






# Metabolite signatures of diverse *Camellia sinensis* tea populations

Xiaomin Yu <sup>1,10</sup>, Jiajing Xiao<sup>2,3,10</sup>, Si Chen<sup>1,10</sup>, Yuan Yu <sup>1</sup>, Jianqiang Ma<sup>4</sup>, Yuzhen Lin<sup>1</sup>, Ruizi Li<sup>1</sup>, Jun Lin<sup>1</sup>, Zhijun Fu<sup>1</sup>, Qiongqiong Zhou<sup>5</sup>, Qianlin Chao<sup>6</sup>, Liang Chen <sup>4✉</sup>, Zhenbiao Yang <sup>7,8✉</sup> & Renyi Liu <sup>1,9✉</sup>

The tea plant (*Camellia sinensis*) presents an excellent system to study evolution and diversification of the numerous classes, types and variable contents of specialized metabolites. Here, we investigate the relationship among *C. sinensis* phylogenetic groups and specialized metabolites using transcriptomic and metabolomic data on the fresh leaves collected from 136 representative tea accessions in China. We obtain 925,854 high-quality single-nucleotide polymorphisms (SNPs) enabling the refined grouping of the sampled tea accessions into five major clades. Untargeted metabolomic analyses detect 129 and 199 annotated metabolites that are differentially accumulated in different tea groups in positive and negative ionization modes, respectively. Each phylogenetic group contains signature metabolites. In particular, CSA tea accessions are featured with high accumulation of diverse classes of flavonoid compounds, such as flavanols, flavonol mono-/di-glycosides, proanthocyanidin dimers, and phenolic acids. Our results provide insights into the genetic and metabolite diversity and are useful for accelerated tea plant breeding.

<sup>1</sup>FAFU-UCR Joint Center for Horticultural Biology and Metabolomics, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, 350002 Fuzhou, China. <sup>2</sup>Shanghai Center for Plant Stress Biology, Chinese Academy of Sciences, 3888 Chenhua Road, 201602 Shanghai, China. <sup>3</sup>University of Chinese Academy of Sciences, 100049 Beijing, China. <sup>4</sup>Key Laboratory of Tea Biology and Resources Utilization, Ministry of Agriculture and Rural Affairs, Tea Research Institute, Chinese Academy of Agricultural Sciences, 310008 Hangzhou, China. <sup>5</sup>College of Horticulture, Henan Agricultural University, 450000 Zhengzhou, China. <sup>6</sup>Wuyi Star Tea Industry Co., Ltd, 354300 Wuyishan, China. <sup>7</sup>Institute of Integrative Genome Biology, University of California at Riverside, Riverside, CA 92521, USA. <sup>8</sup>Department of Botany and Plant Sciences, University of California at Riverside, Riverside, CA 92521, USA. <sup>9</sup>Center for Agroforestry Mega Data Science, Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, 350002 Fuzhou, China. <sup>10</sup>These authors contributed equally: Xiaomin Yu, Jiajing Xiao, Si Chen. ✉email: [liangchen@tricaas.com](mailto:liangchen@tricaas.com); [yang@ucr.edu](mailto:yang@ucr.edu); [ryliu@fafu.edu.cn](mailto:ryliu@fafu.edu.cn)

Plants are a rich source for specialized metabolites that are not essential for their growth, development, and reproduction. Based on chemical structures, they are grouped into three major classes: terpenes, phenolics, and nitrogen-containing compounds. It is estimated that 100,000 to one million specialized metabolites are collectively produced by plants and any single plant produces a subset ranging from 5000 to tens of thousands of these metabolites<sup>1–3</sup>. Recent research also suggests that there exists a high level of qualitative and quantitative variations of metabolism within a plant species<sup>4</sup>. Specialized metabolites not only have key roles in plant adaptation to the environment and resistance to biotic and abiotic stresses but also provide natural products used for treating human diseases and important for human health and food quality<sup>3</sup>. More than two-thirds of small-molecule drugs introduced in the last two decades are either plant extracts or their close derivatives<sup>2</sup>. Due to their importance, intensive efforts have been devoted to the dissection of biosynthesis and genetic regulation of plant-specialized metabolites through reverse and forward genetic approaches<sup>4</sup>. In particular, the recent application of genomic, transcriptomic, and metabolomic profiling data makes it possible to not only explore the metabolite diversity between different species and different accessions of the same species, and understand the underlying evolutionary mechanisms<sup>5–7</sup>, but also identify candidate regulators through association analyses<sup>8,9</sup>. However, most of the identified candidate metabolic quantitative loci (mQTLs) and regulators lack experimental validation<sup>4</sup> and scientists cannot yet answer important questions such as: what are the underlying evolutionary mechanisms for metabolomic diversity between different species and different accessions of the same species? What metabolites are responsible for flavors of plant food products? How are plant metabolites regulated at the transcriptional, translational, and epigenetic levels? What are the functional roles of structurally similar but distinct metabolites?

The tea plant [*Camellia sinensis* (L.) O. Kuntze] is an excellent model system to address these questions due to its high contents and diversity in all three classes of specialized metabolites<sup>10–14</sup>. Tea is the most popular non-alcoholic beverage and offers a plethora of health benefits such as anti-oxidant, anti-cancer, anti-cardiovascular disease and anti-allergic activities<sup>15</sup>. Tea popularity is also attributed to a variety of rich flavors that come from all three classes of specialized metabolites<sup>16</sup>. Among the structurally diverse phytochemicals produced in tea plants, flavonoids such as catechins are best characterized molecularly and biochemically<sup>17–19</sup>. Synthesized through the phenylpropanoid and flavonoid pathways, catechins in tea are a mixture of different enantiomers and their gallic acid conjugates. They are most abundantly detected in tea leaves, among which (–)-epigallocatechin-3-gallate (EGCG) is predominant and the most bioactive<sup>20</sup>. Furthermore, tea plants synthesize a myriad of aroma compounds (e.g., volatile terpenes, fatty acid derivatives, and phenylpropanoids/benzenoids) in response to biotic and abiotic stresses<sup>21</sup>. Last but not the least, caffeine<sup>22</sup> and non-proteinaceous amino acid L-theanine<sup>23,24</sup>, which is particularly abundant in tea plants, also are key contributors to tea quality. An important goal for tea improvement is to breed for the increases of specific target metabolites and/or downregulation of some other target metabolites<sup>25</sup>. Comprehensive evaluation of metabolite contents of representative accessions will not only help us identify metabolite properties and signatures of different accessions, but also help us make wise selections of parental lines for tea breeding. Previous studies have revealed the metabolite content differences among different types of tea, but they mostly focused on processed tea products<sup>26–28</sup>. Because the metabolite types and contents change dramatically during tea processing<sup>29–31</sup>, the metabolite differences among processed products may not correlate to the genetic

backgrounds of corresponding tea accessions. To date only a very limited number of targeted or untargeted metabolite profilings have been performed on a small number of tea accessions to compare the metabolite contents of fresh tea samples<sup>20,32</sup>. No untargeted metabolomic studies have been carried out on fresh tea leaves from diverse tea populations.

China is likely the center of origin for tea plant, and is the top country for tea cultivation and production, accounting for ~40% of total world production in 2017 ([www.fao.org/faostat](http://www.fao.org/faostat)). China is home to more than 3000 tea accessions, and the genetic and metabolite diversity of tea population in this country largely represents the tea diversity in the world<sup>33</sup>. Modern tea cultivars are derived from hybrids within or between two major tea varieties, the large-leaved *C. sinensis* var. *assamica* (CSA) and the small-leaved *C. sinensis* var. *sinensis* (CSS). Molecular markers have been used to illustrate the genetic relationship among cultivated tea accessions. For example, Yao et al.<sup>34</sup> used 96 EST-SSR markers to analyze 450 tea accessions in different tea-producing regions in China and found that the cultivated tea accessions could be classified into five groups, clustered roughly around their growing locations. However, a recent study, using 6,252,201 single-nucleotide polymorphism (SNP) markers obtained from genome-resequencing data, separated 81 collected accessions into three clusters (CSS, CSA, and wild type)<sup>35</sup>. The smaller number of clusters revealed by this study is most likely because only 81 accessions were evaluated, among which only 58 were cultivated accessions. To resolve this discrepancy, a comprehensive evaluation of genetic diversity and population structure of a larger number of representative tea accessions, especially cultivated tea accessions, using genome-wide markers such as SNPs is needed.

Recently, the draft genomes of CSA and CSS have been published<sup>18,19,35,36</sup>, making it feasible to conduct genome-wide large-scale omics analyses. The tea genome is large (~3.1 GB) and complex, containing at least 34,000 protein-coding genes. Similar to other large plant genomes, the majority (>64%) of the tea genome contains various repetitive elements. The genome sequences not only provide a list of genes that are involved in the biosynthesis of three key compounds, catechins, caffeine, and theanine, and evidence for lineage-specific expansions of genes associated with flavonoid biosynthesis, but also help reveal the variations of metabolites and gene expression among different tissues and among different *Camellia* species. Comparison of the CSA and CSS genomes indicated that they diverged ~0.38–1.54 million years ago and analysis of genic collinearity showed that the tea genome resulted from two rounds of whole-genome duplications<sup>19</sup>. However, the genetic and metabolite diversity among different tea accessions remains to be explored.

Here we integrate transcriptomic and metabolomic analyses to study the population structure and phylogenetic relationships among major tea cultivars and association of tea metabolites with populations. We chose to use transcriptomic data rather than genome-resequencing data for this work because transcriptomic data can provide sufficient amount of polymorphism markers that are mostly within or around gene-encoding regions, without wasting sequencing power on intergenic regions and with additional benefit of examining gene expression changes. Deep RNA-sequencing is emerging as an important tool for rapid analysis of phylogenetic relationships among cultivars and evolutionary history of the plant kingdom<sup>37–39</sup>. Our comprehensive analyses showed that these representative cultivated tea accessions could be classified into five major groups and each group had unique gene expression and metabolite signatures. Our results provide molecular and metabolic markers for tea breeding, insights into the relationship between tea populations and specialized metabolites, and a foundation for the elucidation of mechanisms

underlying the diversity, high contents, and dynamics of specialized metabolites in tea plants.

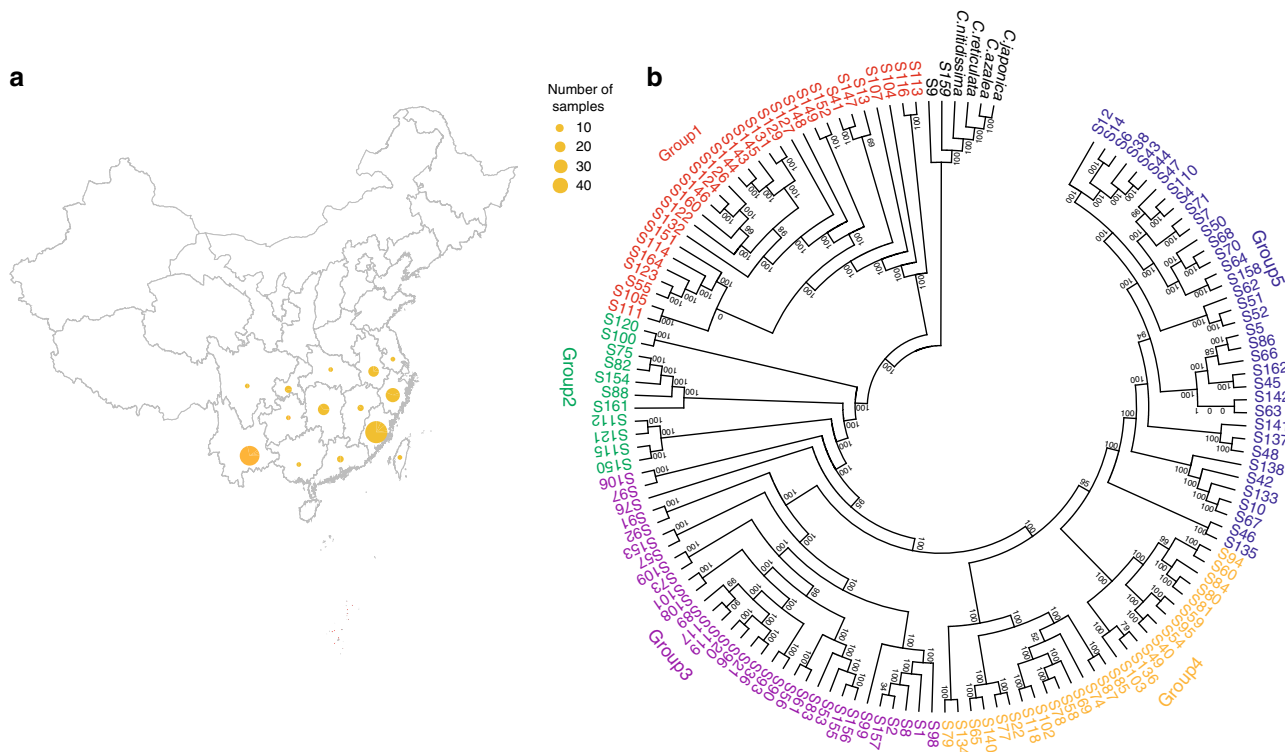
**Results**

**Phylogenetic relationships among representative Chinese tea accessions.** Over thousands of years, diverse tea cultivars have been derived for various tea flavors and adaptation to the environment via a combination of breeding and natural variation, and are mostly propagated through cuttings. However, their origins and phylogenetic relationship remained poorly understood. Here we used RNA-sequencing (RNA-seq) to study the phylogenetic relationships among 136 accessions (128 cultivars) collected from different growing regions, covering all major tea-producing provinces/regions in China (Fig. 1a), such as Yunnan, Fujian, Hunan, Anhui and Zhejiang. Accessions from Yunnan also included a close tea relative (*C. taliensis*). The second leaf samples (three biological replicates from each accession) were collected and subjected to RNA-seq analysis. On average, more than 5 GB of RNA-seq data were generated in each sample after adapter sequence and low-quality bases were removed. After aligning clean reads to the reference genome<sup>19</sup>, a total of 925,854 high-quality SNPs were identified, including 320,946 SNPs that are located in protein-coding regions.

We analyzed the phylogenetic relationship and evolutionary history of the 136 collected accessions by using a maximum likelihood-based phylogenetic tree constructed from 45,162 SNPs that are located on fourfold-degenerate sites, using the tea relative accession S159 and four other close relatives of tea plants (RNA-seq data for these tea relatives were collected from published data) as the outgroup (Fig. 1b). According to this phylogenetic tree, the cultivated accession “Chaoyang” (S9) is most closely related to the

tea relatives, consistent with the fact that its parental line, “Chongqing Pipacha”, was derived from a wild tea plant from Sichuan Province. The other 134 accessions could then be classified into five major groups, with group 1 containing exclusively CSA accessions or hybrid accessions with a dominant CSA genetic background, such as “Yunkang 10” (S55) and “Yinghong 1” (S116). CSA cultivars are mostly distributed in Yunnan province and are used to mainly process black tea and dark tea such as Pu’er tea. The other four groups contain middle/small-leaved accessions (Fig. 1b).

All accessions included in group 2 were adapted from wild tea plants and propagated asexually. Most of them originated from Hunan, Guangdong, and Chongqing, which may be the natural hybrid zones between CSA and CSS lineages. Group 3 contained mainly hybrid accessions that were generated by breeders or through natural hybridization. “Fuding Dabaicha” and “Fuding Dahaocha”, two premium cultivars used to process white and green tea, were included in this group. Moreover, 15 out of 32 accessions in this group descended from “Fuding Dabaicha”. For example, “Zhenong 113” (S96) is the hybrid accession derived from the crossing between “Yunnan Dayecha” and “Fuding Dabaicha”. Most tea accessions falling into group 4 initiated from geographically close regions in China, such as Zhejiang, Anhui, Jiangxi and Jiangsu, and are most suitable for manufacturing high-quality green tea. Representative examples were “Longjing 43” (S118), “Xicha 5” (S80), “Shifocui” (S103), and “Suchazao” (S78 and S58). Thirty out of 36 tea accessions in group 5 were mainly grown in Fujian and are typically used to process oolong tea. For example, Anxi Tieguanyin, the most well-known oolong tea from Southern Fujian, is processed using the “Tieguanyin” (S44) tea cultivar. Some hybrid accessions such as



**Fig. 1 Geographic origins and phylogenetic relationships of 136 representative tea plant accessions in China.** **a** Geographic origins of the tea plant accessions examined in this study. The map of China was generated using the R package “chinamap” (<https://github.com/GuangchuangYu/chinamap>). **b** An approximate Maximum Likelihood-based phylogenetic tree constructed using 45,162 fourfold-degenerate SNPs that were identified from mapped RNA-sequencing data. The tree was rooted using five tea relative species (in black) as outgroup. Numbers at the branch points represent support values (percentage) based on 1000 bootstrapping replicates. Five main clades were identified and indicated in different colors: group 1 (red), group 2 (green), group 3 (purple), group 4 (yellow), group 5 (blue). Source data underlying **b** are provided as a Source Data file.

**Table 1** Number of signature SNP sites on which major alleles were different between a pair of tea groups.

	Group 1	Group 2	Group 3	Group 4	Group 5
Group 1					
Group 2	6838				
Group 3	11,362	2617			
Group 4	13,534	2817	132		
Group 5	11,762	3791	674	132	

“Chuntaoxiang” (S68), “Huangguanyin” (S4), “Huangmeigui” (S110), “Mingke 1” (S6), and “Jinmudan” (S14) were derived from a cross between “Tieguanyin” (S44) and “Huangdan” (S47) and hence were clustered with their parental lines in the phylogenetic tree. Also included in this group were many cultivars like “Rougui” (S133), “Shuijingui” (S137), “Bantianyao” (S138), and “Baijiguan” (S135), which are traditionally used for producing Wuyi Rock tea, a well-known oolong tea from Northern Fujian.

Next, we examined each of the 925,854 high-quality SNP sites and identified signature SNPs that could be used to separate different groups of tea accessions. As shown in Table 1, the largest difference was observed between group 1 and the other groups, differing in 6838–13,534 SNP locations. In contrast, the numbers of SNP sites that separated the remaining groups were much smaller, reflecting a closer relationship among middle/small-leaved tea accessions. Notably, on 8187 SNP sites, group 1 possessed different major alleles from the other four groups, representing genetic divergence between CSA and CSS lineages. Signature SNPs were also found on genes that are known to be involved in the biosynthesis of characteristic metabolites in tea. For example, two genes involved in catechin and caffeine biosynthesis, namely, *LAR* (TEA027582, encoding a leucoanthocyanidin reductase) and *TCS* (TEA015791, encoding a caffeine synthase), contained several SNP sites on which the major alleles varied among different groups, with some being non-synonymous (Supplementary Fig. 1). Non-synonymous mutations cause changes in protein sequence, and could change protein function/enzyme activity. *LAR* is an important enzyme in the catechin biosynthesis pathway and is responsible for converting leucocyanidin/leucodelphinidin to C/GC. Previous evolutionary analysis showed that extensive sequence diversity exists among *LAR* genes in different plants as well as among three *LAR* homologs in tea<sup>40</sup>. TEA026582 exhibited more than 10-fold higher expression than the other two *LAR* homologs in our samples and thus likely had a major functional role. Although the two non-synonymous SNPs that we discovered here are not located within the three well-conserved *LAR*-specific motifs (RFLP, ICCN, and THD) or at the known substrate-binding sites<sup>40</sup>, they may still cause change in enzyme activity and thus are worth further investigation. Tea caffeine synthase (*TCS*) is a *N*-methyltransferase that converts 7-methylxanthine to theobromine (theobromine synthase or TS activity) and theobromine to caffeine (caffeine synthase or CS activity) and has a major role in the caffeine synthesis pathway<sup>41</sup>. There are 11 *TCS* homologs in the tea genome, among which TEA015791 (*TCS1*) has the highest expression level and has a predominant role. It has been shown that there are at least six natural *TCS1* alleles (named *TCS1a* to *TCS1f*, respectively) in different tea accessions<sup>41</sup>. Furthermore, different *TCS1* alleles exhibit significant difference in TS and CS activities. While *TCS1a* and *TCS1d-f* have both TS and CS activities, *TCS1a* has a lower CS/TS ratio than *TCS1d-f*<sup>41</sup>. *TCS1b-c* have only TS activity, resulting high level of theobromine and almost no caffeine in the two “caffeine-free” wild tea accessions “Hongyacha” and Cocoa tea (*C. pilophylla* Chang)<sup>42</sup>. Amino acid

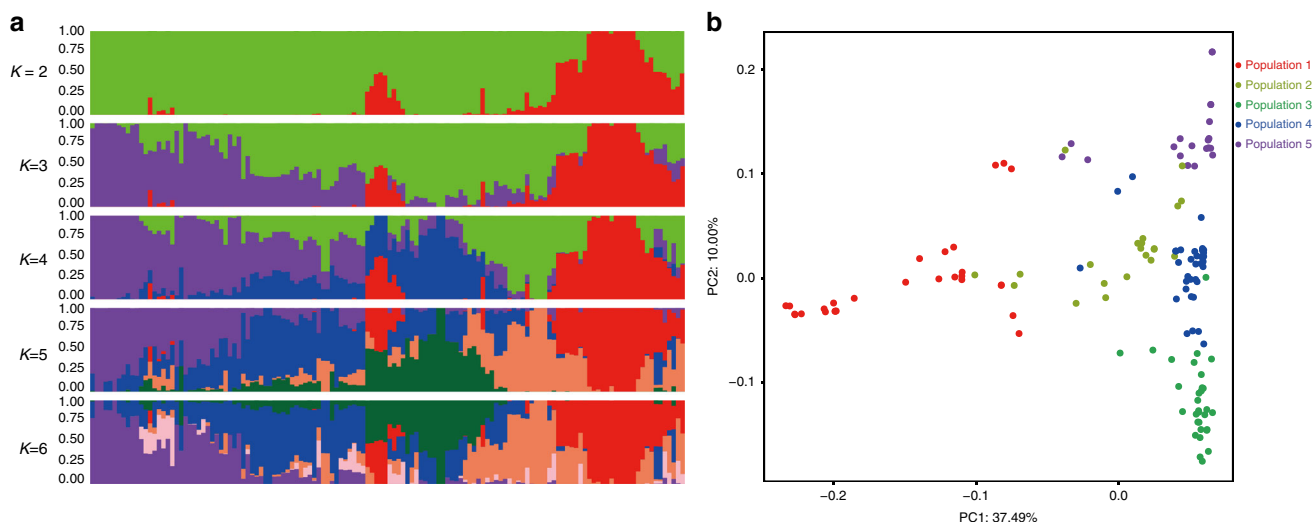
residue variations on our non-synonymous SNP sites indicate that in CSS accessions *TCS1a* is the dominant *TCS1* allele, whereas in CSA accessions, a *TCS1d-f*-like allele is the major allele. The enzyme activity difference of the major *TCS1* allele may affect the theobromine and caffeine accumulation levels in different tea accessions.

**Population structure of tea accessions in China.** To further understand the genetic diversity of tea accessions in China, a Bayesian inference of population structure was conducted using the STRUCTURE software<sup>43</sup>. The optimal number of clusters (*K*) was determined to be 5 by Harvester<sup>44</sup> (Fig. 2a), indicating that the sampled tea accessions in China could be grouped into five populations, in agreement with the phylogenetic analysis. With *K* = 2, 134 collected accessions were classified into a large-leaved tea population and a middle/small-leaved tea population, indicating the apparent genetic divergence between CSA and CSS. With the optimal *K* being 5, population 1 mainly consisted of accessions in group 1 and several other accessions that would be classified into other groups mainly due to controlled crossing and/or selection from natural hybridization. Population 2 mainly contained accessions from group 2 and several other accessions that came from natural hybridization between CSA and CSS. Population 3 only included accessions in group 5, most of which were derived from “Tieguanyin” and “Huangdan”. Population 4 mainly contained accessions clustered in group 4 and several oolong tea accessions. Population 5 mainly contained accessions in group 3 derived from the hybridization breeding process between CSA and CSS. Therefore, results from population structure analysis agreed well with those from the phylogenetic analysis: the representative tea accessions in China could be classified into five groups/populations. Principal component analysis (PCA) of the genetic distance of these accessions using the SNP data also illustrated that accessions in each of the five groups/populations were naturally clustered together (Fig. 2b).

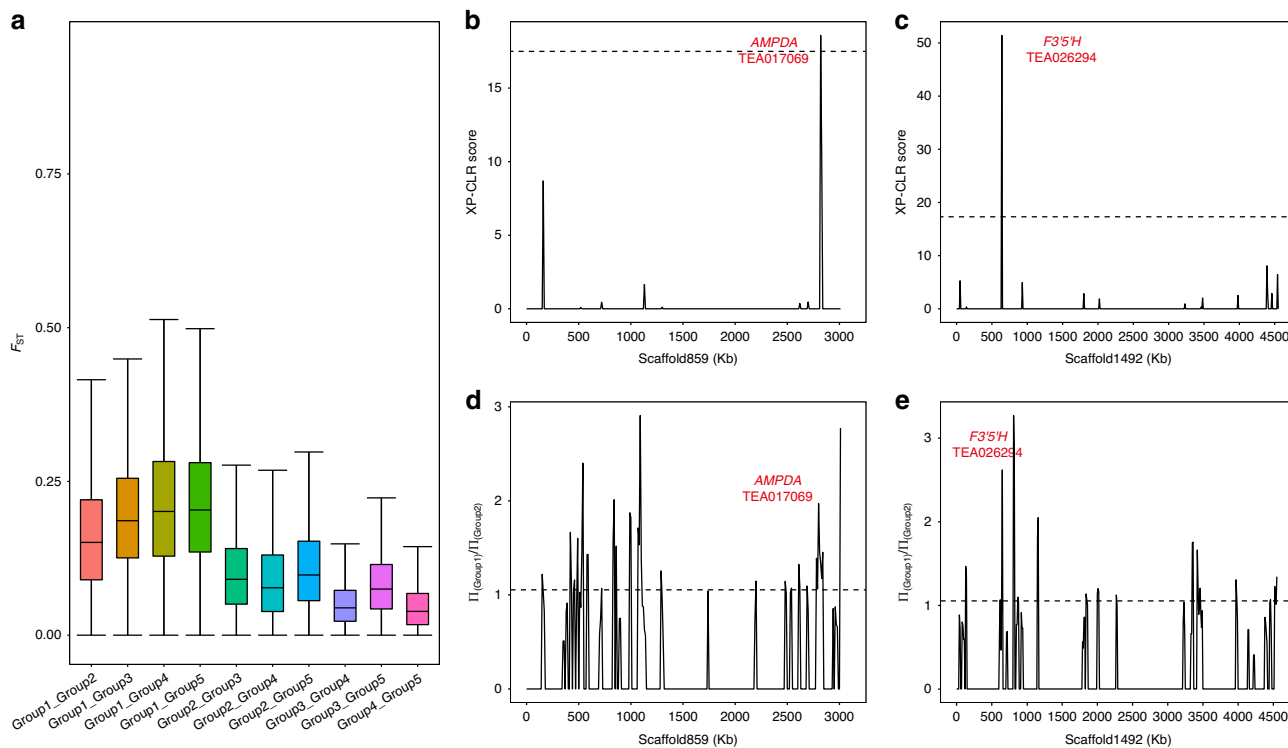
**Identification of regions that were potentially subject to selective sweeps.** To illustrate the genetic differentiation among five tea groups, we calculated the average *F<sub>st</sub>* values in pairwise comparisons of tea groups (Fig. 3a). The genetic differentiation between CSA and any of the other four groups were significantly (*p* < 0.001) higher than those between any pair of subpopulations within CSS, supporting the apparent genetic divergence between CSA and CSS. The lowest genetic differentiation was detected between group 4 and group 5, indicating the relatively close genetic relationship between green tea and oolong tea accessions.

During the evolution and domestication of tea plants, some genomic regions may have been subject to selective sweep because such regions contain genes that were related to traits selected by natural environments or by artificial breeding. The XP-CLR software<sup>45</sup> was used to identify potential selective sweep regions by comparing non-overlapping 10 kb regions along the tea genome between any two of the aforementioned five groups of tea accessions that we identified through phylogenetic analysis





**Fig. 2 Population structure of 134 representative tea plant accessions in China.** **a** Model-based clustering analysis with different K-values (number of clusters). The optimal K-value is 5 as determined by the Harvester software. **b** PCA analysis using SNP allele data showing the genetic distances among tea plant accessions in five groups identified in **a**. Source data are provided as a Source Data file.



**Fig. 3 Population differentiation and selective sweep regions across five groups of tea plant accessions.** **a** Boxplot of  $F_{st}$  values calculated from pairwise comparisons of five groups of tea plant accessions (identified in Fig. 1a).  $F_{st}$  was calculated on 100 kb sliding windows with a step size of 10 kb along all scaffolds in the tea reference genome. ( $n = 180,916, 181,969, 182,127, 182,256, 181,842, 181,136, 181,436, 181,989, 182,235, 181,300$  sliding windows for plots ordered from left to right). Boxes = interquartile ranges, middles = medians, whiskers =  $1.5 \times$  the interquartile range. **b–e** Plot of selective sweep (**b**, **c**) and nucleotide diversity values (**d**, **e**) along scaffold859 and scaffold1492 that contain selective sweep regions. The dotted line in **b**, **c** indicates a threshold value of 18.40 and the dotted line in **d**, **e** indicates a threshold value of 1.054. Source data are provided as a Source Data file.

(Fig. 1b). We identified 833 potential selective sweep regions that contained 1132 genes.

Two examples of selective sweep regions were shown in Fig. 3b–e. A selective sweep region on scaffold859 contained the *AMPDA* gene (TEA017069, encoding an adenosine monophosphate deaminase) that is involved in caffeine biosynthesis. As one of the early steps in caffeine biosynthesis, AMPDA converts

adenosine monophosphate (AMP) to inosine monophosphate (IMP), freeing an ammonia molecule in the process. Another gene, *F3'5'H* (TEA026294, encoding a flavonoid 3',5'-hydroxylase), is involved in the biosynthesis of catechin, and was recently identified to have a role in governing the ratio of di/tri-hydroxylated catechins and catechin contents<sup>46</sup>. These results suggested that some metabolic pathways may have been subject to

**Table 2 Total number of detected, differentially accumulated, and signature metabolites identified in this study.**

Mode	Total <sup>a</sup>	DAMs <sup>b</sup>	Signature metabolites in group 1 <sup>c</sup>	Signature metabolites in group 2 <sup>c</sup>	Signature metabolites in group 3 <sup>c</sup>	Signature metabolites in group 4 <sup>c</sup>	Signature metabolites in group 5 <sup>c</sup>
POS	2672	129	15	3	0	0	1
NEG	1997	199	21	9	0	1	4

<sup>a</sup>Total number of metabolic features detected<sup>b</sup>Number of differentially accumulated metabolites (DAMs).<sup>c</sup>Number of metabolites that showed significantly higher accumulation in one group of tea accessions than any of the other four tea groups.

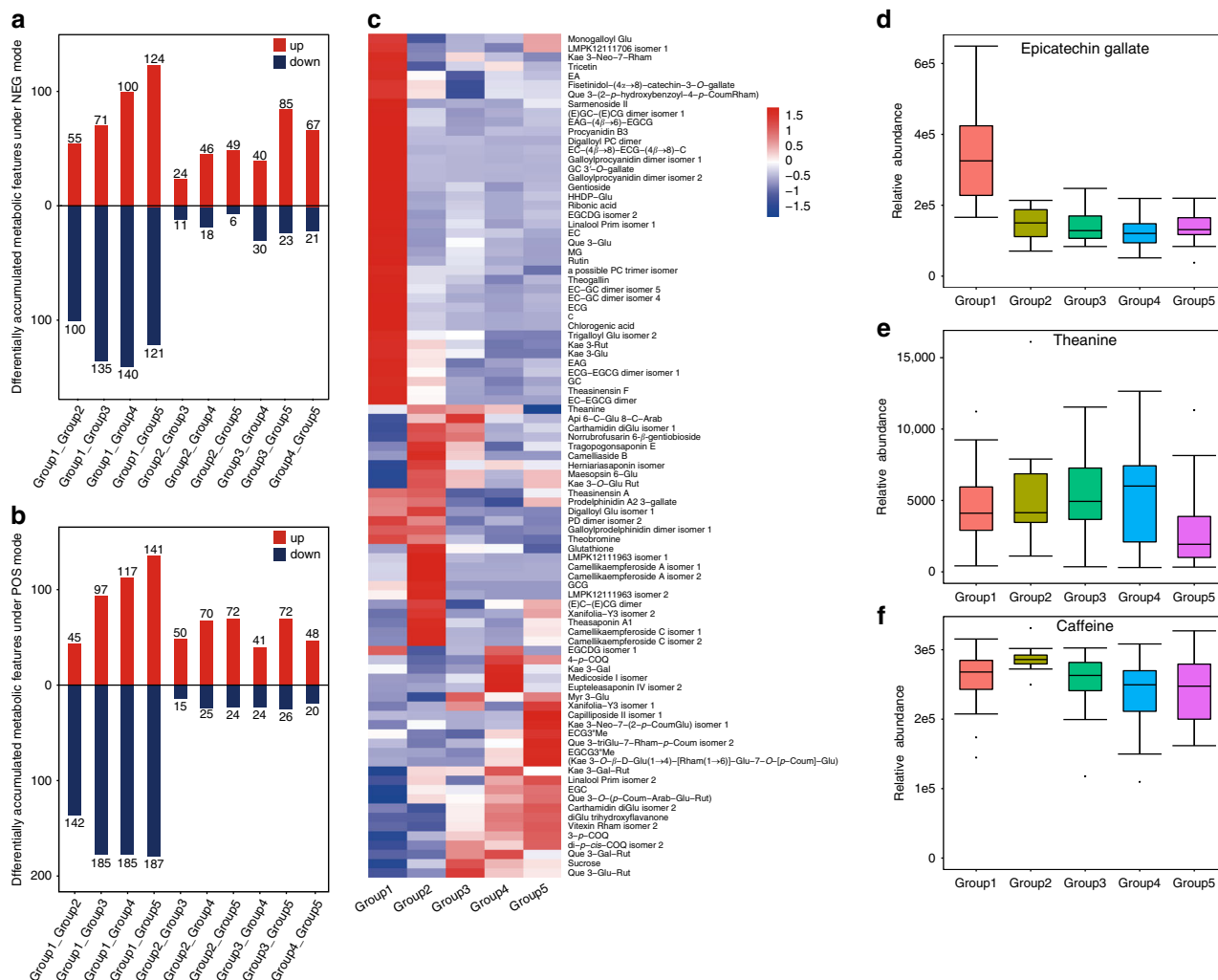
strong selection during the evolution and domestication of tea and some specialized metabolites may have been key traits for breeding and domestication.

**Identification of tea metabolites via untargeted metabolomic analyses of different tea populations.** As a first step in assessing how tea metabolites are related to diverse genetic backgrounds, we sought to expand our knowledge of tea metabolites by untargeted metabolomic analysis. Previous untargeted analyses were limited to processed tea products or fresh leaves of a small group of closely related tea cultivars<sup>20,26,28,30,47,48</sup>. We anticipate that many tea metabolites were missed in these analyses. Hence we analyzed the metabolite contents of fresh tea leaves in the 136 representative tea accessions using untargeted metabolomics analysis of the second leaf samples by ultra-performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UPLC-QTOF MS). In total, 2672 and 1997 mass/retention time features were detected in positive (POS) and negative (NEG) electrospray ionization (ESI), respectively. To remove ultra-low abundance signals, ions with a normalized relative abundance lower than 500 in all accessions were filtered, leaving 752 and 503 metabolic features in respective modes for further analysis (Table 2 and Supplementary Data 1, 2). Putative metabolite identity was assigned for tea leaf constituents in accordance with databases and literatures, and with comparison with authentic standards (Supplementary Data 3 and 4).

Like other plant species, tea plants possess their own specialized metabolome comprised of many isomeric compounds (Supplementary Data 3 and 4). Given their similar/same MS/MS fragmentation behaviors and small differences in retention time, resolving these structural isomers by the MS-based metabolomics approach alone still remains challenging and may necessitate the use of nuclear magnetic resonance (NMR) for further unambiguous structural elucidation. Nevertheless, we found that flavonol glycosides (in the RT window of 5.1–11.3 min and mostly between 7 and 11 min) and proanthocyanidins (in the RT window of 2.4–10.1 min and mostly between 4 and 9 min) were the two most structurally diverse classes of specialized metabolites detected in the methanol extracts of fresh tea leaves. For example, three compounds (RT = 9.00, 9.33, and 9.59 min) showing  $[M-H]^-$  at  $m/z$  635.1609, 635.1615, and 635.1608, respectively, were all putatively identified as kaempferol acetylhexose deoxyhexose. Two doubly charged ions (RT = 10.41 and 11.07 min) with their corresponding singly charged ions around  $m/z$  1031.30 were found to be another pair of isomers. They were putatively characterized as isomers of kaempferol 3-(*p*-coumaroyl-rhamnosyl)rutinoside-7-rhamnoside, whose occurrence in tea plants has not been documented. Six compounds (RT = 4.11, 4.35, 4.44, 4.80, 4.93, and 5.78 min) that were predicted to have the same formula  $C_{30}H_{26}O_{13}$  shared similar fragment ions and were putatively assigned as isomers of EC-GC dimer. Likewise, in ESI<sup>-</sup>, two metabolites eluted at 5.23 and 5.75 min, respectively, were shown to have the same formula  $C_{52}H_{42}O_{25}$ . Given the same MS/MS fragmentation ions generated, we believe that again they

were isomers, and based on the match in the Dictionary of Natural Products database (<http://dnp.chemnetbase.com>), we tentatively assigned them as two galloylated trimeric proanthocyanidins. Metabolites with the same formula have not been previously reported in tea plants. The RT window between 11 and 17 min was occupied primarily by triterpenoid saponins and terpenoid glycosides (albeit in low abundance), many of which have rarely been described as constituents in fresh tea leaves. Another metabolite that immediately catches our attention eluted at 5.26 min and had  $m/z$  at 225.0972 in ESI<sup>+</sup>. The deduced formula was  $C_9H_{12}N_4O_3$ , matching that of theacrine, a caffeine-like xanthine alkaloid that has so far only been reported in *C. assamica* var. *Kucha*<sup>49</sup>. The biosynthesis of theacrine has sparked a lot of research interest since theacrine is non-stimulatory and hence may guide the development of decaffeinated drinks<sup>49</sup>. To our great interest, out of all the tea resources that we screened, only “Nannuoshan Dayecha 3” (S131) was found to produce a large amount of theacrine. This tea accession adds another genetic resource, besides *Kucha*, to facilitate mechanistic investigations into how caffeine is transformed into theacrine.

**Metabolite signatures for the five different tea phylogenetic groups.** Global clustering analysis of the tea metabolite profiles described above suggests that genetic background had a stronger effect on metabolite contents than environment factors because tea accessions in the same phylogenetic group were more likely to cluster together than those from the same growing location (Supplementary Fig. 2a, b). As described above, different tea plant phylogenetic groups/populations more or less represent their processing suitability (e.g., group 4 accessions are commonly used for making green tea, whereas group 5 accessions are typically used for making oolong tea). Hence, we were interested in understanding the metabolic basis of the groupings. Specialized metabolites are the primary factors in determining health benefits, tastes and aroma of tea products. To identify metabolite signatures that would ultimately assist in breeding, we performed pairwise comparisons to detect metabolic features that were differentially accumulated among five groups of tea accessions. 409 and 325 metabolic features were found to be differentially accumulated in at least one pairwise comparison under POS and NEG modes, respectively. Not surprisingly, the highest number was found in group 1 under both modes (Fig. 4a, b), indicating that CSA tea accessions had a quite different metabolite profile from that of CSS tea accessions, consistent with the result from clustering analysis. After careful inspection of raw spectral data, 280 (68%) and 126 (39%) features under POS and NEG modes, respectively, were found to be fragment ions and removed from further analysis. The remaining 129 and 199 features were classified as differentially accumulated metabolites (DAMs), with 75 DAMs being detected in both modes. Among these DAMs, 108 (84%) and 171 (86%) in the respective mode could be confidently (with reference to authentic standards) or putatively assigned to known metabolites (Supplementary Data 3–6).



**Fig. 4** Metabolites that showed significant changes in concentration in pairwise comparisons of five groups of tea accessions. **a, b** Number of metabolic features that were detected under NEG (**a**) and POS (**b**) modes, respectively, and were identified as differentially accumulated in pairwise comparisons of five groups of tea accessions. Red and blue bars indicate the numbers of metabolic features showing increase and decrease in concentrations, respectively. **c** Heatmap showing the abundance patterns of annotated metabolites with an average relative abundance greater than 500 in at least one tea group and with significant changes in abundance in at least one comparison. **d–f** Box plots of abundance of epicatechin gallate, theanine and caffeine in different tea groups. ( $n = 29$  for Group 1, 11 for Group 2, 32 for Group 3, 26 for Group 4, and 36 for Group 5). Boxes = interquartile ranges, middles = medians, whiskers =  $1.5 \times$  the interquartile range, single points = outliers. Source data underlying **c–e**, and **f** are provided as a Source Data file.

Clustering analysis of the annotated DAMs showed that 40 metabolites, mostly a wide range of flavonoids, were highly accumulated in group 1 but in general were lowly accumulated in other groups. Many flavanols (C, EC, GC, ECG, epigallocatechin digallate, gallic catechin 3'-O-gallate, epiafzelechin 3-gallate, and epiafzelechin) were more abundantly accumulated in group 1 (Fig. 4c and Supplementary Data 7). For instance, almost all accessions in group 1 exhibited a significantly higher content of ECG. The mean abundance of this compound in group 1 was at least 2.2-fold that of the other groups (Fig. 4d). This result partly agrees with a previous targeted analysis of catechin contents in representative Chinese tea germplasms, where the accumulation levels of total catechin, ECG, EC, and C were found to be significantly higher in CSA tea accessions than in CSS tea accessions<sup>50</sup>. In addition, mono-/di-/triglycosides of quercetin and kaempferol, proanthocyanidin dimers, hydrolysable tannins and quinic acid derivatives had higher abundance in group 1. These results suggest that the genes involved in the phenylpropanoid/flavonoid pathways may be upregulated in the CSA lineage. Theobromine, the second most abundant purine alkaloid

in fresh tea leaves, was also enriched in this group, with a similar mean content observed in accessions from group 2 (Fig. 4c).

Higher accumulation of kaempferol glucosylrutinoside, two acylated kaempferol tetraglycosides, four acylated kaempferol triglycosides and five triterpenoid saponins were found in group 2. Myricetin 3-glucoside and quercetin 3-O-glucosylrutinoside were more enriched in group 3. Kaempferol 3-O-galactoside, kaempferol 3-O-galactosyl rutinoside, quercetin 3-O-galactosyl rutinoside and two triterpenoid saponins appeared to be enriched in group 4, although the identities of these two metabolites need to be further confirmed by spectroscopic methods. Finally, methylated catechins (EGCG<sup>3</sup>Me and ECG<sup>3</sup>Me) and coumaroyl derivatives of quercetin and kaempferol tri-/tetraglycosides were specifically present in higher levels in group 5. However, the level of theanine was apparently lower in group 5 than in other groups (Fig. 4e). Some metabolites may be accumulated at a higher level in more than one groups. For examples, one galloylprodelphinidin dimer, prodelphinidin A2 3-gallate and theasinensin A were present at higher levels in groups 1 and 2. One carthamidin diglucoside, norrubrofusarin 6-β-gentiobioside

as well as apigenin 6-*C*-glucoside 8-*C*-arabinoside occurred more abundantly in groups 2 and 3. Some metabolites such as EGC, diglucopyranosyl trihydroxyflavanone and vitexin rhamnoside were found to have a higher accumulation in tea accessions from groups 4 and 5 (Fig. 4c). On the other hand, the accumulation levels of caffeine did not show significant difference among different groups (Fig. 4f), consistent with the fact that all these tea accessions have been domesticated and/or selected for tea production. In addition, the accumulation levels of EGCG, some proanthocyanidins (one galloylprodelphinidin dimer, procyanidin B2, and one procyanidin trimer), and some other metabolites (e.g., 5-coumaroylquinic acid and linalool oxide primeveroside) appeared to be stable among different tea groups.

Next, we set to detect the signature metabolites whose concentration in one tea group was significantly higher than that in any of the other tea groups. A total of 40 annotated signature metabolites were found. Among them, 36 metabolites were annotated with matches to metabolite databases or authentic standards and the elemental compositions for the remaining four were calculated based on the accurate mass values (Table 3). Group 1 had the highest number of signature metabolites under both POS and NEG modes, whereas group 3 had no signature metabolite, probably because tea accessions in group 3 resulted from hybrid breeding between the two major types of tea (CSA and CSS). Among the annotated signature metabolites detected under the NEG mode (Table 3), five flavonol glycosides (rutin, sarmenoside II, quercetin 3-*O*-glucoside, kaempferol 3-neoheperidoside-7-rhamnoside, and kaempferol 3-(4''-(*E*)-*p*-coumaroylrobinobioside)-7-rhamnoside isomer), five proanthocyanidins (two EC-GC dimers, procyanidin B3, galloylprocyanidin dimer, and digalloylprocyanidin dimer), three phenolic acids (theogallin, methylgallate and chlorogenic acid), four flavanols (C, GC, EC, and ECG), one terpenoid glycoside (linalool primeveroside isomer), one ellagitannin (hexahydroxydiphenoyl-glucose), one hydrolysable tannin (trigalloylglucose) and one sugar acid derivative (likely ribonic acid) were highly enriched in CSA accessions (group 1). In contrast, group 2 consistently exhibited the enrichment of four complex coumaroylated kaempferol glycosides (two camelliakaempferoside C isomers and two camelliakaempferoside A isomers), two triterpenoid saponins (theasaponin A1 and tragopogonsaponin E), and one hydrolysable tannin (digalloylglucose). The signature metabolite for group 4 was a putative triterpenoid saponin (eupteleasaponin IV isomer). Group 5 had high accumulation of two methylated catechins (EGCG3''Me and ECG3''Me) and one triterpenoid saponin (xanifolia-Y3 isomer). The signature metabolites detected under the POS mode in general agreed with those under the NEG mode.

We also identified three (8.07 min<sub>367.0126</sub> *m/z*, 13.40 min<sub>467.1343</sub> *m/z*, and 16.41 min<sub>1189.5400</sub> *m/z*) and two signature metabolites (13.40 min<sub>457.1377</sub> *m/z* and 16.89 min<sub>731.4146</sub> *m/z*) that were unannotated in NEG and POS modes, respectively. For example, a metabolite eluted at 16.41 min with *m/z* 1189.5400 in NEG mode was much enriched in group 2 tea accessions, with an abundance at least 2.8-fold of that in other groups. Another unannotated metabolite eluted at 8.07 min with *m/z* 367.0126 in NEG mode was only detected in groups 2 and 3 and its average abundance in group 2 was 14-fold higher than that in group 3. In POS mode, the concentration of an unknown metabolite (*m/z* = 731.4146, RT = 16.89 min) in group 2 was at least 2.5-fold of that in other groups of accessions. These unannotated signature metabolites deserve further investigation.

It is well-known that the accumulation of specialized metabolites is highly impacted by environmental factors. Thus, the aforementioned existence of signature metabolites for each group of tea accessions collected from diverse locations and

environments is quite significant, as it suggests that these signature metabolites are predominantly determined by genetic factors.

**Differentiation of gene expression profiles among different tea populations.** To assess whether the metabolite signatures described above are associated with the transcription of specific genes, we analyzed the difference in gene expression in the second leaves among different tea accessions. We first performed *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) analysis of the global gene expression profiles. The results showed that the expression profiles of tea accessions in group 1 tended to cluster together, but not for tea accessions in the other four groups (Supplementary Fig. 3a), suggesting that on the one hand, CSA and CSS lineages have significant divergences in gene expression; on the other hand, genetic background is not the sole deciding factor. In particular, the gene expression profiles of tea samples collected from Jiamu Yeyatang tea plantation, which is located in a mountainous region in Yunnan Province with high elevation (~2000 m) and low temperature (average annual temperature is ~17 °C, and average temperature in the coolest month, January, is ~9 °C), form its own cluster (Supplementary Fig. 3b), regardless of their genetic backgrounds, indicating that environmental factors may greatly influence the gene expression profile in tea plants. Clustering analysis using the expression profiles of genes involved in the biosynthesis of catechins, caffeine and theanine also gave similar results, indicating the biosynthetic genes of these metabolites may show differential expression depending on both genetic backgrounds and growing environments. To further illustrate this point, we compared the overall gene expression profiles of five sets of sample groups that are of the same genotype but were grown in different locations. We found that the four “Yunkang 10” samples (S55, S114, S123, and S164) did not cluster together in the *t*-SNE figures (Supplementary Fig. 3a). Twenty-eight out of 33 genes that showed high correlations with the first dimension in *t*-SNE analysis, significantly changed their expression in at least one pairwise comparison in these four samples. In comparison with the other three tea accessions, three genes, namely TEA016601 (*FLS*), TEA023333 (*CHS*), and TEA023790 (*F3'H*), were significantly downregulated in tea sample S164 collected from Jiamu Yeyatang tea plantation, suggesting the downregulation of these three genes in S164 was mainly caused by environmental factors (high elevation and low temperature).

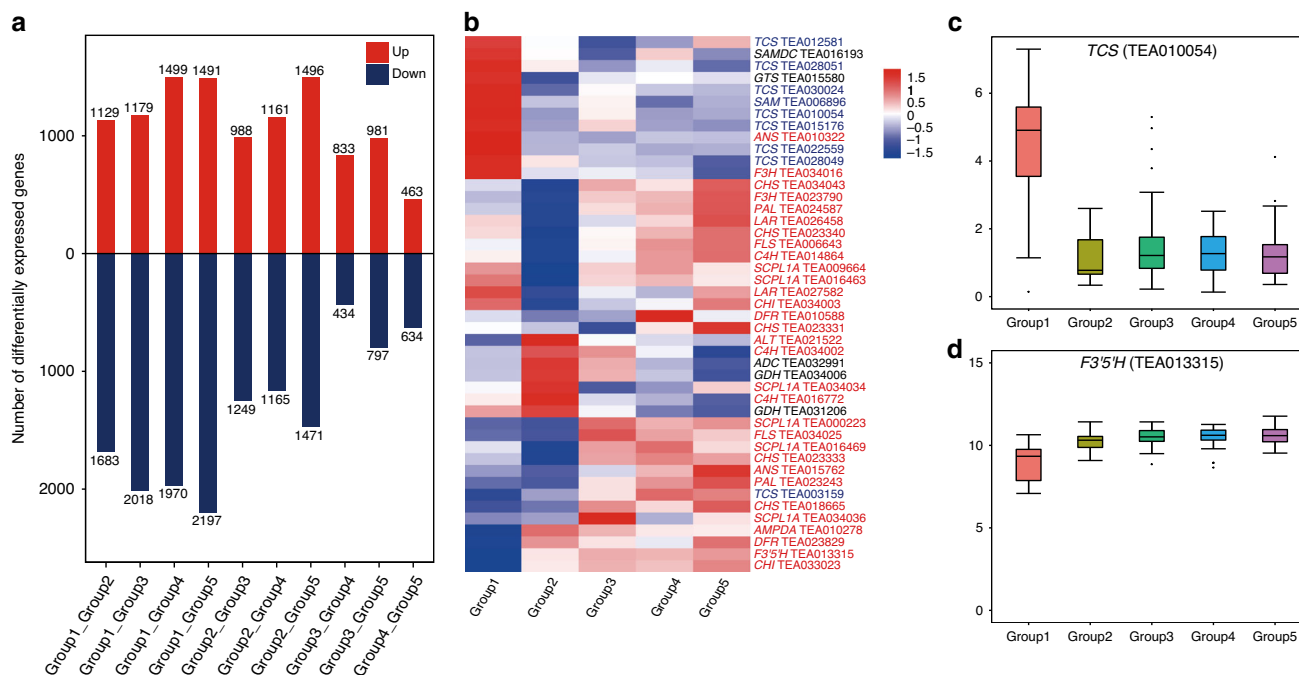
To further compare the gene expression profiles in five groups of tea plants, we performed pairwise comparisons of gene expression levels in different tea groups and identified differentially expressed genes (DEGs), after removing accessions collected from the Jiamu Yeyatang tea plantation. In total, 7674 DEGs were found in at least one comparison. As shown in Fig. 5a, the number of DEGs was apparently lower in the pairwise comparisons among groups 3, 4, and 5 than when any of them was compared to group 1 or 2, suggesting that the gene expression profiles in groups 3, 4, and 5 were more similar to each other. Among the identified DEGs, 31 genes involved in catechin biosynthesis were found to be differentially expressed in at least one pairwise comparison. Similarly, nine and five DEGs were found in the caffeine and theanine biosynthetic pathways, respectively (Fig. 5b), indicating the expression of structural genes involved in the biosynthesis of these metabolites were dynamically regulated and may, at least in part, be attributed to the genetic backgrounds of tea accessions. Clustering analysis of the expression levels of these genes indicated that the expression levels of seven caffeine biosynthetic genes were in general higher in group 1, while many catechin biosynthetic genes were highly expressed in oolong tea cultivars (group 5) (Fig. 5b). For example,



**Table 3 List of annotated signature metabolites of five tea groups.**

Signature group	Metabolites detected	Tentative metabolite identification	Metabolite class	Mean abundance in group 1	Mean abundance in group 2	Mean abundance in group 3	Mean abundance in group 4	Mean abundance in group 5
Group 1	1.93_48110621 m/z	HHDP-glucose	Ellagitannins	667.5	18.8	107.9	28.5	51.4
Group 1	2.06_167.0549 m/z	Ribonic acid	Sugar acids and derivatives	1131.7	30.8	71.5	23.2	44.7
Group 1	2.89_344.0743n	Theogallin	Phenolic acids	31146.8	14,975.9	12,871.7	8404.9	12,089.9
Group 1	3.83_306.0740n	Gallocatechin	Flavanols	16166.9	7079.3	6195.0	5533.4	6741.1
Group 1	4.80_593.197 m/z	EC-GC dimer isomer 4	Proanthocyanidins	1176.5	491.2	377.7	360.7	454.4
Group 1	4.93_593.1293 m/z	EC-GC dimer isomer 5	Proanthocyanidins	1600.1	566.2	396.9	323.0	412.8
Group 1	5.11_577.1348 m/z	Procyanidin B3	Proanthocyanidins	1201.0	307.4	235.4	351.5	341.6
Group 1	5.34_290.0791n	Catechin	Proanthocyanidins	24,258.8	4032.7	3625.6	3542.0	3670.3
Group 1	5.36_183.0298 m/z	Methylgallate	Phenolic acids	4633.4	1208.2	1888.1	1306.2	1000.6
Group 1	5.51_354.0948n	Chlorogenic acid	Phenolic acids	6272.5	803.9	265.5	265.9	134.6
Group 1	5.83_747.1552 m/z <sup>a</sup>	EC-EGC dimer	Proanthocyanidins	3602.3	907.5	512.1	588.6	517.6
Group 1	6.25_290.0794n	Epicatechin	Flavanols	116,557.4	48,408.2	57,998.7	57,396.7	52,306.4
Group 1	6.67_551.1402 m/z	Gentioside	Xanthone glycosides	735.6	137.8	248.7	242.1	171.1
Group 1	6.63_318.0476n	Trigalloylglucose isomer 2	Hydrolysable tannins	9744.1	3327.0	3954.0	1512.2	1156.8
Group 1	7.69_609.1438 m/z	Rutin	Flavonol glycosides	11,904.4	1726.1	2665.6	1859.8	1404.2
Group 1	6.78_729.1455 m/z	Galloylprocyanidin dimer isomer 1	Proanthocyanidins	8493.2	1435.0	1174.9	1070.7	1265.0
Group 1	7.85_442.0901n	Epicatechin 3-O-gallate	Flavanols	334,280.4	151,298.2	137,580.1	121,634.2	137,860.9
Group 1	7.86_303.0485 m/z <sup>a</sup>	Tricetin	Flavones	4889.7	818.6	1984.6	2437.5	1774.5
Group 1	8.00_464.0948n	Quercetin 3-O-glucoside	Flavonol glycosides	7216.1	1158.4	2866.5	1918.1	1710.0
Group 1	8.26_739.2086 m/z	Kaempferol 3-neohesperidoside-7-rhamnoside	Flavonol glycosides	1035.9	1.2	356.9	20.5	0.4
Group 1	8.42_594.1630n <sup>a</sup>	Kaempferol 3-O-rutinoside	Flavonol glycosides	3885.6	1526.3	874.5	431.7	497.0
Group 1	9.37_441.0818n	Digalloylprocyanidin dimer	Proanthocyanidins	504.5	116.2	117.1	85.6	93.7
Group 1	10.24_451.1235n	Sarmentoside II	Flavonol glycosides	1481.6	0.0	22.2	0.5	98.1
Group 1	10.55_443.1260n	coumaroylrobinobioside)-7-rhamnoside isomer 1	Flavonol glycosides	612.2	5.4	6.0	0.4	93.6
Group 1	11.20_448.2300n	Linolool primveroside isomer 1	Terpenoid glycosides	1807.1	549.0	765.8	559.3	691.2
Group 2	5.01_242.0423n	Digalloylglucose isomer 1	Hydrolysable tannins	2155.5	9033.6	718.2	368.6	218.5
Group 2	10.48_517.1448n	Camellikaempferoside C isomer 1	Flavonol glycosides	113.8	4182.1	449.6	327.3	193.9
Group 2	10.74_517.1448n	Camellikaempferoside C isomer 2	Flavonol glycosides	5.7	1459.3	116.9	90.2	358.9
Group 2	10.90_436.1180n	Camellikaempferoside A isomer 1	Flavonol glycosides	113.9	1616.3	59.5	45.3	54.1
Group 2	11.13_871.2298 m/z	Camellikaempferoside A isomer 2	Flavonol glycosides	23.0	518.4	7.8	5.0	8.3
Group 2	11.33_1189.5616 m/z	Theasaponin A1	Triterpenoid saponins	10.9	623.5	73.6	52.3	193.7
Group 2	16.41_1189.5400 m/z	C <sub>60</sub> H <sub>86</sub> O <sub>24</sub>	Unknown	14.9	677.3	72.4	21.0	237.3
Group 2	16.80_1131.5352 m/z	Trigopogonsaponin E	Triterpenoid saponins	122.9	1685.5	375.2	34.5	350.9
Group 2	16.89_731.4146 m/z <sup>a</sup>	C <sub>44</sub> H <sub>58</sub> O <sub>2</sub>	Unknown	74.1	1173.1	431.9	461.7	467.7
Group 2	16.36_1157.5720 m/z	Eupileasaponin IV isomer 2	Triterpenoid saponins	0.0	1.6	52.9	940.7	19.7
Group 4	7.42_472.1004n	EGCG3'Me	Flavanols	577.4	436.9	199.8	3086.1	8095.4
Group 5	8.89_456.1053n	ECG3'Me	Flavanols	227.9	187.4	46.2	784.1	1666.9
Group 5	13.40_451.1377 m/z <sup>a</sup>	C <sub>25</sub> H <sub>22</sub> O <sub>8</sub>	Flavanols	65.0	158.5	307.6	376.0	821.8
Group 5	13.40_467.1343 m/z	C <sub>25</sub> H <sub>24</sub> O <sub>8</sub>	Unknown	7.2	144.3	360.7	464.8	1051.2
Group 5	16.41_1125.5460 m/z	Xanthifolia-Y3 isomer 1	Triterpenoid saponins	13.3	127.7	152.8	7.4	617.5

<sup>a</sup>Detected in ESI<sup>+</sup>. Other metabolites were detected in ESI<sup>-</sup>.



**Fig. 5** Differentially expressed genes identified in pairwise comparisons of five groups of tea accessions. **a** Number of differentially expressed genes in pairwise comparisons of five groups of tea accessions. Numbers of up- and downregulated genes are indicated in red and blue, respectively. **b** Heatmap showing the expression patterns of genes that are known to be involved in the biosynthesis of catechins (in red), caffeine (in blue), and theanine (in black), and show significant changes in expression in at least one pairwise comparison. **c, d** Box plots of gene expression values (log<sub>2</sub>-transformed counts per million) of TCS and F3'5'H in different tea groups. ( $n = 25$  for Group 1, 9 for Group 2, 28 for Group 3, 26 for Group 4, and 34 for Group 5). Boxes = interquartile ranges, middles = medians, whiskers =  $1.5 \times$  the interquartile range, single points = outliers. Source data underlying **b, c**, and **d** are provided as a Source Data file.

TCS (TEA010054), which encodes an *S*-adenosyl-L-methionine (SAM)-dependent *N*-methyltransferase that catalyzes the methylation step to synthesize theobromine and caffeine, had a much higher expression level in group 1 than in other groups (Fig. 5c). However, no significant expression change was detected for the predominate TCS (TEA015791). TEA013315, with the highest expression level among all F3'5'H genes in tea plants, was expressed at the lowest level in the CSA lineage, and may have contributed to the apparent differential accumulations of catechin-derived metabolites between CSS and CSA (higher levels of C/EC/ECG in CSA, and a higher level of EGC in CSS) (Fig. 5d). However, the expression levels of genes related to caffeine or theanine biosynthesis did not show direct correlation with the concentration of respective metabolites (seven TCS were highly expressed in group 1, but group 1 did not have the highest caffeine concentration). Similarly, three genes involved in theanine synthesis, including arginine decarboxylase and two glutamate dehydrogenase genes, were highly expressed in group 2, yet group 2 did not have the highest theanine concentration (Figs. 4e, f and 5b), suggesting that much of the regulation of metabolite levels may occur post-transcriptionally.

## Discussion

The rich constituents of specialized metabolites in the growing tea leaf are believed to be essential for the flavor and quality of tea products<sup>12,18,19,35,36,51</sup>. Therefore, tea plant offers a good model to study the molecular and genetic basis underpinning the abundance, diversity, and regulation of specialized metabolites in plants. By analyzing transcriptomic and metabolomic data from 136 representative tea accessions in China, we were able to classify these accessions into five phylogenetic groups/populations, identify over 8000 polymorphic markers that can be used

for marker-assisted breeding, explore the dynamic variations in metabolite compositions and gene expression, and identify dozens of signature metabolites that are highly accumulated in one group of tea accessions but not in other groups. Our results show that there exists a high level of metabolite diversity in different tea populations and accessions, which can be explored to investigate the underlying regulatory mechanisms and guide molecular breeding for tea improvement.

With the transcriptomic data, we were able to identify 925,854 high-quality SNPs, providing a rich set of molecular markers that may be useful for marker-assisted breeding. Phylogenetic and population structure analyses showed that tea cultivars in China may be grouped into five populations, which is in general agreement with the results from previous studies using chloroplast DNA and nuclear microsatellite markers<sup>34</sup>. Our phylogenetic and population structure analyses showed that accessions from similar geographical origin, with similar morphological characteristics (e.g., large-leaved vs small/middle-leaved), or similar breeding/domestication history tended to cluster together nicely, as expected (Fig. 1b). With the genome-wide SNP markers that are mostly located within or near gene-encoding regions, our study not only provides reliable evaluation of genetic relationships and distances, but also offers a list of markers that may change the encoded protein sequences and markers with major alleles that are specific to certain tea groups, which are ideal markers for breeding practice. In a recent study, Xia et al.<sup>35</sup> used SNP markers genome-resequencing data separated 81 accessions into three clades (CSS, CSA, and wild tea), which disagrees with our conclusion or results using SSR markers<sup>34</sup>. This is likely because only a limited number of accessions (81) are analyzed, among which only 58 were cultivated accessions and thus the population did not fully represent the genetic diversity of natural tea populations. Another recent study used SNP markers from

transcriptome data to separate more than 212 accessions into five subpopulations<sup>51</sup>, which is consistent with our phylogenetic analysis and results using SSR markers. However, many accessions were not grouped together by geographical origin or morphological characteristics such as leaf size and there was no distinction between wild tea accessions and cultivated accessions<sup>51</sup>. These results are not consistent with our data. This is likely caused by a small number of SNP markers, derived from a smaller amount of transcriptome data, being used for phylogenetic analysis<sup>51</sup>.

Our untargeted metabolomics data show that thousands of metabolic features can be detected in fresh tea leaves, in addition to the well-known catechins, caffeine, and theanine. However, the majority of these detected metabolic features were lowly accumulated, with only 25–28% having a relative abundance higher than 500 in at least one examined accession. After careful annotation and manual curation of highly accumulated metabolic features (753 and 503), we found that 74% and 45% of them were fragment ions that were generated by the mass spectrum fragmentation process under POS and NEG mode, respectively, and thus were not natural metabolites in tea plants. After removing fragment ions, the identities for 179 and 258 abundantly detected tea metabolites could be confidently assigned or putatively assigned, representing the largest metabolite identification effort made in tea plants to date. While flavanols are the most abundantly occurring group of phenolic compounds in fresh tea leaves, flavonol glycosides and proanthocyanidins emerge as the most diverse ones. In particular, dozens of kaempferol and quercetin derivatives, with some not being reported in fresh tea leaves previously, are found to vary both in structures and concentrations across different groups of tea accessions. It is increasingly recognized that tailoring enzymes such as those catalyzing glycosylation, acylation, and methylation make a greater contribution to the structural variations of flavonoid metabolites. Their structural diversities in turn determine their biological activities, which are often implied in conferring tolerance to various stresses<sup>52,53</sup>. The natural variations of flavonol glycosides in tea leaves, many of which are observed to be heavily decorated by various sugars as well as coumaric acid, are presumed to be ascribed to the differential activities of specific UDP-dependent glycosyltransferases and acyltransferases yet to be functionally characterized<sup>52,54</sup>. Unraveling candidate genes responsible for flavonol decoration and teasing out which enzymes are functionally important will shed light on flavonoid biosynthesis in tea plants.

Our comparisons of metabolites from different tea groups suggest that the CSA tea type has a distinct metabolite profile from that of CSS tea type, resulted from natural and/or agronomic selection. It is well-known that tea plants contain high level of catechins, among which EGCG and EGC have the highest accumulation, followed by ECG, EC, EC, and C<sup>20,55,56</sup>. Our comparison of metabolite contents in different groups of tea plants show that the CSA tea accessions have higher accumulation of diverse classes of flavonoids (e.g., C, EC, GC, ECG, flavanols, flavonol glycosides and procyanidin dimers) and derivatives of gallic acid and quinic acid, with relatively lower level of EGCG. These results are in agreement with a previous targeted metabolomic analysis of catechin contents in 403 Chinese tea germplasms<sup>50</sup>. Green tea accessions contain lower levels of catechin compounds and relatively higher levels of two triterpenoid saponins and galactosylated derivatives of kaempferol/quercetin glycosides. During green tea processing, major catechins from young leaves of *C. sinensis* remain unoxidized. It is generally believed that a lower ratio of total polyphenols to amino acids in fresh leaves is essential to balance the astringent and the mellow tastes, and hence a prerequisite for producing premium

green teas<sup>57,58</sup>. Oolong tea accessions are enriched with two methylated catechins and complex kaempferol/quercetin glycoside derivatives acylated with a coumaroyl group. Interestingly, the study by Lv et al.<sup>59</sup> also suggests that oolong tea cultivars may be a good source for finding tea cultivars with higher methylated catechins. The extensive variations of catechins and some other metabolites that were revealed by this study suggest that metabolic profiles may be used to distinguish tea cultivars and metabolic markers may be used to assist tea breeding. On the other hand, environmental factors are known to greatly affect the accumulation levels of specialized metabolites. Future studies should determine whether signature metabolites are quantitatively affected by a specific environmental factor, as this information may help to determine growth conditions that optimize the production of a desired metabolite.

What are the underlying molecular mechanisms for the apparent differential accumulations of catechin compounds in different tea groups? Although we detected 31 structural genes in the catechin biosynthesis pathway that were differentially expressed in different tea groups (Fig. 5b), direct correlation between gene expression level and metabolite level is not obvious. For example, the anthocyanin reductase (ANR) is responsible for converting delphinidin to EGC and cyanidin to EC, but the ANR gene is not expressed at a higher level in CSA than in CSS tea accessions. Additionally, many structural genes have multiple copies in the genome and different copies in the same gene family may display different expression patterns. For example, the anthocyanidin synthase (ANS) is responsible for converting leucocyanidin to cyanidin and leucodelphinidin to delphinidin and we found that two ANS genes (TEA010322 and TEA015762) displayed opposite expression patterns with TEA010322 being highly expressed in CSA tea accessions and TEA015762 being highly expressed in green tea and oolong tea accessions (Fig. 5b). Similar pattern was also observed for the two LAR genes (TEA026458 and TEA027582) that encode the leucoanthocyanidin reductases that are responsible for converting leucocyanidin to C and leucodelphinidin to GC. On the other hand, caffeine did not show significant change in accumulation levels in different tea groups (Fig. 4f), indicating that it is an integral part of metabolites for any tea products and traditional breeding by crossing different tea cultivars is not effective in changing their concentration. Nevertheless, we found that nine genes in the caffeine biosynthetic pathways that were differentially expressed (Fig. 5b), again suggesting no direct correlation between gene expression and metabolite level. These results suggest that much of the regulation of metabolite levels may not occur at the transcriptional level. Further studies are needed, probably through genome-wide correlation analyses between metabolite concentrations and gene expression levels or molecular markers, to identify key regulators for metabolite production<sup>9</sup>.

## Methods

**Sample collection.** We collected the fully expanded second leaves from the young shoots (one bud with two leaves) of 136 representative tea accessions (belonging to 128 cultivars) grown in major tea-growing regions (e.g., Fujian, Zhejiang, and Yunnan Provinces) in China from April 13th to 25th, 2018 (Fig. 1 and Supplementary Table 1). For each tea accession, three biological replicates were prepared for RNA-sequencing and five biological replicates were prepared for metabolomics analysis with each replicate representing a pool of leaf samples collected from 15–20 individual tea plants of the same accession. Fresh tea leaves were immediately frozen in liquid nitrogen, brought back to the laboratory and stored at –80 °C until further analysis.

**RNA-sequencing and RNA-seq data analysis.** Total RNA was extracted using the CTAB (BBI Life Sciences, Shanghai, China) and PBIOZOL (Bioer, Hangzhou, China) reagents according to the manufacturer's protocol. RNA concentration and integrity were examined with the Agilent Bioanalyzer 2100 system (Agilent, CA, USA). Oligo (dT) beads were used to isolate poly(A)-containing mRNAs, which

were fragmented into ~250 bp fragments. cDNA libraries were constructed according to the standard protocol from Beijing Genomics Institute (Shenzhen, China) and paired-end 100 bp reads were generated on a BGISEQ-500 platform with a depth of approximately 5 GB clean data per sample. Transcriptomes from four wild relatives of tea plants in genus *Camellia*, including *C. japonica*<sup>60</sup>, *C. azalea*<sup>61</sup>, *C. nitidissima*<sup>62</sup>, and *C. reticulata*<sup>63</sup> were downloaded from the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>) to be used in the current study.

Raw RNA-seq reads were processed with SOAPnuke<sup>64</sup> to remove low-quality regions and adapter sequences. Clean reads were mapped to the CSS reference genome (downloaded from <http://pcsb.ahau.edu.cn:8080/CSS>) using hisat2<sup>65</sup> and gene expression levels were summarized by HTseq-count<sup>66</sup>. Raw counts were then normalized to counts per million (CPM) and genes with CPM < 1 in 90% samples were regarded as lowly-expressed genes and were removed from further analysis. Normalized gene expression was log<sub>2</sub>-transformed and used for clustering analysis with *t*-SNE in R version 3.5.1. Differentially expressed genes among five groups of tea varieties were identified by performing pairwise comparisons using edgeR<sup>67</sup> with significance thresholds of false discovery rate (FDR) < 0.05 and fold-change > 2. Approximately 40–120 million reads for each sample were uniquely mapped to the reference genome. To minimize the effect of library size on quantification of genome expression patterns, the total uniquely mapped reads larger than 80 million were then down-sampled to 80 million reads with GATK v4.0.4.0<sup>68</sup>.

**Evolutionary analyses.** To identify SNPs among the collected tea varieties, clean reads were further processed to filter PCR duplicates, and retained reads were used to call variants following the mapping process with GATK v4.0.4.0<sup>68</sup>. The HaplotypeCaller function was then used to generate a GVCF file for each accession, followed by population variant calling with the function GenotypeGVCFs. Hard filtering was applied to the raw variant set using GATK, with parameters “QD < 2.0 | FS > 60.0 | MQ < 60.0 | MQRankSum < -12.5 | ReadPosRankSum < -8.0” to obtain high-quality SNPs. Redundant SNPs were discarded such that candidate SNP loci were more than 5 bp away from each other. Only biallelic SNPs with a minor allele frequency larger than 0.05 and missing rate less than 20% in all samples were retained as final candidate SNPs for further analysis. Candidate SNPs in coding regions were further classified into synonymous SNPs and non-synonymous SNPs with ANNOVAR<sup>69</sup>. A major allele for an SNP in each tea group is defined as the allele with a frequency of at least 0.75 in the group.

Only non-missing SNPs at the fourfold-degenerate sites were selected to estimate genetic distances across all samples. An approximate maximum likelihood phylogenetic tree was constructed with 45,162 fourfold-degenerate SNPs using FastTree<sup>70</sup> with 1000 bootstrap replications. The wild tea (S159), together with the aforementioned wild relatives of tea plants, was used as the outgroups for rooting the tree.

The genetic relationship of 134 accessions was estimated using PCA performed by using PLINK v1.9<sup>71</sup>. Population structure was inferred with STRUCTURE<sup>44</sup>. To determine the optimal number of populations, STRUCTURE was run 10 times and with 20,000 MCMC reps for each *K* (*K* = 2–9). The optimal *K* was estimated to be 5 with Harvester<sup>44</sup>.

Based on the high-quality SNPs identified in tea accessions, selective sweep regions were detected among five groups of tea varieties with XP-CLR<sup>45</sup>. XP-CLR estimated each scaffold in non-overlapping 10-kb windows with a 10-kb sliding-step to detect allele frequency differentiation between each two populations across each reference genome region. Adjacent windows with the highest XP-CLR scores (5%) were grouped into a single region and regions with the top 1% XP-CLR scores were considered as potentially selected sweeps. Nucleotide divergence ( $\pi$ ) in each group was also calculated in 10-kb sliding windows with 1-kb steps across the reference genome which aided in improving prediction accuracy. Only potential selective sweep regions that were identified by XP-CLR and had a top 50%  $\pi$  ratio were kept as candidate sweeps.

**Metabolomics analysis.** Metabolite extracts were prepared by adding 750  $\mu$ L of 70% methanol to 30 mg ( $\pm 0.5$  mg) of the ground and pre-lyophilized leaf samples as previously described<sup>20</sup>. The samples were spiked with 250  $\mu$ L of 200  $\mu$ g/m 2',7'-dichlorofluorescein as an internal standard. A 10  $\mu$ L aliquot was further diluted 100-fold with 70% methanol and filtered through a 0.22  $\mu$ m polyvinylidene fluoride (PVDF) filter (Millipore, Billerica, MA, USA).

Data acquisitions were performed using an LC-MS system, which is a Waters Acquity UPLC system coupled in tandem to a Waters photodiode array (PDA) detector and a SYNAPT G2-Si HDMS QTOF mass spectrometer (Waters, Manchester, UK). Gradient elution was achieved on a Waters Acquity UPLC HSS T3 column (100  $\times$  2.1 mm, 1.8  $\mu$ m) with water containing 0.1% formic acid (solvent A) and acetonitrile containing 0.1% formic acid (solvent B) at a flow rate of 0.3 mL/min. The column temperature was maintained at 40 °C. The gradient elution program was as follows: 1–7% B (0–2 min), 7–40% B (2–13 min), 40–60% B (13–17 min), immediately elevated to 99% B (17 min), held at 99% B (17–22 min) and allowed to equilibrate for a further 3 min prior to the next injection. The last 8 min of the chromatogram solutions were discarded. The injection volume was 1  $\mu$ L. MS data were recorded using a QTOF mass spectrometer with an ESI source and operated in both the positive and the negative modes. The MS data were acquired in continuum mode using ramp collision energy from 10 to 50 eV. The following

MS parameters were applied: capillary voltage, 2.5 kV (ESI<sup>+</sup>) and 2.0 kV (ESI<sup>-</sup>); cone voltage, 40 eV; collision energy, 4 eV; source temperature, 120 °C; desolvation temperature, 450 °C; cone gas flow, 50 L/h; desolvation gas flow, 800 L/h; *m/z* range, 50–1200 Da. Quality control (QC) samples were prepared by pooling the equal amount of all second leaf samples and were injected every ten samples throughout the analytical run to check instrument performance. The instrument was operated under the control of the MassLynx software (ver 4.1, Waters, Milford, MA, USA).

Components eluting between 1 and 17 min from the UPLC-QTOF MS system were processed in Progenesis QI (v2.1, Nonlinear Dynamics, Newcastle upon Tyne, UK) for data preprocessing with default settings, except that each sample was normalized to the internal standard. Subsequent multivariate analyses, such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), were carried out by Progenesis QI extension EZinfo, following Pareto scaling. After manual inspection to remove outliers, the datasets including mass features and normalized peak area (relative abundance) were exported to Microsoft Office Excel for subsequent statistical analysis. Compound information obtained from Progenesis QI was used as the start point for manual metabolite identification. First, metabolites, where authentic standards were available, were verified by comparisons of their retention time and MS/MS fragmentations. When no authentic standards were found, tentative identification was made by comparing the mass spectra with those in online spectral databases of Metlin<sup>72</sup>, MassBank<sup>73</sup>, HMDB<sup>74</sup>, KnapSack<sup>75</sup>, and ReSpec<sup>76</sup> and verified with the literature information on similar compounds, especially for those that had been reported in tea. Collision-induced dissociation (CID) fragmentation of selected ions, if needed, was performed to confirm the structural assignment. UV spectra were used for identification whenever possible.

Outlier metabolite data were detected and discarded based on the median absolute deviation (MAD) method in the five replicates of each sample. Metabolites with relative abundance < 500 in all samples were regarded as lowly accumulated metabolites and were removed from further analyses. One-tailed Student's *t*-test was performed to identify differentially accumulated metabolites (DAMs) and the Benjamini–Hochberg (BH) correction was used to adjust *p*-values due to multiple comparisons. The metabolites with an adjusted *p*-value less than 0.05 and fold-change larger than 2 were regarded as DAMs. Using the total panel of metabolite values as the reference control, all data were normalized and then log<sub>2</sub>-transformed for *t*-SNE clustering analysis in R version 3.5.1.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

RNA-sequencing data that support the findings of this study have been deposited to the Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI) and are accessible with project number “PRJNA562973”. Metabolomics data have been deposited to the MetaboLights database<sup>77</sup> at the EMBL-European Bioinformatics Institute (EBI) with project number “MTBLS1405”. All other relevant data are available from the corresponding author on request. Source data are provided with this paper.

Received: 28 November 2019; Accepted: 14 October 2020;

Published online: 04 November 2020

## References

1. Fernie, A. R., Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5**, 763–769 (2004).
2. Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *Plant J.* **90**, 764–787 (2017).
3. Fang, C., Fernie, A. R. & Luo, J. Exploring the diversity of plant metabolism. *Trends Plant Sci.* **24**, 83–98 (2019).
4. Fernie, A. R. & Tohge, T. The genetics of plant metabolism. *Annu. Rev. Genet.* **51**, 287–310 (2017).
5. Biais, B. et al. Metabolic acclimation to hypoxia revealed by metabolite gradients in melon fruit. *J. Plant Physiol.* **167**, 242–245 (2010).
6. Chen, W. et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* **7**, 12767 (2016).
7. Tohge, T. et al. Characterization of a recently evolved flavonol-phenylacetyltransferase gene provides signatures of natural light selection in Brassicaceae. *Nat. Commun.* **7**, 12399 (2016).
8. Chan, E. K. F., Rowe, H. C., Corwin, J. A., Joseph, B. & Kliebenstein, D. J. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biol.* **9**, e1001125 (2011).



9. Zhu, G. et al. Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**, 249–261.e12 (2018).
10. Ashihara, H. & Kubota, H. Patterns of adenine metabolism and caffeine biosynthesis in different parts of tea seedlings. *Physiol. Plant* **68**, 275–281 (1986).
11. Chen, D. et al. Tea polyphenols, their biological effects and potential molecular targets. *Histol. Histopathol.* **23**, 487–496 (2008).
12. Stodt, U. W. & Engelhardt, U. H. Progress in the analysis of selected tea constituents over the past 20 years. *Food Res. Int.* **53**, 636–648 (2013).
13. Kowalsick, A. et al. Metabolite profiling of *Camellia sinensis* by automated sequential, multidimensional gas chromatography/mass spectrometry reveals strong monsoon effects on tea constituents. *J. Chromatogr. A* **1370**, 230–239 (2014).
14. Turkozu, D. & Şanlıer, N. L.-theanine unique amino acid of tea, and its metabolism, health effects, safety. *Crit. Rev. Food Sci.* **57**, 1681–1687 (2017).
15. Higdon, J. V. & Frei, B. Tea catechins and polyphenols: Health effects, metabolism, and antioxidant functions. *Crit. Rev. Food Sci.* **43**, 89–143 (2003).
16. Chacko, S. M., Thambi, P. T., Kuttan, R. & Nishigaki, I. Beneficial effects of green tea: a literature review. *Chin. Med.* **5**, 13 (2010).
17. Li, C.-F. et al. Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). *BMC Genomics* **16**, 560 (2015).
18. 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).
19. 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).
20. Chen, S. et al. Metabolite profiling of 14 Wuyi Rock tea cultivars using UPLC-QTOF MS and UPLC-QqQ MS combined with chemometrics. *Molecules* **23**, 104 (2018).
21. Zeng, L., Watanabe, N. & Yang, Z. Understanding the biosyntheses and stress response mechanisms of aroma compounds in tea (*Camellia sinensis*) to safely and effectively improve tea aroma. *Crit. Rev. Food Sci.* **59**, 2321–2334 (2019).
22. Kato, M. et al. Caffeine biosynthesis in young leaves of *Camellia sinensis*: In vitro studies on *N*-methyltransferase activity involved in the conversion of xanthosine to caffeine. *Physiol. Plant* **98**, 629–636 (2006).
23. Sasaoka, K., Kito, M. & Onishi, Y. Some properties of the theanine synthesizing enzyme in tea seedlings. *Agric. Biol. Chem.* **29**, 984–988 (1965).
24. Yu, Z. & Yang, Z. Understanding different regulatory mechanisms of proteinaceous and non-proteinaceous amino acid formation in tea (*Camellia sinensis*) provides new insights into the safe and effective alteration of tea flavor and function. *Crit. Rev. Food Sci.* **60**, 844–858 (2019).
25. Lai, X. J. & Schnable, J. C. Harnessing the potential of the tea tree genome. *Mol. Plant* **10**, 788–790 (2017).
26. Fraser, K. et al. Analysis of metabolic markers of tea origin by UHPLC and high resolution mass spectrometry. *Food Res. Int.* **53**, 827–835 (2013).
27. Lee, J. et al. Geographical and climatic dependencies of green tea (*Camellia sinensis*) metabolites: a <sup>1</sup>H NMR-based metabolomics study. *J. Agric. Food Chem.* **58**, 10582–10589 (2010).
28. Ji, H. et al. Metabolic phenotyping of various tea (*Camellia sinensis* L.) cultivars and understanding of their intrinsic metabolism. *Food Chem.* **233**, 321–330 (2017).
29. Tan, J. et al. Study of the dynamic changes in the non-volatile chemical constituents of black tea during fermentation processing by a non-targeted metabolomics approach. *Food Res. Int.* **79**, 106–113 (2016).
30. Dai, W. et al. Characterization of white tea metabolome: Comparison against green and black tea by a nontargeted metabolomics approach. *Food Res. Int.* **96**, 40–45 (2017).
31. Wang, Y. et al. Novel insight into the role of withering process in characteristic flavor formation of teas using transcriptome analysis and metabolite profiling. *Food Chem.* **272**, 313–322 (2019).
32. Li, C. F. et al. Comprehensive dissection of metabolic changes in albino and green tea cultivars. *J. Agric. Food Chem.* **66**, 2040–2048 (2018).
33. Chen, L., Apostolides, Z. & Chen, Z.-M. Global Tea Breeding: Achievements, Challenges and Perspectives (Springer-Zhejiang University Press, Hangzhou, China, 2012).
34. Yao, M.-Z., Ma, C.-L., Qiao, T.-T., Jin, J.-Q. & Chen, L. Diversity distribution and population structure of tea germplasms in China revealed by EST-SSR markers. *Tree Genet. Genomes* **8**, 205–220 (2012).
35. Xia, E. et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into genome evolution and adaptation of tea plants. *Mol. Plant* **13**, 1013–1026 (2020).
36. Chen, J.-D. et al. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Hortic. Res.* **7**, 63 (2020).
37. Montgomery, S. B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
38. Xu, C. et al. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. *Nat. Commun.* **8**, 15275 (2017).
39. Edger, P. P. et al. Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547 (2019).
40. Wang, P. et al. Evolutionary and functional characterization of leucoanthocyanidin reductases from *Camellia sinensis*. *Planta* **247**, 139–154 (2018).
41. Jin, J.-Q., Yao, M.-Z., Ma, C.-L., Ma, J.-Q. & Chen, L. Natural allelic variations of *TCSI1* play a crucial role in caffeine biosynthesis of tea plant and its related species. *Plant Physiol. Biochem.* **100**, 18–26 (2016).
42. Jin, J.-Q. et al. Hongyacha, a naturally caffeine-free tea plant from Fujian, China. *J. Agric. Food Chem.* **66**, 11311–11319 (2018).
43. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
44. Earl, D. A. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
45. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
46. Jin, J.-Q., Ma, J.-Q., Yao, M., 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 (2016).
47. Fang, S. et al. Geographical origin traceability of Keemun black tea based on its non-volatile composition combined with chemometrics. *J. Sci. Food Agric.* **99**, 6937–6943 (2019).
48. Zhou, P. et al. UPLC-Q-TOF/MS-based untargeted metabolomics coupled with chemometrics approach for Tieguanyin tea with seasonal and year variations. *Food Chem.* **283**, 73–82 (2019).
49. Zhang, Y. H. et al. Identification and characterization of *N9*-methyltransferase involved in converting caffeine into non-stimulatory theacrine in tea. *Nat. Commun.* **11**, 1473 (2020).
50. Jin, J., Ma, J., Ma, C., Yao, M. & Chen, L. Determination of catechin content in representative Chinese tea germplasms. *J. Agric. Food Chem.* **62**, 9436–9441 (2014).
51. Zhang, W. et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
52. Tohge, T., de Souza, L. P. & Fernie, A. R. Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. *J. Exp. Bot.* **68**, 4013–4028 (2017).
53. Peng, M. et al. Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. *Nat. Commun.* **8**, 1975 (2017).
54. Wang, S., Alseekh, S., Fernie, A. R. & Luo, J. The structure and function of major plant metabolite modifications. *Mol. Plant* **12**, 899–919 (2019).
55. Ashihara, H., Deng, W.-W., Mullen, W. & Crozier, A. Distribution and biosynthesis of flavan-3-ols in *Camellia sinensis* seedlings and expression of genes encoding biosynthetic enzymes. *Phytochemistry* **71**, 559–566 (2010).
56. Wei, K. et al. Catechin contents in tea (*Camellia sinensis*) as affected by cultivar and environment and their relation to chlorophyll contents. *Food Chem.* **125**, 44–48 (2011).
57. Li, X. et al. Brassinosteroids improve quality of summer tea (*Camellia sinensis* L.) by balancing biosynthesis of polyphenols and amino acids. *Front. Plant Sci.* **7**, 1304 (2016).
58. Zhang, Q., Liu, M. & Ruan, J. Integrated transcriptome and metabolic analyses reveals novel insights into free amino acid metabolism in Huangjiunya tea cultivar. *Front. Plant Sci.* **8**, 291 (2017).
59. Lv, H. et al. Analysis of naturally occurring 3''-methyl-epigallocatechin gallate in 71 major tea cultivars grown in China and its processing characteristics. *J. Funct. Foods* **7**, 727–736 (2014).
60. Li, Q. et al. RNA-seq based transcriptomic analysis uncovers alpha-linolenic acid and jasmonic acid biosynthesis pathways respond to cold acclimation in *Camellia japonica*. *Sci. Rep.* **6**, 36463 (2016).
61. Fan, Z. et al. Genome-wide transcriptome profiling provides insights into floral bud development of summer-flowering *Camellia azalea*. *Sci. Rep.* **5**, 9729 (2015).
62. Zhou, X. et al. De novo assembly of the *Camellia nitidissima* transcriptome reveals key genes of flower pigment biosynthesis. *Front. Plant Sci.* **8**, 1545 (2017).
63. Yao, Q. Y., Huang, H., Tong, Y., Xia, E. H. & Gao, L. Z. Transcriptome analysis identifies candidate genes related to triacylglycerol and pigment biosynthesis and photoperiodic flowering in the ornamental and oil-producing plant, *Camellia reticulata* (Theaceae). *Front. Plant Sci.* **7**, 163 (2016).
64. Chen, Y. et al. SOAPnucke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, 1–6 (2018).
65. Perte, M., Kim, D., Perte, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650 (2016).

66. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
67. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
68. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
69. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
70. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
71. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
72. Tautenhahn, R. et al. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat. Biotechnol.* **30**, 826–828 (2012).
73. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
74. Wishart, D. S. et al. HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
75. Afendi, F. M. et al. KNApSACK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **53**, e1 (2012).
76. Sawada, Y. et al. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* **82**, 38–45 (2012).
77. Haug, K. et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786 (2013).

## Acknowledgements

This work was supported by the Fujian Agriculture and Forestry University (FAFU) Construction Project for Technological Innovation and Service System of Tea Industry Chain (K1520005A02) and other funds from FAFU to X.Y., Y.Y. and R.L. We wish to thank Dr. Keke Chen, Dr. Changsong Chen, Shixian Chao, and Yinbi Cai for their assistance in collecting tea samples.

## Author contributions

Z.Y. and R.L. designed and coordinated the study. X.Y., Y.Y., S.C., J.M., J.L., Z.F., Q.Z., Q.C., and L.C. collected the tea leaf samples. X.Y., S.C., Y.L., and R.L. performed the metabolomics analyses. J.X., X.Y., S.C., and R.L. analyzed the data. R.L., J.X., X.Y., and Y.Y. wrote the manuscript with input from all authors. R.L., L.C., and Z.Y. edited the manuscript. All authors read and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-19441-1>.

**Correspondence** and requests for materials should be addressed to L.C., Z.Y. or R.L.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020, corrected publication 2021