennan EA, Belov K, Hogg CJ, Grueber CE. From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. BMC Genomics.2019; 20:453.

49.Zhao Y, Wang K, Wang WL, Yin TT, Dong WQ, Xu CJ. A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics.2019; 20:160.

50.Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics.2005; 21:2128–2129.

51.Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics.2000; 155(2):945–959.

52.Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol.2005; 14(8):2611–2620.

53.Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016; 33:1870–1874.

54.Liu SR, Khan MRG, Li YP, Zhang JZ, Hu CG. Comprehensive analysis of CCCH-type zinc finger gene family in citrus (Clementine mandarin) by genome-wide characterization. Mol Genet Genomics.2014; 289:855–872.

55.Liu S, Mi X, Zhang R, An Y, Zhou Q, Yang T, Xia X, Guo R, Wang X, Wei C. Integrated analysis of miRNAs and their targets reveals that miR319c/TCP2 regulates apical bud burst in tea plant (Camellia sinensis). Planta.2019.

56.Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T *et al.* Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics.2011; 12:131.

57.Wu ZJ, Li XH, Liu ZW, Xu ZS, Zhuang J. De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in Camellia sinensis. BMC Plant Biol.2014; 14:277.

58.Li CF, Zhu Y, Yu Y, Zhao QY, Wang SJ, Wang XC, Yao MZ, Luo D, Li X, Chen L *et al.* Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (Camellia sinensis). BMC Genomics.2015; 16:560.

59.Wang YS, Xu YJ, Gao LP, Yu O, Wang XZ, He XJ, Jiang XL, Liu YJ, Xia T. Functional analysis of Flavonoid 3',5'-hydroxylase from Tea plant (Camellia sinensis): critical role in the accumulation of catechins. BMC Plant Biol.2014; 14:347.

Tables

Table 1 Characteristics of 48 newly-developed InDel markers

Marker ID	Scaffold location	Fragment size (bp)	*Na*	MAF	*Ho*	*He*	PIC
CsInDel01	Scaffold 5: 236696	139-156	3	0.787	0.383	0.361	0.327
CsInDel02	Scaffold 5: 1208833	186-205	4	0.489	1.000	0.633	0.555
CsInDel03	Scaffold 12: 195263	332-354	3	0.500	0.489	0.577	0.478
CsInDel04	Scaffold 30: 3820588	214-242	5	0.532	0.532	0.636	0.576
CsInDel05	Scaffold 39: 128636	236-264	4	0.479	0.979	0.556	0.448
CsInDel06	Scaffold 41: 2074123	280-295	3	0.808	0.180	0.319	0.273
CsInDel07	Scaffold 46: 249178	176-189	3	0.734	0.362	0.405	0.336
CsInDel08	Scaffold 51: 314982	206-215	6	0.394	0.638	0.691	0.627
CsInDel09	Scaffold 51: 760768	201-248	7	0.532	0.660	0.679	0.645
CsInDel10	Scaffold 52: 469482	288-306	3	0.745	0.255	0.394	0.329
CsInDel11	Scaffold 60: 843530	292-332	6	0.383	0.213	0.748	0.701
CsInDel12	Scaffold 60: 843632	240-275	5	0.426	0.660	0.704	0.645
CsInDel13	Scaffold 64: 151635	270-289	3	0.404	0.617	0.643	0.559
CsInDel14	Scaffold 66: 500052	203-232	4	0.436	0.064	0.621	0.535
CsInDel15	Scaffold 77: 505984	185-207	2	0.500	1.000	0.505	0.375
CsInDel16	Scaffold 89: 1202911	231-248	2	0.819	0.149	0.300	0.252
CsInDel17	Scaffold 98: 664107	306-354	6	0.395	0.256	0.731	0.677
CsInDel18	Scaffold 114: 416691	283-326	6	0.489	0.809	0.703	0.661
CsInDel19	Scaffold 129: 540746	180-214	6	0.422	1.000	0.652	0.579
CsInDel20	Scaffold 154: 767901	285-297	5	0.266	0.979	0.763	0.709
CsInDel21	Scaffold 225: 80286	191-204	2	0.649	0.362	0.461	0.352
CsInDel22	Scaffold 1000: 52494	216-288	3	0.532	0.404	0.612	0.537
CsInDel23	Scaffold 1001: 123324	236-326	6	0.628	0.489	0.568	0.526
CsInDel24	Scaffold 1001: 149678	190-199	2	0.798	0.021	0.326	0.271
CsInDel25	Scaffold 1001: 155681	195-218	2	0.649	0.319	0.461	0.352
CsInDel26	Scaffold 1001: 1251845	341-363	3	0.583	0.833	0.511	0.399
CsInDel27	Scaffold 1001: 1261469	273-290	3	0.777	0.064	0.359	0.306
CsInDel28	Scaffold 1001: 1400899	213-253	6	0.660	0.383	0.537	0.501
CsInDel29	Scaffold 1001: 1491192	182-226	4	0.457	1.000	0.586	0.489
CsInDel30	Scaffold 1001: 1691928	238-258	4	0.745	0.362	0.411	0.363
CsInDel31	Scaffold 1001: 1982826	284-316	4	0.489	0.915	0.619	0.539
CsInDel32	Scaffold 1452: 285463	272-299	3	0.596	0.426	0.511	0.406
CsInDel33	Scaffold 1539: 196438	271-280	2	0.798	0.404	0.326	0.271
CsInDel34	Scaffold 1541: 138532	265-286	3	0.564	0.851	0.523	0.413
CsInDel35	Scaffold 1543: 253456	172-207	2	0.915	0.128	0.157	0.144
CsInDel36	Scaffold 1551: 196819	157-237	3	0.606	0.745	0.499	0.391
CsInDel37	Scaffold 1553: 529121	211-237	4	0.564	0.511	0.547	0.451
CsInDel38	Scaffold 1555: 5209	109-340	14	0.298	0.489	0.869	0.849
CsInDel39	Scaffold 1579: 1466247	261-272	2	0.606	0.787	0.483	0.363
CsInDel40	Scaffold 1592: 672899	276-329	7	0.596	0.979	0.666	0.489
CsInDel41	Scaffold 1593: 1022219	172-187	2	0.957	0.085	0.082	0.078
CsInDel42	Scaffold 1594: 195199	184-206	3	0.691	0.426	0.454	0.380
CsInDel43	Scaffold 1611: 1270988	226-254	5	0.426	0.319	0.684	0.619
CsInDel44	Scaffold 2220: 166816	292-328	3	0.543	0.575	0.521	0.402
CsInDel45	Scaffold 15285: 211487	281-321	5	0.333	0.952	0.752	0.699
CsInDel46	Scaffold 15433: 302840	190-253	2	0.638	0.468	0.467	0.355
CsInDel47	Scaffold 15579: 267174	176-186	2	0.957	0.043	0.082	0.078
CsInDel48	Scaffold 15650: 137667	228-266	6	0.489	0.596	0.671	0.614
Average	-	-	4.02	0.585	0.524	0.528	0.457

Note: *Na*, the number of alleles; MAF, major allele frequency; *Ho*, observed heterozygosity; *He*, expected heterozygosity; PIC, polymorphism information content.

Table 2 Statistics on SNPs and InDels within catechin biosynthesis-related genes

Gene name	Gene ID	SNP		InDel		Gene name	Gene ID	SNP		InDel	
		DNA	CDs	DNA	CDs			DNA	CDs	DNA	CDs
PAL	TEA014056.1	2	2	0	0	F3H	TEA004906.1	0	0	1	0
	TEA034008.1	6	6	0	0		TEA010326.1	1	1	0	0
	TEA003137.1	16	16	0	0		TEA032907.1	3	3	1	1
	TEA023243.1	3	3	0	0		TEA028622.1	75	1	3	0
	TEA024587.1	3	3	0	0		TEA009737.1	4	4	0	0
	TEA003374.1	2	2	1	0		TEA000753.1	1	1	0	0
C4H	TEA034001.1	16	8	1	1		TEA023937.1	1	1	0	0
	TEA016772.1	5	1	1	0		TEA016601.1	4	2	1	0
	TEA034002.1	6	6	0	0		TEA023790.1	10	3	1	0
4CL	TEA018887.1	1	1	0	0		TEA000474.1	8	1	0	0
	TEA034012.1	9	4	1	1		TEA026443.1	1	1	0	0
	TEA019275.1	14	10	0	0		TEA004898.1	1	1	0	0
	TEA027829.1	12	3	1	0		TEA006643.1	15	15	0	0
	TEA025906.1	2	1	0	0		TEA014951.1	29	8	2	0
	TEA009431.1	42	10	4	2	DFR	TEA032730.1	2	0	1	0
	TEA018045.1	22	3	4	0		TEA023829.1	13	1	0	0
	TEA006577.1	6	1	0	0		TEA021807.1	2	0	0	0
	TEA031627.1	11	8	0	0		TEA021815.1	2	2	0	0
	TEA022274.1	2	1	0	0	ANS	TEA010322.1	1	1	0	0
	TEA010681.1	8	4	0	0		TEA015762.1	1	1	0	0
	TEA002100.1	13	0	1	0		TEA015769.1	1	0	0	0
CHS	TEA018665.1	1	1	0	0	ANR	TEA030023.1	1	1	0	0
	TEA034046.1	34	10	0	0		TEA022960.1	6	2	0	0
	TEA034011.1	6	4	0	0		TEA007646.1	1	0	1	0
	TEA034045.1	1	1	0	0		TEA003247.1	1	1	0	0
	TEA023331.1	2	2	0	0	LAR	TEA021535.1	1	1	0	0
	TEA023340.1	3	3	2	0		TEA027582.1	0	0	2	0
	TEA034013.1	2	2	0	0		TEA009266.1	3	3	1	0
	TEA034043.1	31	7	0	0	SCPLA1	TEA034031.1	4	2	0	0
	TEA034019.1	3	3	0	0		TEA034032.1	11	5	0	0
	TEA034014.1	1	1	0	0		TEA010715.1	6	5	0	0
	TEA011908.1	6	1	0	0		TEA034056.1	33	1	0	0
	TEA019029.1	4	4	0	0		TEA009664.1	4	0	0	0
CHI	TEA034003.1	10	2	1	0		TEA016469.1	2	0	0	0
	TEA033023.1	127	4	10	0		TEA016463.1	9	1	0	0
	TEA033031.1	2	1	0	0		TEA034055.1	59	1	0	0
F3’H	TEA016718.1	2	2	0	0		TEA034034.1	4	0	0	0
	TEA010133.1	5	2	0	0		TEA034036.1	1	1	0	0
	TEA006847.1	14	10	1	1		TEA023444.1	3	0	0	0
F3’5’H	TEA013315.1	12	12	0	0		TEA034039.1	31	2	0	0
	TEA034021.1	6	1	0	0		TEA023451.1	4	1	0	0
	TEA034051.1	32	4	4	0		TEA000223.1	4	0	0	0

Figures

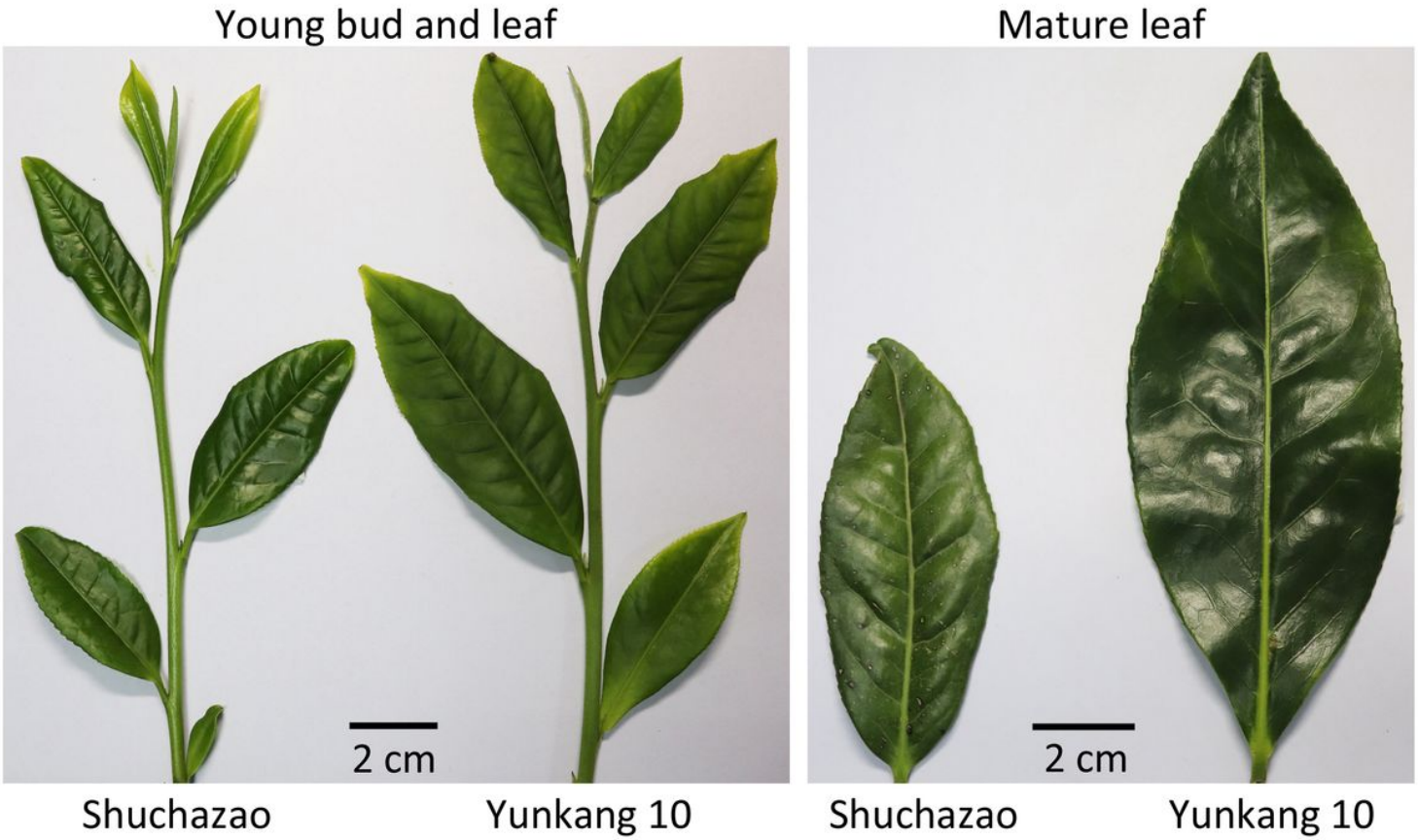


Figure 1

Comparison of bud and leaf size between ‘Shuchazao’ and ‘Yunkang 10’. Young bud and leaf were collected on April 2019, while mature leaves were collected from branches of last-year autumn.

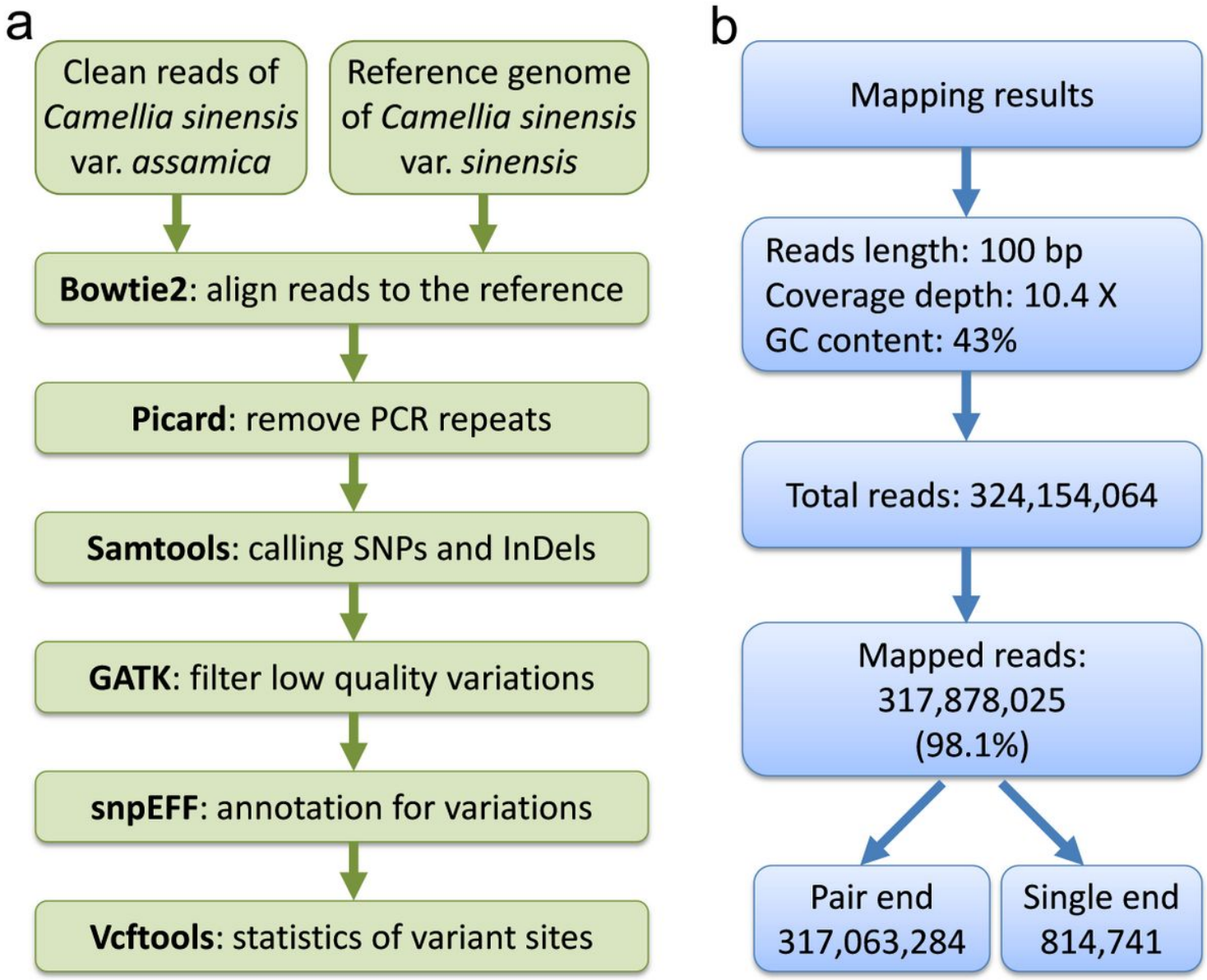


Figure 2

Identification of genome-wide genetic variations and mapping results. a Flowchart diagram for identifying genome-wide genetic variations between 'Shuchazao' and 'Yunkang 10'. b The obtained results by mapping clean reads to the reference genome sequences.

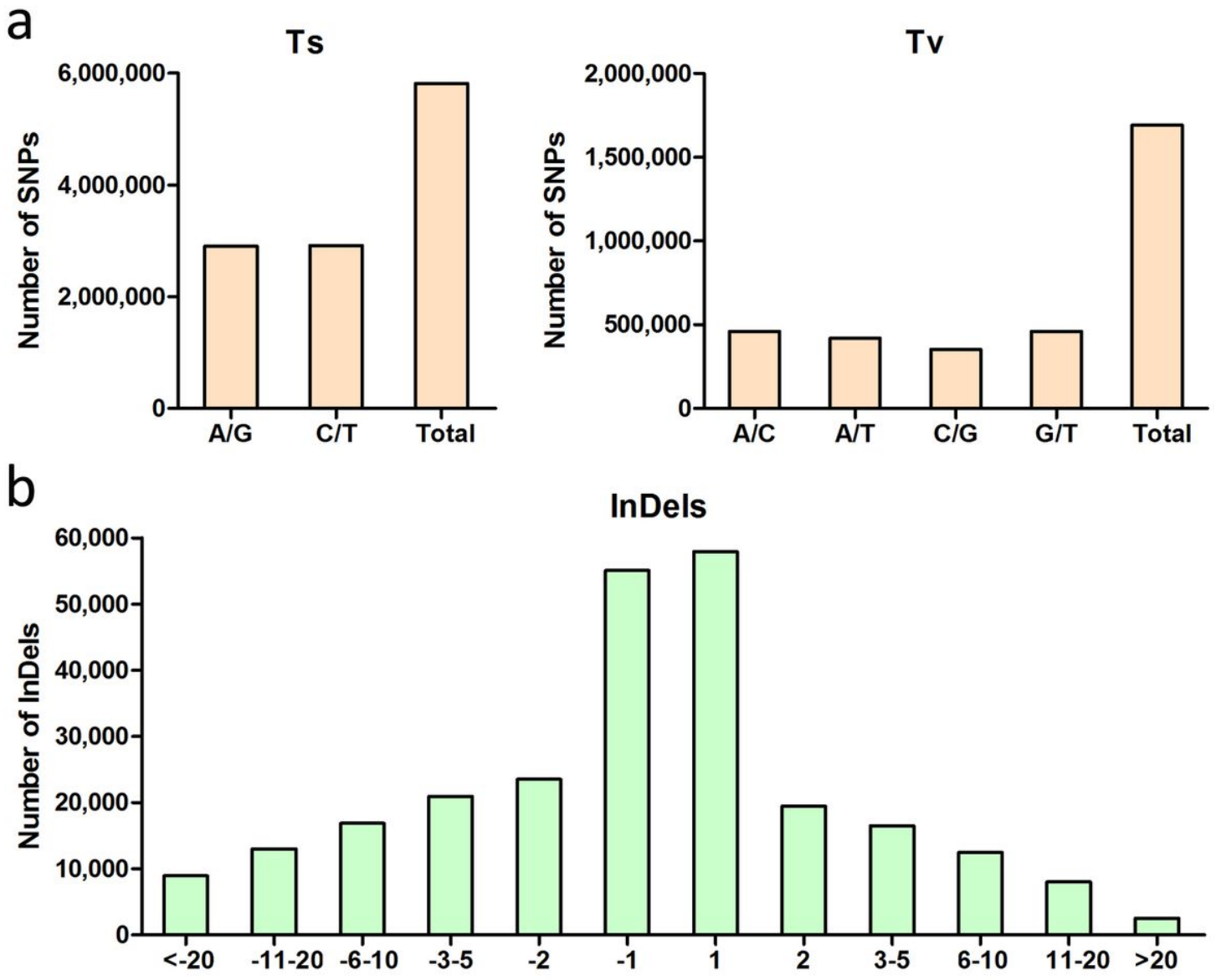


Figure 3

Classification and distribution of identified SNPs/InDels in 'Yunkang 10'/ 'Shuchazao' comparison. a Frequency of different substitution types in the identified SNPs; the x-axis and y-axis represent the types and number of SNPs, respectively. b Distribution of the length of InDels identified between the two tea cultivars; the x-axis shows the number of nucleotides of InDels, the y-axis represents the number of InDels at each length.

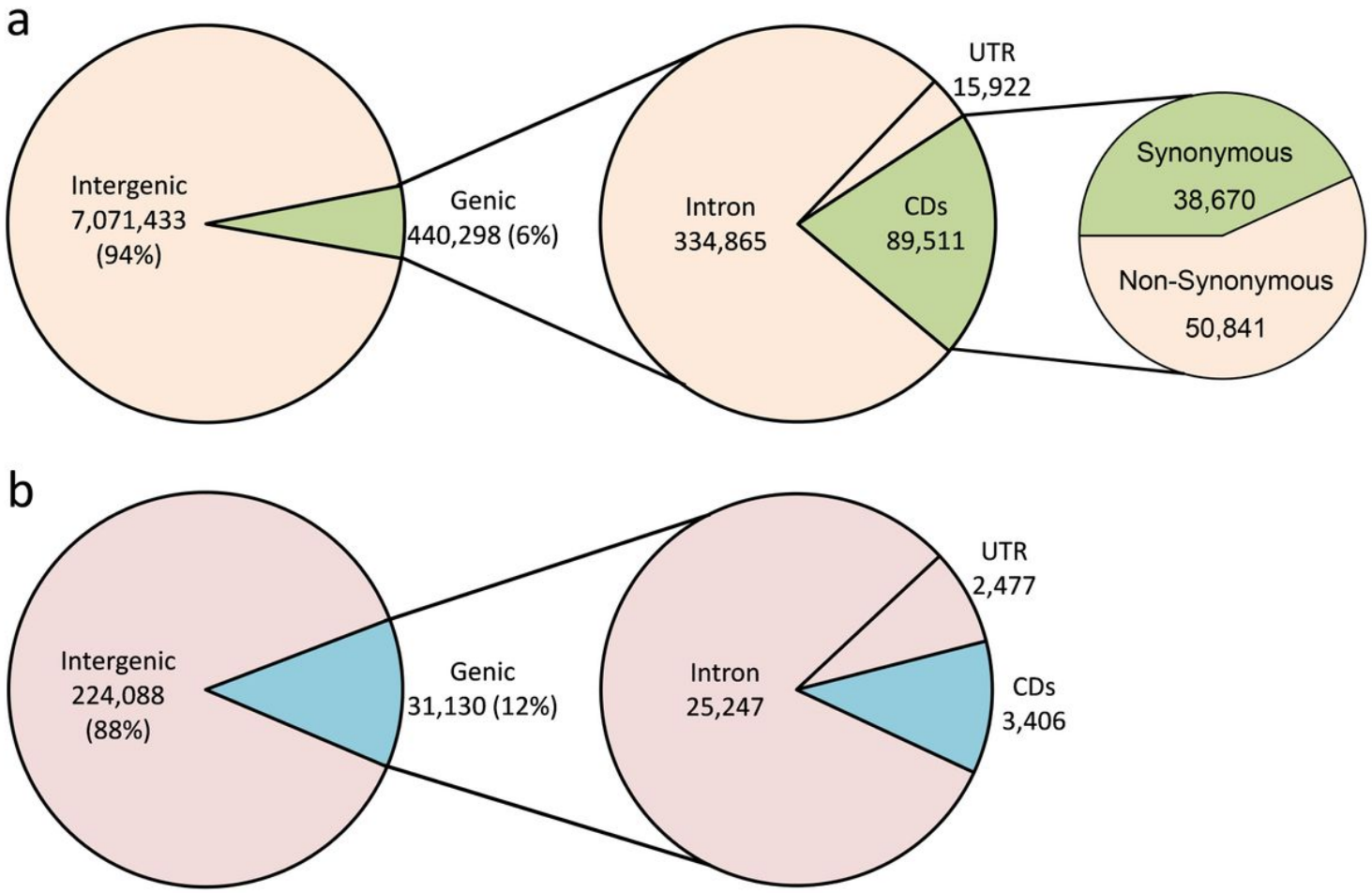


Figure 4

Annotation of SNPs and InDels identified between ‘Shuchazao’ and ‘Yunkang 10’. a Annotation of SNPs. b Annotation of InDels. SNPs and InDels were classified as intergenic and genic on the ‘Shuchazao’ reference genome, and locations within the gene models were annotated.

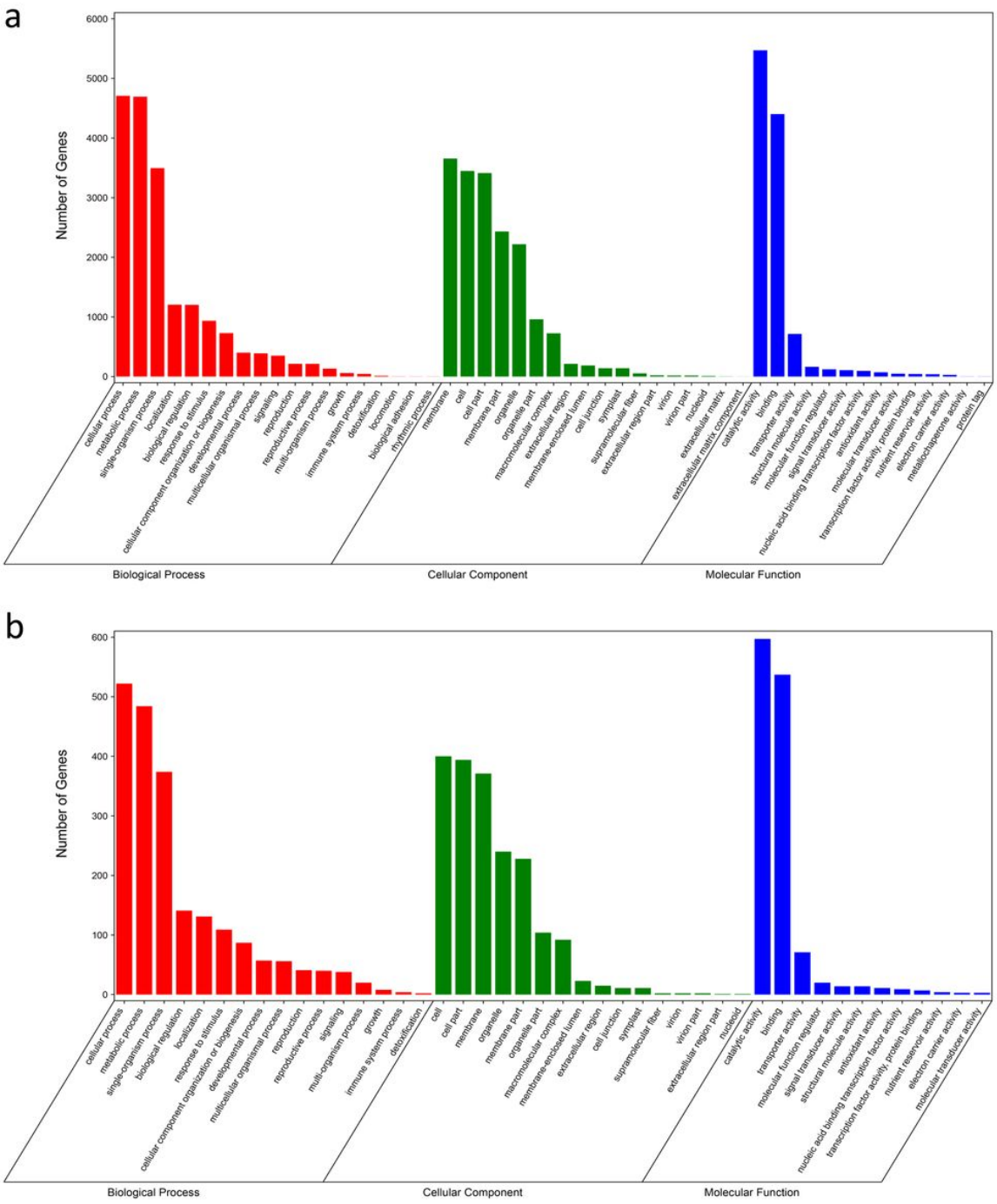


Figure 5

Functional categorization of the genes containing genetic variations within CDs region. a Functional annotation of genes containing SNPs within in CDs region. b Functional annotation of genes containing InDels within in CDs region. These genes were categorized based on GO annotation, and the number of each category is showed based on biological process, cellular component and molecular function.

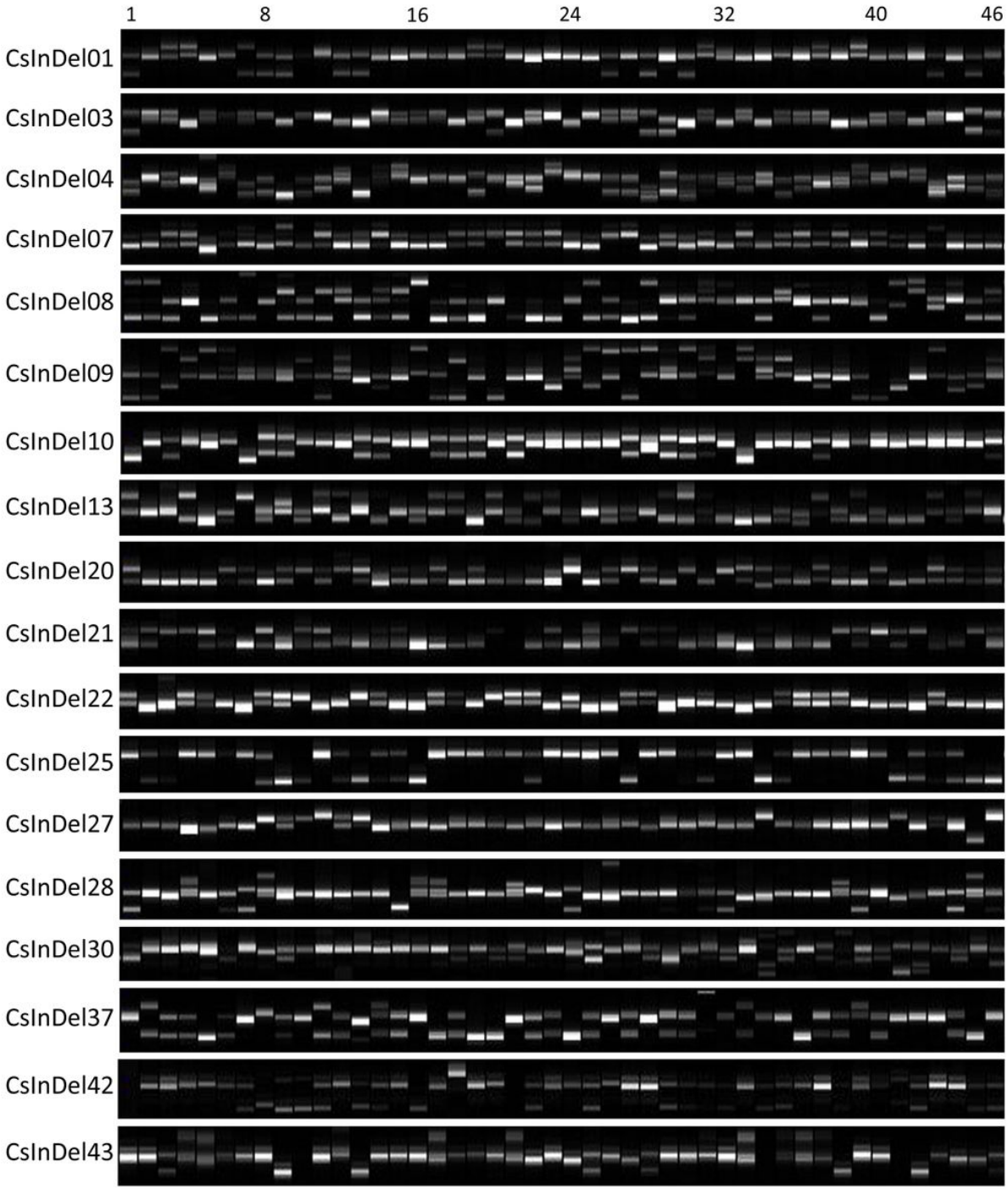


Figure 6

Exhibition of transferability and polymorphism detected by 18 out of 48 InDel markers among forty-six tea cultivars.

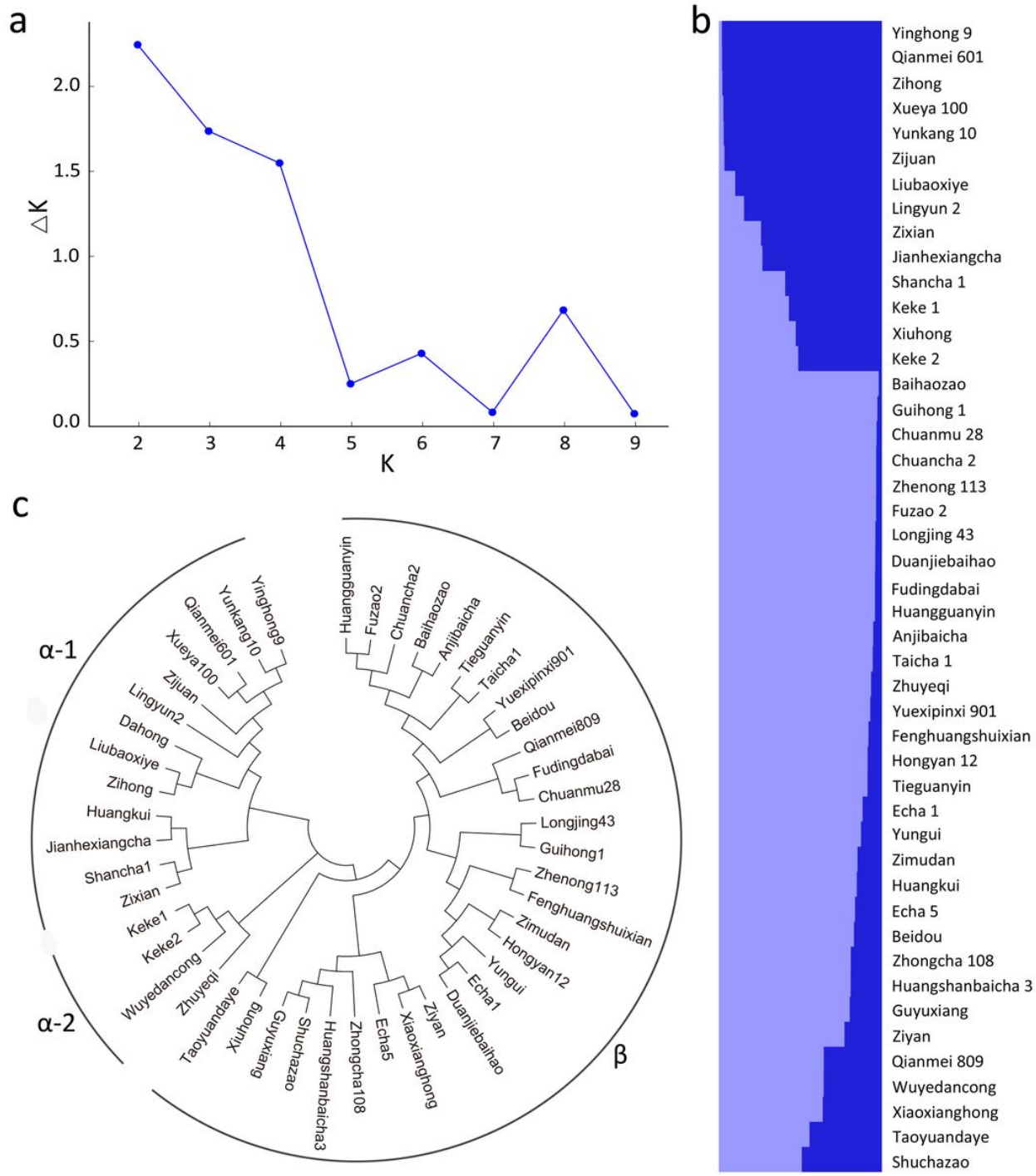


Figure 7

Population structure and phylogenetic relationship analysis based on 48 InDel markers. a Estimation of the optimal group number through ΔK, the number of K was set from 2 to 9. b Q-plot of the population structure when K=2. Each tea cultivar is represented by a horizontal bar. C The dendrogram was constructed based on genotypes using neighbor-joining algorithm with 1000 bootstrap replicates.

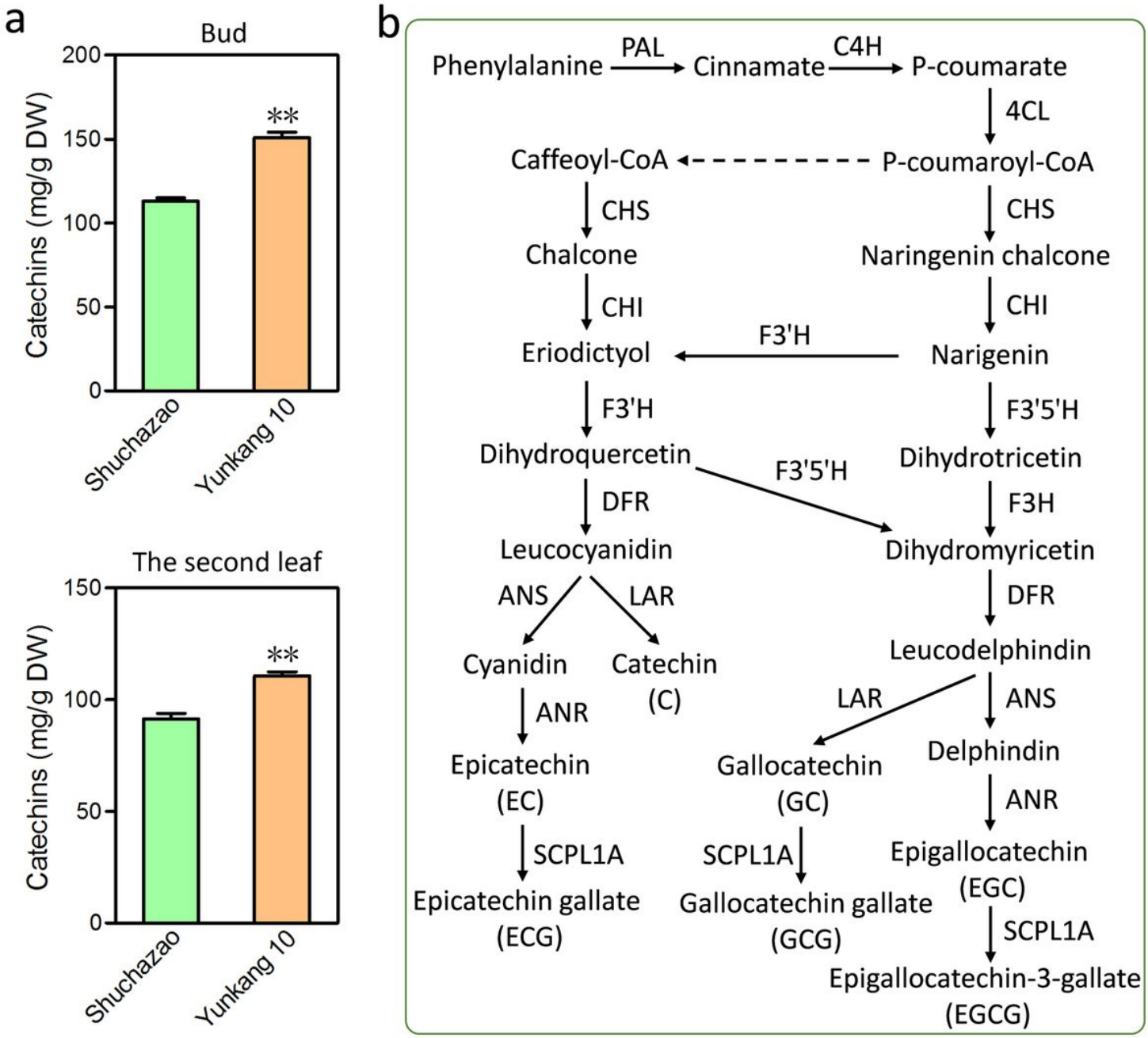
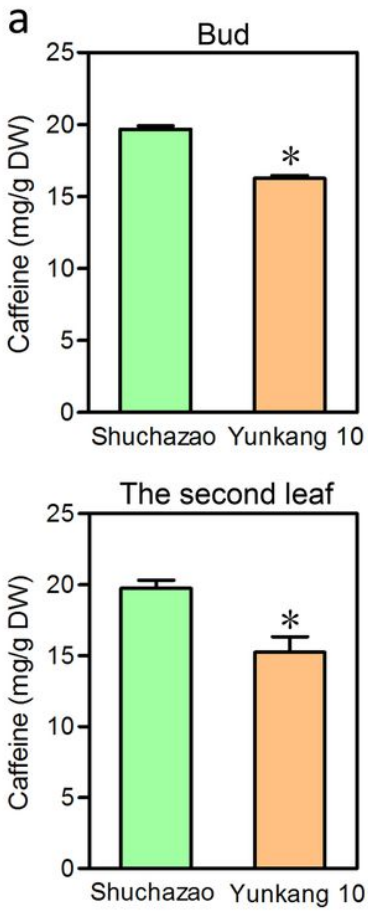
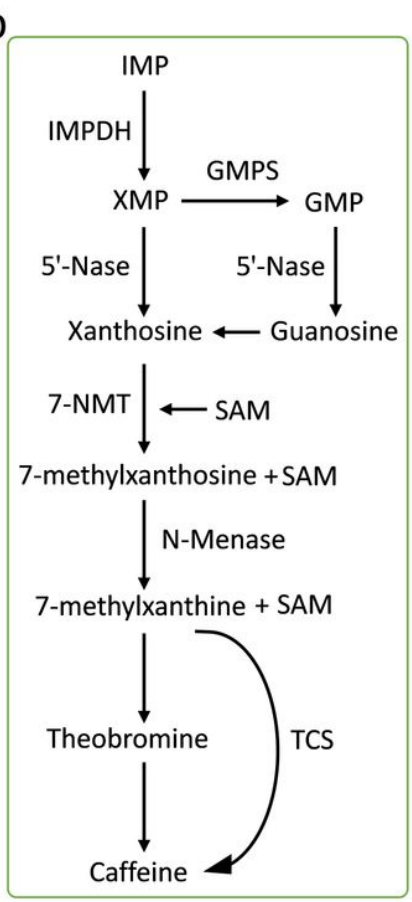


Figure 8

Detection of catechin content and genetic variations within catechin biosynthesis-related genes. a Detection of catechin content of the bud and leaf of both ‘Shuchaza’ and ‘Yunkang 10’. T-test was employed for significant analysis and two asterisks represent the $p<0.01$, each sample was tested with three independent biological replicates and two technical replicates. b The flavonoid biosynthesis pathway. PAL, phenylalanine ammonia-lyase; C4H, cinnamic acid 4-hydroxylase; 4CL, 4-coumarate-CoAligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; F3',5'H, flavonoid 3',5'-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; ANR, anthocyanidin reductase; LAR, leucocyanidin reductase; SCLP1A, subclade 1A of serine carboxypeptidase-like acyltransferases. c SNPs and InDels calling in catechin biosynthesis-related genes between ‘Yunkang 10’ and ‘Shuchazao’.





c

Gene name	Gene ID	SNP		InDel	
		DNA	CDs	DNA	CDs
IMPDH	TEA025393.1	105	2	0	0
	TEA028111.1	60	3	2	0
GMPS	TEA003403.1	26	1	2	0
	TEA006444.1	3	0	0	0
	TEA031622.1	18	7	3	1
	TEA011157.1	6	4	0	0
5'-Nase	TEA004804.1	3	0	0	0
	TEA005058.1	17	1	3	0
	TEA011804.1	14	3	2	0
	TEA023055.1	28	1	4	0
	TEA025000.1	10	7	3	1
	TEA025001.1	11	3	0	0
	TEA027082.1	19	4	2	0
TCS	TEA015176.1	6	1	2	0
	TEA010054.1	26	2	1	0
	TEA028050.1	3	2	0	0
	TEA024636.1	2	2	0	0
	TEA030024.1	2	0	0	0
	TEA010123.1	71	2	1	0
	TEA012581.1	1	1	0	0
	TEA028049.1	1	1	0	0
	TEA028052.1	1	1	0	0

Figure 9

Detection of caffeine content and genetic variations within caffeine biosynthesis-related genes. a. Detection of catechin content of the bud and leaf of both ‘Shuchaza’ and ‘Yunkang 10’. T-test was employed for significant analysis and one asterisk represents the $p<0.05$, each sample was tested with three independent biological replicates and two technical replicates. b. The caffeine biosynthesis pathway. IMP, Inosine monophosphate; XMP, Xanthosine monophosphate; GMP, Guanosine monophosphate; IMPDH, IMP dehydrogenase; GMPS, Guanosine synthase; 5'-Nase, 5'-nucleotidase; 7-NMT, 7-methylxanthosine synthase; SAM, S-Adenosyl-L-methionine; N-MeNase, N-methylnucleotidase; TCS, tea caffeine synthase. c. SNPs and InDels calling in caffeine biosynthesis-related genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.docx
- TableS2.docx