Transcriptome-wide association analysis of flavonoid biosynthesis genes and their correlation with leaf phenotypes in hawk tea (*Litsea coreana* var. *sinensis*)

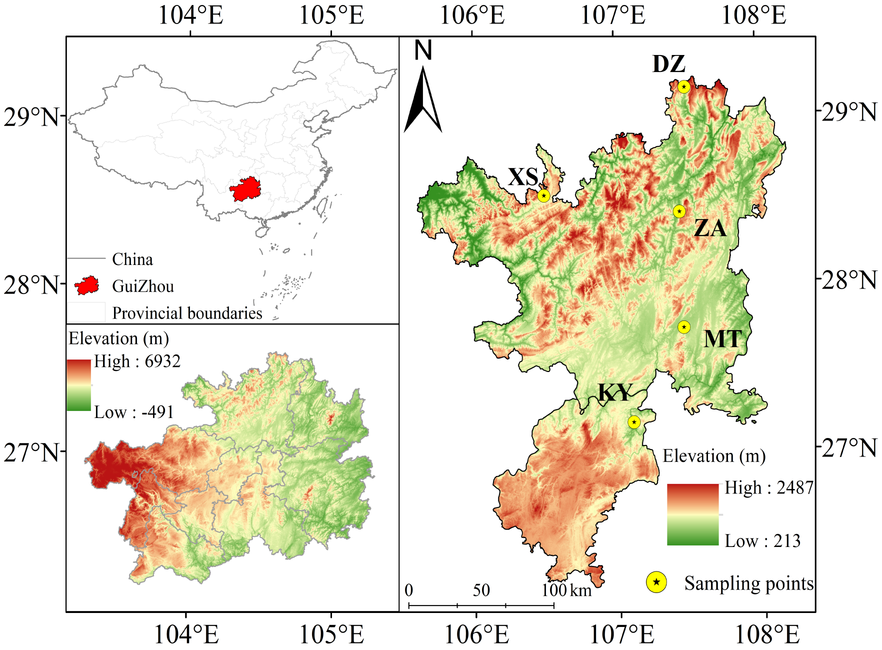
Lan Yang[1], Huie Li[1], Na Xie[1], Gangyi Yuan[2], and Qiqiang Guo[3]

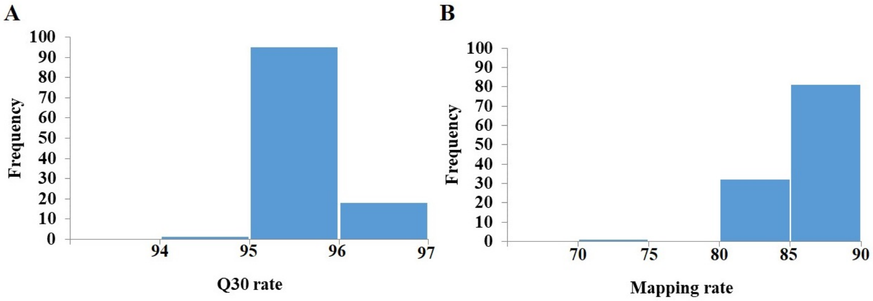
[1]Guizhou University
[2]Yunnan Province
[3]Institute for Forest Resources and Environment of Guizhou, College of Forestry, Guizhou University

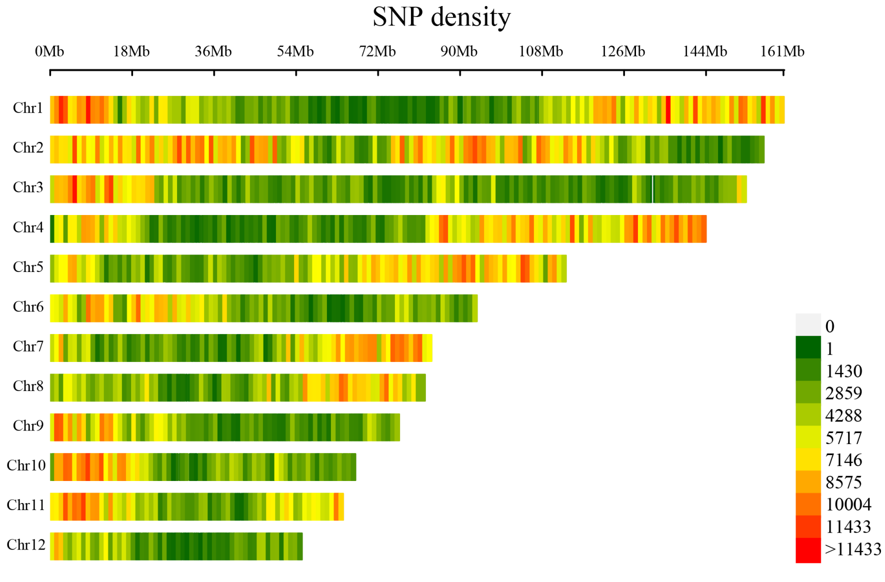
August 30, 2024

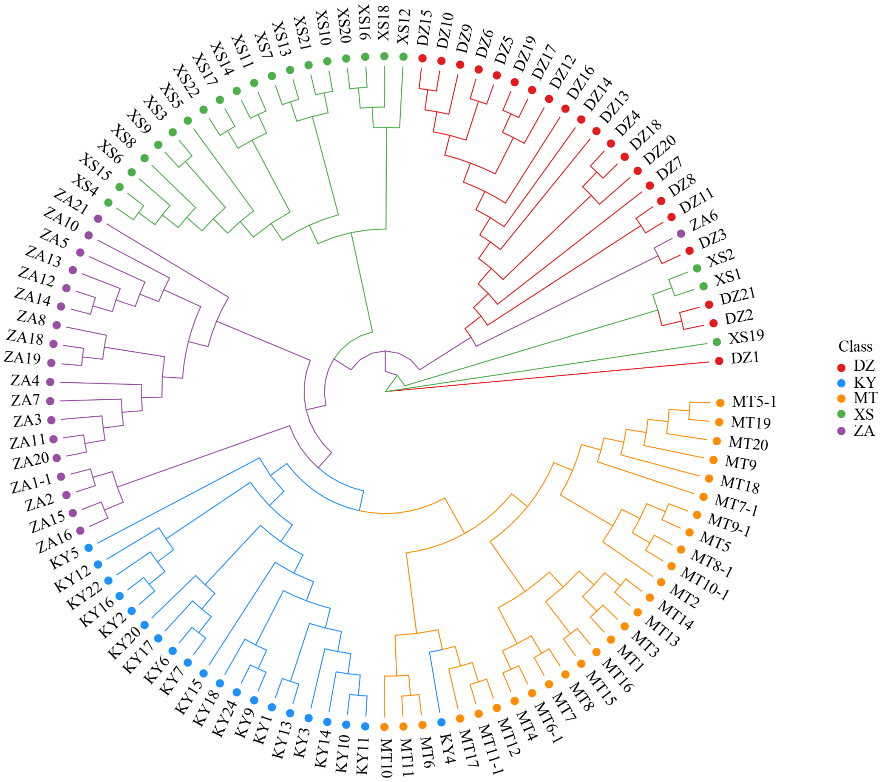
Abstract

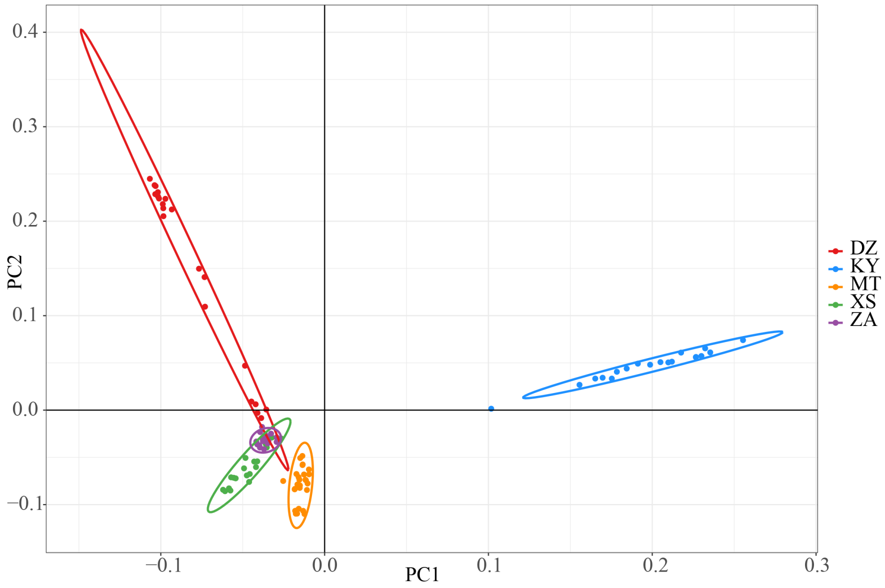
Hawk tea (*Litsea coreana* var. *sinensis*), derived from the tender shoots or leaves, rich in flavonoids that can promote health-care for humans. The primary flavonoid are kaempferol-3-O-β-D-glucoside, kaempferol-3-O-β-D-galactoside, quercetin-3-O-β-D-glucoside, and quercetin-3-O-β-D-galactoside. Is there an association between leaf phenotype and flavonoid content? And the mechanisms of flavonoid biosynthesis are not fully understood. In this study, 109 samples were analyzed to determine the correlation and genetic variability in leaf phenotype and flavonoid content. Furthermore, a transcriptome-wide association study identified candidate loci implicated in the biosynthesis of four key flavonoids. The study revealed that genetic variability in leaf traits and flavonoid concentrations is predominantly attributed to inter-population differences. Flavonoid accumulation may correlate with tree diameter at breast height (DBH), indicative of age-related traits. Transcriptome-wide association analysis identified 84 significant SNPs associated with flavonoid content, with only 13 located within gene regions. The majority of these genes are implicated in metabolic processes and secondary metabolite biosynthesis. Notably, structural genes within these regions are directly involved in pathways known to regulate flavonoid metabolism, exerting a pivotal influence on flavonoid biosynthesis. These results lay a solid theoretical groundwork for subsequent explorations into the genetic determinants influencing flavonoid accumulation of hawk tea.

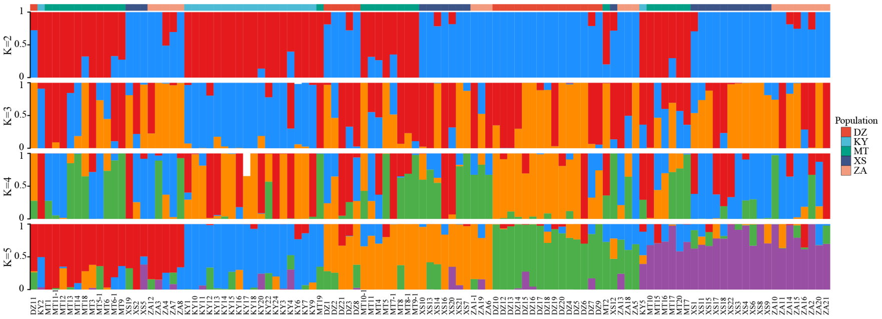

104°E
105°E
106°E
107°E
108°E
29°N
28°N
27°N
N
DZ
XS
ZA
MT
KY
China
GuiZhou
Provincial boundaries
Elevation (m)
High : 6932
Low : -491
Elevation (m)
High : 2487
Low : 213
Sampling points
0
50
100 km

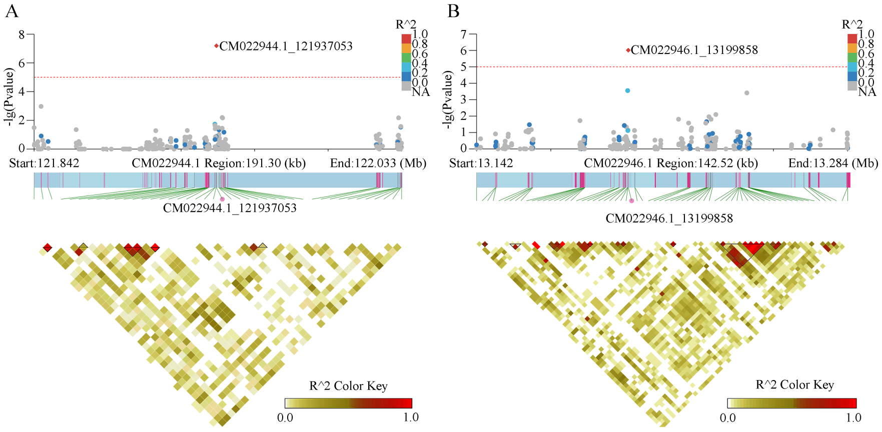

A
Frequency
100
90
80
70
60
50
40
30
20
10
0
94
95
96
97
Q30 rate
B
Frequency
100
90
80
70
60
50
40
30
20
10
0
70
75
80
85
90
Mapping rate

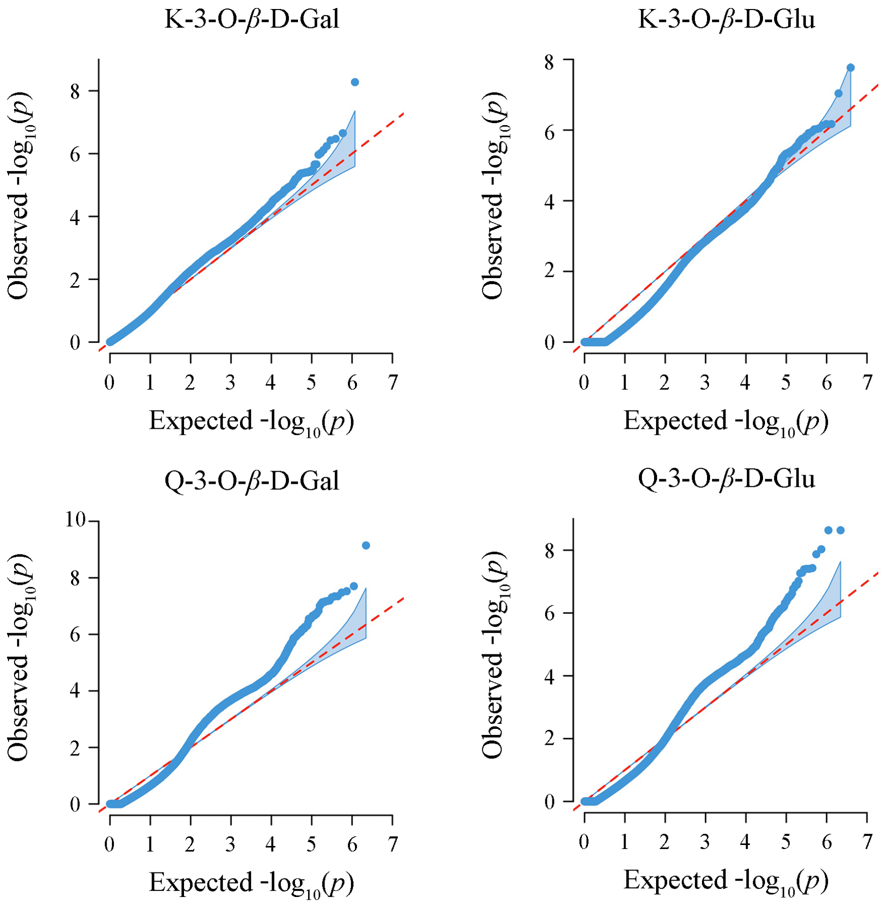

SNP density
0Mb
18Mb
36Mb
54Mb
72Mb
90Mb
108Mb
126Mb
144Mb
161Mb
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
0
1
1430
2859
4288
5717
7146
8575
10004
11433
>11433

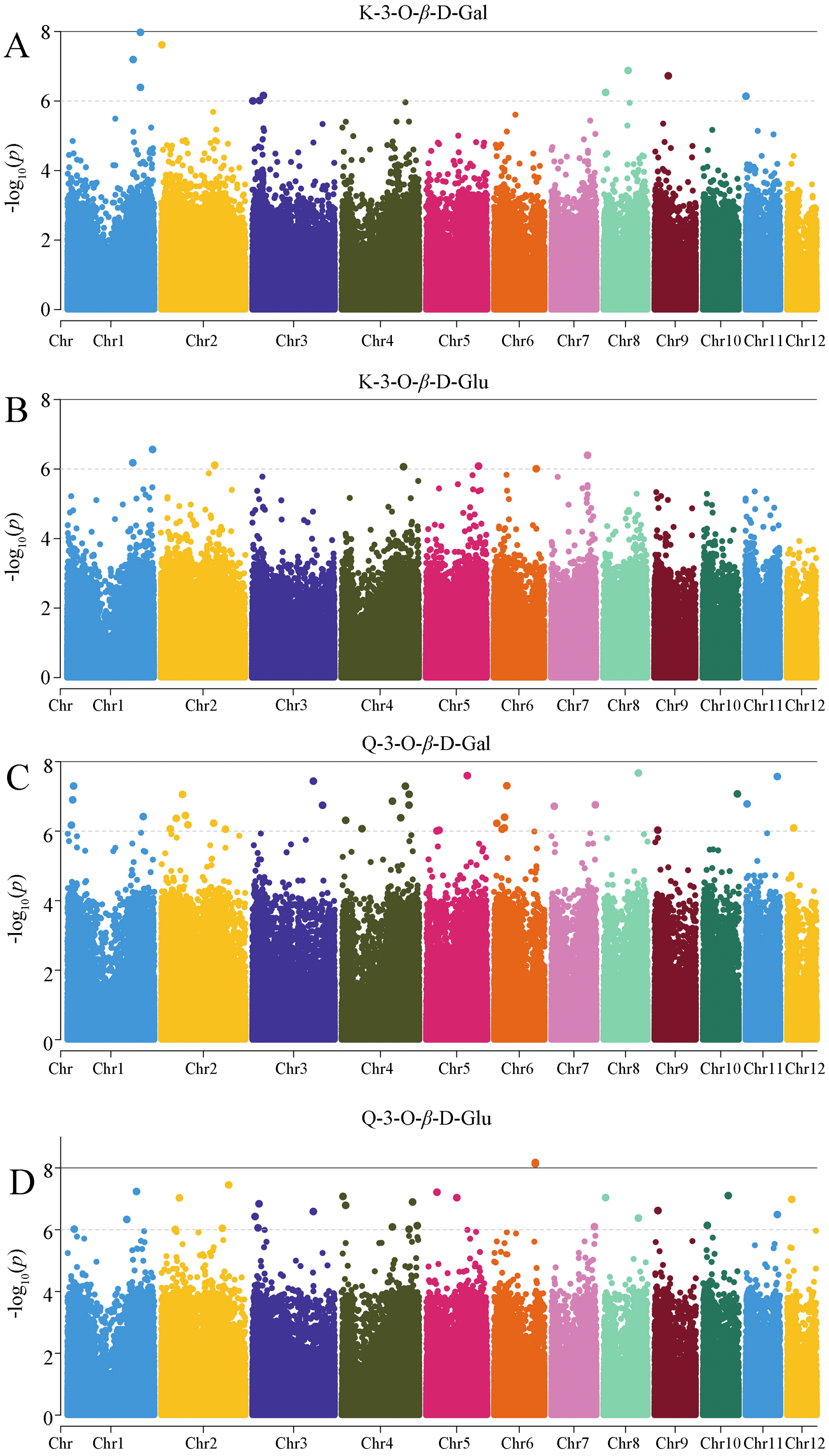

XS12
DZ15
DZ10
DZ9
DZ6
DZ5
DZ19
DZ17
DZ12
DZ16
DZ14
DZ13
DZ4
DZ18
DZ20
DZ7
DZ8
DZ11
ZA6
DZ3
XS2
XS1
DZ21
DZ2
XS19
DZ1
MT5-1
MT19
MT20
MT9
MT18
MT7-1
MT9-1
MT5
MT8-1
MT10-1
MT2
MT14
MT13
MT3
MT1
MT16
MT15
MT8
MT7
MT6-1
MT4
MT12
MT11-1
MT17
KY4
MT6
MT11
MT10
KY11
KY10
KY14
KY3
KY13
KY1
KY9
KY24
KY18
KY15
KY7
KY6
KY17
KY20
KY2
KY16
KY22
KY12
KY5
ZA16
ZA15
ZA2
ZA1-1
ZA20
ZA11
ZA3
ZA7
ZA4
ZA19
ZA18
ZA8
ZA14
ZA12
ZA13
ZA5
ZA10
ZA21
XS4
XS15
XS6
XS8
XS9
XS3
XS5
XS22
XS17
XS14
XS11
XS7
XS13
XS21
XS10
XS20
XS16
XS18
Class
DZ
KY
MT
XS
ZA


0.4
0.3
0.2
0.1
0.0
−0.1
PC2
−0.1
0.0
0.1
0.2
0.3
PC1
DZ
KY
MT
XS
ZA


K=2
K=3
K=4
K=5
1
0.5
0
Population
DZ
KY
MT
XS
ZA
DZ11
KY2
MT1
MT11-1
MT12
MT13
MT14
MT18
MT3
MT5-1
MT6
MT6-1
MT9
XS19
XS2
XS5
ZA12
ZA3
ZA4
ZA7
ZA8
KY1
KY10
KY11
KY12
KY13
KY14
KY15
KY16
KY17
KY18
KY20
KY22
KY24
KY3
KY4
KY6
KY7
KY9
MT19
DZ1
DZ2
DZ21
DZ3
DZ8
MT10-1
MT11
MT4
MT5
MT7-1
MT8
MT8-1
MT9-1
XS10
XS13
XS14
XS16
XS20
XS21
XS7
ZA1-1
ZA19
ZA6
DZ10
DZ12
DZ13
DZ14
DZ15
DZ16
DZ17
DZ18
DZ19
DZ20
DZ4
DZ5
DZ6
DZ7
DZ9
MT2
XS12
ZA13
ZA18
ZA5
KY5
MT10
MT15
MT16
MT17
MT20
MT7
XS1
XS11
XS15
XS17
XS18
XS22
XS3
XS4
XS6
XS8
XS9
ZA10
ZA11
ZA14
ZA15
ZA16
ZA2
ZA20
ZA21


A
-lg(Pvalue)
8
6
4
2
0
CM022944.1_121937053
R^2
1.0
0.8
0.6
0.4
0.2
0.0
NA
Start:121.842
CM022944.1 Region:191.30 (kb)
End:122.033 (Mb)
CM022944.1_121937053
R^2 Color Key
0.0
1.0
B
-lg(Pvalue)
7
6
5
4
3
2
1
0
CM022946.1_13199858
R^2
1.0
0.8
0.6
0.4
0.2
0.0
NA
Start:13.142
CM022946.1 Region:142.52 (kb)
End:13.284 (Mb)
CM022946.1_13199858
R^2 Color Key
0.0
1.0


K-3-O-β-D-Gal
K-3-O-β-D-Glu
Q-3-O-β-D-Gal
Q-3-O-β-D-Glu
Observed -log$_{10}$(p)
Expected -log$_{10}$(p)
0
1
2
3
4
5
6
7
8
10


A
K-3-O-β-D-Gal
B
K-3-O-β-D-Glu
C
Q-3-O-β-D-Gal
D
Q-3-O-β-D-Glu
-log$_{10}$(p)
0
2
4
6
8
Chr
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12

Transcriptome-wide association analysis of flavonoid biosynthesis genes and their correlation with leaf phenotypes in hawk tea (*Litsea coreana* var. *sinensis*)

Lan Yang[1]·Huie Li[2]·Na Xie[1]·Gangyi Yuan[3]·Qiqiang Guo[1*]

[1] Institute for Forest Resources and Environment of Guizhou, Key Laboratory of Forest Cultivation in Plateau Mountain of Guizhou Province, College of Forestry, Guizhou University, Guiyang 550025, People's Republic of China.

[2] College of Agriculture, Guizhou University, Guiyang, 550025, People's Republic of China.

[3] The People's Government of Yongshan County, Yunnan Province, 657000, People's Republic of China.

[*]**Correspondence**: Qiqiang Guo, Institute for Forest Resources and Environment of Guizhou, Key Laboratory of Forest Cultivation in Plateau Mountain of Guizhou Province, College of Forestry, Guizhou University, Guiyang 550025, People's Republic of China. Email: hnguoqiqiang@126.com.

Funding information:National Natural Science Foundation of China, Grant/Award Number: 32060349; China Scholarship Council, Grant/Award Number: 2021(15).

Abstract: Hawk tea (*Litsea coreana* var. *sinensis*), derived from the tender shoots or leaves, rich in flavonoids that can promote healthcare for humans. The primary flavonoid are kaempferol-3-O-β-D-glucoside, kaempferol-3-O-β-D-galactoside, quercetin-3-O-β-D-glucoside, and quercetin-3-O-β-D-galactoside. Is there an association between leaf phenotype and flavonoid content? And the mechanisms of flavonoid biosynthesis are not fully understood. In this study, 109 samples were analyzed to determine the correlation and genetic variability in leaf phenotype and flavonoid content. Furthermore, a transcriptome-wide association study identified candidate loci implicated in the biosynthesis of four key flavonoids. The study revealed that genetic variability in leaf traits and flavonoid concentrations is predominantly attributed to inter-population differences. Flavonoid accumulation may correlate with tree diameter at breast height (DBH), indicative of age-related traits. Transcriptome-wide association analysis identified 84 significant SNPs associated with flavonoid content, with only 13 located within gene regions. The majority of these genes are implicated in metabolic processes and secondary metabolite biosynthesis. Notably, structural genes within these regions are directly involved in pathways known to regulate flavonoid metabolism, exerting a pivotal influence on flavonoid biosynthesis. These results lay a solid theoretical groundwork for subsequent explorations into the genetic determinants influencing flavonoid accumulation of hawk tea.

Keywords: Antioxidant compound; SNP; GWAS; structural genes

1. INTRODUCTION

Hawk tea (*Litsea coreana* var. *sinensis*), an ancient tea species endemic to China, has been cultivated and consumed for millennia in the southwest region (Jia et al. 2017). The tea is primarily derived from tender shoots and leaves, rich in flavonoids, amino acids, volatile oils, and other bioactive compounds (Ye et al. 2012). Research has highlighted that hawk tea's predominant polyphenols are flavonol glycosides, distinguishing it as a caffeine-free beverage (Liang et al. 2007). Flavonols, a subset of flavonoids characterized by a hydroxyl flavone backbone, vary due to the phenolic hydroxyl groups' substitution patterns (Singh et al. 2013). Among the most prevalent flavonoids in vegetation, quercetin and kaempferol stand out as hawk tea's principal flavonols, undergoing glycosylation predominantly at the carbon ring's position 3 (Liu et al. 2020). In addition to their plant-based roles, flavonol glycosides exhibit significant antioxidative activities and stability against light, heat, and oxygen, offering the potential to scavenge free radicals (Fan et al. 2022), inhibit oxidase activity, and provide preventive benefits against cardiovascular, cerebrovascular diseases, and cancer (Bondonno et al. 2019). Their antioxidative properties are intricately linked to anti-aging, with flavonol glycosides playing a crucial role in delaying aging processes, protecting against Alzheimer's disease, and boosting immunity (Yao et al. 2004). In an era marked by growing chronic disease prevalence and a booming food industry, the focus on food health and safety has intensified, spotlighting the development of green health foods and natural additives (Carmela et al. 2022). Hawk tea's inherent health benefits and natural properties underscore its promising future in the food sector.

Current research on hawk tea primarily concentrates on the isolation and characterization of its flavonoid compounds (Yan et al. 2020) and its pharmacological properties (Jia et al. 2017). The flavonoid content has emerged as a critical parameter for assessing the quality of hawk tea germplasm resources. Investigations have revealed significant variability in leaf morphology across different germplasm resources of the same species, serving as a potential criterion for germplasm identification (Khan et al. 2018). This variability may also, to some extent, indicate differences in flavonoid content among these resources (Song et al. 2022). Previous research has uncovered the composition of the main flavonol components in hawk tea, predominantly consisting of kaempferol-3-O-β-D-glucoside (K-3-O-β-D-glu), kaempferol-3-O-β-D-galactoside (K-3-O-β-D-gal), quercetin-3-O-β-D-galactoside (Q-3-O-β-D-gal), and quercetin-3-O-β-D-glucoside (Q-3-O-β-D-glu) (Tan et al. 2022). Recent research offers scant insights into whether leaf morphological traits in hawk tea germplasm resources serve as indicators of flavonoid content. Additionally, diameter at breast height (DBH) has been proposed by Wu et al. (2019) as a growth attribute for identifying superior hawk tea germplasm, particularly when flavonoid content is the primary trait of interest.

Association analysis aims to identify quantitative trait loci through the linkage disequilibrium between different alleles on chromosomes (Liao et al. 2021). A genome-wide association study (GWAS) can serve as a method to investigate genes associated with quantitative traits (e.g., flavonols) in hawk tea. GWAS employs a vast array of high-density single nucleotide polymorphisms (SNPs) throughout the genome as molecular genetic markers for conducting genome-wide correlation analyses (Bhinder et al. 2022). This involves assessing the correlation significance between each variant locus and the target trait, thereby identifying specific gene locus variations that influence the complex trait (Li et al. 2018). However, the complete genome of hawk tea has not been published yet, and possessing a reference genome is a fundamental prerequisite for GWAS analysis (Luo et al. 2019). The continuous advancements in transcriptome sequencing technology coupled with decreasing sequencing costs have facilitated the development of transcriptome-wide association analysis methods. These methods are particularly suited for species whose genomes have not yet been sequenced (Maeda et al. 2019). Utilizing transcriptome sequencing (mRNA-Seq) data to derive gene expression or structural variations and their correlation with phenotypic variations was initially implemented in *Brassica chinensis* (Harper et al. 2012). Compared to GWAS, transcriptome-wide association analysis can identify new candidate genes that, upon functional validation, are capable of regulating target traits, thus demonstrating the reliability of the results obtained through this method (Kim et al. 2011). Given that the full genome data for hawk tea remains unpublished, full-length transcriptome sequencing has become increasingly significant for this species.

The relationship between the flavonoid content and its leaf phenotypic traits, as well as the genetic foundation of its biosynthesis, remains uncharted territory necessitating further research and thorough investigation. Therefore, in this study, the genetic and phenotypic differentiation coefficients of leaf character, DBH, and flavonoid content of one leaf and two buds in 109 samples of hawk tea from five regions were calculated, and conducted correlation analysis. Furthermore, transcriptome sequencing was conducted on 109 samples, utilizing second and third-generation sequencing technologies. Subsequently, transcriptome-wide association analysis was conducted, leveraging data on flavonoid content and a high-quality SNP dataset. The aim of the study was to investigate whether the variation in leaf character, DBH, and flavonoid content in hawk tea primarily originates between or within populations, identify variables highly associated with flavonoid content, and ascertain SNPs with high correlations to flavonoid biosynthesis in hawk tea. This study is anticipated to offer theoretical insights for advancing research on the natural variation and associated genetic structure of hawk tea. Additionally, it could provide direction for future endeavors in breeding and transgenic research aimed at enhancing the flavonoid content in hawk tea.

2. MATERIAL AND METHODS

2.1 Leaf character, DBH, and flavonoid content determination and analysis

Hawk tea is classified as a diploid organism (Ha et al. 2022). In May 2021, samples of the same species were collected from five sites in Kaiyang County (KY), Xishui County (XS), Meitan County (MT), Daozhen County (DZ), and Zheng'an County (ZA) in Guizhou Province, China (Fig. 1). The five sites feature a subtropical humid monsoon climate, characterized by distinct local microclimates and significant vertical climate variations. The average annual temperature ranges from 13.19 to 15.59 °C, with annual precipitation between 1, 080 and 1, 255 mm (Yuan et al., 2023). Hawk tea was systematically investigated and sampled at the designated site, with adult plants being specifically targeted for sampling. To mitigate the impact of kinship relations, a minimum distance of 30 meters was maintained between each sampled individual. One hundred and nine samples were collected totally, including twenty-one samples from DZ County, they were primarily found in open areas near the river and on the hillside, with limited seedling regeneration, the slope ranged from 7 to 18 degrees, facing southeast. Twenty-two samples from XS County, they were primarily distributed in evergreen broadleaved forests surrounding cultivated land and on nearby slopes, seedling regeneration is observed under the forest canopy, with slopes ranging from 10 to 15 degrees and facing southwest. Nineteen samples from ZA County, they were primarily found in secondary evergreen broad-leaved forests or bamboo forests with high canopy density, adult individuals are few and mostly located in open areas, the slope ranges from 3 to 8 degrees and faces southwest. Twenty samples from KY County, they were primarily found in mountain orchards near villages, characterized by a low canopy and slopes ranging from 8 to 15 degrees, facing southwest. And twenty-seven samples from MT County, they were primarily distributed in open mountains near cultivated land and around the reservoir, no seedling regeneration was observed, the slopes range from 8 to 16 degrees and face south.

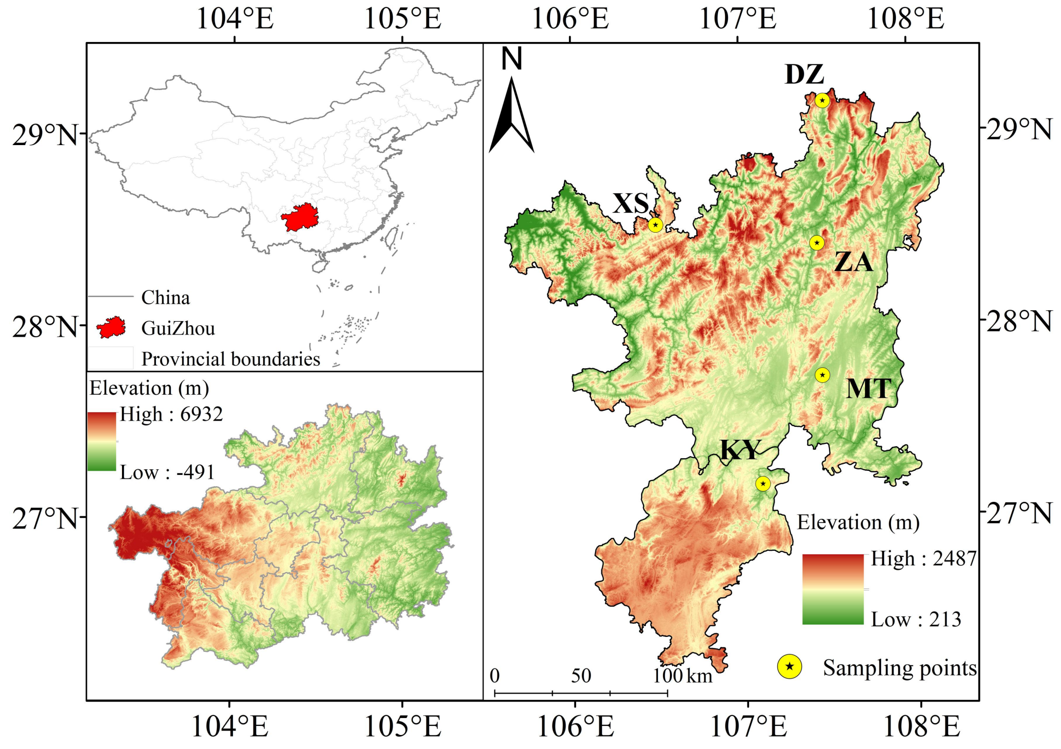


Fig. 1. A map showing the natural distribution and the location of study areas in Guizhou province, SW China. (KY: Kaiyang County, XS: Xishui County, MT: Meitan County, DZ: Daozhen County, ZA: Zheng'an County)

The DBH of each tree was recorded, and mature leaf samples free from pests and diseases were individually collected from the cardinal directions-southeast and northwest. Following labeling, the samples were secured in ziplock bags, stored at 4°C in a portable refrigerator, and transported to the laboratory on the same day for assessment of leaf phenotypic indicators. In addition, one leaf and two buds were collected to wrapped carefully in tin foil, labeled, immediately frozen in liquid nitrogen, and stored in a -80°C refrigerator for further analysis.

Leaf length (LL), leaf width (LW), leaf area (LA), leaf thickness (LT), and leaf perimeter (LP) were quantified using a portable leaf area meter (AM350, ADC, UK), and the leaf shape index (LS) (leaf length/leaf width) was calculated. Leaf petiole length (LPL) was measured with an electronic digital caliper to an accuracy of 0.01 mm, and the relative chlorophyll content (SPAD) was noted using a chlorophyll meter (SPAD-502). The fresh weight of the leaves was determined using an electronic balance accurate to 0.01g. Leaves were then dried at 80°C for 48 hours until reaching a constant weight, at which point the dry weight was measured. The leaf dry matter content (LDMC) and specific leaf area (SLA) were calculated, representing the ratio of dry weight to fresh weight and the ratio of leaf area to dry weight, respectively.

The contents of flavonoid from one and two buds, including K-3-O-β-D-gal, K-3-O-β-D-glu, Q-3-O-β-D-gal, and Q-3-O-β-D-glu, were determined and extracted using high-performance liquid chromatography (HPLC) following the methodology outlined by Liang et al. (2005).

2.2 Data analysis

Phenotypic data underwent descriptive statistical analysis utilizing R software, version 3.6. Variance analysis for all traits was conducted employing a linear model, articulated as: $X_{ijk}=\mu+P_i+C_{j(i)}+\varepsilon_{ijk}$, where X represents phenotypic individual observations, μ represents the population average, and P_i represents effect at place i (i=1, 2, 3, 4, 5), $C_{j(i)}$ represents the effect of the j clone in the i origin (j=1,2,..., 20), ε_{ijk} represents residual.

The effects within and between origin clones, excluding the overall mean, were treated as random variables. ANOVA analysis was conducted using PROC GLM in SAS software (SAS Institute, Inc., SAS/STAT software, v8) to investigate differences both between and within the origin clones. The variance components, namely σ_{p^2} (between the origins), $\sigma_{c(p)^2}$ (within the origins), and σ_{e^2} (residual), were estimated based on the previously mentioned linear model. The coefficient of variation (CV) was calculated using the following formula: $CV=\delta_p/\mu$, where δ_p represents the standard deviation of the phenotype, and μ represents the mean value of the phenotype. The genetic correlation matrix and phenotypic correlation matrix between the two traits were calculated, and significance tests were conducted using the asreml software package.

2.3 RNA-seq

RNA extraction was conducted from one leaf and two buds of each hawk tea clone sample utilizing the RNA rapid extraction kit (Beijing, China). For quality control, each sample purity of OD260/280 between 2.0-2.2 and RIN value of ≥8.0. Subsequently, equal amounts of total RNA from each sample were pooled, and the task of conducting transcriptome sequencing was entrusted to Hangzhou Kaitai Biotechnology Co., LTD. In this process, the utmost accuracy in our transcriptome sequencing results was ensured by utilizing high-quality transcript assembly, which combined second-generation transcriptome sequencing with third-generation full-length transcriptome sequencing.

The raw image data generated from the second-generation high-throughput sequencing instrument, Illumina NovaSeq 6,000, were subjected to base calling to convert them into sequence data, resulting in the acquisition of raw reads. It is important to note that these raw reads may potentially contain adapters or low-quality base reads, which have the potential to adversely affect subsequent analyses. Therefore, it is imperative to perform data filtering to ensure the integrity of the information analysis process. In the context of quality control sequencing, the quality of the bases plays a critical role in achieving high sequencing accuracy (Li et al. 2004). Q20 serves as a primary criterion for assessing data quality. An attainment of Q20 greater than 85% signifies that over 85% of the bases exhibit a sequencing accuracy rate of 99% (Baid et al. 2023). To achieve this, the data is disconnected from the sequencing platform, and a multi-step data filtering process is subsequently executed, as detailed below:

a. Reads with joint contamination greater than 5bp were excluded from the dataset. In the case of double-ended sequencing, both ends of the reads were discarded if one end exhibited splicing contamination.

b. Reads with a quality score (Q) below 15, encompassing more than 30% of their length, were eliminated. In the context of double-ended sequencing, if one end contained low-quality reads, both ends were removed.

c. Reads that contained more than 5% of the base 'N' were filtered out. In the case of double-ended sequencing, if one end contained more than 5% 'N' bases, that specific end was excluded from the analysis.

Three generations of sequencing data were acquired utilizing Oxford Nanopore Technologies (ONT) sequencers. ONT sequencing boasts extended read lengths and high throughput, making it particularly advantageous in genome assembly, transcriptome assembly, epigenetic modification studies, and various other research domains (Zhang et al. 2023). The data filtering process was executed as follows: initial data assessment and statistics were performed using NanoPlot, followed by joint processing using Porechop. Subsequently, mass filtration was conducted with Nanofilt. Finally, NanoPlot was employed once more for comprehensive data statistics and evaluation of the resulting clean data. The merge assembly approach was employed to consolidate multiple samples into an initial transcriptome set. In cases where the sample size exceeded 20 samples, a random selection method was adopted, grouping them into sets of three, ensuring the inclusion of a total of 15 samples in the subsequent assembly process. The integration of NGS data and ONT data was accomplished using the default parameters of rnaspades v3.15.2, with the resulting transcripts fasta file serving as the foundation for subsequent analyses. To gauge the quality of assembly, reads were aligned to the assembled transcripts fasta using bowtie2 v2.4.2 to calculate the mapping rate, where a higher mapping rate is generally indicative of superior assembly integrity (Hyten et al. 2010). Assessment of transcript assembly integrity was carried out using BUSCO v5.0.0.

2.4 SNP calling

STAR2.3 was employed for the comparison, and GATK4 was utilized for SNP calling. With *Litsea cubeba* as the reference, read mapping was conducted using STAR, information was appended to the BAM files using "Add or Replace Read Groups," and repeated reads were annotated using "Mark Duplicates." The BAM files were subsequently subjected to validation using "Validate SamFile," while splice reads underwent processing through "Split NCigar Reads." SNP calling was executed with "Haplotype Caller," and VCF merging was accomplished with "MergeVcfs." Variation filtration was applied using "Variant Filtration," and variants were extracted using "Select Variants," retaining only those reads that passed the filtration criteria. Mutation statistics were generated with Vcftools, and data visualization was performed using R packages.

Data conversion was carried out with vcf2phy to ensure that 90% of individuals possessed base information at the same site. Evolutionary trees were constructed using IQTrees, and for phylogenetic tree visualization, ggtree was employed. Subsequently, data conversion was conducted using plink, followed by PCA analysis utilizing Smartpca, and the results were visualized through ggplot2. Structural analysis was performed using admixture, with K values ranging from 2 to 5 chosen for display. Furthermore, Gmap was utilized to forecast the mapping of three generations of transcriptome sequences (CDS) onto the reference genome, determining their structural positions.

2.5 Transcriptome association analysis

Plink was employed for data transformation, and the association between SNP sites and flavonols was analyzed using a general linear model (GLM). The filtering criterion was set to -Log10(*p*)>6.0. LD Block Show and Show LD SVG were utilized to construct LD blocks within the GWAS locus. The top 10 most significant loci were selected for each trait. The regions of interest extended 100kb base pairs upstream and downstream of each significant association site, resulting in the analysis of a total of 200kb base pair regions.

3. RESULTS

3.1 Genetic variation in DBH, leaf traits, and flavonoid content in hawk tea.

The results of variance analysis (Table 1) indicate that, except for DBH, the origin significantly influenced leaf traits and flavonoid content (P<0.001). The origin's impact accounted for 0.01% to 57.83% of the total variation in leaf traits and 0.57% to 31.69% of the total variation in flavonoid content. Among leaf traits, the clonal population in the KY area exhibited the highest values for LW, LA, LP, SPAD, and SLA, which were 4.57, 35.35, 31.51, 48.22, and 66.16, respectively (Table 2).

Table 1. Variance analysis and genetic parameter estimation of leaf traits, DBH, and flavonoid content

Traits	Mean±SD	CV	σ_p^2	$\sigma_{c(p)}^2$	σ_e^2
DBH	10.93±1.83	31.64	13.07	62.44	61.75
LL	11.51±1.29	15.55	41.59***	1.61	2.99
LW	3.53±0.68	19.26	8.70***	0.32	0.45
LPL	1.15±0.27	23.48	0.40***	0.06	0.07
LT	0.26±0.89	342.31	0.14***	0.00	0.01
LA	25.11±3.38	29.39	57.73***	12.17	54.97
LP	25.59±0.83	18.87	48.52***	6.73	23.49
SPAD	45.80±1.18	11.31	57.83***	21.31	24.75
LDMC	0.52±0.04	7.69	0.01***	0.00	0.00
LS	3.34±0.65	19.46	3.89***	0.27	0.35
SLA	48.7±1.97	10.20	33.32***	65.51	223.94
1	0.64±0.45	70.31	1.40***	0.15	0.18
2	2.39±1.75	73.22	23.31***	2.20	2.56
3	0.48±0.30	62.50	0.57***	0.07	0.08
4	8.88±0.20	69.82	31.69***	25.71	28.93

DBH: Diameter at breast height (cm); LL: Leaf length (cm); LW: Leaf width (cm); LA: Leaf area (cm^2); LT: Leaf thickness (cm); LP: Leaf perimeter (cm); LS: Leaf shape index; LPL: Leaf petiole length (cm); SPAD: The relative chlorophyll content; LDMC: Leaf dry matter content; SLA: Specific leaf area; 1: K-3-O-β-D-gal (mg/g dry weight); 2: K-3-O-β-D-glu (mg/g dry weight); 3: Q-3-O-β-D-gal (mg/g dry weight); 4: Q-3-O-β-D-glu (mg/g dry weight); CV: coefficient of variation; σ_p^2: variation between the origin; $\sigma_{c(p)}^2$: variation within the origin; σ_e^2: residual.

*: *p*<0.05, **: *p*<0.01, ***: *p*<0.001.

Conversely, the LL, LW, LPL, LA, LP, and SPAD of clonal populations in the ZA area were the smallest, measuring 10.25, 2.79, 0.94, 17.52, 21.05, and 41.29, respectively. The maximum LL observed in the clonal population was 13.57 in the XS area (Table 2).

Table 2. Average leaf traits, DBH, and flavonoids in 5 areas

	DZ	KY	MT	ZA	XS
DBH	11.03±1.71	10.66±1.03	13.05±1.42	9.68±1.87	10.25±0.98
LL	10.65±1.35c	12.48±1.23b	10.61±1.02c	10.25±1.42c	13.57±1.28a
LW	3.21±0.37c	4.57±0.39a	3.50±0.38b	2.79±0.29d	3.59±0.27b
LPL	1.13±0.20b	1.20±0.33ab	1.33±0.25a	0.94±0.18c	1.15±0.23ab
LT	0.21±0.04b	0.22±0.03b	0.24±0.03b	0.41±0.08a	0.21±0.03b

LA	20.84±2.98c	35.35±2.91a	22.07±2.52c	17.52±2.84d	29.76±1.66b
LP	23.00±2.76c	31.51±2.78a	22.98±1.68c	21.05±2.70c	29.40±2.87b
SPAD	46.84±1.52a	48.22±2.80a	47.69±0.36a	41.29±5.45b	44.94±3.60a
LDMC	0.54±0.03a	0.54±0.02a	0.50±0.04bc	0.52±0.04b	0.49±0.03c
LS	3.37±0.63b	2.75±0.33c	3.08±0.51b	3.72±0.66a	3.79±0.39a
SLA	38.95±1.42d	66.16±0.70a	44.15±0.91c	33.53±0.39e	60.87±0.34b
1	0.21±0.11b	0.66±0.27a	0.91±0.47a	0.62±0.34a	0.79±0.57a
2	0.70±0.39c	2.54±0.87ab	3.16±1.43ab	2.12±1.14b	3.44±2.60a
3	0.18±0.06b	0.53±0.16a	0.58±0.33a	0.49±0.34a	0.60±0.32a
4	2.36±1.18c	9.61±3.06b	9.82±2.38b	8.77±2.15b	13.81±1.89a

DBH: Diameter at breast height (cm); LL: Leaf length (cm); LW: Leaf width (cm); LA: Leaf area (cm^2); LT: Leaf thickness (cm); LP: Leaf perimeter (cm); LS: Leaf shape index; LPL: Leaf petiole length (cm); SPAD: The relative chlorophyll content; LDMC: Leaf dry matter content; SLA: Specific leaf area; 1: K-3-O-β-D-gal (mg/g dry weight); 2: K-3-O-β-D-glu (mg/g dry weight); 3: Q-3-O-β-D-gal (mg/g dry weight); 4: Q-3-O-β-D-glu (mg/g dry weight).

Values with different superscripts in the same column significantly differ at the 0.05 level.

Regarding flavonoid content, Q-3-O-β-D-glu, K-3-O-β-D-gal, and K-3-O-β-D-glu exhibited the highest values in clonal populations from the XS area, measuring 3.44, 0.60, and 13.81, respectively. Conversely, the contents of these four flavonoids in the clonal population from the DZ area were the smallest, measuring 0.21, 0.70, 0.18, and 2.36, respectively.

For DBH, the clonal variation between and within origins did not reach a significant level ($p > 0.05$). In contrast, for leaf traits and flavonoid content, the primary source of genetic variation stemmed from the variation between populations.

3.2 Correlations among DBH, leaf traits, and four types of flavonoids.

The phenotypic and genetic correlations among DBH, leaf traits, and the four flavonoids are presented in Table 3. The results indicated that the phenotypic and genetic correlation coefficients between K-3-O-β-D-glu, DBH, and leaf traits were not statistically significant, with coefficients ranging from 0.2064 to 0.4086.

Table 3. Genetic correlation (upper triangle) and phenotypic correlation (lower triangle) among DBH, leaf traits, and four flavonoids

	DBH	LL	LW	LPL	LT	LA	LP	SPAD	LDMC	LS	SLA	1	2	3	4
DBH	1	0.1230	1.8832***	54.8223***	853.1259***	0.0014	0.0055	0.0076	18903.3056***	3.1337***	0.0002	11.0104***	0.1745*	37.9788***	0.0036
LL	0.7481***	1	0.0762	2.3373***	34.0007***	0.0000	0.0001	0.0004	887.0409***	0.1194	0.0000	0.50050***	0.0076	1.7298***	0.0002
LW	0.6058***	0.5105***	1	0.7693***	10.8626***	0.0000	0.0000	0.0001	278.1474***	0.0261	0.0000	0.1603	0.0025	0.5546***	0.0001
LPL	0.5417***	0.7112***	0.8442***	1	8.7696***	0.0000	0.0001	0.0001	202.7070***	0.0337	0.0000	0.1113	0.0013	0.3957***	0.0000
LT	0.2461**	0.1585	0.2921***	0.1928*	1	0.0000	0.0000	0.0000	36.5254***	0.0061	0.0000	0.0214	0.0003	0.0729	0.0000
LA	0.1759*	0.0256	0.0546	0.0201	0.2038**	1	0.0002	0.0011	2971.2349***	0.4635***	0.0000	1.6884***	0.0259	5.8420***	0.0005
LP	0.2030**	0.0762	0.2120**	0.1207	0.1941*	0.1151	1	0.0008	2025.2952***	0.3358***	0.0000	1.1516***	0.0177	3.9258***	0.0004
SPAD	0.0098	0.1532	0.0440	0.0879	0.2387**	0.4069***	0.0835	1	3783.7029***	0.5233***	0.0000	2.1893***	0.0350	7.4346***	0.0007
LDMC	0.3086***	0.1894*	0.2406**	0.1550	0.8505***	0.0591	0.0736	0.1198	1	0.0043	0.0000	0.0150	0.0002	0.0517	0.0000
LS	0.2194**	0.1444	0.2389**	0.1521	0.9586***	0.2447**	0.0535	0.2498**	0.88700***	1	0.0000	0.2186**	0.0035	0.7544***	0.0001
SLA	0.0625	0.1115	0.0242	0.0012	0.4208***	0.0943	0.1051	0.4139***	0.4523***	0.4338***	1	3.5279***	0.0533	12.3344***	0.0000
1	0.1999**	0.1661	0.2472**	0.2337**	0.2789***	0.0225	0.0576	0.0673	0.3132***	0.2760***	0.2383**	1	0.0008	0.2895***	0.0000
2	0.0893	0.1134	0.1709*	0.1231	0.8207***	0.6548***	0.1057	0.3804***	0.6624***	0.8660***	0.3857***	0.2576**	1	1.6567***	0.0001
3	0.3318***	0.1697*	0.2763***	0.1730*	0.7374***	0.4304***	0.0450	0.0243	0.8623***	0.7318***	0.3992***	0.3228***	0.3594***	1	0.0000
4	0.0023	0.0006	0.0010	0.0004	0.0471	0.0544	0.0808	0.0269	0.0267	0.0367	0.1286	0.1324	0.0094	0.0933	1

DBH: Diameter at breast height (cm); LL: Leaf length (cm); LW: Leaf width (cm); LA: Leaf area (cm2); LT: Leaf thickness (cm); LP: Leaf perimeter (cm); LS: Leaf shape index; LPL: Leaf petiole length (cm); SPAD: The relative chlorophyll content; LDMC: Leaf dry matter content; SLA: Specific leaf area; 1: K-3-O-β-D-gal (mg/g dry weight); 2: K-3-O-β-D-glu (mg/g dry weight); 3: Q-3-O-β-D-gal (mg/g dry weight); 4: Q-3-O-β-D-glu (mg/g dry weight).

*: p<0.05, **: p<0.01, ***: p<0.001.

On the other hand, Q-3-O-β-D-gal and K-3-O-β-D-gal exhibited significant positive correlations with LL, LA, LP, SPAD, LS, and SLA. Additionally, K-3-O-β-D-gal was positively correlated with LW, and LPL showed a significant positive correlation. Furthermore, Q-3-O-β-D-gal, Q-3-O-β-D-glu, and K-3-O-β-D-gal displayed significant positive correlations with DBH, suggesting that DBH can serve as an indirect selection indicator for hawk tea flavonoids.

Regarding the correlations between the four flavonoid contents, Q-3-O-β-D-gal and Q-3-O-β-D-glu (p<0.01), Q-3-O-β-D-gal and K-3-O-β-D-gal, Q-3-O-β-D-glu and K-3-O-β-D-gal (p<0.001) exhibited statistically significant phenotypic correlations (Table 3). Additionally, significant genetic correlations were observed between Q-3-O-β-D-gal and K-3-O-β-D-gal, as well as between Q-3-O-β-D-glu and K-3-O-β-D-gal ($p < 0.001$).

3.3 Second and third-generation sequencing and SNP statistics

Based on the Clean Data statistics for each hawk tea sample, the data utilization rate falls within the range of 93.15% to 98.81%. The distribution of GC content ranges from 46.37% to 49.36%. Furthermore, more than 94% of the bases exhibit a Q30 quality score (Fig. 2A). These observations collectively indicate that the sequencing data possesses high quality and is suitable for sequence fragment assembly and subsequent analysis. The raw reads have been deposited in NCBI and are accessible under BioProject PRJNA992466.

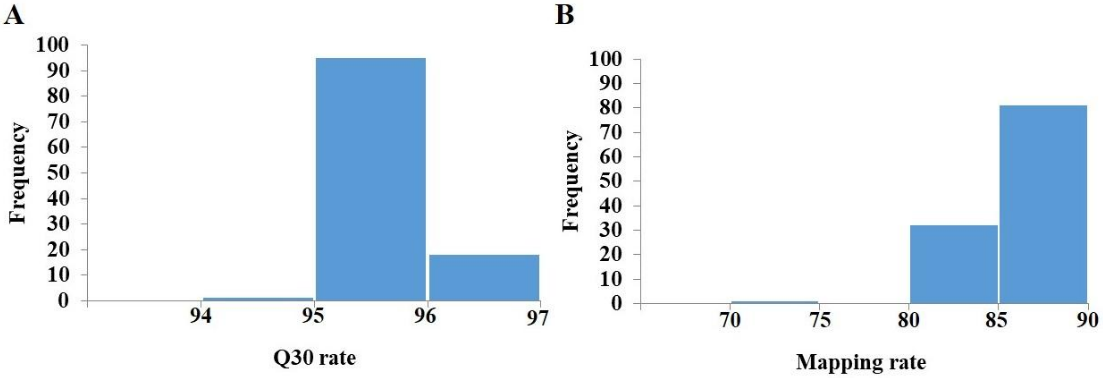


Fig. 2. Statistics of sequencing results. (A) The distribution of Q30 rates. (B) The distribution of alignment rates.

Following the assembly and splicing process, a total of 349,993 transcripts were obtained, comprising 449,816,814 bases. The average transcript length was 1,285bp, with an N50 length of 2,494bp. Notably, transcripts falling within the 200-500bp range constituted a relatively substantial portion, accounting for 43.35% of the total transcripts (Table 4).

Table 4. Statistical distribution of transcription length sequence

Unigene length	Total Number
200-500bp	151731(43.35%)
500-1000bp	63231(18.06%)
1000-2000bp	61033(17.42%)
2000-3000bp	33484(9.56%)
>3000bp	39791(11.37%)

Total Number	349993
Total Length	449816814
N50 Length	2494
Mean Length	1285

With the exception of KY13, all other samples exhibited mapping values exceeding 94%, and the transcript integrity was notably high at 96.3%, as assessed by BUSCO software. In general, assembled results typically fall within the range of 70% to 98% and are deemed suitable for subsequent analysis (Kishi et al. 2022).

The valid data obtained were compared with the *Litsea cubeba* genome, yielding an average alignment rate of 85.37%, falling within a confidence interval of 72.53% to 88.59% (Fig. 2B). Subsequently, SNP calling was conducted using GATK, resulting in each sample containing more than 600,000 SNPs (Fig. 3). Notably, the Phred values for the majority of these sites exceeded 1,000. Fig. 3 illustrates the distribution of these SNPs across chromosomes, revealing that, apart from chromosome 12, each of the other chromosomes harbored more than 10,000 SNPs.

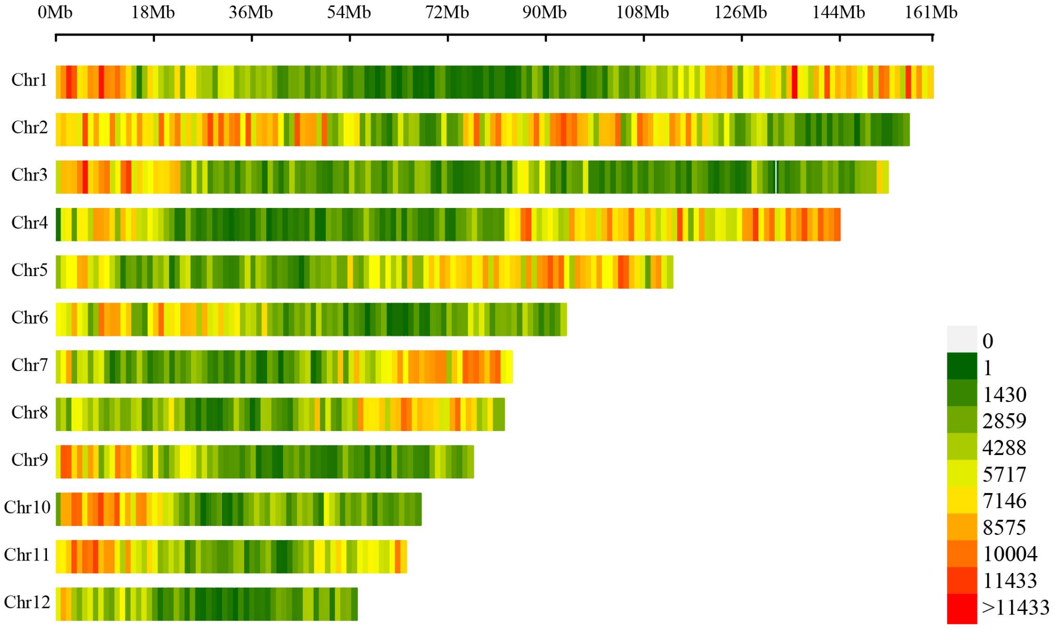


Fig. 3. Distribution of SNP density across chromosomes. (Different colored regions indicate varying SNP counts across chromosomes)

3.4 Genetic evolutionary analysis

Based on the phylogenetic tree constructed using the neighbor-joining clustering method, which was based on genetic distance, the results (Fig. 4) revealed the division of 109 hawk tea clones from 5

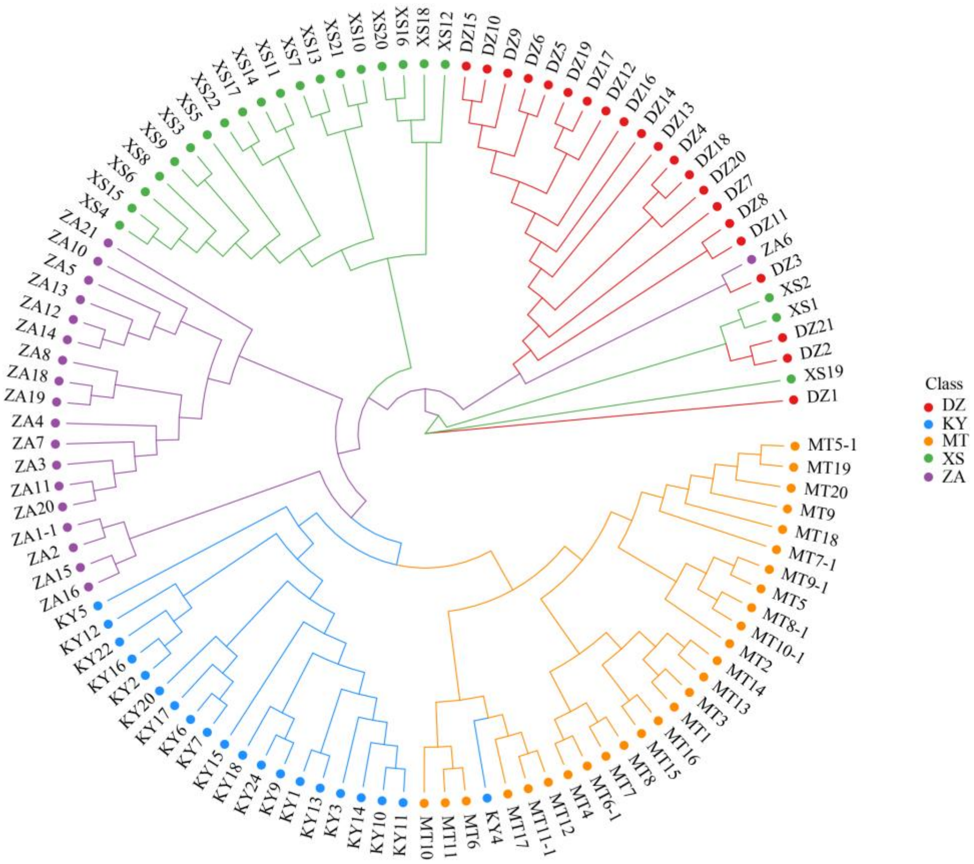


Fig. 4. Phylogenetic tree of hawk tea populations constructed based on genetic distance. (Red represent DZ area, blue represent KY area, orange represent MT area, green represent XS area, purple represent ZA area.)

different regions into 5 distinct subgroups. The first subgroup primarily consisted of clones from DZ, XS, and ZA, while the second subgroup was predominantly composed of clones from XS. Clones from the ZA provenance dominated the third subgroup, whereas the fourth subgroup was mainly comprised of clones from KY. The fifth subgroup predominantly consisted of clones from MT and KY.

Further insights into the clustering patterns of all samples were obtained through PCA analysis of the transformed data (Fig. 5). This analysis categorized the samples into 5 distinct groups, with DZ, KY, MT, and XS forming 4 separate categories, while ZA clustered together with DZ and XS. To delve into the population structure of the studied materials, Admixture software was employed (Fig. 6). The results indicated that when K=5, the 109 hawk tea clones were classified into five subgroups, with the lowest cross-verification error rate observed at this value.

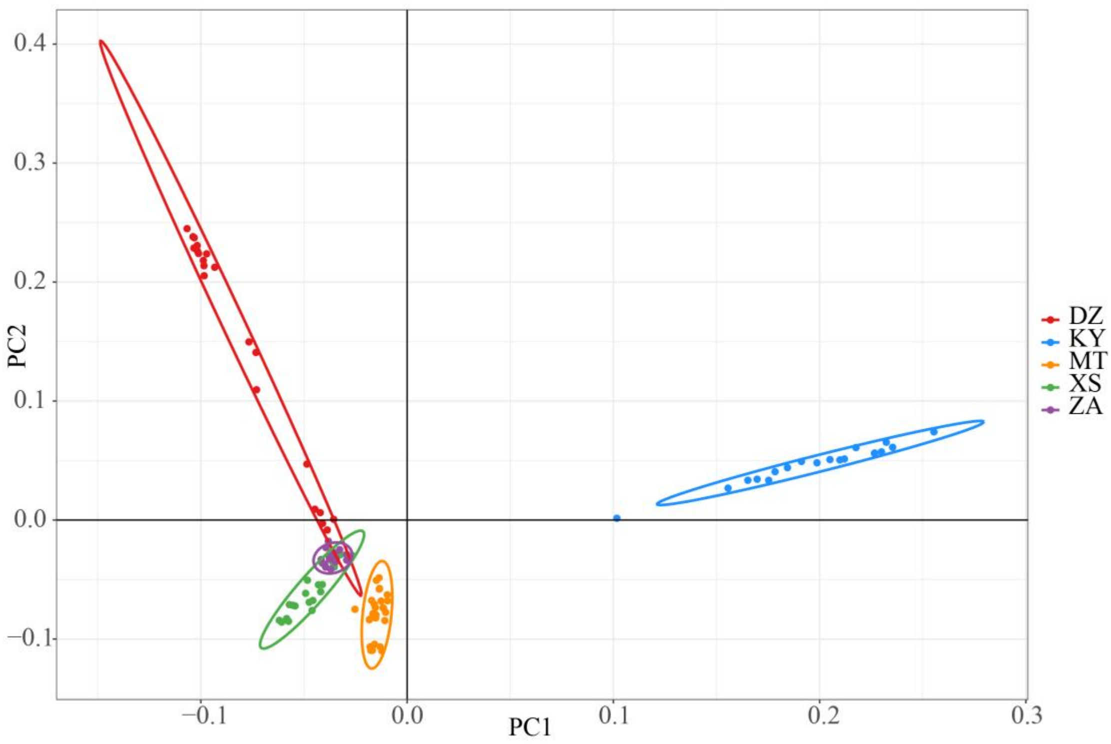


Fig. 5. Principal component analysis of hawk tea. (Red represent DZ area, blue represent KY area, orange represent MT area, green represent XS area, purple represent ZA area.)

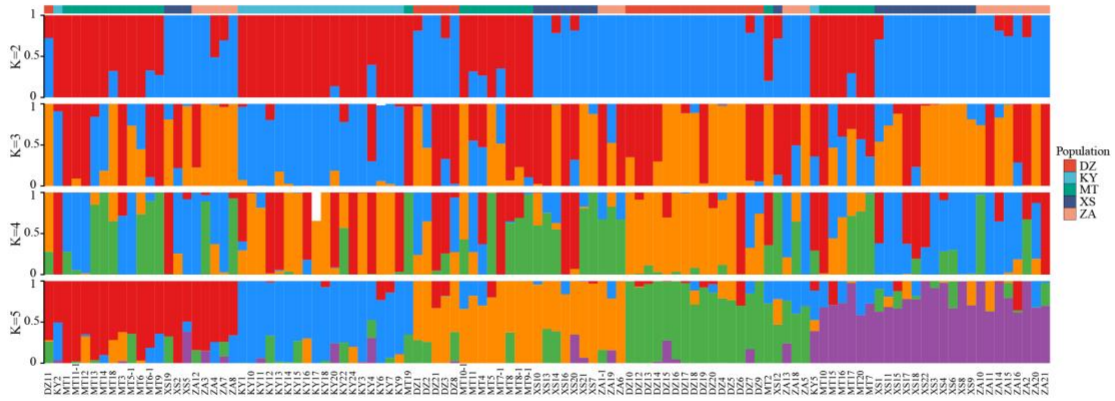


Fig. 6. Results of the Bayesian clustering analysis conducted using STRUCTURE. (Highlighting the clustering patterns of genetic components across 2-5 groupings.)

The population was stratified into 5 subgroups through the application of admixture software, the neighbor-joining clustering method based on genetic distance, and principal component analysis. It can be inferred that the clustering outcomes obtained from these three methods exhibited analogous trends, thereby indicating a relatively high level of reliability in the clustering results.

3.5 Transcriptome-wide association analysis of flavonoids

The primary content of the four flavonoids in the tender shoots of hawk tea has been determined, and significant variations in flavonoid content among different cultivation regions and clones of hawk tea have been observed (Table 1). In practical applications, the selection of superior traits within the hawk tea species is a matter of great urgency. Therefore, the exploration of genetic loci linked to these crucial traits is deemed of substantial importance. Consequently, the inaugural transcriptome-wide association analysis in hawk tea has been undertaken in this study, with the objective of identifying significant SNPs associated with the four flavonoids.

After filters were applied based on criteria such as marker missing rate, sample missing rate, and minor allele frequency (MAF), a total of 235 high-quality SNPs associated with flavonoids were identified, of which 84 demonstrated statistical significance. Among these SNPs, 66 (78.57%) were found to be situated in intergenic regions. Further breakdown reveals that 10 werelocated in upstream regions, 23 in introns, 9 in downstream regions, 15 represented missense variants, and 9 were synonymous variants. Moreover, functional annotations were available for 44 of these SNPs (Table S1). In hawk tea's tender shoots, significant SNPs associated with the four flavonoids were identified, with totals of 11, 7, 30, and 36 for each respective flavonoid. It is important to mention that only a limited number of SNPs were localized within gene regions (Table 5).

Given the unavailability of the hawk tea genome, our investigation was constrained to genes exhibiting significant SNPs. A total of 44 protein-coding genes presenting *p*-values below 0.0001 were discerned (Table S1). Three genes, associated with K-3-O-β-D-gal content, were categorized into three

Table 5. Summary of the significant SNPs by associated analysis

Traits	Significant SNPs	SNPs in genic region	Associated genes	Chromosome	SNP position	-log10 P valuae	Annotation	KEGG pathways	KO
K-3-O-β-D-Gal	11	8	3	CM022944.1	121937053	7.1918	Cytochrome P450 CYP4/CYP19/CYP26 subfamilies	CYP86B1; fatty acid omega-hydroxylase	K09590
				CM022946.1	13199858	6.0132	Selenium-binding protein	SELENBP1; methanethiol oxidase	K17285
				CM022952.1	25558941	6.7258	UDP-glucuronosyl and UDP-glucosyl transferase	UGT74B1; N-hydroxythioamide S-beta-glucosyltransferase	K11820
K-3-O-β-D-Glu	7	4	1	CM022944.1	121937053	6.1803	Cytochrome P450 CYP4/CYP19/CYP26 subfamilies	CYP86B1; fatty acid omega-hydroxylase	K09590
Q-3-O-β-D-Gal	30	12	3	CM022945.1	39681060	7.0556	Predicted importin 9	IPO9, RANBP9; importin-9	K20224
				CM022945.1	28136888	6.3667	Serine/threonine protein phosphatase 2A, regulatory subunit	PPP2R5; serine/threonine-protein phosphatase 2A regulatory subunit B'	K11584
				CM022947.1	7598503	6.3111	Dihydrolipoamide acetyltransferase	DLAT, aceF, pdhC; pyruvate dehydrogenase E2 component (dihydrolipoyllysine-residue acetyltransferase)	K00627
K-3-O-β-D-Glu	36	20	6	CM022953.1	47477583	7.1035	Scaffold/matrix specific factor hnRNP-U/SAF-A, contains SPRY domain	DLD,1pd,pdhD; dihydrolipoy1 dehydrogenase	K00382
				CM022947.1	2558160	7.0773	-	ppc; phosphoenolpyruvate carboxylase	K01595
				CM022946.1	12277330	6.8386	Sterol O-acyltransferase/ Diacylglycerol O-acyltransferase	P4HA; prolyl4-hydroxylase	K00472
				CM022947.1	7529974	6.7878	-	DGAT1; diacylglycerol O-acyltransferase 1	K11155
				CM022950.1	79888055	6.0975	Mitogen-activated protein kinase	LEU1; 3-isopropylmalate dehydratase	K01702
				CM022944.1	12714282	6.0191	-	HPR2-3; glyoxylate/hydroxypyruvate reductase	K15919

distinct functional classes: the cytochrome P450 subfamilies CYP4/CYP19/CYP26, selenoproteins, and uridine diphosphate glucose transferases (Fig. 7). Regarding K-3-O-β-D-glu, a solitary gene from the cytochrome P450 subfamilies CYP4/CYP19/CYP26 was identified. In contrast, Q-3-O-β-D-gal content was linked to five genes, inclusive of those coding for dihydroceramide transferases. Moreover, six genes correlated with K-3-O-β-D-glu content were also pinpointed (Table 5).

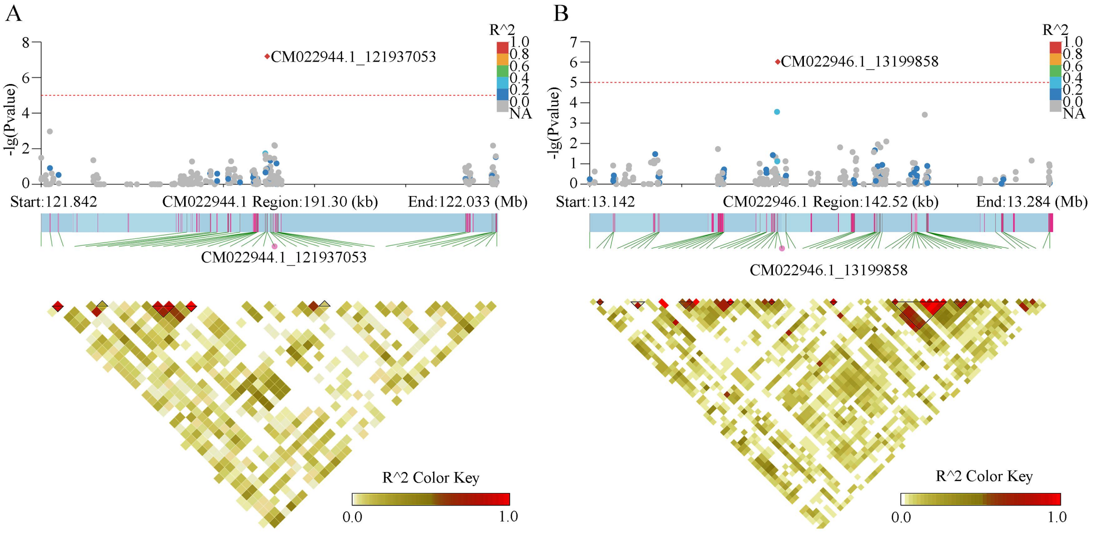


Fig. 7. Association screening for the SNP locus grounded on kaempferol-3-O-β-D-galactoside content. (A) Location of SNP locus 1219377053 on the chromosome CM022944.1. (B) Location of SNP locus 13199858 on the chromosome CM022946.1.

The examination of *p*-value distributions from GLM association analyses for flavonol traits, K-3-O-β-D-gal, K-3-O-β-D-glu, Q-3-O-β-D-gal, and Q-3-O-β-D-glu (Fig. 8), results showed that certain phenotypes are subject to the effects of population stratification and genetic relatedness. Manhattan plots illustrating the *p*-values from the association analyses for these four flavonoid traits are presented in Fig. 9.

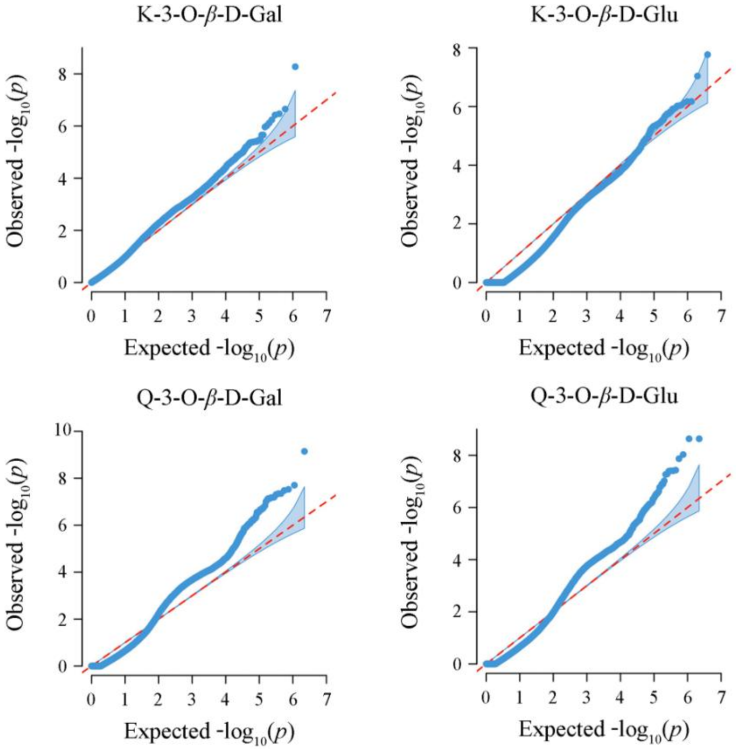


Fig. 8. QQ map of P-value distribution of SNP associated with flavonol-related traits of hawk tea.

Within the SNP sites linked to K-3-O-β-D-gal, 34 were identified, with 11 showing significant associations (Fig. 9A). The polymorphism of SNPs primarily arises from transition (C-T, G-A) and transversion (C-A, C-G, G-T, A-T) mutations. Among these sites, transitions constitute 55.88% and transversions make up 44.12% (Table S1). For K-3-O-β-D-glu, 22 SNP sites were found, 7 of which were significantly associated (Fig. 9B). In this context, transitions represent 27.27%, whereas transversions account for 72.73% (Table S1). Regarding Q-3-O-β-D-gal, 104 SNP sites were identified, with 36 being significantly associated (Fig. 9C). Here, transition mutations comprise 86.54%, and transversion mutations 13.46% (Table S1). Lastly, for Q-3-O-β-D-glu, 75 SNP sites were noted, with 30 showing significant associations (Fig. 9D). Among these, transition mutation sites are 69.33%, and transversion mutation sites are 30.67% (Table S1)

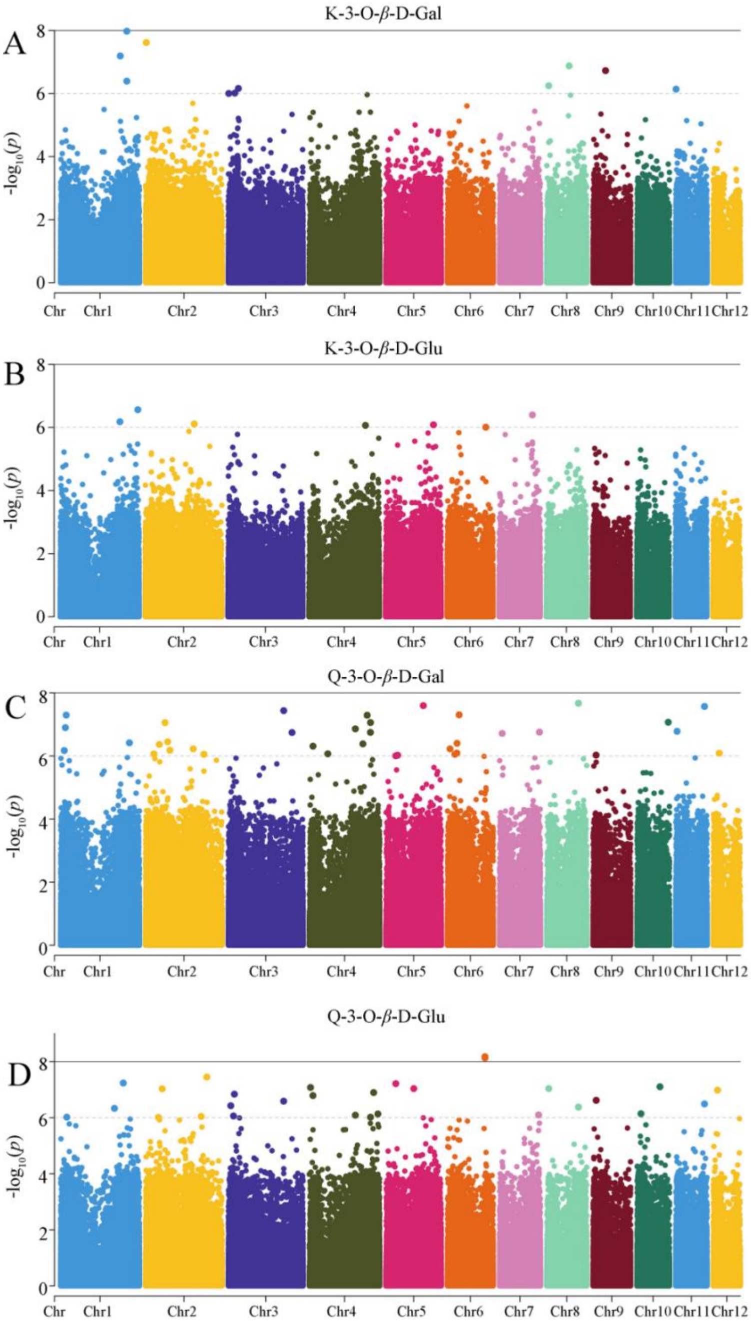


Fig. 9. Manhattan plot of transcriptome-wide association analysis for flavonoid-related traits in hawk tea. The Bonferroni-adjusted suggestive and significant thresholds are illustrated by black and gray dotted horizontal lines (−log10[*p*] values of 8 and 6, respectively.) The X-axis displays the chromosome numbers.

4. DISCUSSION

4.1 Genetic variation of DBH, leaf traits, and flavonoid content of Hawk tea

Guizhou Province, situated in southwest China, is distinguished by its extensive distribution of carbonate rocks and karst landforms (Zhang et al. 2022). This region stands out globally due to its intricate geographical features that cultivate a variety of microclimates, potentially leading to variations in plant characteristics and the concentration of active compounds (Xiong et al. 2023). Our research revealed that the differences in DBH across and within Guizhou regions were not statistically significant, suggesting uniform growth patterns for hawk tea across the province. The diversity in leaf traits and flavonoid content primarily stemmed from the distinct habitats, highlighting that hawk tea's growth and development exhibit variation in response to the unique microclimatic conditions prevalent in Guizhou. This observation aligns with the findings of Hsiung et al. (2017), who noted that minor geoclimatic shifts can induce morphological and anatomical adaptations in leaves, facilitating plant survival and establishment in novel environments. This has profound implications for our understanding of plant survival, adaptation, and evolution. Factors such as temperature, sunlight intensity, and rainfall not only serve as fundamental prerequisites for plant growth but also significantly influence the composition of plant active components (Yu et al. 2015). Consequently, variations in the microclimate of different areas may also reflect in the regional differences in flavonoid content.

4.2 Correlation between DBH, leaf traits, and four kinds of flavonoids

The correlation coefficient serves as a crucial statistical tool for quantifying the relationship between two variables (Baak et al. 2020). In our analysis, significant positive correlations were observed between both Q-3-O-β-D-gal and K-3-O-β-D-gal with LL, LA, LP, SPAD values, LS, and SLA. Moreover, K-3-O-β-D-gal also showed a significant positive correlation with LW and LPL. Given that the flavonol content influences the taste of hawk tea, our findings suggest that leaves with superior quality are more desirable for processing hawk tea. The significant positive correlation of Q-3-O-β-D-gal, Q-3-O-β-D-glu, and K-3-O-β-D-gal with DBH implies that DBH could serve as an indirect selection criterion for hawk tea content, hinting at a link between flavonol accumulation and tree age (Wang et al. 2022). The interrelations among the four flavonol contents indicate that their accumulation in hawk tea is contingent upon the planting environment and genetic factors. The genetic background determines the capacity of plants to adapt to environmental conditions. Differences in metabolite production have been observed between samples of the same species grown under varying environmental conditions. Specific environmental factors have been identified as major sources of variation in intraspecies metabolism. For instance, abiotic factors such as soil nutrients and water availability can induce significant differences in the amount of compounds accumulated by plants in different regions (Liang et al. 2005). Plant traits emerge from the prolonged interplay between genetic attributes and environmental conditions (Florez et al. 2009). Optimal temperatures and altitudes can foster enhanced flavonol growth (Marotti et al. 2020). The presence of genetic traits within and among plant populations could facilitate a quicker adaptation to environmental shifts, allowing plants to survive, adapt, and evolve in new settings and consequently produce various flavonol classes (Agostini-Costa 2022).

4.3 Second and third-generation sequencing data and SNP statistics

Both second and third-generation transcriptome sequencing techniques were utilized. By employing the "three + two" model, the third-generation full-length transcriptome data was refined with the help of parameter-free assembly data from the second generation, leading to the acquisition of high-quality transcripts. The proportion of Q30 bases exceeded 94% (Fig. 2A), underscoring the high quality of the sequencing data. Moreover, the mapping rates for the sequencing samples were predominantly above 94% (with the exception of KY13), signifying excellent data fidelity. The completeness of the transcripts, as assessed by BUSCO, reached an impressive 96.3%. The data generated were then aligned with the genome of *Litsea cubeba*, a species closely related, achieving an average mapping rate of 85.37% and identifying over 600,000 SNPs per sample (Fig. 3). In conclusion, the "three + two" model implemented has proven to be an effective strategy for generating high-quality transcripts for further analysis in this study.

4.4 Analysis of population genetic structure of hawk tea

The determinants of association analysis outcomes are primarily governed by factors such as the quantity of SNPs, the diversity and scale of population materials, and the choice of statistical techniques (Kim et al. 2022). A notable challenge in association analysis is the potential for population structure to spuriously link target traits with unrelated genes, elevating the rate of false positives (Iwata et al. 2007). The efficacy of association analysis is maximized in populations with simple structures, where the likelihood of erroneous links is minimized (Kaler et al. 2020). Conversely, intricate population structures amplify linkage disequilibrium across the population, increasing the incidence of false associations between traits and gene polymorphisms (Iwata et al. 2007). Implementing population structure analyses can mitigate the rate of false associations, with strategies such as structural association analysis, principal component analysis, genomic control, and multidimensional scaling addressing the impact of population structure on association studies (Hu and Ziv 2008). Three methodologies were employed to examine the genetic structure of hawk tea populations. The initial approach involved constructing a cluster model from multi-locus genotype data, applying a mixed population model to depict genetic structure, calculating the K value to represent allelic variation frequency types, and determining the potential subpopulation count using the K value. The second approach constructed phylogenetic trees from allele frequency data by evaluating genetic distances among individuals within the population. The third approach utilized allele frequencies for genotype virtual variable transformation and PCA analysis to map individual-level spatial sequencing relationships, facilitating the investigation of genetic structure and differentiation at the population level. The outcomes from these three methodologies were consistent, classifying 109 clones into five subgroups, thereby enabling their correlation with quantitative traits.

4.5 Association analysis of flavonols

The combined analysis of expression profiles, metabolic profiles, and transcriptome association studies stands as a crucial approach for investigating quantitative traits within complex metabolic systems (Robinson et al. 2007). In the case of hawk tea, flavonols represent the primary constituents. Nonetheless, the intricate nature and extensive labor required for qualitative and quantitative assessments have limited research into the SNP sites associated with anabolic metabolism and its genetic underpinnings. Metabolic data, transcriptome expression profiles, and high-density variant findings derived from "three + two" mode sequencing were leveraged in conducting a quantitative analysis of four flavonols in 109 hawk tea samples from various regions. Through transcriptome association analysis, SNPs linked to the biosynthesis of four flavonol glycosides were identified within the hawk tea transcriptome. This discovery lays the groundwork for future efforts to pinpoint genes related to hawk tea.

Initially, a population consisting of 109 individual trees from five regions in Guizhou, China, was constructed for the study. Through deep sequencing, each sample exhibited over 600,000 SNPs (Fig. 3), indicating high genetic diversity within this group. Transcriptome association analysis revealed a set of candidate genes related to the content of four types of flavonols. Based on the correction for multiple testing and setting the *p*-value threshold at $p < 0.0001$, 13 SNPs were identified as significant for functional annotation (Table 3). Functional annotation showed that these genes mainly belong to categories such as metabolic pathways, biosynthesis of secondary metabolites, and transport of secondary metabolites. Notably, among the candidate genes associated with K-3-O-β-D-gal, one was annotated as UGT74B1. Jiang (2018) et al. found that *UGT* genes might be related to the biosynthesis of K-3-O-β-D-gal and K-3-O-β-D-glu, while Zhang (2021) et al. found that Q-3-O-β-D-glu has a certain inhibitory effect on recombinant *UGT1A* subtypes in vitro. Moreover, as indicated by the data presented in Table 5, the structural genes (cytochrome P450 enzyme, selenium-binding protein, glycoside glycosyltransferase, phosphoenolpyruvate carboxylase, diacylglycerol acyltransferase) were found to be directly engaged in established pathways governing flavonoid metabolism, thus holding pivotal significance in flavonol biosynthesis.

A natural population comprising 109 samples characterized by a limited diversity of samples from various regions and possessing a relatively complex structure, impacted the outcomes of the association analysis, generally yielding a low association signal. Nonetheless, the considerable sequencing depth and comprehensive transcriptome coverage achieved in this study, coupled with the high density and reliability of the identified loci within the transcriptome, safeguarded the accuracy of the association signals.

Although candidate genes associated with flavonol content were not further analysis and verification in this study, it represents the inaugural effort to perform an association analysis of hawk tea at the transcriptome level. This pioneering research holds significant implications for advancing our understanding of the genes and genetic mechanisms underlying the important secondary metabolites in hawk tea.

5. CONCLUSIONS

To summarize, results reveal no significant regional variation in DBH in hawk tea across Guizhou, highlighting that the diversity in leaf traits and flavonol levels primarily originates from habitat differences. Flavonol content emerged as a crucial determinant of hawk tea taste, exhibiting a notable correlation with tree age. Leaves of superior quality, distinguished by their flavonol levels, proved optimal for hawk tea production. Integrating second and third generation transcriptome sequencing technologies enhances the generation of high-quality transcripts, proving to be an efficacious strategy. Through transcriptome association analysis, thirteen significant SNPs were identified to link to flavonol content, situated within gene regions. Notably, structural genes (including cytochrome P450 enzyme, selenium-binding protein, glycoside glycosyltransferase, phosphoenolpyruvate carboxylase, diacylglycerol acyltransferase) were pointed as integral components of known pathways directly regulating flavonoid metabolism and playing pivotal roles in flavonol biosynthesis. The findings lay a robust theoretical groundwork for the subsequent implementation of effective selection and breeding strategies in hawk tea.

AUTHOR CONTRIBUTIONS

Lan Yang: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Software (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Huie Li:** Conceptualization (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Na Xie:** Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal). **Gangyi Yuan:** Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal). **Qiqiang Guo:** Conceptualization (equal); Data curation (equal);

Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal).

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (32060349) and China Scholarship Council ([2021]15).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Raw reads have been deposited in the National Center for Biotechnology Information (NCBI; BioProject accession number PRJNA992466, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA992466).

REFERENCES

Agostini-Costa, T.D., 2022. Genetic and environment effects on bioactive compounds of *Opuntia cacti*–a review. J. Food Compos. Anal. 109, 104514. https://doi.org/10.1016/j.jfca.2022.104514

Baak, M., Koopman, R., Snoek, H., Klous, S., 2020. A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. Comput. Stat. Data Anal. 152, 107043. https://doi.org/10.1016/j.csda.2020.107043

Baid, G., Cook, D.E., Shafin, K., Yun, T., Llinares-López, F., Berthet, Q., Belyaeva, A., Töpfer, A., Wenger, A.M., Rowell, W.J., Yang, H., Kolesnikov, A., Ammar, W., Vert, J.P., Vaswani, A., McLean, C.Y., Nattestad, M., Chang, P.C., Carroll, A., 2023. Deep consensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat. Biotechnol. 41, 232–238. https://doi.org/10.1038/s41587-022-01435-7

Bhinder, G., Sharma, S., Kaur, H., Akhatar, J., Mittal, M., Sandhu, S., 2022. Genomic regions associated with seed meal quality traits in *Brassica napus* germplasm. Front. Plant Sci. 13, 882766. https://doi.org/10.3389/fpls.2022.882766

Bondonno, N.P., Dalgaard, F., Kyro, C., Murray, K., Bondonno, C.P., Lewis, J.R., Croft, K.D., Gislason, G., Scalbert, A., Cassidy, A., Tjonneland, A., Overvath, K., Hodgson, J.M., 2019. Flavonoid intake is associated with lower mortality in the Danish diet cancer and health cohort. Nat. Commun. 10, 3651. https://doi.org/10.1038/s41467-019-11622-x

Carmela, G. Giovanna, G., 2022. Flavonoids from plants to foods: from green extraction to healthy food ingredient. Molecules. 27(9), 2633. https://doi.org/10.3390/molecules27092633

Fan, X.L., Fan, Z.Q., Yang, Z.Y., Huang, T.T., Tong, Y.D., Yang, D.Y., Mao, X.P., Yang, M.Y., 2022. Flavonoids–natural gifts to promote health and longevity. Int. J. Mol. Sci. 23(4), 2176. https://doi.org/10.3390/ijms23042176

Florez, A., Pujolà, M., Valero, J., Centelles, E., Almirall, A., Casañas, F., 2009. Genetic and environmental effects on chemical composition related to sensory traits in common beans (*Phaseolus vulgaris* L.). Food Chem. 113(4), 950–956. https://doi.org/10.1016/j.foodchem.2008.08.036

Ha, D.L., Shi, G.H., Liu, Z., Xiong, B., 2022. Estimation of genome size and genomic characteristics of the main source plants for making Hawk-tea by flow cytometry and k-mer analysis. J. Plant Genet. Resour. 23(4), 1166–1174. (in Chinese with English abstract) https://doi.org/10.13430/j.cnki.jpgr.20211222004

Harper, A.L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P., Bancroft, I., 2012. Associative transcriptomics of traits in the polyploidy crop species *Brassica napus*. Nat. Biotechnol. 30(8), 798–802. https://doi.org/10.1038/nbt.2302

Hsiung, H.Y., Huang, B.H., Chang, J.T., Huang, Y.M., Huang, C.W., Liao, P.C., 2017. Local climate heterogeneity shapes population genetic structure of two undifferentiated insular *Scutellaria* species. Front. Plant Sci. 8, 159. https://doi.org/10.3389/fpls.2017.00159

Hu. D., Ziv. E., 2008. Confounding in genetic association studies and its solutions. Methods Mol. Biol. 448, 31–39. https://doi.org/10.1007/978-1-59745-205-2_3

Hyten, D.L., Cannon, S.B., Song, Q.J., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., Cregan, P.B., 2010. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics, 11, 38. https://doi.org/10.1186/1471-2164-11-38

Iwata, H., Ebana, K., Fukuoka, S., Jannink, J.L., Hayashi, T., 2007. Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor. Appl. Genet. 114, 1437–1449. https://doi.org/10.1007/s00122-008-0945-6

Jia, X.J., Li, P., Wan, J.B., He, C.W., 2017. A review on phytochemical and pharmacological properties of *Litsea coreana*. Pharm. Biol. 55(1), 1368–1374. https://doi.org/10.1080/13880209.2017.1302482

Jiang, X.L., Shi, Y.F., Dai, X.L., Zhuang, J.H., Fu, Z.P., Zhao, X.Q., Liu, Y.J., Gao, L.P., Xia, T., 2018. Four flavonoid glycosyltransferases present in tea overexpressed in model plants *Arabidopsis thaliana* and *Nicotiana tabacum* for functional identification. J. Chromatogr. B. 1100, 148–157. https://doi.org/10.1016/j.jchromb.2018.09.033

Kaler, A.S., Gillman, J.D., Beissinger, T., Purcell, L.C., 2020. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. Front. Plant Sci. 10, 1794. https://doi.org/10.3389/fpls.2019.01794

Khan, W.A., Hou, X.L., Han, K., Khan, N., Dong, H.J., Saqib, M., Zhang, Z.S., Naseri, E., Hu, C.M., 2018. Lipidomic study reveals the effect of morphological variation and other metabolite interactions on the lipid composition in various cultivars of *Bok choy*. Biochem. Biophys. Res. Commun. 506(3), 755–764. https://doi.org/10.1016/j.bbrc.2018.04.112

Kim, H., Bi, Y.T., Pal, S., Gupta, R., Davuluri, R.V., 2011. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. BMC Bioinformatics. 12, 305. https://doi.org/10.1186/1471-2105-12-305

Kim, J.M., Lyu, J.I., Kim, D.G., Hung, N.N., Seo, J.S., Ahn, J.W., Lim, Y.J., Eom, S.H., Ha, B.K., Kwon, S.J., 2022. Genome wide association study to detect genetic regions related to isoflavone content in a mutant soybean population derived from radiation breeding. Front. Plant Sci.13, 968466. https://doi.org/10.3389/fpls.2022.968466

Kishi-Kaboshi, M., Tanaka, T., Sasaki, K., Noda, N., Aida, R., 2022. Combination of long-read and short-read sequencing provides comprehensive transcriptome and new insight for *Chrysanthemum morifolium* ray-floret colorization. Sci. Rep. 12, 17874. https://doi.org/10.1038/s41598-022-22589-z

Li, M., Nordborg, M., Li, L.M., 2004. Adjust quality scores from alignment and improve sequencing accuracy. Nucleic Acids Res. 32(17), 5183–5191. https://doi.org/10.1093/nar/gkh850

Li, Y.L., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T.D., Pang, J.Y., Siddique, K.H.M., Sutton, T., 2018. Investigating drought tolerance in Chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. Front. Plant Sci. 9, 00190. https://doi.org/10.3389/fpls.2018.00190

Liang, H.L., Liang, Y.R., Dong, J.J., Lu, J.L., Xu, H.R., Wang, H., 2007. Decaffeination of fresh green tea leaf (*Camellia sinensis*) by hot water treatment. Food Chem. 101(4), 1451–1456. https://doi.org/10.1016/j.foodchem.2006.03.054

Liang, Q.R., Qian, H., Yao, W.R., 2005. Identification of flavonoids and their glycosides by high-performance liquid chromatography with electrospray ionization mass spectrometry and with diode array ultraviolet detection. Eur. J. Mass Spectrom. 11(1), 93–101. https://doi.org/10.1255/ejms.710

Liao, G.L., Zhong, M., Jiang, Z.Q., Tao, J.J., Jia, D.F., Qu, X.Y., Huang, C.H., Liu, Q., Xu, X.B., 2021. Genome-wide association studies provide insights into the genetic determination of flower and leaf traits of *Actinidia eriantha*. Front. Plant Sci. 12, 730890. https://doi.org/10.3389/fpls.2021.730890

Liu, Y., Luo, Y.K., Zhang, L., Luo, L.Y., Xu, T., Wang, J., Ma, M.J., Liang, Z., 2020. Chemical composition, sensory qualities, and pharmacological properties of primary leaf hawk tea as affected using different processing methods. Food Biosci. 36, 100618. https://doi.org/10.1016/j.fbio.2020.100618

Luo, B.W., Ma, P., Nie, Z., Zhang, X., He, X., Ding, X., Feng, X., Lu, Q.X., Ren, Z.Y., Lin, H.J., Wu, Y.Q., Shen, Y., Zhang, S.Z., Wu, L., Liu, D., Pan, G.T., Rong, T.Z., Gao, S.B., 2019. Metabolite profiling and genome-wide association studies reveal response mechanisms of phosphorus deficiency in maize seedling. Plant J. 97, 947–969. https://doi.org/10.1111/tpj.14160

Maeda, H., Akagi, T., Onoue, N., Kono, A., Tao, R., 2019. Evolution of lineage-specific gene networks underlying the considerable fruit shape diversity in persimmon. Plant Cell Physiol. 60(11), 2464–2477. https://doi.org/10.1093/pcp/pcz139

Marotti, I., Whittaker, A., Benvenuti, S., Benedettelli, S., Ghiselli, L., Dinelli, G., Bosi, S., 2020. Temperature-associated effects on flavonol content in field-grown *Phaseolus vulgaris* L. zolfino del pratomagno. Agronomy. 10(5), 682. https://doi.org/10.3390/agronomy10050682

Robinson, A.R., Ukrainetz, N.K., Kang, K., Mansfield, S.D., 2007. Metabolite profiling of douglas-fir (*Pseudotsuga menziesii*) field trials reveals strong environmental and weak genetic variation. New Phytol. 174, 762–773. https://doi.org/10.1111/j.1469-8137.2007.02046.x

Singh, S., Vishwakarma, R.K., Kumar, R.J.S., Sonawane, P.D., Ruby, Khan, B.M., 2013. Functional Characterization of a Flavonoid Glycosyltransferase Gene from *Withania somnifera* (Ashwagandha). Appl. Biochem. Biotechnol. 170, 729–741 2013. https://doi.org/10.1007/s12010-013-0230-2

Song, L.L., Tian, Q., Li, G., Li, Z.X., Liu, X.Y., Gui, J., Li, Y.C., Cui, Q., Zhao, Y., 2022. Variation in characteristics of leaf functional traits of alpine vegetation in the Three-River Headwaters Region, China. Ecol. Indic. 145, 109557. https://doi.org/10.1016/j.ecolind.2022.109557

Tan, L.H., Zhang, D., Wang, G., Yu, B., Zhao, S.P., Wang, J.W., Yao, L., Cao, W.G., 2016. Comparative analyses of flavonoids compositions and antioxidant activities of Hawk tea from six botanical origins. Ind. Crops Prod. 80, 123–130. https://doi.org/10.1016/j.indcrop.2015.11.035

Wang, Q.J., Jiang, Y., Mao, X.Y., Yu, W.W., Lu, J.K., Wang, L., 2022. Integration of morphological, physiological, cytological, metabolome and transcriptome analyses reveal age inhibited accumulation of flavonoid biosynthesis in *Ginkgo biloba* leaves. Ind. Crops Prod. 187, 115405. https://doi.org/10.1016/j.indcrop.2022.115405

Wu, Y.Q., Ma, X.Y., Zhou, Q., Xu, L.A., Wang, T.L., 2019. Selection of crown type provides a potential to improve the content of isorhamnetin in *Ginkgo biloba*. Ind. Crops Prod. 143, 111943. https://doi.org/10.1016/j.indcrop.2019.111943

Xiong, Y., Zhou, Z.F., Ding, S.J., Zhang, H., Huang, J., Gong, X.H., Su, D., 2023. Spatiotemporal variation characteristics and influencing factors of karst cave microclimate environments: a case study in Shuanghe cave, Guizhou province, China. Atmosphere. 14(5), 813. https://doi.org/10.3390/atmos14050813

Yao, L.H., Jiang, Y.M., Shi, J., Tomas-Barberan, F. A., Datta, N., Singanusong, R., Chen, S.S., 2004. Flavonoids in food and their health benefits. Plant Foods Hum. Nutr. 59, 113–122. https://doi.org/10.1007/s11130-004-0049-7

Ye, M., Liu, D., Zhang, R., Yang, L., Wang, J., 2012. Effect of hawk tea (Litsea coreana L.) on the numbers of lactic acid bacteria and flavour compounds of yoghurt. Int. Dairy J. 23(1), 68–71. https://doi.org/10.1016/j.idairyj.2011.09.014

Yu, F.L., Wang, Q.L., Wei, S.L., Wang, D., Fang, Y.Q., Liu, F.B., Zhao, Z.G., Hou, J.L., Wang, W.Q., 2015. Effect of genotype and environment on five bioactive components of cultivated licorice (*Glycyrrhiza uralensis*) populations in northern China. Biol. Pharm. Bull. 38(1), 75–81. https://doi.org/10.1248/bpb.b14-00574

Yuan, G.Y., Guo, Q.Q., Zhang, Y.Q., Gui, Q., Xie, N., Luo, S.Q., 2023. Geographical differences of leaf traits of the endangered plant *Litsea coreana* Levl. var. *sinensis* and its relationship with climate. J. For. Res. 34, 125–135. https://doi.org/10.1007/s11676-022-01588-w

Zhang, M., Yang, W., Yang, M.X., Yan, J., 2022. Guizhou karst carbon sink and sustainability–an overview. Sustainability. 14(18), 11518. https://doi.org/10.3390/su141811518

Zhang, R., Wei, Y., Yang, T.Y., Huang, X.X., Zhou, J.P., Yang, C.X., Zhou, J., Liu, Y., Shi, S.J., 2021. Inhibitory effects of quercetin and its major metabolite quercetin 3-O-β-D glucoside on human UDP glucuronosyltransferase 1A isoforms by liquid chromatography tandem mass spectrometry. Exp. Ther. Med. 22(2), 842. https://doi.org/10.3892/etm.2021.10274

Zhang, T.Y., Li, H.Z., Ma, S.L., Cao, J., Liao, H., Huang, Q.Y., Chen, W.L., 2023. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl. Environ. Microbiol. 89(10), e0060523. https://doi.org/10.1128/aem.00605-23

SUPPORTING INFORMATION

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