Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2024) **22**, pp. 1001–1016 doi: 10.1111/pbi.14241

Depicting the genetic and metabolic panorama of chemical diversity in the tea plant

Haiji Qiu[1,2,3,†], Xiaoliang Zhang[1,†], Youjun Zhang[4,5,†], Xiaohui Jiang[1], Yujia Ren[1], Dawei Gao[1], Xiang Zhu[6], Björn Usadel[7,8], Alisdair R. Fernie[4,5,*] and Weiwei Wen[1,2,3,*]

[1]*National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Key Laboratory of Horticultural Plant Biology (MOE), College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, China*
[2]*Shenzhen Institute of Nutrition and Health, Huazhong Agricultural University, Wuhan, China*
[3]*Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China*
[4]*Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany*
[5]*Center of Plant Systems Biology and Biotechnology, Plovdiv, Bulgaria*
[6]*Thermo Fisher Scientific, Shanghai, China*
[7]*Institute of Bio- and Geosciences, IBG-4: Bioinformatics, CEPLAS, Forschungszentrum Jülich, Jülich, Germany*
[8]*Institute for Biological Data Science, Heinrich Heine University, Düsseldorf, Germany*

Received 6 July 2023;
revised 11 September 2023;
accepted 12 November 2023.
**Correspondence* (Weiwei Wen: Tel +86 (0) 27-87282010; fax +86 (0)27-87282010; email wwwen@mail.hzau.edu.cn; Alisdair R. Fernie: Tel +49 (0)331 5678259; fax +49 (0) 331 5678408; email fernie@mpimp-golm.mpg.de)
[†]Equal contribution.

Keywords: tea plant, chemical diversity, mGWAS, UDP-glycosyltransferase, caffeoyl-CoA *O*-methyltransferase.

Summary

As a frequently consumed beverage worldwide, tea is rich in naturally important bioactive metabolites. Combining genetic, metabolomic and biochemical methodologies, here, we present a comprehensive study to dissect the chemical diversity in tea plant. A total of 2837 metabolites were identified at high-resolution with 1098 of them being structurally annotated and 63 of them were structurally identified. Metabolite-based genome-wide association mapping identified 6199 and 7823 metabolic quantitative trait loci (mQTL) for 971 and 1254 compounds in young leaves (YL) and the third leaves (TL), respectively. The major mQTL (i.e., $P < 1.05 \times 10^{-5}$, and phenotypic variation explained (PVE) > 25%) were further interrogated. Through extensive annotation of the tea metabolome as well as network-based analysis, this study broadens the understanding of tea metabolism and lays a solid foundation for revealing the natural variations in the chemical composition of the tea plant. Interestingly, we found that galloylations, rather than hydroxylations or glycosylations, were the largest class of conversions within the tea metabolome. The prevalence of galloylations in tea is unusual, as hydroxylations and glycosylations are typically the most prominent conversions of plant specialized metabolism. The biosynthetic pathway of flavonoids, which are one of the most featured metabolites in tea plant, was further refined with the identified metabolites. And we demonstrated the further mining and interpretation of our GWAS results by verifying two identified mQTL (including functional candidate genes *CsUGTa*, *CsUGTb*, and *CsCCoAOMT*) and completing the flavonoid biosynthetic pathway of the tea plant.

Introduction

Tea is one of the most popular beverages around the world and more than 2 billion cups of tea are consumed per day (Brody, 2019). Records of *Camellia sinensis*, the leaves of which are major sources for processing tea, date back to some 3000 years ago in China. Currently, as an important and economic crop, the tea plant is cultivated in many countries worldwide (Xia *et al.*, 2017), with plantations covering 5.08 million hectares and with 6.4 million tons tea being produced in 2019 (http://www.fao.org). As it is frequently consumed worldwide, different fields of research related to tea, including plant science, food science, chemistry and medicine have been extensively conducted (Inoue-Choi *et al.*, 2022).

The fresh leaves of tea plant are rich in diverse metabolites, including amino acids, flavonoids, phenolic acids, volatiles and their derivatives (Chen *et al.*, 2020b; Fu *et al.*, 2021; Jing *et al.*, 2019; Li *et al.*, 2022b). The metabolites in fresh leaves are the important foundation of tea quality and flavours (e.g., astringency, bitterness etc.) and bioactive health beneficial properties after tea processing. Besides, these metabolites are associated with plant growth and development as well as protection against biotic and abiotic stresses (e.g., cold, drought, pathogens, ultraviolet radiation etc; Zeng *et al.*, 2020). However, little research concerning the comprehensive profiling of the tea metabolome is currently available, despite the fact that this would almost certainly deeply enrich our insight into the composition of its metabolites as well as providing understanding into the underlying biological mechanisms of their bioactivities.

With the development of mass spectrometry (MS) platforms and sequencing technology, metabolome-based genome-wide association studies (mGWAS) has been successfully applied to exploit natural genotypic variation and identify metabolic quantitative trait loci (mQTL) responsible for metabolomic variation in humans (Hagenbeek *et al.*, 2020; Qin *et al.*, 2012; Zhang *et al.*, 2015), animals (Hartiala *et al.*, 2016) and plants

(Alseekh *et al.*, 2018; Francisco *et al.*, 2016; Slaten *et al.*, 2020; Wang *et al.*, 2021b; Wen *et al.*, 2014; Zhu *et al.*, 2018). This approach is effective and highly complementary to QTL mapping based on bi- or multi-parental derived breeding populations (Li *et al.*, 2019; Shi *et al.*, 2020). Indeed, such studies have greatly deepened our understanding of the mechanisms underlying resistance of and tolerance to biotic and abiotic stresses, respectively. They have additionally revealed metabolic changes associated with changes in flavour and organ size during the domestication or crop improvement processes (Alseekh *et al.*, 2021; Kettunen *et al.*, 2012; Zhu *et al.*, 2018). Given the importance of tea and the metabolic significance for the broad use of tea plant, it is vital to uncover the mystery of tea metabolic diversity as well as the genetic basis which underlies it (Qiu *et al.*, 2020; Zhao *et al.*, 2020). The diverse tea genetic resources, coupled with metabolite detection methodologies, enable the identification of mQTL which would further lead to better understanding of the natural variation of metabolites in tea plant. However, previous studies using tea germplasm to map mQTL have merely been pursued on a handful of metabolites or inadequate genotypic data (Fang *et al.*, 2021; Hazra *et al.*, 2021; Huang *et al.*, 2022; Yamashita *et al.*, 2020; Yu *et al.*, 2020; Zhang *et al.*, 2020b, 2021). As such a large-scale combined metabolomics and population genetics study in tea research can yield significant insights into the genetic and metabolic landscape of tea chemical diversity.

In this study, for the first time, a large-scale targeted metabolomic study with the combination of high resolution and accurate quantification was performed using a diverse tea germplasm collection containing 215 genotypes. Following the characterization of thousands of metabolites and metabolomic quantification for each tea genotype, we performed mGWAS using data collected from two tissues of the diverse tea germplasm. Additionally, we constructed a metabolic network and identified genome-wide mQTL to uncover the metabolic diversity of the tea plant. The flavonoid biosynthesis of tea plant was further updated with the identified metabolites and verified functional candidate genes.

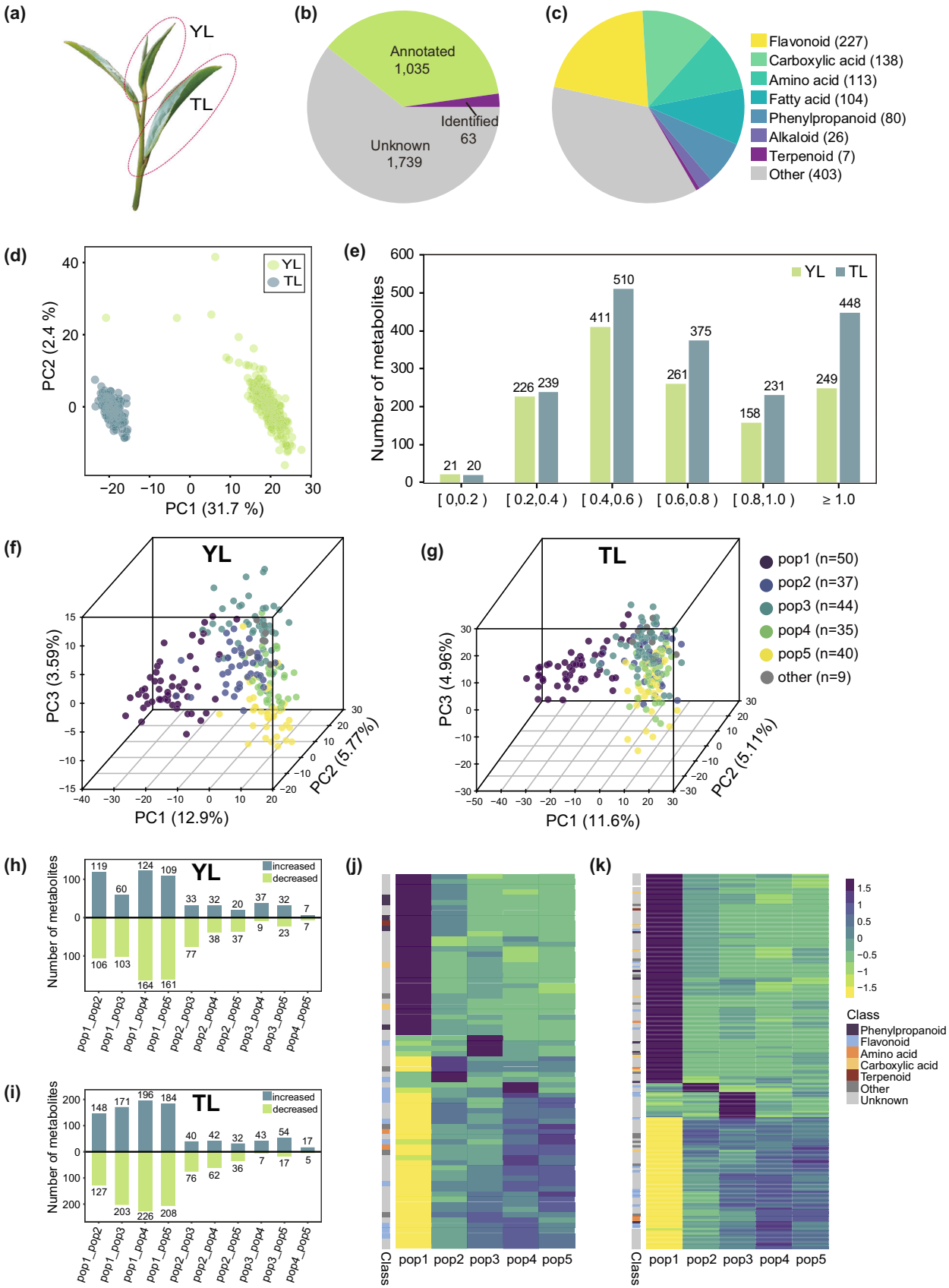
Results

Metabolic variation of different types of tea leaves of diverse tea germplasm

Two different leaf samples (YL and TL, Figure 1a) were collected from 215 tea genotypes and metabolic profiling performed using a pooled sample led to the detection of 2837 metabolites. Among these, 1098 were structurally annotated and categorized as follows: 227 flavonoids, 138 carboxylic acids, 113 amino acids, 80 phenylpropanoids, 104 fatty acids, 26 alkaloids, 7 terpenoids and 403 miscellaneous compounds. A total of 63 of these compounds were structurally identified by comparison to authentic standards, including amino acids, flavonoids, and coumarins (Figure 1b,c; Data S1). 1326 metabolites in YL and 1823 in TL were quantified across 215 samples with 1206 metabolites detected in both tissues. A high correlation between YL and TL was found in the content of these 1206 metabolites with an average Pearson correlation coefficient (PCC) of 0.515 (Figure S1a). The content of phenylpropanoids and flavonoids display higher correlation between tissues than that of amino acids, fatty acids, carboxylic acids, and others (Figure S1b). Partial least squares discriminant analysis (PLS-DA) revealed marked tissue specificity of metabolite content, and 34.1% metabolic variation was explained by the top two principal components (Figure 1d). The distributions of coefficient of variation (CV) across all metabolites uncovered a CV of over 0.4 for more than 81.37% and 85.79% of the metabolites in YL and TL, respectively (Figure 1e), indicating considerable natural variation among the panel in both tissues.

According to our previous study, these tea accessions were divided into five subpopulations based on the phylogenetic and population structure analyses (Zhang *et al.*, 2020b). The 3D scatter plot from the PLS-DA score matrix indicated that 22.26% and 21.67% variation in metabolite abundance were explained by the top three principal components in YL and TL, respectively (Figure 1f,g). The top two principal components (PC1 and PC2) significantly (ANOVA-Tukey's test, $P < 0.05$) distinguished subpopulation one (pop1), pop2, pop3, pop4-5 from each other and the third principal component (PC3) can significantly separate pop4 from pop5 in YL (Figure 1f; Figure S2a). In TL, PC1 significantly (ANOVA-Tukey's test, $P < 0.05$) distinguished five subpopulations into three groups (pop1, pop2, pop3-4-5), with PC3 separating pop3 from pop4-5 and only PC4 distinguishing pop5 from the other subpopulations (Figure 1g; Figure S2b). To identify metabolites differently accumulated in different subpopulations, pairwise comparisons between five subpopulations were conducted. In summary, 475 and 681 metabolites exhibited differential accumulation in any two subpopulations in YL and TL, respectively, with 286 differentially accumulated metabolites (DAMs) overlapping in YL and TL (Figure 1h,i). Upon clustering analysis of the DAMs, 72 and 163 signature metabolites (i.e., metabolites significantly high or low accumulated in one subpopulation in comparison with the other four subpopulations) were found in YL and TL, respectively, and 41 signature metabolites were found in both tissues. Interestingly, 88.89% (64/72) of these metabolites in YL and 90.80% (148/163) of them in TL were significantly accumulated in pop1 (Figure 1j,k). Five signature metabolites, two in YL and three in TL, were identified by authentic standards. Caffeic acid and (+)-catechin exhibited high accumulation in pop1, whereas eriodictyol-7-*O*-glucoside, vitexin-2″-*O*-rhamnoside and naringenin 7-*O*-neohesperidoside displayed low accumulation (Figure S3).

Figure 1 Metabolic profiling in tea plant population. (a) The phenotype of the young leaves (YL) and third leaves (TL) used in the study. (b) Pie graph of identified, annotated and unknown metabolites detected in this study. (c) The number and classification of annotated and identified metabolites. (d) Score plot of PLS-DA results of 1206 metabolic profiles in both YL (green) and TL (indigotin) among 215 samples, each point represents one sample. (e) Coefficient of variation of metabolites detected in YL and TL. (f, g) 3D scatter plot of PLS-DA results of metabolic profile in YL (f) and TL (g), the points in the f-g indicate the accessions of tea and different colours display different subgroups. (h, i) Number of differential metabolites quantified in YL (h) and TL (i) in pairwise comparisons among any two sub-populations mentioned above. (j, k) Heatmap of metabolites significantly high or low accumulated in one sub-population than the other four sub-populations in YL (j) and TL (k), respectively, (One-way ANOVA with Tukey's test, $P < 0.05$).



Metabolite network construction to broaden the horizon of tea metabolome

To gain more insights into the tea metabolome, the MS/MS relationship between each metabolite pair was further investigated. On one hand, 702 metabolite isomer pairs pass the threshold (extracted mass ≤2 ppm and five fragment pairs ≤ 2 ppm) (Figure 2a), corresponding to 840 metabolites. The distribution of the PCC between isomer metabolites range from −0.37 to 0.99 with an average of 0.54, and the absolute value of the PCC was highly correlated with the number of fragments within 2 ppm between metabolite pairs (Figure 2b). Metabolite pairs share the same metabolite were considered as an isomer group, resulting in the formation of 343 isomer groups corresponding to 840 metabolites. The number of metabolites in one isomer group varies from 2 to 7 (Data S2). 46 unknown metabolites obtained putative annotation based on the metabolite isomer analysis, including 6 carboxylic acids, 2 amino acids, 11 flavonoids, 4 fatty acids, 5 phenylpropanoids and 18 other metabolites (Figure 2a).

On the other hand, the relationship of any metabolites could be the substrates or products through known enzymatic reactions; thus, a candidate substrate-product pairs (CSPP) network was constructed in this study. A total of 2079 CSPPs were obtained, corresponding to 1242 metabolites (Figure 2c). Any two CSPPs with overlapping metabolites were clustered into a group. The network encompassed 192 groups, with the number of groups ranging from two to 140 (Data S3). The chemical conversions within these CSPPs comprised 313 galloylations, 300 hydroxylations, 154 hydrations, 129 hexosylations, 119 methylations, 116 methoxylations, 84 reductions, 61 catecholizations, 60 pentosylations, 59 rhamnosylations, 52

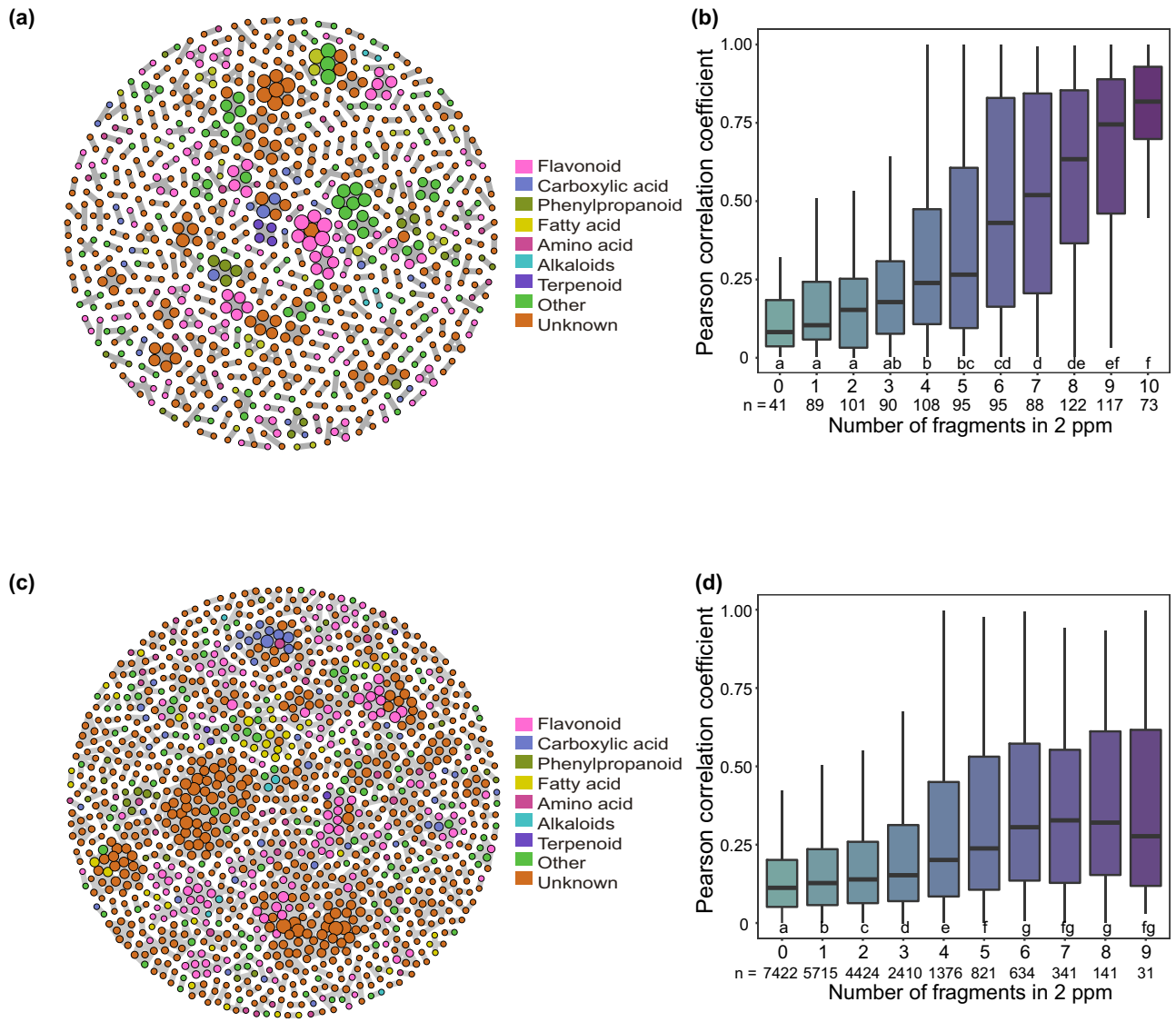


Figure 2 Putative isomers network and candidate substrate-product pairs network. (a) Putative isomers network based on no less than five fragments shared by metabolite pairs in 2 ppm. (b) Boxplot of pairwise Pearson correlation coefficient between putative isomer pairs, the number under the axis line indicates the number of fragments shared by the metabolite pairs in 2 ppm, different letters indicate significant difference in One-way ANOVA with Tukey's test ($P < 0.05$). (c) Candidate substrate-product pairs network based on no less than five fragments shared by metabolites pairs in 2 ppm. (d) Boxplot of pairwise Pearson correlation coefficient between candidate substrate-product pairs. The number under the axis line indicates the number of fragments shared by the metabolite pairs in 2 ppm. Different letters below indicate significant difference in One-way ANOVA with Tukey's test ($P < 0.05$).

coumaroylations (Table S1). Consequently, the largest six chemical conversions are galloylations, hydroxylations, hydrations, hexosylations, methylations and methoxylations. In total, 303 unknown metabolites were further annotated based on the CSPP network, including eight alkaloids, 11 amino acids, 29 carboxylic acids, 29 fatty acids, 116 flavonoids, 29 phenylpropanoids and 81 other metabolites. Metabolites within the same class tend to group together (Figure 2c). The PCC between CSPPs ranges from −0.55 to 0.99 with an average of 0.31, and unsurprisingly, the absolute value of the PCC was highly correlated with the number of fragments within 2 ppm between metabolite pairs (Figure 2d). The metabolite isomer and CSPP analysis aid in the annotation of unknown metabolites, thereby broadening the horizon of tea metabolome.

Identification of metabolome associated QTL across the tea genome

Genome-wide association studies were performed using the 1326 metabolites and 1823 metabolites mentioned above in YL and TL, respectively. In total, we obtained 6199 mQTL corresponding to 971 metabolites in YL and 7823 mQTL for 1254 metabolites in TL (Figure 3a,b; Data S4). The number of QTL associated with each metabolic trait ranged from 1 to 95 in YL and from 1 to 92 in TL, respectively. On average, there were 6.38 significant loci per trait in YL and 6.23 significant loci per trait in TL (Table 1). Most of the metabolites (68.69% in YL and 68.02% in TL) were associated with multiple QTL (Figure 3c), suggesting complex genetic basis for the metabolomic variation. Phenotype variation explanation of mQTL ranged from 8.7% to 83.61% with an average of 11.65% in YL. And 8.7% to 75.9% metabolic variation was explained by the mQTL in TL with an average of 11.6% (Table 1; Data S4). PVE of 31 mQTL (0.5%) in YL and 60 mQTL (0.77%) in TL were more than 25%. There were 0 to 73 mQTL in YL and 0 to 77 mQTL in TL in 1 Mb sliding windows with a step of 100 Kb, respectively (Figure S4). After 1000 random permutations by assigning mQTL to the tea genome, we defined a hotspot region as a 1 Mb sliding window with the number of mQTL greater than 8 in YL and 9 in TL ($P < 0.01$). Combining the overlapped or adjacent window, 185 and 200 hotspots were detected in YL (Figure 3d) and TL (Figure 3e), respectively. There were 33 hotspots detected in both YL and TL. These 33 hotspots included 680 mQTL associated with 346 metabolites in YL and 866 mQTL associated with 408 metabolites in TL (Data S4). Notably, 55 flavonoids and 16 phenylpropanoids were associated with these 33 hotspots (Data S4). Besides, 32 amino acid derivatives, 27 carboxylic acid and derivatives, six fatty acid and derivatives, four alkaloids, two terpenoids, 63 others and 388 unknown metabolites were associated with the 33 hotspots. KEGG enrichment analysis showed that genes in seven (7/33) hotspots were enriched in flavonoid biosynthesis. Furthermore, the relationship between CSPPs and the genes in the mQTL of associated metabolites was explored. Within the CSPPs, 20 metabolites related to glycosylation associated with mQTL harbouring UDP-glycosyltransferases (Figure S5a); 11 metabolites related to methylation associated with mQTL harbouring *O*-methyltransferases (Figure S5b); 52 metabolites related to acylation associated with mQTL harbouring acyltransferase (Figure S5c) and 45 metabolites related to hydroxylation associated with mQTL harbouring CYP450 (Figure S5d). The mQTL with large phenotypic variation explanation would be further investigated in priority for candidate gene identification (Data S4). We then mined candidate genes from the loci with PVE more than 25% and 24 candidate genes were assigned, including two UDP-glycosyltransferases, two *O*-methyltransferases and 20 transcription factors. Interestingly, the UGTs and *O*-methyltransferases were associated with several flavonoids, indicating that the genes play an important role in flavonoid biosynthesis (Table 2; Data S4).

Toward complete understanding of flavonoid biosynthesis in tea plant

Flavonoids are the largest class of annotated metabolites in this study (Figure 1). A proposed pathway of flavonoid biosynthesis was constructed based on the identified metabolites in this study (Figure 4). In summary, various flavonoids were formed through simple enzymatic reactions based on skeleton metabolites (e.g., phenylalanine, naringenin chalcone, naringenin, apigenin, eriodictyol, kaempferol, and leucopelargonidin), involving glycosylations, hydroxylations, galloylations, methylations, and other modifications. Four flavonoids were newly identified in fresh leaves of tea plant and 33 flavonoid associated candidate genes were identified by mGWAS, which provide new genetic clue for further exploring the flavonoid biosynthesis (Figure 4). Additionally, a conversion network was constructed using 37 metabolites involved in flavonoid biosynthesis as bait. This network comprised 19 subgroups, 115 nodes, and 108 CSPPs, encompassing glycosylations, acylations, hydroxylations, and methylations, thereby expanding the understanding of flavonoid biosynthesis in the tea plant (Figure S6).

A remarkable QTL on chromosome 1 was significantly associated ($P < 1.05 \times 10^{-5}$) with the natural variation of multiple flavonoids (Figure 5a,b; Data S4). The locus contained five genes and two of them were UDP-glycosyltransferases (W01g002927 and W01g002928, named as *CsUGTa* and *CsUGTb* here after; Figure 5c,d) belonging to the UGT708C subfamily (Figure 5e). Both *CsUGTa* and *CsUGTb* exhibited high expression levels in root, bud and the first leaf (Figure S7a, b). There are five non-synonymous SNPs in the coding sequence of *CsUGTa* among the association panel corresponding to seven alleles (C/A/A/C/A, T/G/AG/G/A, T/G/G/G/A, T/G/G/G/AC, TC/A-G/AG/G/A, TC/AG/AG/GC/A and TC/AG/AG/GC/AC). The C/A/A/C/A allele (*CsUGTaH*) corresponds to samples with high metabolite levels, while the T/G/G/G/A allele (*CsUGTaL*) corresponds to those with low metabolite levels (Figure 5f; Figure S8a). There are two non-synonymous SNPs in the coding sequence of *CsUGTb* among the association panel corresponding to four alleles (A/C, C/G, CA/G and CA/GC). The A/C allele (*CsUGTbH*) corresponds to samples with high metabolite levels, while the C/G allele (*CsUGTbL*) corresponds to those with low metabolite levels (Figure 5g, Figure S8b). We selected two alleles with high (allele H) and low (allele L) levels of metabolic contribution for further functional validation, respectively.

Functional validation of *CsUGTa* and *CsUGTb* were carried out in both tobacco and *in vitro* enzyme assay by using an *ex vivo* metabolomics approach (Feussner and Feussner, 2020; Zhang *et al.*, 2020b). Using the GFP as a negative control, both alleles of each gene were transiently expressed in tobacco leaves. Both alleles of *CsUGTa* and *CsUGTb* affected the *in vivo* metabolite profiles of the over-expressed transgenic tobacco lines when compared with the GFP control *in vivo* (Table S2). In addition, the *CsUGTa* and *CsUGTb* proteins were purified by affinity purification and incubated with tea extract for 30 min in an *ex vivo* metabolomics approach (Table S3).

As a result, several metabolites were significantly changed by the *CsUGTa* and *CsUGTb* by overexpression in tobacco

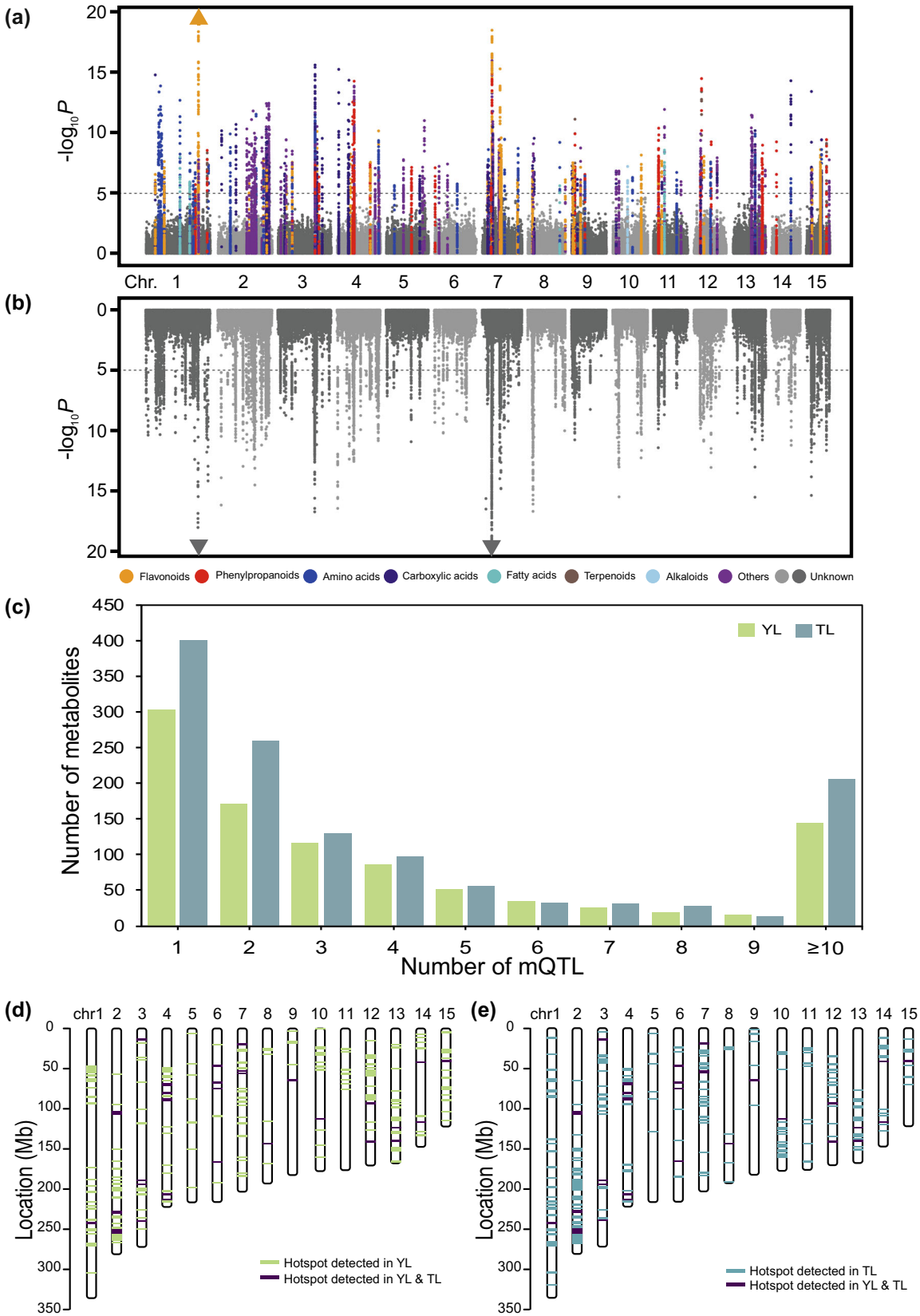


Figure 3 Summary of mQTL in YL and TL. (a, b) Manhattan plots of mGWAS results. The upper plot indicates the mGWAS results of known metabolites, and the bottom plot indicates the mGWAS results of unknown metabolites. Dots in different colours represent different metabolite classes. (c) Distribution of the number of mQTL associated with a metabolite. (d, e) Genome-wide distribution of mQTL hotspots detected in YL and TL, respectively.

Table 1 Summary of genome-wide significant associations in this study

Item	YL	TL
Number of traits	971	1254
Number of loci	6199	7823
Average loci per trait	6.38	6.23
Maximum explained variation (%)	83.61%	75.9%
Average explained variation per SNP (%)	11.65%	11.60%

leaves, including increased di-p-coumaroyl-di-glucoside, naringin dihydrochalcone, 2,3-dihydroxybenzoic acid 5-*O*-D-glucoside, lisorhamnetin-3-*O*-rutinoside, dihydroxybenzoic acid, trans-caffeic acid derivative, *p*-coumaroyl-cinnamoyl-glucoside, galloylglucose, 4-coumaroylshikimate, kaempferol-tri sugars, ferulic acid 4-*O*-glucuronide, feruloyl derivative, quercitrin derivative, and decreased 3-caffeoyl-quinic acid, caffeic acid glucoside, kaempferol, kaempferol-tri sugars, feruloylgalactarate (Table S2). Moreover, both UGTa and UGTb converted several flavonoid related compounds in *ex vivo* experiments (Table S3). Notably, caffeic acid glucoside, hydroxycinnamoyl-tyrosine, *p*-coumaroyl-cinnamoyl-glucoside, epicatechin, caffeoylshikimic acid, kaempferol-tri sugars and kaempferol derivative are significantly decreased in both *in vivo* and *ex vivo* assay. Quercetin-3-D-galactoside, quercetin-3-glucuronide, methylsyringin and galloylglucose significantly increased in both experiments (Figure 5h).

Table 2 List of candidate genes of major QTL ($P < 1.05 \times 10^{-5}$, PVE > 25%) assigned from mGWAS

QTL	Tissue	Metabolite ID	Class	Chr.	Region (bp)	-Log10 (P)	PVE	QTL group	Gene
1	TL	cmp1712	Flavonoid	1	268 348 770–268 654 646	17.57	29.98%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
2	TL	cmp1832	Flavonoid	1	268 348 770–268 654 646	18.35	31.06%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
3	TL	cmp1887	Flavonoid	1	268 347 540–268 654 646	30.87	47.22%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
4	TL	cmp2015	Flavonoid	1	268 348 770–268 654 646	20.44	34.13%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
5	TL	cmp2017	Unknown	1	268 348 770–268 654 646	26.19	41.67%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
6	TL	cmp2114	Flavonoid	1	268 348 770–268 654 646	25.25	40.49%	TL_hot_22	W01g002927(UGT);W01g002928(UGT)
7	TL	cmp1887	Flavonoid	1	269 133 796–269 874 436	19.45	32.65%	TL_hot_22	W01g002940(WRKY);W01g002948(MYB); W01g002954(MYB)
8	TL	cmp2015	Flavonoid	1	269 141 010–269 874 436	14.76	25.63%	TL_hot_22	W01g002940(WRKY);W01g002948(MYB); W01g002954(MYB)
9	TL	cmp2017	Unknown	1	269 325 166–269 874 436	15.41	26.62%	TL_hot_22	W01g002940(WRKY);W01g002948(MYB); W01g002954(MYB)
10	TL	cmp2114	Flavonoid	1	269 141 010–269 874 436	16.08	27.64%	TL_hot_22	W01g002940(WRKY);W01g002948(MYB); W01g002954(MYB)
11	TL	cmp835	Unknown	2	20 918 785–21 180 559	16.11	27.95%		W02g003425(WD40)
12	TL	cmp1207	Unknown	2	193 610 977–194 397 707	14.45	25.11%	TL_hot_37	W02g004824(HD-ZIP);W02g004827(WRKY)
13	YL	cmp2842	Other	2	233 540 133–233 779 262	14.51	25.24%	YL_hot_36	W02g005460(GATA)
14	TL	cmp1087	Unknown	5	185 405 127–185 605 212	15.3	27.67%		W05g012965(CAMTA)
15	TL	cmp1296	Unknown	7	19 373 965–19 574 651	14.57	25.35%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
16	TL	cmp1304	Flavonoid	7	19 373 965–19 574 651	14.64	26.00%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
17	YL	cmp1298	Flavonoid	7	19 373 965–19 574 651	18.19	30.82%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
18	TL	cmp1298	Flavonoid	7	19 373 965–19 574 651	17.88	30.37%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
19	TL	cmp1306	Unknown	7	19 373 965–19 574 651	14.57	25.91%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
20	TL	cmp1345	Unknown	7	19 373 965–19 574 651	15.4	27.97%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
21	TL	cmp1640	Unknown	7	19 373 965–19 574 651	14.71	25.52%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
22	YL	cmp1651	Flavonoid	7	19 373 965–19 574 651	15.37	26.57%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
23	TL	cmp1651	Flavonoid	7	19 373 965–19 574 651	15.1	26.13%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
24	YL	cmp1668	Unknown	7	19 373 965–19 574 651	17.07	29.15%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
25	TL	cmp1668	Unknown	7	19 373 965–19 574 651	16.44	28.20%	both_hot_20	W07g015551(OMT);W07g015552(OMT)
26	TL	cmp641	Unknown	7	155 279 042–155 479 140	25.46	40.75%	TL_hot_121	W07g017164(MYB_related)
27	YL	cmp2385	Unknown	7	155 279 042–155 479 048	44.21	60.54%		W07g017164(MYB_related)
28	YL	cmp2279	Unknown	9	16 349 280–16 738 187	14.39	25.02%	YL_hot_116	W09g019626(M-type_MADS)
29	YL	cmp2320	Unknown	9	16 349 203–16 738 187	15.78	27.18%	YL_hot_116	W09g019626(M-type_MADS)
30	YL	cmp2247	Unknown	10	31 482 376–32 366 717	15.43	26.65%	YL_hot_121	W10g021581(TALE)
31	YL	cmp413	Unknown	13	112 938 009–114 243 850	15.46	26.71%	YL_hot_156	W13g027373(MYB);W13g027379(C3H); W13g027382(bHLH);W13g027391(bHLH)
32	TL	cmp1384	Unknown	15	26 572 476–27 541 995	15.31	26.46%	TL_hot_196	W15g031583(HD-ZIP)
33	YL	cmp2094	Unknown	15	82 505 621–83 017 841	15.8	27.22%	YL_hot_181	W15g032428(bHLH);W15g032429(bHLH); W15g032432(C2H2);W15g032433(C2H2)

Chr, Chromosome; PVE (%), Phenotypic variation explained by the QTL. More information is listed in Data S4.

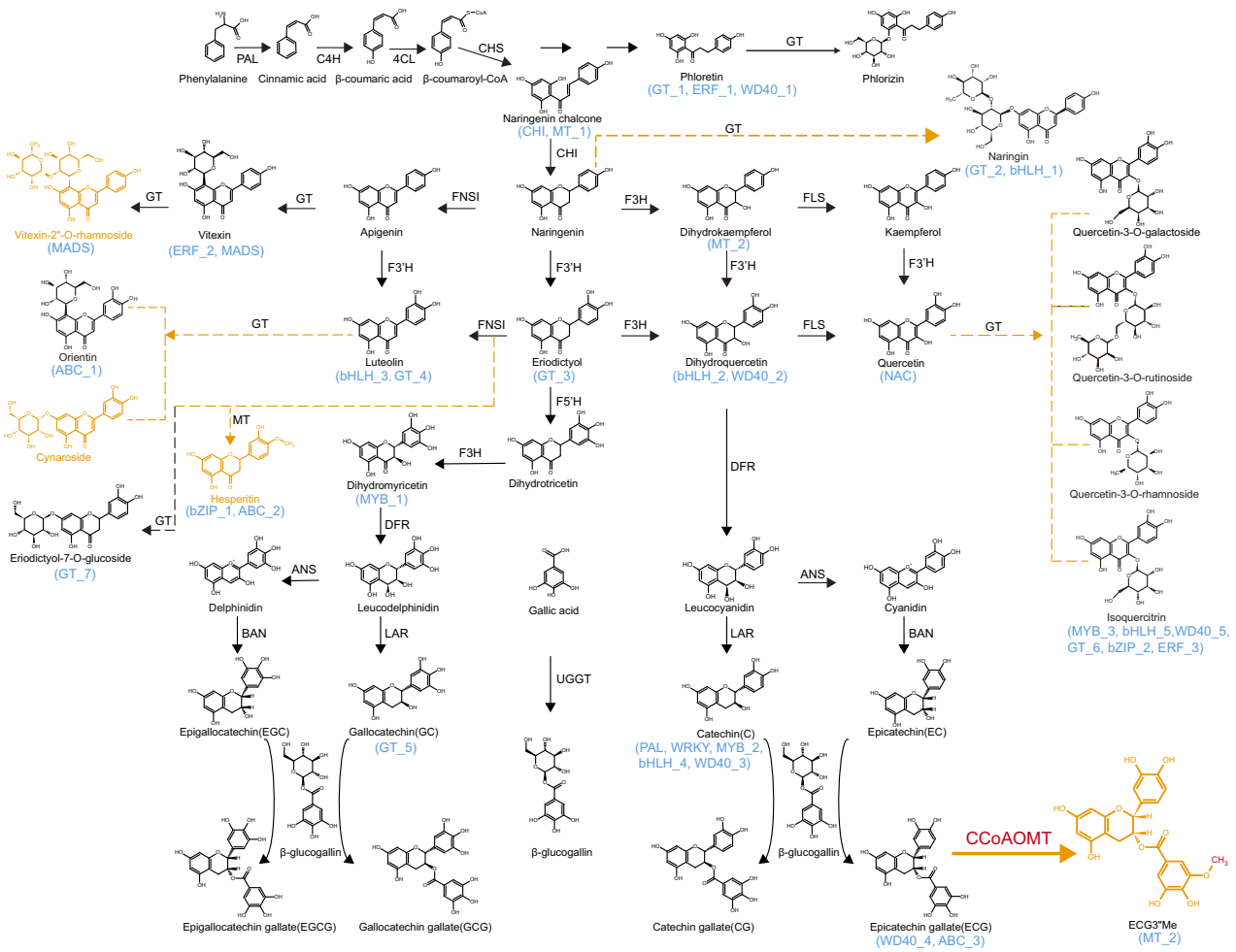


Figure 4 Proposed pathway of flavonoid biosynthesis in tea plant. Candidate genes identified by mGWAS are coloured in blue under the associated metabolites. The candidate genes selected for additional functional validation are highlighted in red colour. The newly identifying metabolites in tea plant are highlighted in orange colour. The dashed lines represent putative metabolic routes. CHI, chalcone isomerase (W06g013933); PAL, phenylalanine ammonia lyase (W12g025660); MADS, MADS box domain containing protein (W01g001727); NAC, NAC domain containing protein (W01g002770); WRKY, WRKY transcription factors (W10g022573); MT, methyl transferase (MT_1, W06g013934, MT_2, W07g015551); GT, glycosyl transferase (GT_1, W06g013993, GT_2, W06g0142690, GT_3, W02g005991, GT_4, W02g005682, GT_5, W01g000668, GT_6, W11g023625, GT_7, W05g012457); ERF, ethylene-responsive transcription factor (ERF_1, W03g006507, ERF_2, W02g005726, ERF_3, W04g011169); WD40, WD40 transcription factor (WD40_1, W03g006631, WD40_2, W02g005814, WD40_3, W05g012824, WD40_4, W02g005894, WD40_5, W04g011291); bHLH, bHLH transcription factor (bHLH_1, W01g000546, bHLH_2, W02g005816, bHLH_3, W02g005691, bHLH_4, W06g014048, bHLH_5, W01g001052); ABC, ABC transporter (ABC_1, W07g017413, ABC_2, W01g002785, ABC_3, W03g008531); bZIP, bZIP transcription factor (bZIP_1, W01g001743, bZIP_2, W04g011167); MYB, MYB transcription factor (MYB_1, W09g019405, MYB_2, W13g027669, MYB_3, W04g011288).

Another interesting QTL on chromosome 7 is significantly associated with 20 metabolites (Table 2; Data S4; Figure S9). It is worth noting that four of these 20 metabolites contain methylation chemical conversion, including putative epicatechin 3-(3″-*O*-Methyl) gallate (ECG3″Me) ($P = 1.39 \times 10^{-15}$) (Figure 6a). The putative ECG3″Me have similar MS^2 fragments with epicatechin gallate (ECG), and it was finally validated by co-elution with an authentic standard (Figure 6b). Two genes located in the QTL and one of them, W07g015551, was functionally annotated as caffeoyl-CoA *O*-methyltransferase (*CsCCoAOMT*). Two non-synonymous SNPs were found in the coding sequence (Figure 6c), resulted in four genotypes (A/C, G/C, G/G and GA/C) among the association panel. The accessions with *CsCCoAOMT*-A/C genotypes had significantly higher ECG3″Me levels than the ones with genotype *CsCCoAOMT*-G/G (Figure 6e; Figure S8c). *CsCCoAOMT* is close to *ClCCoAOMT* (*Camellia lanceoleosa*) in the phylogenetic tree (Figure 6d), and it exhibited high expression levels in root, the second leaf and petal of the tea plant (Figure S7c). A strong *cis*-eQTL was detected for the expression of *CsCCoAOMT* ($P = 6.97 \times 10^{-13}$) (Figure S10). There is a 15 bp insertion/deletion variation at the UTR of *CsCCoAOMT*, 45 bp upstream of ATG. Tea accessions with presence of the 15 bp showed significantly higher expression level of *CsCCoAOMT* and metabolite abundance of ECG3″Me (Figure 6f,g). *In vitro* enzymatic assays showed *CsCCoAOMT*-A/C can catalyse methylation of ECG into ECG3″Me, while *CsCCoAOMT*-G/G could not (Figure 6h; Figure S11).

Discussion

As one of the most cultivated woody horticultural crops, the tea plant is mainly used for producing various types of tea beverage (Jiang *et al.*, 2022). The rich taste, flavour and health benefits of

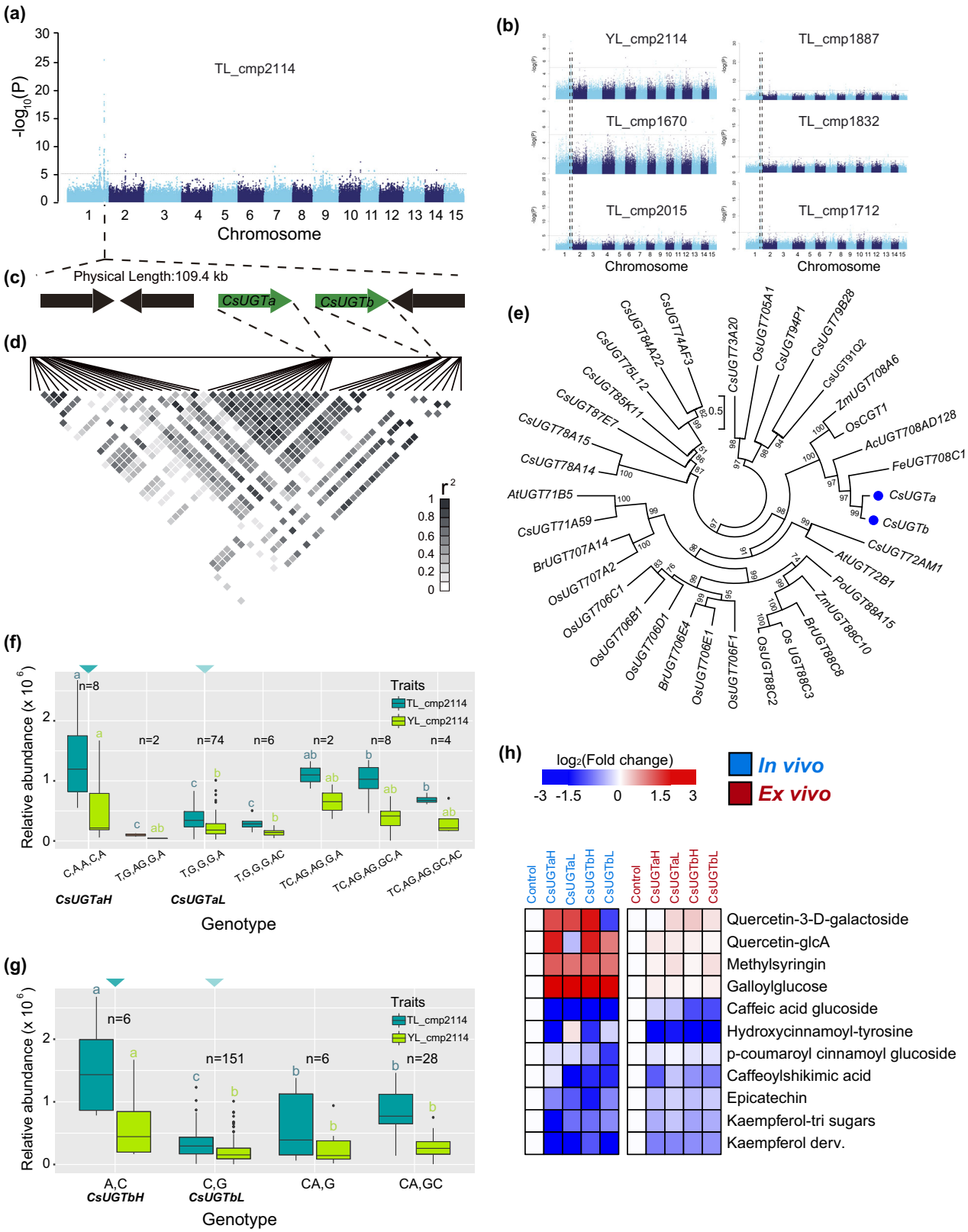


Figure 5 Identification and validation of *CsUGTa* and *CsUGTb*. (a-b) Manhattan plots displaying the locus on Chr.1 were associated with seven putative flavonoids. (c) Gene model in the most significant mQTL. (d) Pairwise r^2 values (a measure of LD) among all polymorphic sites within the most significant mQTL. (e) Phylogenetic analysis of *CsUGTa* and *CsUGTb*. (f-g) Boxplots of the genotype (each SNP site is separated by a comma, and NN indicates heterozygous SNPs) analysis using non-synonymous SNPs in the candidate genes for *CsUGTa* and *CsUGTb*, respectively. Different letters above the boxplot indicate significant difference in One-way ANOVA with Tukey's test ($P < 0.05$). (h) Heatmap of significantly changed metabolites in both *in vivo* (left) and *ex vivo* metabolomic (right) experiments with the presence of *CsUGTa* and *CsUGTb* ($P < 0.05$; *t* test; two sided).

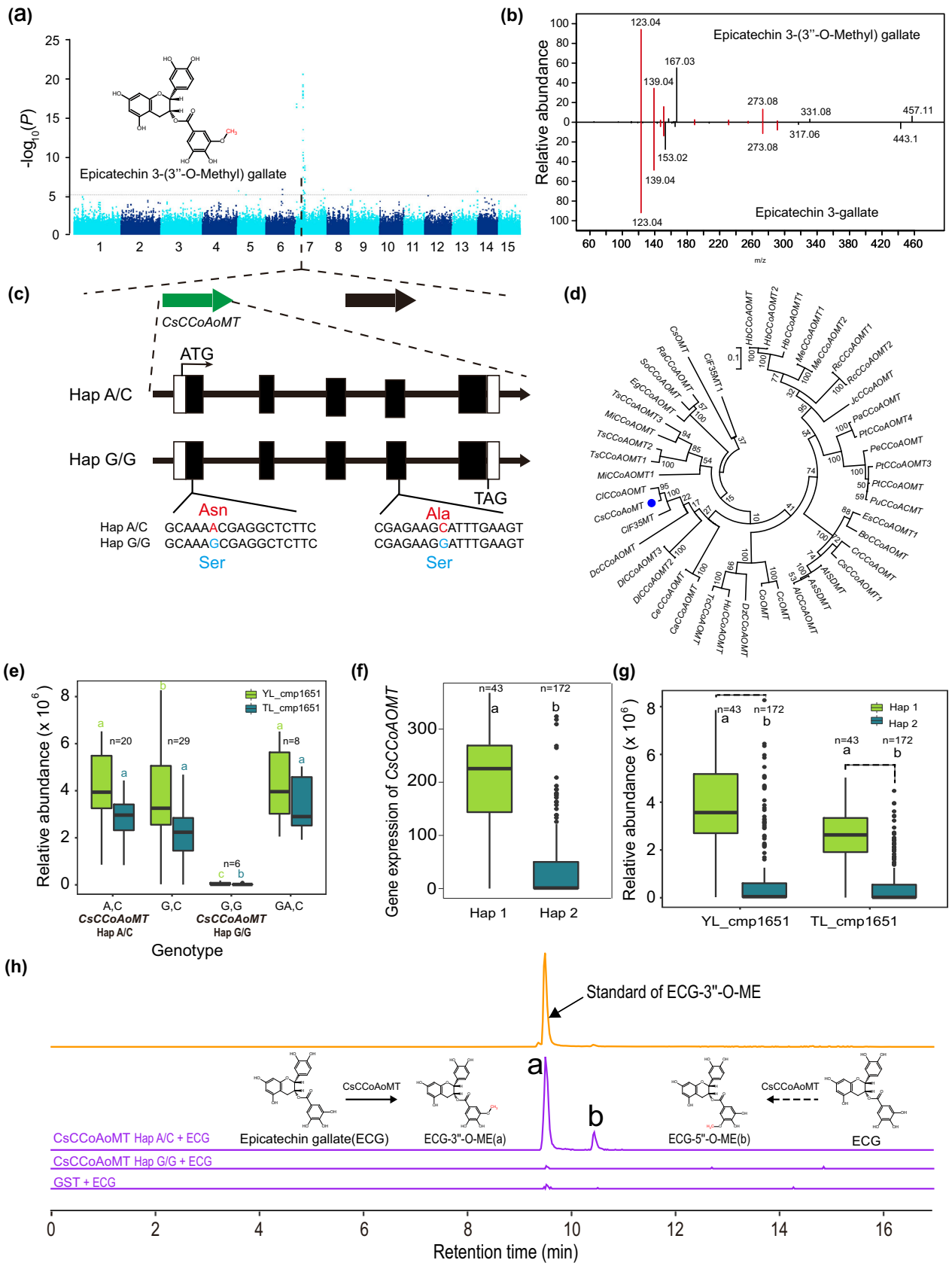


Figure 6 Functional investigation of *CsCCoAOMT*. (a) Manhattan plot displaying the GWAS result of (−)-Epicatechin 3-(3″-O-Methyl) gallate (cmp1651) in TL. (b) LC–MS/MS spectra alignment between (−)-Epicatechin 3-(3″-O-Methyl) gallate and (−)-Epicatechin 3-gallate. (c) genes in the QTL region and gene (colour in green) model of different alleles of *CsCCoAOMT*. (d) Phylogenetic analysis of *CsCCoAOMT*. (e) Boxplots of the genotype analysis of *CsCCoAOMT* corresponding to cmp1653 in YL and TL. (f-g) Boxplot of gene expression of *CsCCoAOMT* (f) and relative abundance of cmp1653 (g) in YL and TL corresponding to 15 bp InDel in UTR of *CsCCoAOMT*. Hap 1 represents sample with the 15 bp insertion, and Hap 2 represents the samples without the 15 bp insertion. Different letters below indicate significant difference in One-way ANOVA with Tukey's test ($P < 0.05$). (h) The in vitro enzymatic assays of CsCCoAOMT by using epicatechin gallate (ECG) as substrate.

tea, coupled with its regional cultural significance, have made it a highly popular beverage. Metabolites produced by crops are critical for their own growth and development and also for meeting various demands of humans (Brotman *et al*., 2021; Kc *et al*., 2021; Li *et al*., 2022a; Sreenivasulu and Fernie, 2022; Tiozon *et al*., 2022; Vollmer *et al*., 2018). The fresh leaves of tea plants contain a variety of metabolites, as reported previously and investigated in this study, which are key components and important underpinning to determine the final quality of tea (i.e., various tastes, flavours and health benefits) after being processed into consumed tea (Wei *et al*., 2018; Yang *et al*., 2013; Yu *et al*., 2020). In addition, huge metabolic diversity has been revealed among diverse tea germplasm resources in several recent studies (Fang *et al*., 2021; Huang *et al*., 2022; Yu *et al*., 2020; Zhang *et al*., 2020b, 2021). A better understanding of the chemical diversity of fresh leaves of tea plant is benefitial to produce a more flavorful and nutritious beverage supply. However, systematic metabolomics studies performed in tea germplasm coupled with genetic studies on the metabolic variation of tea remain limited. Owing to the advances of sequencing technology and bioinformatic tools, the genomic research on tea plant, which possesses a large and complex genome, currently present an unparalleled opportunity. Up to ten chromosome-level tea plant genomes have been assembled, and hundreds of diverse tea genotypes have been subject to genome resequencing or RNA-sequencing (Chen *et al*., 2020a; Wang et al., , 2020, 2021a, 2022; Wei *et al*., 2018; Xia *et al*., 2020; Yu *et al*., 2020; Zhang *et al*., 2020a,b, 2021). Taking advantage of the advances in tea genomic research and the diverse germplasm, we have collected and genotyped, here we conducted large-scale comprehensive metabolomic analysis and genome-wide association studies. The findings of this study provide valuable resources and toolbox for the genetic improvement, especially quality improvement of the tea plant, and deepen our understanding of chemical diversity and the underlying genetics in tea plants.

First, this study provides the richest resource for tea chemical diversity research compiled to date. Datasets of the entire tea metabolome and the metabolome of two tissue types of each tea accession released here are detailed information for further studies such as genetic and biochemical analyses, and tea quality evaluation and improvement practices. More insights into the metabolic contribution to the tea quality could benefit from this large-scale metabolite profiling. As young leaves (usually contains bud and the first leaf) are used for making green tea while mature leaves are collected for producing black tea, the comparative metabolome data between tissues is especially important for uncovering the biological basis of tea plant genetic improvement and tea processing. However, it is important to note that the tissue-specificity of metabolome has been well documented in different species (Mou *et al*., 2021; Wen *et al*., 2015; Zhang *et al*., 2020b). In addition, metabolite profiling of the sub-populations based on the phylogenetic and population structure analyses, the signature metabolites, and differently accumulated metabolites found in this study will facilitate the breeding practices targeting specific metabolites through parental selection based on the corresponding genetic background and metabolic performance.

Second, based on tremendous efforts in the annotation of the tea metabolome including both manual curation of data and authentic standard identification, as well as network construction-based analysis, this study broadens the understanding of tea metabolism and lays a solid foundation for revealing the natural variations in chemical composition of tea plant. Chemical conversions based on core metabolite skeletons contribute to the production of diverse metabolites (Wang *et al*., 2019). Interestingly, we found that galloylations, rather than hydroxylations or glycosylations, were the largest class of conversions within the tea metabolome. Galloylation involves the addition of a gallic group to the substrate, and its propensity renders tea unusual since hydroxylations and glycosylations are normally the most prominent conversions of plant specialized metabolites (Wang *et al*., 2019). Tea is rich in gallic acid and the content ranges from 0.07 to 2.69 mg/g (Bai *et al*., 2021; Gu *et al*., 2019; Xiao *et al*., 2017), which provides sufficient precursor substances for galloylation (Yao *et al*., 2022). Galloylation protects plants against various stresses, such as ultraviolet radiation, pathogens, and herbivores (Li *et al*., 2022; Zeng *et al*., 2020; Zhang *et al*., 2022). Furthermore, galloylated signatures metabolites were found in tea plant, including epicatechin gallate (ECG), epigallocatechin gallate (EGCG), gallocatechin gallate (GCG), 1,6-digalloylglucose, etc. The total catechins account for a major portion of tea polyphenols, and galloylated catechins account for 75% of the total catechins (6%–18% of the dry leaf mass). Moreover, galloylated catechins play a key role in the astringency and bitter taste which are important quality attribute of tea (Ahmad *et al*., 2020; Ye *et al*., 2021). The link between galloylation of secondary metabolites in the tea plant and its role in adaptation and domestication warrants further investigation.

Thirdly, the mGWAS conducted in this study was efficient in mQTL identification with the results furthering our understanding of the genetic basis of naturally occurred metabolic variation of tea plant. These mQTL and candidate genes offer valuable information for tea plant genetic improvement and the elucidation of metabolic pathways. The tea plant exhibits exceptionally large diversity within species, and rapid linkage disequilibrium decay has been observed (Zhang *et al*., 2020b). In this study, we achieved high mapping resolution, down to a single to a few genes in most cases. Moreover, the structure determination and annotation of metabolites, combined with the functional annotation of genes within the mQTL, facilitated the rapid identification of casual genes. We exemplified the further mining and interpretation of our GWAS results through verifying two of these identified mQTL and thereby completing the flavonoid biosynthetic pathway of the tea plant. For instance, we showed that *CsCCoAOMT* catalyses the methylation of epicatechin 3-

gallate, and the molecular identity of the gene/protein catalysing this reaction is newly identified in the tea plant.

In summary, the combination of large-scale metabolite profiling and genome-wide association study using a panel of diverse tea plant germplasm has allowed us to survey the genetic and metabolic panorama of tea chemical diversity in greater detail and gain novel insights into tea metabolism. Future mining of the rich data resources generated in this study, in conjunction of molecular validations of the analysed results, will contribute to fully uncovering the mechanisms underlying the complex metabolism of this much enjoyed plant.

Experimental procedures

Sample preparation and metabolite profiling

The tea association panel containing 215 genotypes used in this study was grown in the experimental station at Huazhong Agricultural University, Wuhan, China (30°47′ N, 114°36′ E). The young leaves (first leaves with buds; YL) and third leaves (TL) of the panel were collected from eight to ten individual tea tree plants and immediately frozen in liquid nitrogen on the same day in July 2019 (Figure 1). For each genotype, samples from four to five independent plants were pooled to make a biological replicate. Two biological replicates were employed in this study. All the materials were stored at −80 °C until use. 50 mg powder of each frozen dried sample was weighed and extracted using 1.0 mL of 70% methanol for 12 h at 4 °C, centrifuged at 10 000 ***g*** for 10 min at 4 °C. Each extract of all 215 accessions was pooled and analysed using liquid chromatography coupled with Q Exactive Plus mass spectrometry (LC–MS; Thermo Fisher Scientific, California, USA) in full scan and dd-MS^2 acquisition strategy (in both positive and negative modes). Instrumental conditions were set as follows: UHPLC: CORTECS T3 Column (120 Å, 2.7 μm, 2.1 × 100 mm); solvent system: 0.1% formic acid in water (solution A) and methanol (solution B); gradient program: 0–1 min, B 2%; 1–10 min, B 2% to 50%; 10–13 min, B 50% to 95%; 13–14 min, B 95%; 14–14.1 min, B 95% to 2%; 14.1–17 min, B 2%; flow rate: 0.25 mL/min; column temperature: 40 °C; MS detection: spray voltage, +3.8 kV/−3.2 kV; capillary temperature: 300 °C; sheath gas: 40 arb; AUX gas: 10 arb; AUX gas heater temperature: 300 °C; s-lens RF level: 55; scan range: m/z 75–1100; resolution: 70 000 (MS1) and 17 500 (MS^2); stepped normalized collision energy (NCE): 10, 30, and 50%. Mass spectrogram generated from the UHPLC-HRMS runs were processed using commercial software Compound Discoverer 3.2 (Thermo Fisher, San Jose, CA, USA). The extracted mass, retention times, fragments, and peak areas were exported. The obtained information of these peaks from both positive and negative ionization modes was aligned to the mzCloud, mzVault, ChemiSpider, Orbitrap traditional Chinese medicine library (OTCML), Metabolika and KEGG pathway integrated in the Compound Discoverer 3.2 software (Thermo Fisher, San Jose, CA, USA) as well as verified by standards or literature survey. Besides, all peaks were also annotated through the metDNA pipeline (Shen *et al.*, 2019). Targeted compounds were selected as following criteria; a) the peak pattern should be complete, b) the relative abundance (peak area) should be over 1×10^5. More details of targeted compounds are attached in Table S1.

For the large-scale targeted metabolomics, each sample was analysed using triple-quadrupole MS (TSQ Quantis; Thermo Fisher Scientific, California, USA) with the selective reaction monitoring (SRM) scanning mode using the same chromatographic conditions as detailed above. In brief, high resolution mass spectrum of targeted compounds including extracted mass, polarity, retention time and the most intensive fragment (extracted mass (m/z) – fragment (m/z) ≥ 18.015 Da) were exported and transferred to construct a transition list for SRM assay. The fragment of the precursor ion was selected as the SRM fragment, and the corresponding fragment voltage was optimized and used as the optimal CE voltage. This procedure led to an SRM list with 2837 transitions for the triple quadrupole analysis. Quality control was conducted between 10–12 sample data.

Metabolite quantification was performed using the commercial software TraceFinder 4.1 (Thermo Fisher, San Jose, CA, USA) and each metabolite was also verified manually. Quality control-based random forest signal correction algorithm (QC-RFSC) were performed to correct signal drift (Luan *et al.*, 2018). Relative abundance of two biological replicates were averaged for further analysis.

Statistical analysis of metabolites and data visualization of PLS-DA with 200 permutation tests for metabolites in YL and TL was conducted using R package *ropls* (Thévenot *et al.*, 2015). The 3D scatter plot was visualized in R package *scatterplot3d* (Ligges and Mächler, 2003). The one-way analysis of variance (ANOVA) in this study was calculated by *aov* function in R. The Tukey's test for multiple comparisons were conducted in R package *multcomp*. Dot plot, bar plot and box plot in this study were plotted using the R package *ggplot2*.

Construction of isomer metabolite pairs and candidate substrate-product pairs (CSPP)

The extracted mass, polarity, top ten MS^2 fragments, adduct and retention time exported from Compound Discoverer 3.2 were used for constructing the network of isomer metabolite pairs and CSPPs. Isomer metabolite pairs were constructed when the extracted mass between the metabolites in the same polarity differs within 2 ppm, and the masses of no less than five corresponding MS^2 fragments differ within 2 ppm. The metabolites conversion dataset was obtained according to previous studies (Breitling *et al.*, 2006; Morreel *et al.*, 2014; Perez de Souza *et al.*, 2019; Wang *et al.*, 2019; Table S1). The CSPPs were constructed under the criteria, (a) extracted mass of candidate substrate + conversion = extracted mass product ±2 ppm at the same polarity; (b) masses of no less than five corresponding MS^2 fragments differ within 2 ppm. Overlapped metabolites in any two isomer metabolite pairs or CSPPs will be considered as in the same group. The isomer metabolite pair networks were visualized by Gephi 0.9 (https://gephi.org).

Genome-wide association studies

Based on RNA sequencing of the young leaves of the 215 tea genotypes, a total of 208 446 high-quality SNPs and expression data for 33 021 genes in each genotype were obtained. The detailed information was described in our previous study (Zhang *et al.*, 2020b). GWAS was performed using EMMAX software accounting for the population structure and relatedness among the 215 tea genotypes (Kang *et al.*, 2010). The threshold for defining significant SNPs was set as 1.05×10^{-5} (1/independent markers), where 95 036 independent markers were evaluated by GEC software (Li *et al.*, 2012). SNPs with $P < 1.05 \times 10^{-5}$ and within 500 Kb to the lead SNPs and the pairwise LD (r^2) > 0.1 were combined into initial QTL region and each initial QTL region should have at least two significant SNPs. Function of genes within the QTL region was annotated and the PCC between the

corresponding metabolite and gene expression level were calculated in R program.

Phylogenetic analysis

The alignment of amino acid sequences was performed by the ClustalW bundled in MEGA 7 (Kumar *et al*., 2016). The maximum likelihood tree trees were constructed using MEGA 7 software with all default parameters coupled with 1000 bootstrap tests.

cDNA cloning and vector construction

Non-synonymous SNPs of *CsUGTa* and *CsUGTb* were predicted by ANNOVAR. The non-synonymous SNPs were combined into allele, and the contribution of each allele to the metabolic trait was tested by one-way ANOVA. Accessions with genotypes of high- and low- level phenotypic contribution were selected to amplify the cDNA for each gene, respectively. The PCR products were subsequently introduced into T-vectors and sequenced to select the allele for each gene. Primers using for vector construction are listed in Table S4.These fragments were then subcloned into *p*DONR207 donor vector by PCR-based Gateway BP cloning according the protocol (Zhang *et al*., 2019). The stop codon was removed to fuse a nano-GFP tag with the C-terminal of each gene. Expression vectors for transient expression and enzyme assay were constructed using the Gateway LR reaction with *p*K7FWG2.

Transient expression analysis in the tobacco leaves and LC–MS

The above constructs were transformed into Agrobacterium tumefaciens strain AGL1 and transiently expressed in tobacco leaves according to the previous protocol (Zhang *et al*., 2020b). Two days after inoculation, tobacco leaves were checked by confocal microscopy to confirm the expression level of each gene. The tobacco leaves were harvested and crushed into powders in liquid nitrogen and then extracted in 70% methanol for the LC-MS measurement.

Protein expression and enzyme activity assay

All the enzymes were purified from tobacco leaves by the affinity purification with nano-GFP tag (Zhang *et al*., 2019). The tea leaves were extracted using 70% Methanol and dried for the *in vitro* enzyme assay. Both CsUGTa and CsUGTb were incubated with tea extract and 0.1 mM of the co-factor NADPH for 30 min in an end stopped metabolite assay. The empty GFP protein was processed the same treatments as a negative control (Feussner and Feussner, 2020; Zhang *et al*., 2020b).

Prokaryotic expression and protein purification

Different alleles of *CsCCoAOMT* were used to construct *p*GEX-6p-1_VB1 recombinant plasmid (Table S4). The recombinant *p*GEX-6p-1 plasmids were transformed into *E. coli* BL21 (DE3, WEIDI, Shanghai, China). The positive *E. coli* BL21 were cultivated and then added into 200 mL liquid LB medium (with 50 ng/μL ampicillin). When an absorbance of the culture reached 0.6 at OD_{600}, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.05 mmol/L to induce expression of the fusion protein. After 15 h of induction at 18 °C, the bacterial cells were collected, resuspended in GST binding buffer (50 mmol/L Tris-HCl, 2 mol/L NaCl, and 10% glycerine, pH = 8.0) and lysed by sonication. Crude enzymes were transferred into GST affinity column, which contains the Glutathione S-transferase (GST) beads. The GST affinity column was stored at 4 °C for 2 h, then washed by glutathione buffer (pH = 8.0). The eluent was collected in different tubes and the absorbance rate determined at 280 nm. And the selected samples were analysed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (Figure S1).

Enzyme essay

Enzymatic essays of different alleles of *CsCCoAOMT* were conducted using the purified recombinant protein 0.01 mg purified protein was added to a 1.5 mL reaction system including 0.05 mmol/L ECG, 0.16 mmol/L SAM, 0.2 mmol/L $MgCl_2$, and 100 mmol/L Tris–HCl (pH = 7.8). The reaction mixture was incubated at 30 °C for 30 min, before the reaction was terminated by additions of 200 μL methanol to terminate the reaction. After centrifuging at 10 000 ***g*** for 10 min at 4 °C, the supernatant products were analysed by liquid chromatography coupled with Q Exactive Plus mass spectrometry (LC-MS; Thermo Fisher Scientific, California, USA) in full scan and dd-MS^2 acquisition strategy (negative modes) with the same method as described above.

Acknowledgements

The research was financed by the National Key R&D Program of China (2022YFF1003103), the National Natural Science Foundation of China (3211101118) NSFC-DFG collaborative project, Fundamental Research Funds for the Central Universities (2662023PY011) and the HZAU-AGIS Cooperation Fund (SZYJY2021004) to Weiwei Wen, Deutsche Forschungsgemeinschaft (DFG)-Project number 468870408 to Björn Usadel and Alisdair R. Fernie. A.R.F and Y.J.Z would like to thank the funding from the Max-Planck Society, European Union's Horizon 2020 research and innovation programme, project PlantaSYST (SGA-CSA No 664621 and No 739582 under FPA No. 664620).

Conflict of interest statement

The authors declare no conflicts of interest.

Author contributions

WWW conceived and managed the research and wrote the article. HJQ, XLZ, YJZ performed the experiments and data analysis and edited the article. YJR, DWG, XZ performed the experiments, XHJ and BU edited the article. ARF co-supervised the gene functional analysis and edited the paper. All authors read and approved the article.

References

Ahmad, M.Z., Li, P., She, G., Xia, E., Benedito, V.A., Wan, X.C. and Zhao, J. (2020) Genome-wide analysis of serine carboxypeptidase-like acyltransferase gene family for evolution and characterization of enzymes involved in the biosynthesis of galloylated catechins in the tea plant (*Camellia sinensis*). *Front. Plant Sci.* **11**, 848.

Alseekh, S., Wu, S., Brotman, Y. and Fernie, A.R. (2018) Guidelines for sample normalization to minimize batch variation for large-scale metabolic profiling of plant natural genetic variance. *Methods Mol. Biol.* **1778**, 33–46.

Alseekh, S., Scossa, F., Wen, W., Luo, J., Yan, J., Beleggia, R., Klee, H.J. *et al*. (2021) Domestication of crop metabolomes: desired and unintended consequences. *Trends Plant Sci.* **26**, 650–661.

Bai, J., Zhang, Y., Tang, C., Hou, Y., Ai, X., Chen, X., Zhang, Y. *et al*. (2021) Gallic acid: pharmacological activities and molecular mechanisms involved in inflammation-related diseases. *Biomed. Pharmacother.* **133**, 110985.

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M.L. and Barrett, M.P. (2006) Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics*, **2**, 155–164.

Brody, H. (2019) Tea. *Nature* **566**, S1.

Brotman, Y., Llorente-Wiegand, C., Oyong, G., Badoni, S., Misra, G., Anacleto, R., Parween, S. *et al*. (2021) The genetics underlying metabolic signatures in a brown rice diversity panel and their vital role in human nutrition. *Plant J.* **106**, 507–525.

Chen, J.D., Zheng, C., Ma, J.Q., Jiang, C.K., Ercisli, S., Yao, M.Z. and Chen, L. (2020a) The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Hortic. Res.* **7**, 63.

Chen, Y., Guo, X., Gao, T., Zhang, N., Wan, X., Schwab, W. and Song, C. (2020b) UGT74AF3 enzymes specifically catalyze the glucosylation of 4-hydroxy-2,5-dimethylfuran-3(2H)-one, an important volatile compound in *Camellia sinensis*. *Hortic. Res.* **7**, 25.

Fang, K., Xia, Z., Li, H., Jiang, X., Qin, D., Wang, Q., Wang, Q. *et al*. (2021) Genome-wide association analysis identified molecular markers associated with important tea flavor-related metabolites. *Hortic. Res.* **8**, 42.

Feussner, K. and Feussner, I. (2020) *Ex vivo* metabolomics: a powerful approach for functional gene annotation. *Trends Plant Sci.* **25**, 829–830.

Francisco, M., Joseph, B., Caligagan, H., Li, B., Corwin, J.A., Lin, C., Kerwin, R.E. *et al*. (2016) Genome wide association mapping in *Arabidopsis thaliana* identifies novel genes involved in linking allyl glucosinolate to altered biomass and defense. *Front. Plant Sci.* **7**, 1010.

Fu, X., Liao, Y., Cheng, S., Xu, X., Grierson, D. and Yang, Z. (2021) Nonaqueous fractionation and overexpression of fluorescent-tagged enzymes reveals the subcellular sites of L-theanine biosynthesis in tea. *Plant Biotechnol. J.* **19**, 98–108.

Gu, C., Howell, K., Dunshea, F.R. and Suleria, H.A.R. (2019) LC-ESI-QTOF/MS characterisation of phenolic acids and flavonoids in polyphenol-rich fruits and vegetables and their potential antioxidant activities. *Antioxidants (Basel).* **8**, 405.

Hagenbeek, F.A., Pool, R., van Dongen, J., Draisma, H.H.M., Jan Hottenga, J., Willemsen, G., Abdellaoui, A. *et al*. (2020) Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nat. Commun.* **11**, 39.

Hartiala, J.A., Tang, W.H.W., Wang, Z., Crow, A.L., Stewart, A.F.R., Roberts, R., McPherson, R. *et al*. (2016) Genome-wide association study and targeted metabolomics identifies sex-specific association of *CPS1* with coronary artery disease. *Nat. Commun.* **7**, 10558.

Hazra, A., Kumar, R., Sengupta, C. and Das, S. (2021) Genome-wide SNP discovery from Darjeeling tea cultivars - their functional impacts and application toward population structure and trait associations. *Genomics*, **113**, 66–78.

Huang, R., Wang, J.Y., Yao, M.Z., Ma, C.L. and Chen, L. (2022) Quantitative trait loci mapping for free amino acid content using an albino population and SNP markers provides insight into the genetic improvement of tea plants. *Hortic. Res.* **9**, uhab029.

Inoue-Choi, M., Ramirez, Y., Cornelis, M.C., Berrington de González, A., Freedman, N.D. and Loftfield, E. (2022) Tea consumption and all-cause and cause-specific mortality in the UK biobank: a prospective cohort study. *Ann. Intern. Med.* **175**, 1201–1211.

Jiang, X., Zhang, W., Fernie, A.R. and Wen, W. (2022) Combining novel technologies with interdisciplinary basic research to enhance horticultural crops. *Plant J.* **109**, 35–46.

Jing, T., Zhang, N., Gao, T., Zhao, M., Jin, J., Chen, Y., Xu, M. *et al*. (2019) Glucosylation of (Z)-3-hexenol informs intraspecies interactions in plants: a case study in *Camellia sinensis*. *Plant Cell Environ.* **42**, 1352–1367.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354.

Kc, S., Long, L., Liu, M., Zhang, Q. and Ruan, J. (2021) Light intensity modulates the effect of phosphate limitation on carbohydrates, amino acids, and catechins in tea plants (*Camellia sinensis* L.). *Front. Plant Sci.* **12**, 743781.

Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.P., Kangas, A.J. *et al*. (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276.

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

Li, M.X., Yeung, J.M.Y., Cherny, S.S. and Sham, P.C. (2012) Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756.

Li, K., Wen, W., Alseekh, S., Yang, X., Guo, H., Li, W., Wang, L. *et al*. (2019) Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. *Plant J.* **99**, 216–230.

Li, X., Zhang, J., Lin, S., Xing, Y., Zhang, X., Ye, M., Chang, Y. *et al*. (2022) (+)-Catechin, epicatechin and epigallocatechin gallate are important inducible defensive compounds against ectropis grisescens in tea plants. *Plant Cell Environ.* **45**, 496–511.

Li, J., Scarano, A., Gonzalez, N.M., D'Orso, F., Yue, Y., Nemeth, K., Saalbach, G. *et al*. (2022a) Biofortified tomatoes provide a new route to vitamin D sufficiency. *Nat. Plants* **8**, 611–616.

Li, P., Fu, J., Xu, Y., Shen, Y., Zhang, Y., Ye, Z., Tong, W. *et al*. (2022b) *CsMYB1* integrates the regulation of trichome development and catechins biosynthesis in tea plant domestication. *New Phytol.* **234**, 902–917.

Ligges, U. and Mächler, M. (2003) Scatterplot3d - an R package for visualizing multivariate data. *J. Stat. Softw.* **8**, 1–20.

Luan, H., Ji, F., Chen, Y. and Cai, Z. (2018) statTarget: A streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data. *Anal. Chim. Acta*, **1036**, 66–72.

Morreel, K., Saeys, Y., Dima, O., Lu, F., Van de Peer, Y., Vanholme, R., Ralph, J. *et al*. (2014) Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. *Plant Cell* **26**, 929–945.

Mou, J., Zhang, Z., Qiu, H., Lu, Y., Zhu, X., Fan, Z., Zhang, Q. *et al*. (2021) Multiomics-based dissection of citrus flavonoid metabolism using a *Citrus reticulata* × *Poncirus trifoliata* population. *Hortic. Res.* **8**, 56.

Perez de Souza, L., Scossa, F., Proost, S., Bitocchi, E., Papa, R., Tohge, T. and Fernie, A.R. (2019) Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome. *Plant J.* **97**, 1132–1153.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S. *et al*. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60.

Qiu, H., Zhu, X., Wan, H., Xu, L., Zhang, Q., Hou, P., Fan, Z. *et al*. (2020) Parallel metabolomic and transcriptomic analysis reveals key factors for quality improvement of tea plants. *J. Agric. Food Chem.* **68**, 5483–5495.

Shen, X., Wang, R., Xiong, X., Yin, Y., Cai, Y., Ma, Z., Liu, N. *et al*. (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* **10**, 1516.

Shi, T., Zhu, A., Jia, J., Hu, X.M., Chen, J., Liu, W., Ren, X. *et al*. (2020) Metabolomics analysis and metabolite-agronomic trait associations using kernels of wheat (*Triticum aestivum*) recombinant inbred lines. *Plant J.* **103**, 279–292.

Slaten, M.L., Yobi, A., Bagaza, C., Chan, Y.O., Shrestha, V., Holden, S., Katz, E. *et al*. (2020) mGWAS uncovers Gln-glucosinolate seed-specific interaction and its role in metabolic homeostasis. *Plant Physiol.* **183**, 483–500.

Sreenivasulu, N. and Fernie, A.R. (2022) Diversity: current and prospective secondary metabolites for nutrition and medicine. *Curr. Opin. Biotechnol.* **74**, 164–170.

Thévenot, E.A., Roux, A., Xu, Y., Ezan, E. and Junot, C. (2015) Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and opls statistical analyses. *J. Proteome Res.* **14**, 3322–3335.

Tiozon, R.N., Sartagoda, K.J.D., Serrano, L.M.N., Fernie, A.R. and Sreenivasulu, N. (2022) Metabolomics based inferences to unravel phenolic compound diversity in cereals and its implications for human gut health. *Trends Food Sci. Technol.* **127**, 14–25.

Vollmer, M., Esders, S., Farquharson, F.M., Neugart, S., Duncan, S.H., Schreiner, M., Louis, P. *et al*. (2018) Mutual interaction of phenolic compounds and microbiota: metabolism of complex phenolic apigenin-*C*- and kaempferol-*O*-derivatives by human fecal samples. *J. Agric. Food Chem.* **66**, 485–497.

Wang, S., Alseekh, S., Fernie, A. and Luo, J. (2019) The structure and function of major plant metabolite modifications. *Mol. Plant* **12**, 899–919.

Wang, X., Feng, H., Chang, Y., Ma, C., Wang, L., Hao, X., Li, A. *et al*. (2020) Population sequencing enhances understanding of tea plant evolution. *Nat. Commun.* **11**, 4447.

Wang, P., Yu, J., Jin, S., Chen, S., Yue, C., Wang, W., Gao, S. *et al*. (2021a) Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Hortic. Res.* **8**, 107.

Wang, X., Liu, S., Zuo, H., Zheng, W., Zhang, S., Huang, Y., Pingcuo, G. *et al*. (2021b) Genomic basis of high-altitude adaptation in *Tibetan prunus* fruit trees. *Curr. Biol.* **31**, 3848–3860.

Wang, F., Zhang, B., Wen, D., Liu, R., Yao, X., Chen, Z., Mu, R. *et al*. (2022) Chromosome-scale genome assembly of *Camellia sinensis* combined with multi-omics provides insights into its responses to infestation with green leafhoppers. *Front. Plant Sci.* **13**, 1004387.

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., Xia, E. *et al*. (2018) Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4151–E4158.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J. *et al*. (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* **5**, 3438.

Wen, W., Li, K., Alseekh, S., Omranian, N., Zhao, L., Zhou, Y., Xiao, Y. *et al*. (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *Plant Cell* **27**, 1839–1856.

Xia, E.H., Zhang, H.B., Sheng, J., Li, K., Zhang, Q.J., Kim, C., Zhang, Y. *et al*. (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol. Plant*, **10**, 866–877.

Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., Liu, Y. *et al*. (2020) The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol. Plant*, **13**, 1013–1026.

Xiao, W., Zhang, Y., Fan, C. and Han, L. (2017) A method for producing superfine black tea powder with enhanced infusion and dispersion property. *Food Chem.* **214**, 242–247.

Yamashita, H., Uchida, T., Tanaka, Y., Katai, H., Nagano, A.J., Morita, A. and Ikka, T. (2020) Genomic predictions and genome-wide association studies based on RAD-seq of quality-related metabolites for the genomics-assisted breeding of tea plants. *Sci. Rep.* **10**, 17480.

Yang, Z., Baldermann, S. and Watanabe, N. (2013) Recent studies of the volatile compounds in tea. *Food Res. Int.* **53**, 585–599.

Yao, S., Liu, Y., Zhuang, J., Zhao, Y., Dai, X., Jiang, C., Wang, Z. *et al*. (2022) Insights into acylation mechanisms: co-expression of serine carboxypeptidase-like acyltransferases and their non-catalytic companion paralogs. *Plant J.* **111**, 117–133.

Ye, Q.Q., Chen, G.S., Pan, W., Cao, Q.Q., Zeng, L., Yin, J.F. and Xu, Y.Q. (2021) A predictive model for astringency based on *in vitro* interactions between salivary proteins and (−)-Epigallocatechin gallate. *Food Chem.* **340**, 127845.

Yu, X., Xiao, J., Chen, S., Yu, Y., Ma, J., Lin, Y., Li, R. *et al*. (2020) Metabolite signatures of diverse *Camellia sinensis* tea populations. *Nat. Commun.* **11**, 5586.

Zeng, L.T., Zhou, X.C., Liao, Y.Y. and Yang, Z.Y. (2020) Roles of specialized metabolites in biological function and environmental adaptability of tea plant (*Camellia sinensis*) as a metabolite studying model. *J. Adv. Res.* **34**, 159–171.

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X. *et al*. (2015) The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905.

Zhang, Y., Natale, R., Domingues, A.P., Toleco, M.R., Siemiatkowska, B., Fàbregas, N. and Fernie, A.R. (2019) Rapid identification of protein-protein interactions in plants. *Curr. Prot. Plant Biol.* **4**, e20099.

Zhang, Q.J., Li, W., Li, K., Nan, H., Shi, C., Zhang, Y., Dai, Z.Y. *et al*. (2020a) The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Mol. Plant*, **13**, 935–938.

Zhang, W., Zhang, Y., Qiu, H., Guo, Y., Wan, H., Zhang, X., Scossa, F. *et al*. (2020b) Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719–3730.

Zhang, X., Chen, S., Shi, L., Gong, D., Zhang, S., Zhao, Q., Zhan, D. *et al*. (2021) Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1259.

Zhang, X., Ran, W., Li, X., Zhang, J., Ye, M., Lin, S., Liu, M. *et al*. (2022) Exogenous application of gallic acid induces the direct defense of tea plant against ectropis obliqua caterpillars. *Front. Plant Sci.* **13**, 833489.

Zhao, J., Li, P., Xia, T. and Wan, X. (2020) Exploring plant metabolic genomics: chemical diversity, metabolic complexity in the biosynthesis and transport of specialized metabolites with the tea plant as a model. *Crit. Rev. Biotechnol.* **40**, 667–688.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T. *et al*. (2018) Rewiring of the fruit metabolome in tomato breeding. *Cell*, **172**, 249–261.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1 Metabolites detected in this study.
Data S2 Group of isomers.
Data S3 Group of candidate substrate-product pairs.
Data S4 mQTL identified in this study.
Figure S1 Pairwise Pearson correlation coefficient between/-within YL and TL. (a) Distribution of Pearson correlation coefficient among 1206 metabolites both quantified in YL and TL. (b) Boxplot of Pairwise Pearson correlation coefficient among 1206 metabolites attributed to eight classes.
Figure S2 Box-plot of each principal component from PLS-DA in YL and TL. (a) based on the abundance of 1326 metabolites in YL. (b) based on the abundance of 1823 metabolites in TL (One-way ANOVA with Tukey's test, $P<0.05$), different letters represent significant differences, different subpopulations are shown in different colours.
Figure S3 Box-plot of relative abundance of caffeic acid, eriodictyol-7-O-glucoside, (+)-catechin, vitexin-2″-O-rhamnoside, naringenin-7-O-neohesperidoside in each subpopulation. Different letters represent significant differences (One-way ANOVA with Tukey's test, $P < 0.05$).
Figure S4 Barplot of the number of mQTL in 1 Mb sliding window size with step size of 100 Kb across DASZ tea genome in (a) YL and (b) TL.
Figure S5 Venn plot of metabolites with chemical conversion and associated with mQTL harbouring candidate genes encoding (a) glycosyltransferase, (b) methyltransferase, (c) acyltransferase and (d) CYP450.
Figure S6 Candidate substrate-product pairs corresponding to flavonoid biosynthesis in tea plant.
Figure S7 Expression profiles of *CsUGTa* (a), *CsUGTb* (b) and *CsCCoAOMT* (c).
Figure S8 Box-plot of the abundance of flavonoids in different genotypes of *CsUGTa* (a), *CsUGTb* (b) and *CsCCoAOMT* (c). Different letters indicate the abundances are significantly different (One-way ANOVA with Tukey's test, $P < 0.05$).
Figure S9 Manhattan plot and Q-Q plot of the GWAS of metabolites YL_cmp1298, TL_cmp1298 and YL_cmp1658 that mapped to the mQTL harbouring *CsCCoAOMT*.
Figure S10 Manhattan plot and Q-Q plot of GWAS of expression level of *CsCCoAOMT* in tea leaves.

Figure S11 Detecting proteins purified from *E.coli* BL21 in polyacrylamide gels using coomassie brilliant blue.
Table S1 Chemical conversions for candidate substrate-product pairs.
Table S2 List of secondary metabolites that were changed by overexpressing *CsUGTa* and *CsUGTb* in tobacco leaves.
Table S3 List of secondary metabolites that were changed in the *ex vivo* metabolomics analysis utilizing crude tea extracts as substrate.
Table S4 Primers used in this study.