

Genomic analysis of 1,325 *Camellia* accessions sheds light on agronomic and metabolic traits for tea plant improvement

In the format provided by the
authors and unedited

Contents

Pages 1-18: Supplementary Figures 1-31.

Supplementary Figures



Figure S1. Distribution of *Camellia* accessions in four major tea-producing regions of China. The four tea-producing regions were represented by different colors, and the values in the circles represent the number of accessions.

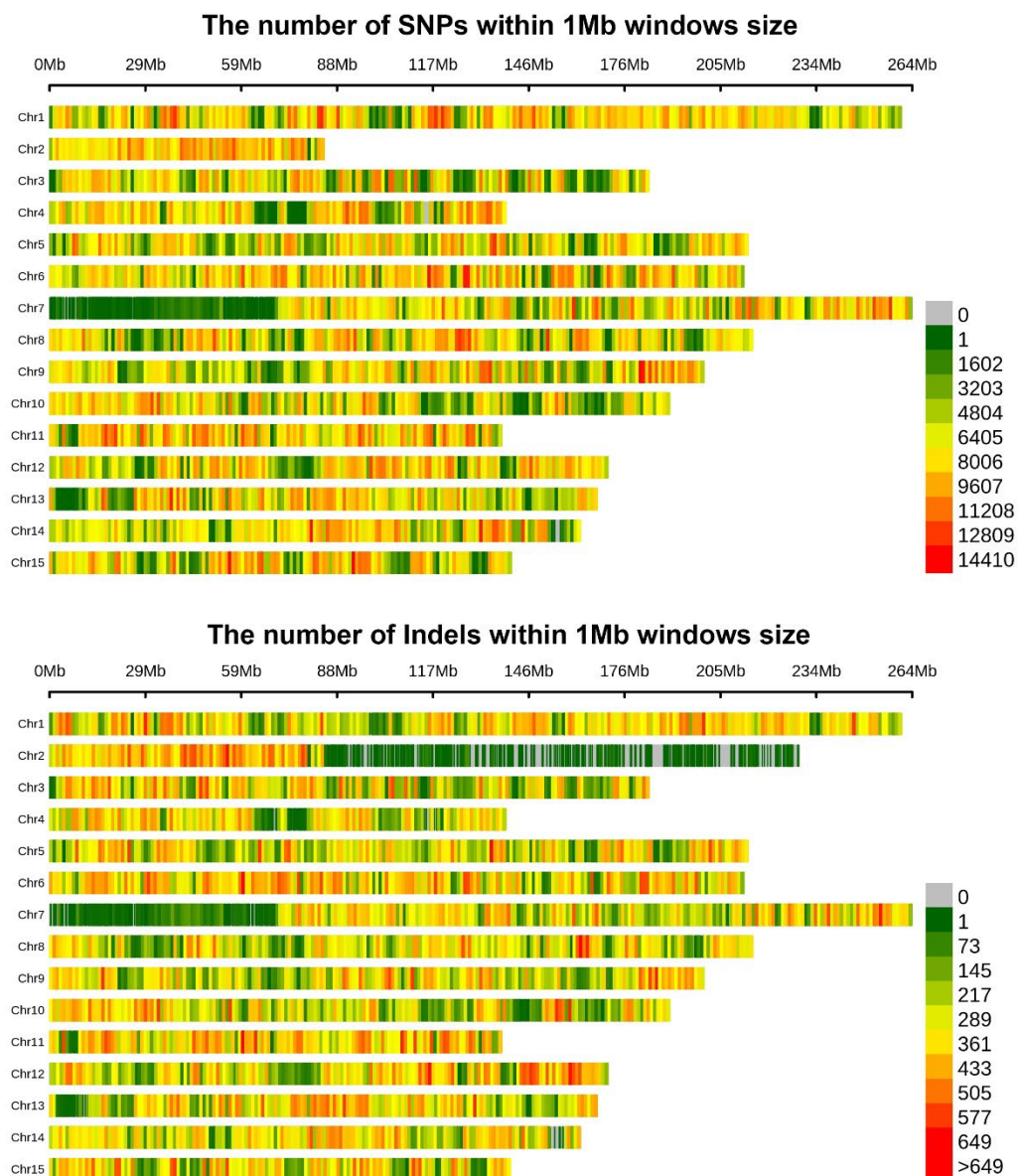


Figure S2. The distribution density of SNPs and Indels on chromosomes.

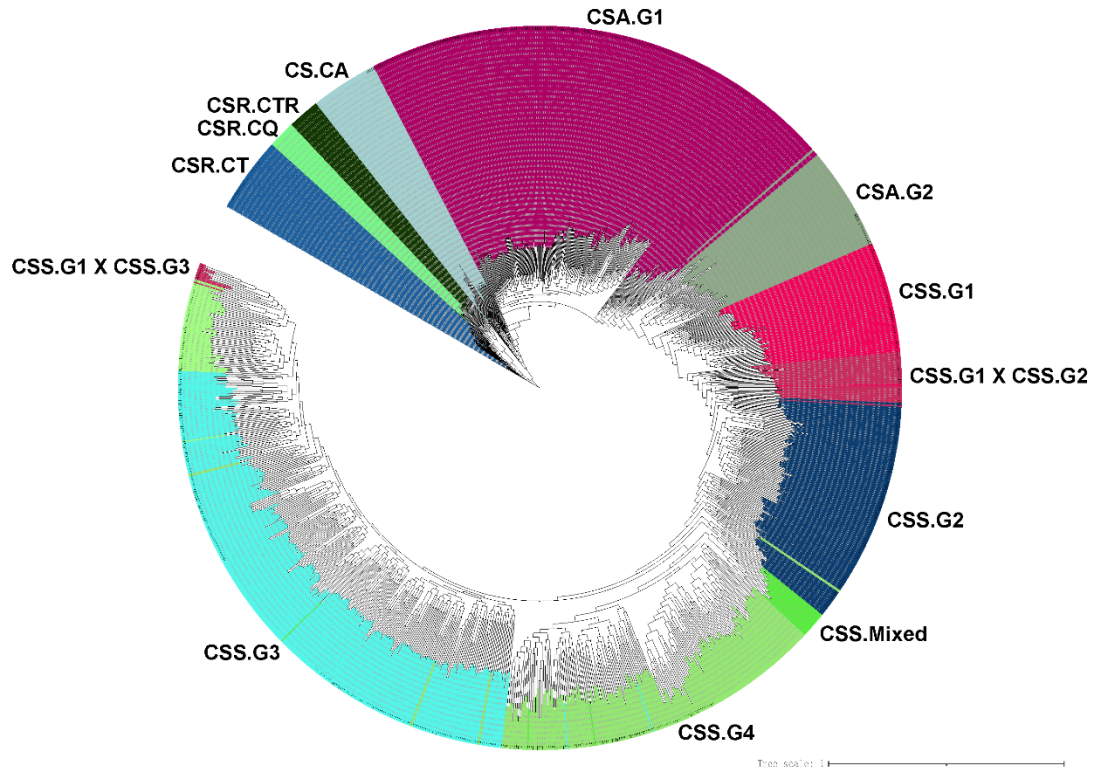


Figure S3. The Maximum Likelihood (ML) phylogenetic tree of 1195 tea plant accessions inferred using all groups shared 542,918 LD filtered SNPs. The 1325 accessions were filtered into 1195 accessions based on Identity By Descent (IBD) value (the samples with IBD greater than 0.98, one was randomly retained). The ML phylogenetic tree was built by using IQ-TREE 2.0 with the GTR+ASC model and 1000 non-parametric bootstrap replicates. Different colors represent different group classifications from Figure 1b. *C. quinquelocularis* (CSR.CQ), *C. tachangensis* var. *remotiserrata* (CSR.CTR), *C. taliensis* (CSR.CT), respectively. CS.CA, CSA (CSA.G1 and CSA.G2 accessions), and CSS (CSS.G1, CSS.G2, CSS.G3, and CSS.G4 accessions) represented the cultivated ancient *C. sinensis*, *C. sinensis* var. *assamica*, and *C. sinensis* var. *sinensis*, respectively. CSS.G1 X CSS.G2, CSA.G1 X CSS.G3, and CSS.Mixed represented three inter-variatal derived subpopulations inferred based on ADMIXTURE ancestry proportions of Figure 1b.

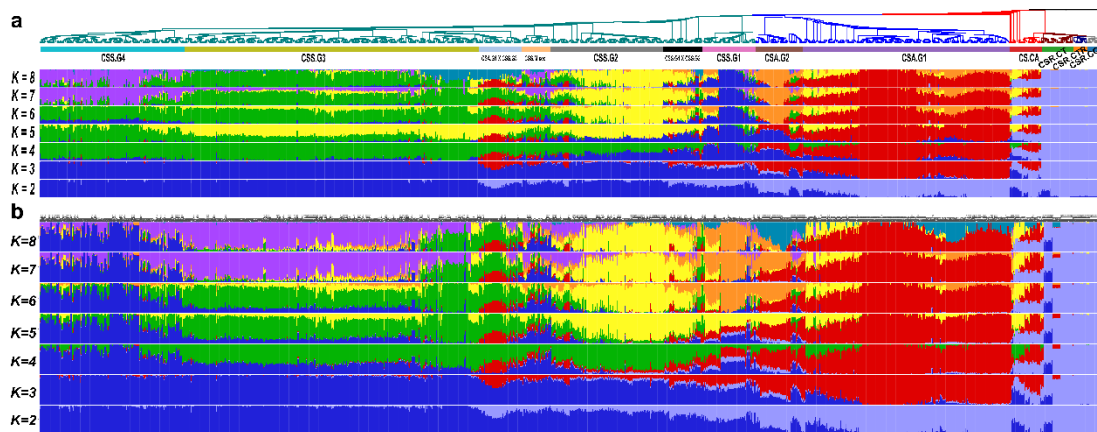


Figure S4. The ADMIXTURE ancestry proportions of all 1325 *Camellia* accessions, were inferred using 93,212 four-fold degenerated SNPs (a) and all group shared 542,918 LD filtered SNPs (b). The individual ancestry estimation of population was also analyzed by ADMIXTURE with K values ranging from 2 to 8.

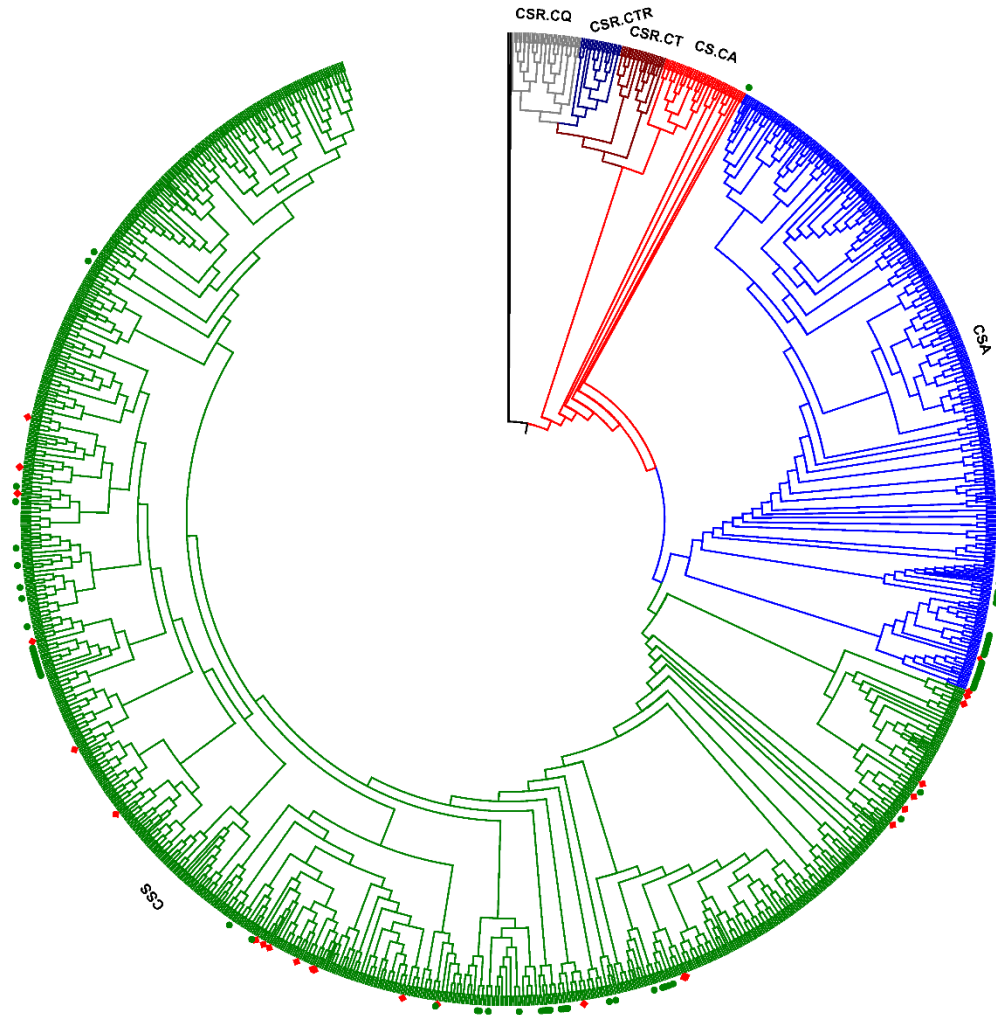


Figure S5. The rooted Maximum Likelihood (ML) phylogenetic tree of all 1325 *Camellia* accessions based on 93,212 four-fold degenerated SNPs (The tree structure is completely the same as Fig. 1b). Samples marked in red diamonds are 25 *C. sinensis* var. *pubilimba* (CSP) accessions. Samples marked with green circles were collected in foreign countries.

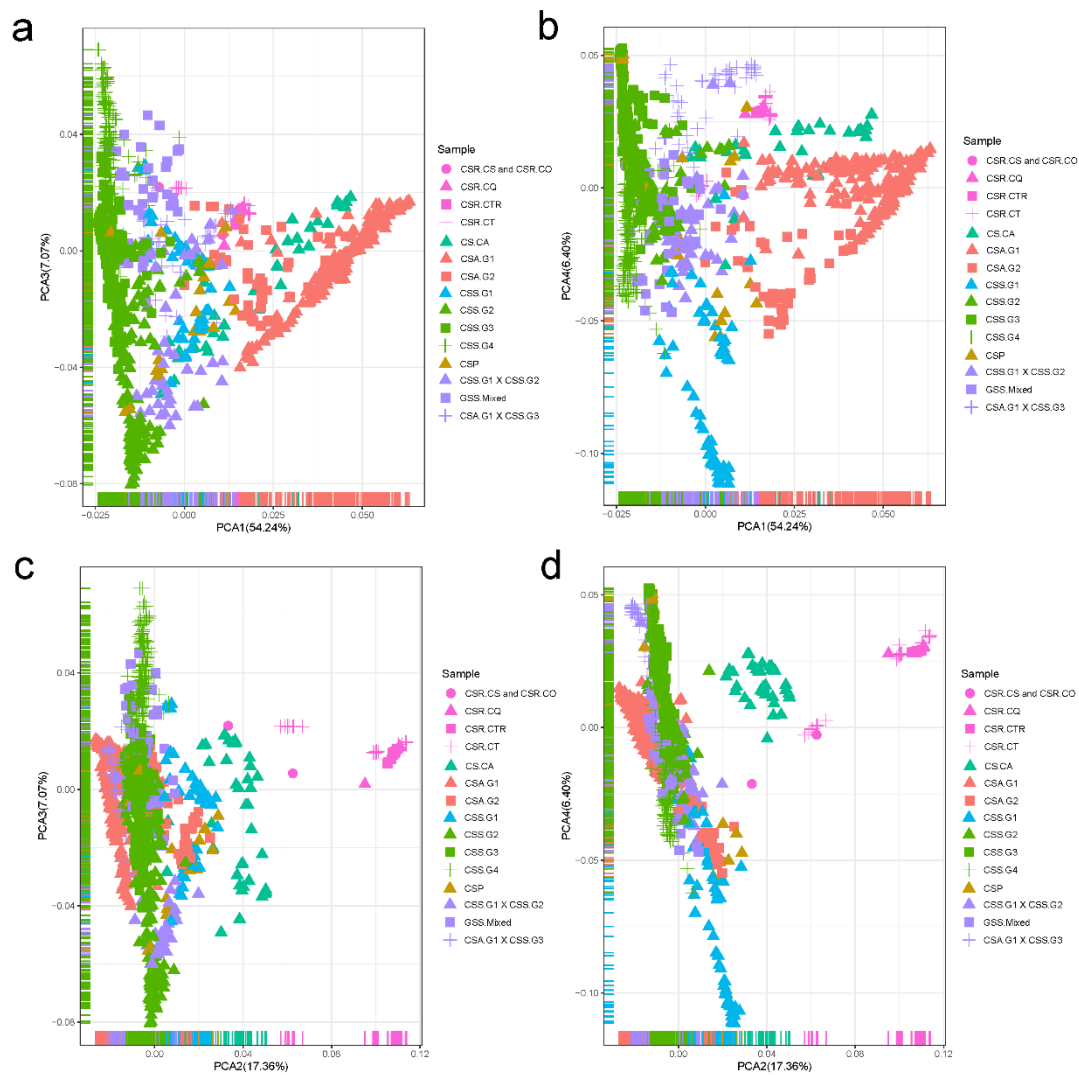


Figure S6. The Principal Component Analysis (PCA) plot of all Camellia accessions.

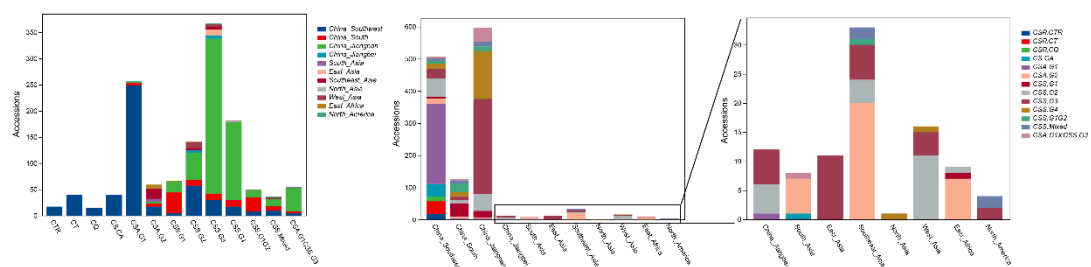


Figure S7. Geographical distribution of different groups of tea plants.

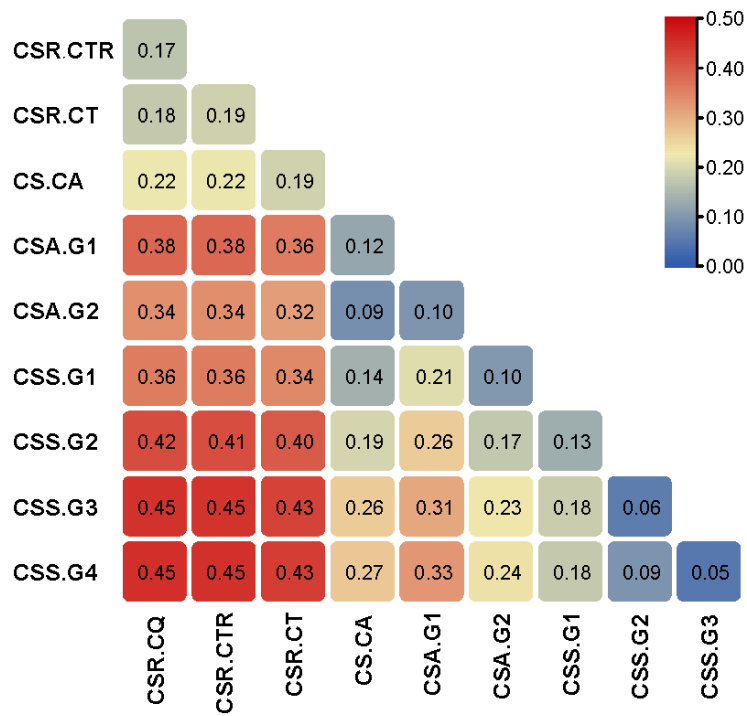


Figure S8. The population divergence (F_{st}) between groups.

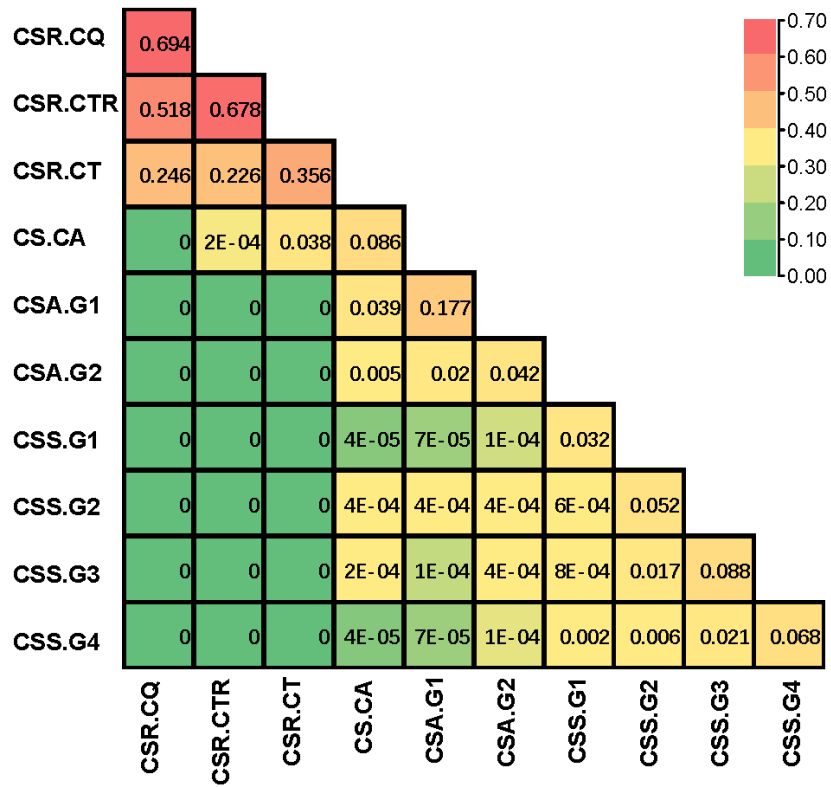


Figure S9. The identify-by-descent (IBD) values within and between groups.

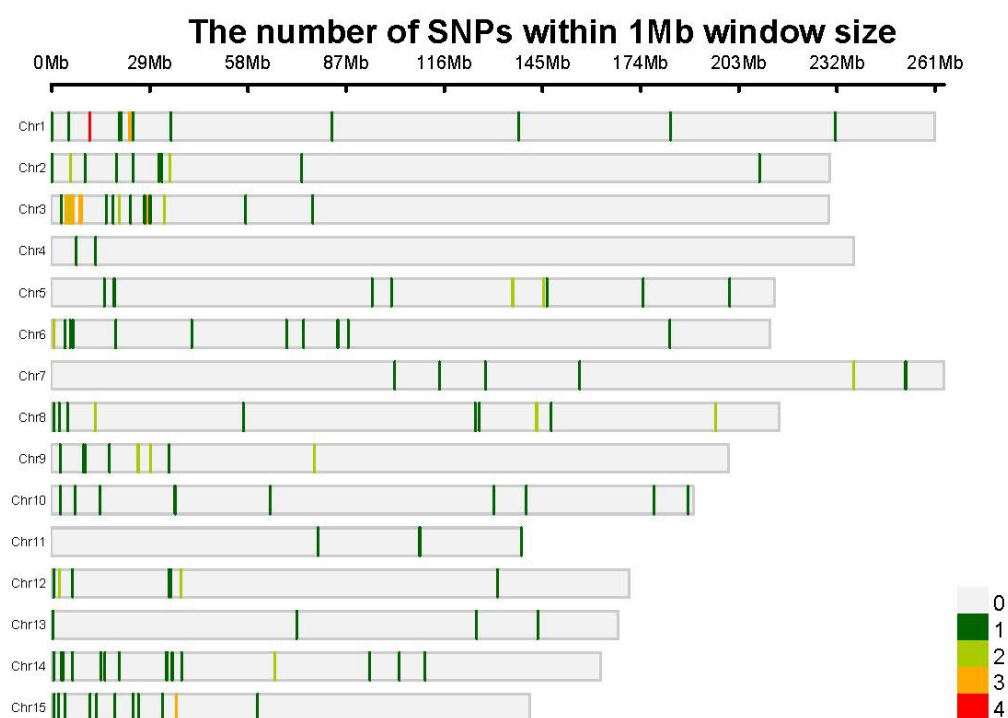


Figure S10. Distribution density of deleterious SNPs on 15 chromosomes.

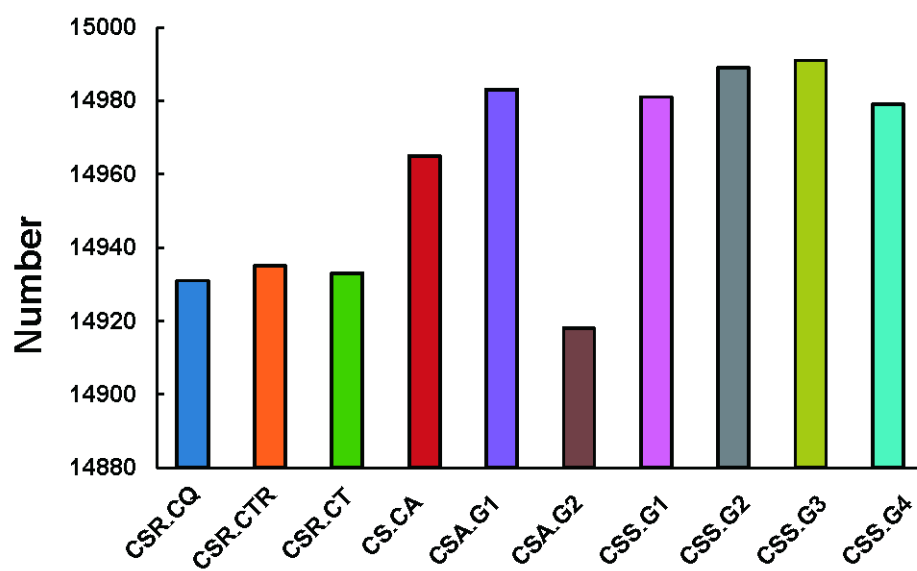


Figure S11. Number of deleterious mutant SNPs in different groups.

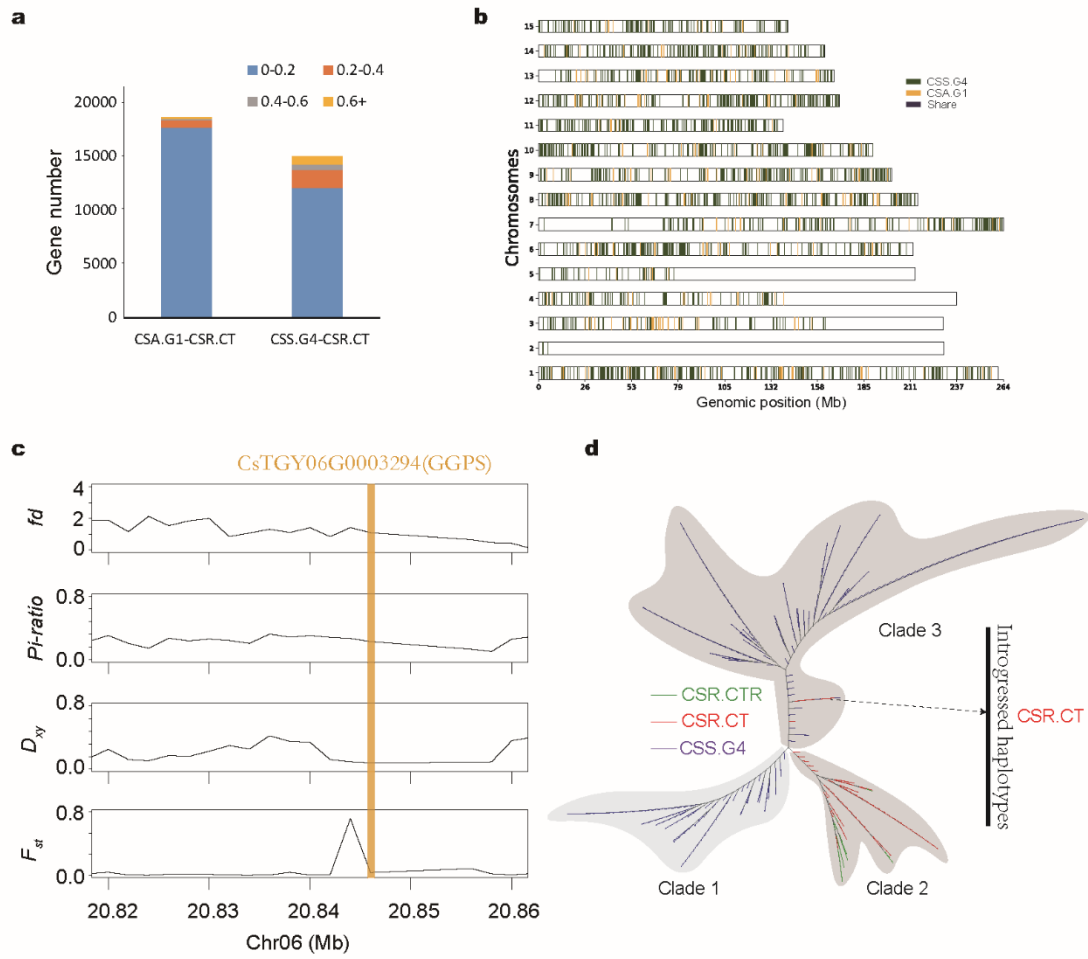


Figure S12. **a**, The gene number in regions within bins of f_d values for the CSA.G1 or CSS.G4 group. **b**, Map of the lengths and distributions of introgression segments on tea plant chromosomes according to the results of the $f_d > 0.6$. The lengths of the colored columns are proportional to the physical lengths of the introgression segments. **c**, A visualization of the region that have undergone introgression between CSR.CT and CSS.G4 group. The *geranylgeranyl diphosphate synthase* (GGPS) gene was highlighted in yellow background. F_{st} , D_{xy} , and Pi -ratio were calculated per 20-kb window between the CSR.CT and CSS.G4 group. **d**, The phylogenetic tree of GGPS gene. green, red, and purple lines denoted the CSR.CTR, CSR.CT, and CSS.G4 group.

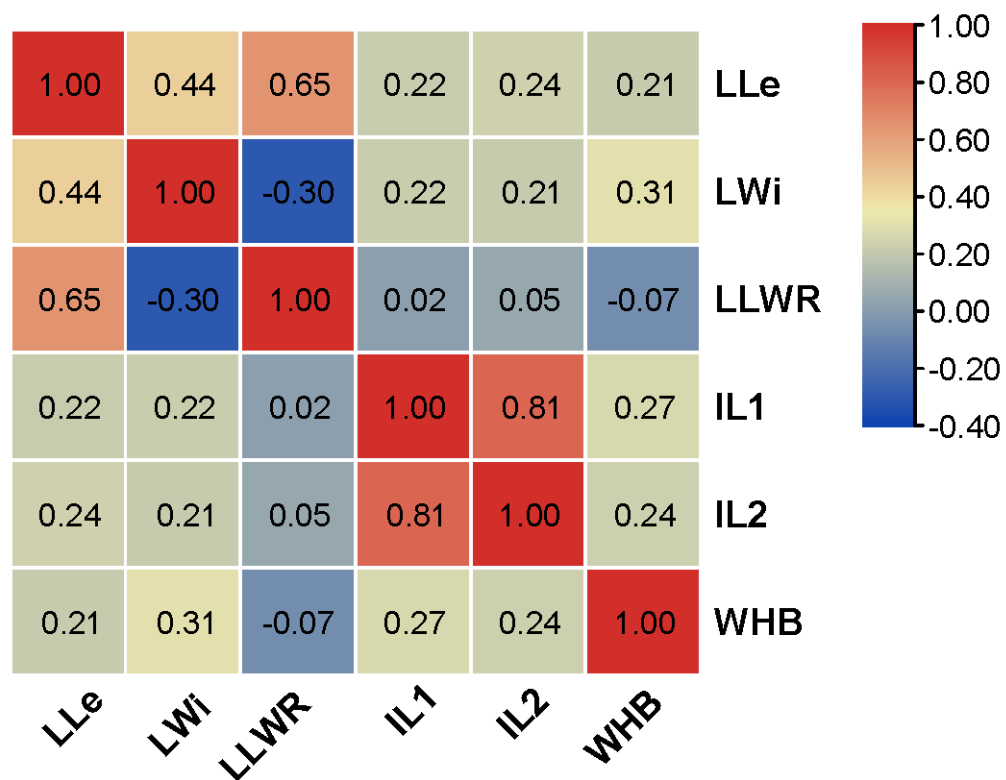


Figure S13. The Pearson correlation of six important agronomic traits including leaf length (LLe), leaf width (LWi), leaf length-width ratio (LLWR), the weight of hundred buds (WHB), the length of first internode (including first bud, IL1), the length of second internode (IL2).

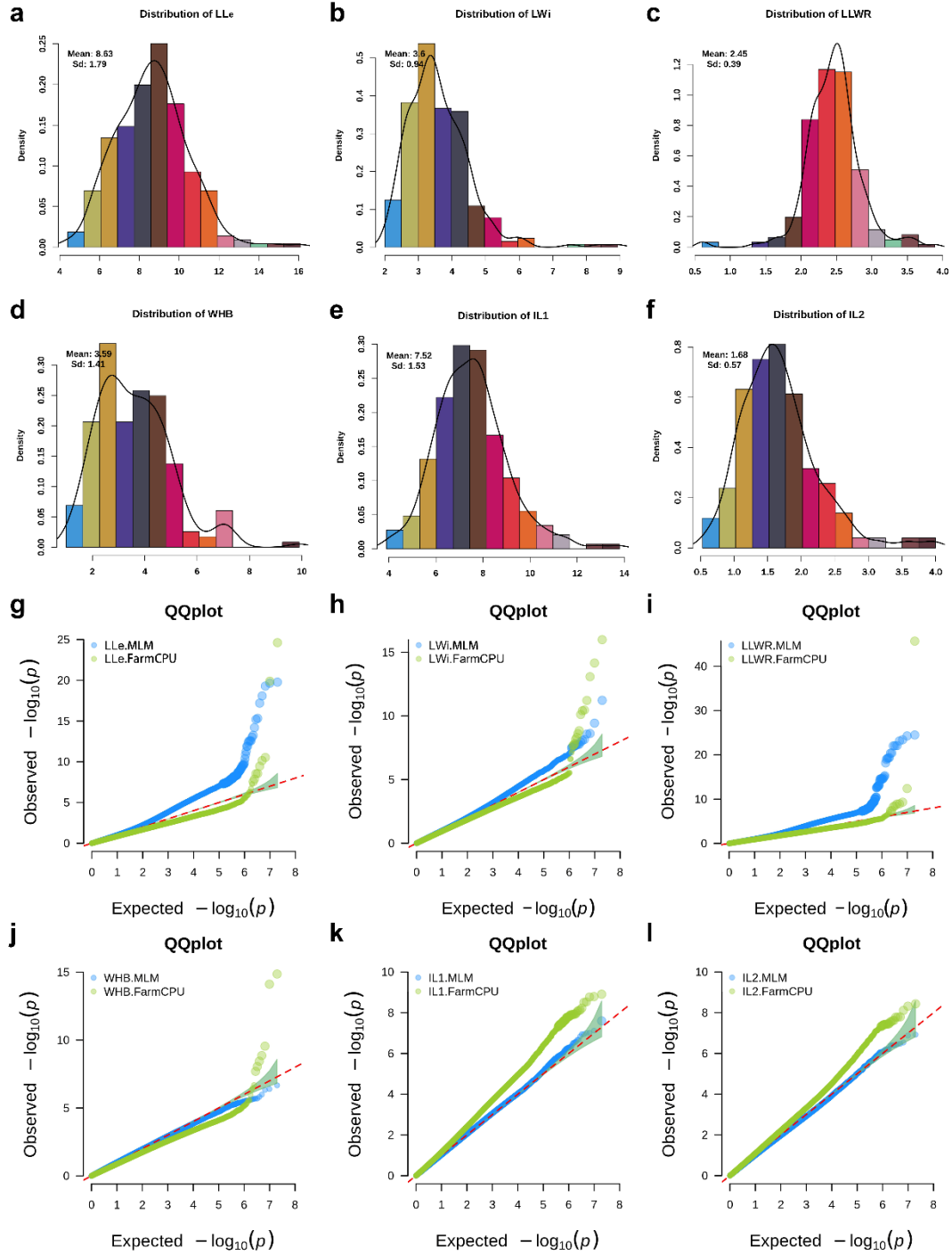


Figure S14. Phenotypic variance (a-f) and QQ plots for GWAS signals (g-l) of agronomic traits in young bud, including leaf length (LLe), leaf width (LWi), leaf length-width ratio (LLWR), the weight of hundred buds (WHB), the length of first internode (including first bud, IL1), and the length of second internode (IL2). In g-l, the statistical analysis was performed with Bonferroni correction.

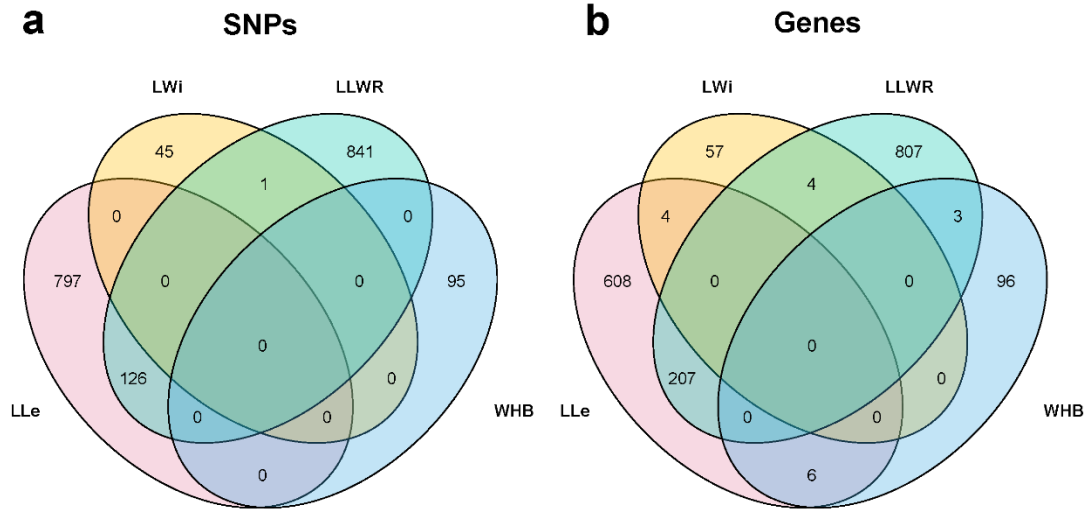


Figure S15. Significant snps related to leaf length (LL), leaf width (LW), leaf length-width ratio (LLWR), and the weight of hundred buds (WHB).

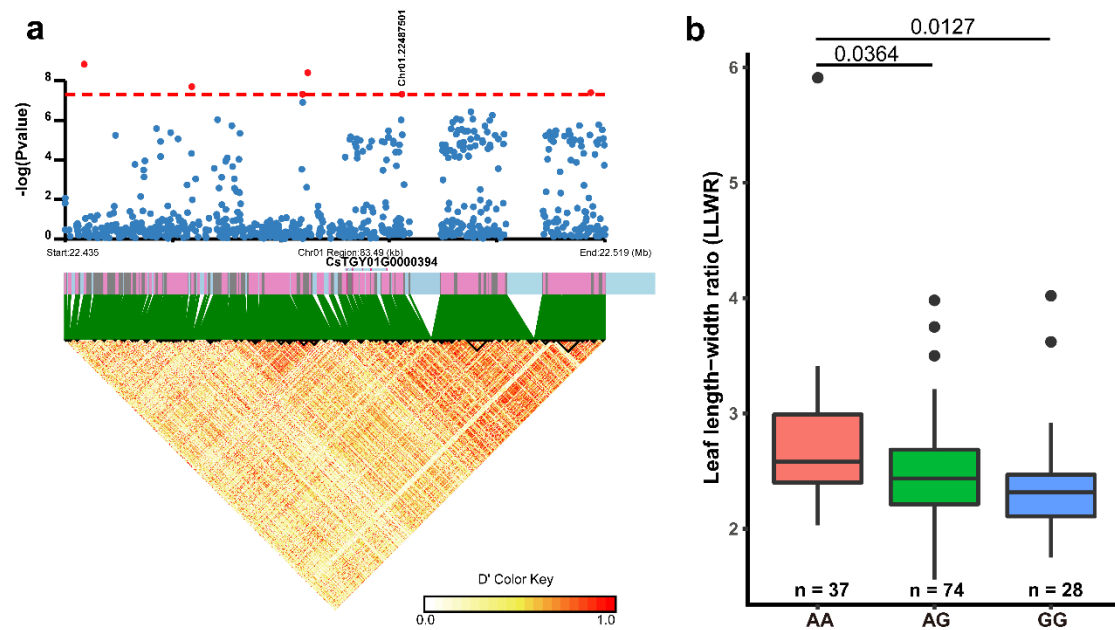


Figure S16. a, Significant snps related to *CstGY01G0000394* (a homolog of *REVOLUTA* in *Arabidopsis*). The horizontal dashed line represents the significant threshold ($P < 5.08 \times 10^{-8}$, Bonferroni correction). **b**, Accessions with different genotypes of *Chr01.22487501* exhibited significant LLWR differences. Significance (p values) of differences between haplotypes was determined by two-sided Student's t-test. n represents the sample size in each genotype. In the box plots, the upper and lower edges represent the 75% and 25% quartiles, the central line denotes the median, the whiskers extend to 1.5 times the interquartile range and the points outside indicate the outliers.

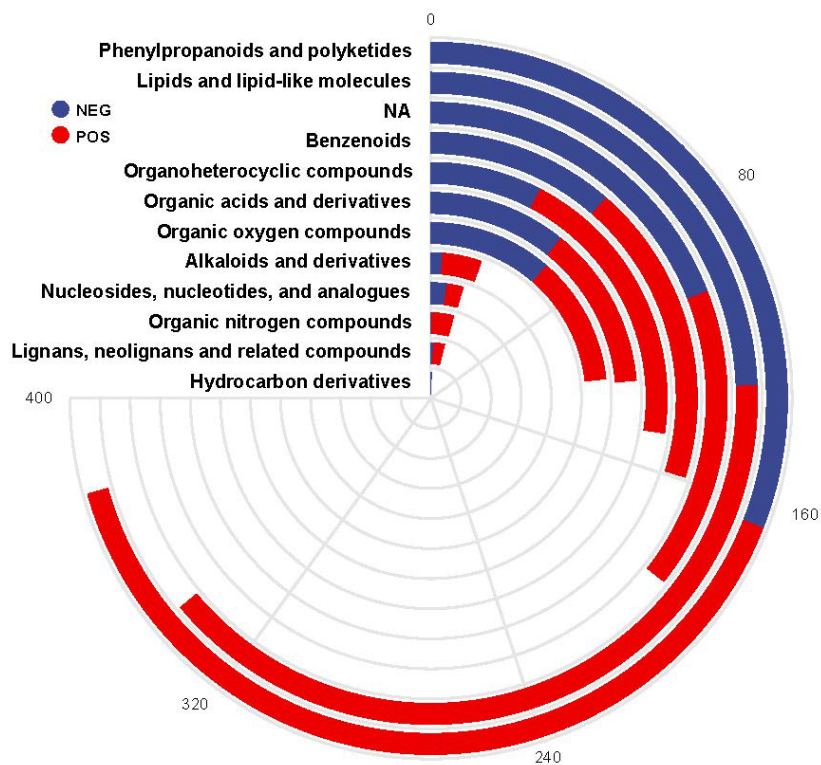


Figure S17. The number and super-classes of annotated metabolites detected in positive (POS) and negative (NEG) electrospray ionization (ESI) modes.

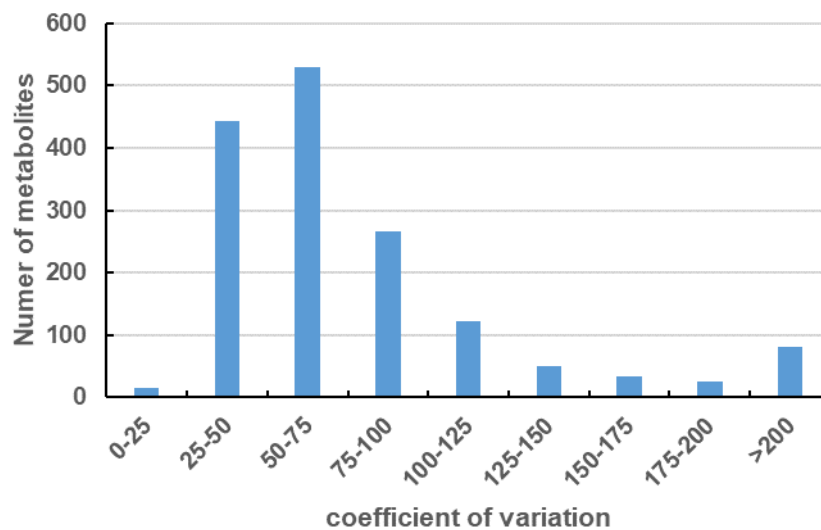


Figure S18. Coefficients of variation (CV) of 1562 annotated metabolites.

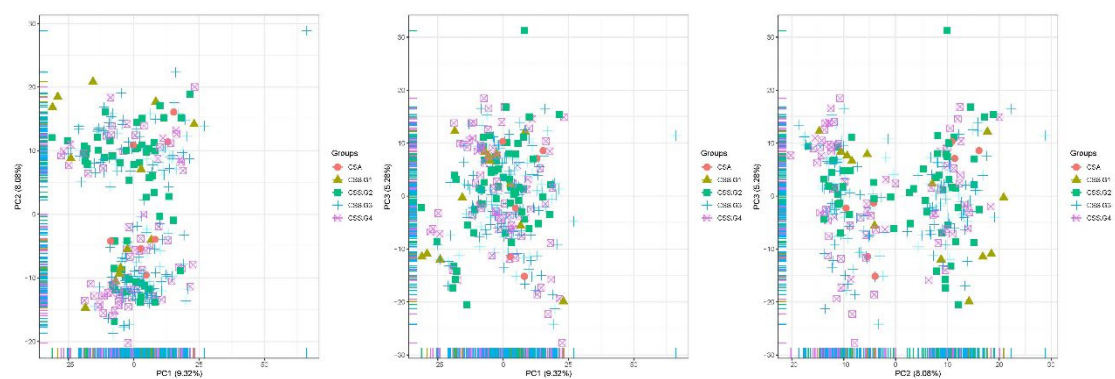


Figure S19. The PCA plot result of all accessions based on all 1562 annotated metabolites.

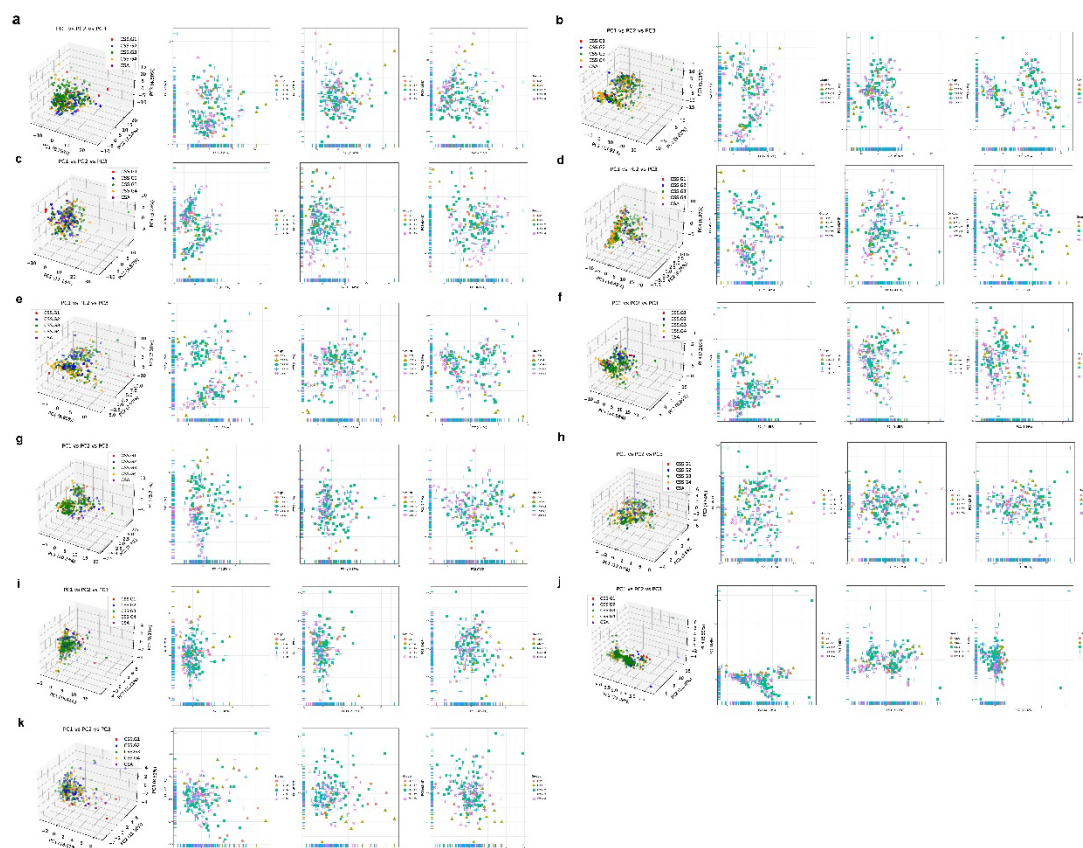


Figure S20. The PCA plot result of all accessions based on 11 super-class metabolites. **a**, Phenylpropanoids and polyketides; **b**, Lipids and lipid-like molecules; **c**, NA; **d**, Benzenoids; **e**, Organoheterocyclic compounds; **f**, Organic acids and derivatives; **g**, Organic oxygen compounds; **h**, Alkaloids and derivatives; **i**, Nucleosides nucleotides and analogues; **j**, Organic nitrogen compounds; **k**, Lignans neolignans and related compounds.

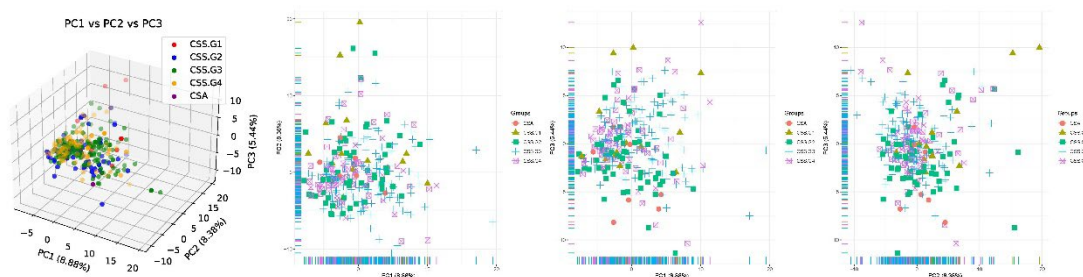


Figure S21. The 3D and 2D plot of PCA result of all accessions based on flavonoids.

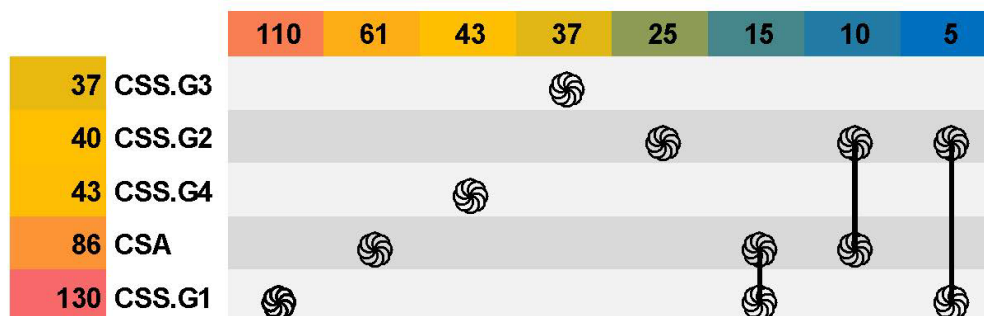


Figure S22. Number of group-specific enriched metabolites.

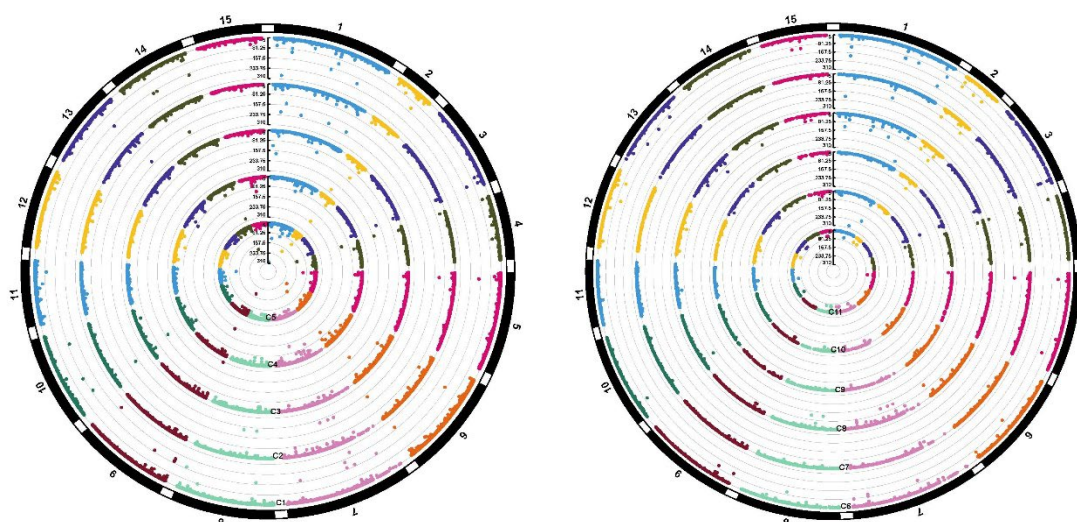


Figure S23. Genome-wide significant SNPs (sSNPs) of 11 metabolite super-classes. **a**, Phenylpropanoids and polyketides; **b**, Lipids and lipid-like molecules; **c**, NA; **d**, Benzenoids; **e**, Organoheterocyclic compounds; **f**, Organic acids and derivatives; **g**, Organic oxygen compounds; **h**, Alkaloids and derivatives; **i**, Nucleosides nucleotides and analogues; **j**, Organic nitrogen compounds; **k**, Lignans neolignans and related compounds.

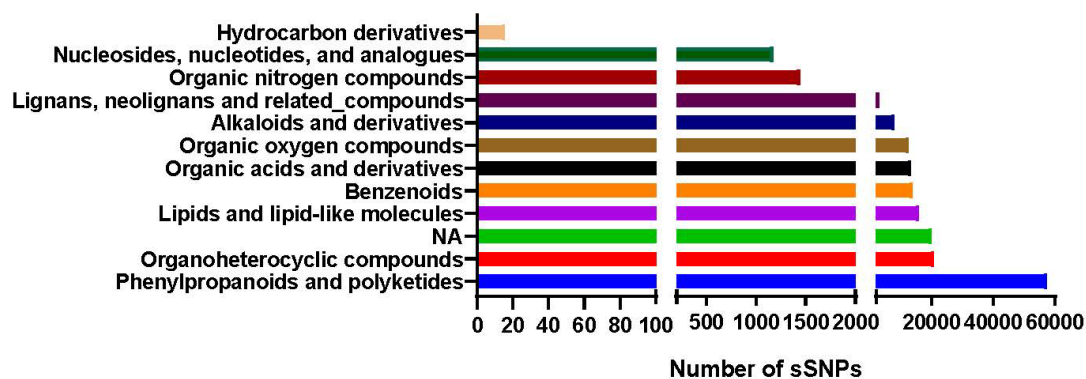


Figure S24. Genome-wide significant SNPs (sSNPs) of 11 metabolite super-classes.

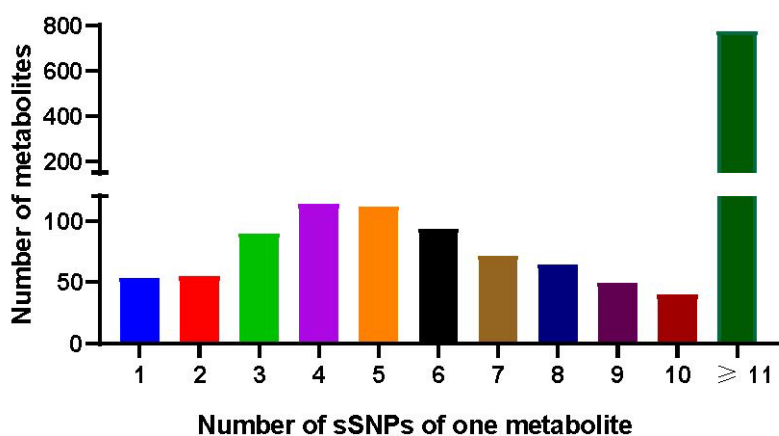


Figure S25. The number of sSNPs of one metabolite.

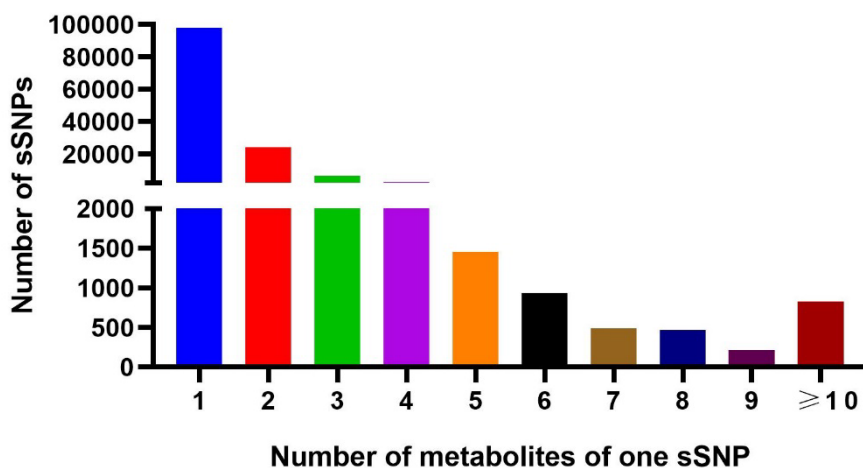


Figure S26. The number of all multi-effect sSNPs.

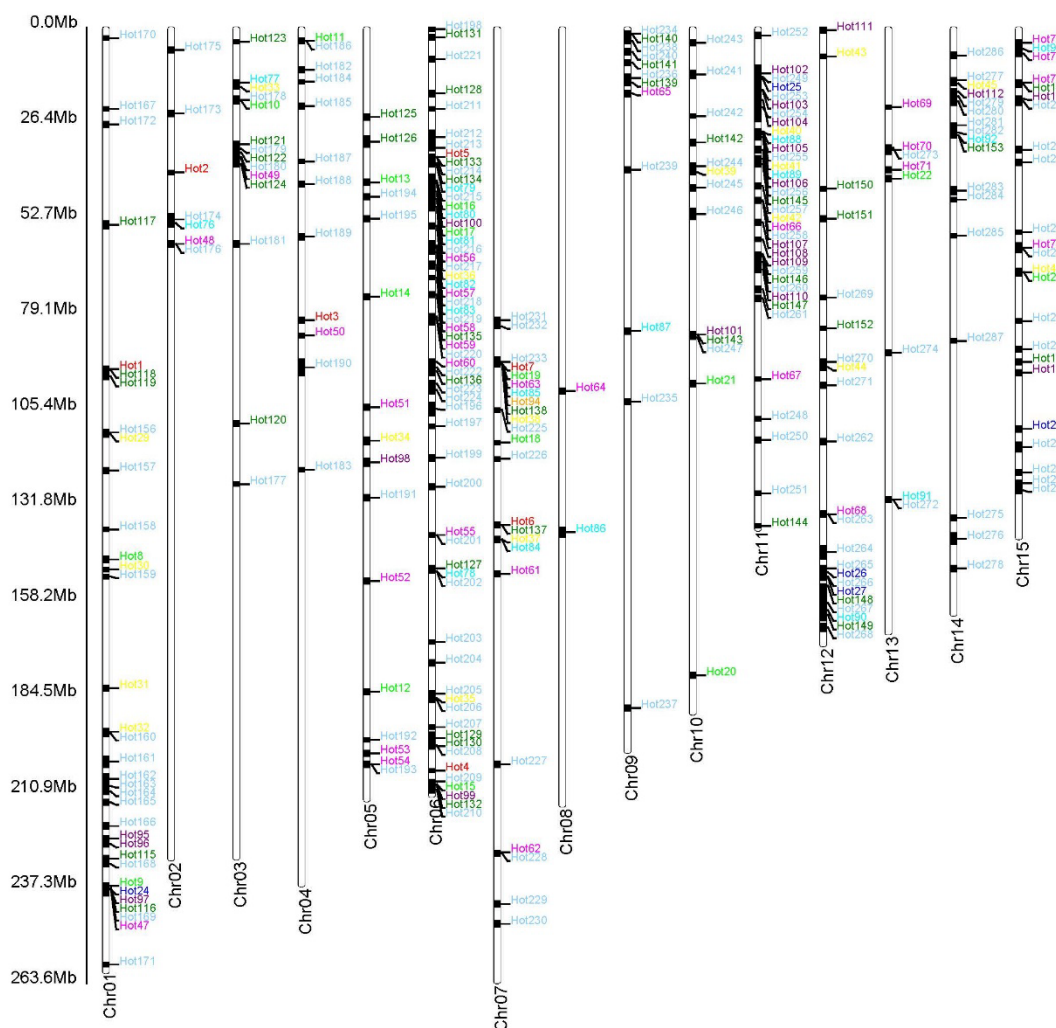


Figure S27. 298 hotspots, including seven, 16, five, 18, 29, 18, one, 20, 41, 143 hotspots related to alkaloids and derivatives, benzenoids, lignans neolignans and related compounds, lipids and lipid-like molecules, NA, organic acids and derivatives, organic nitrogen compounds, organic oxygen compounds, organoheterocyclic compounds, and phenylpropanoids and polyketides super-classes.

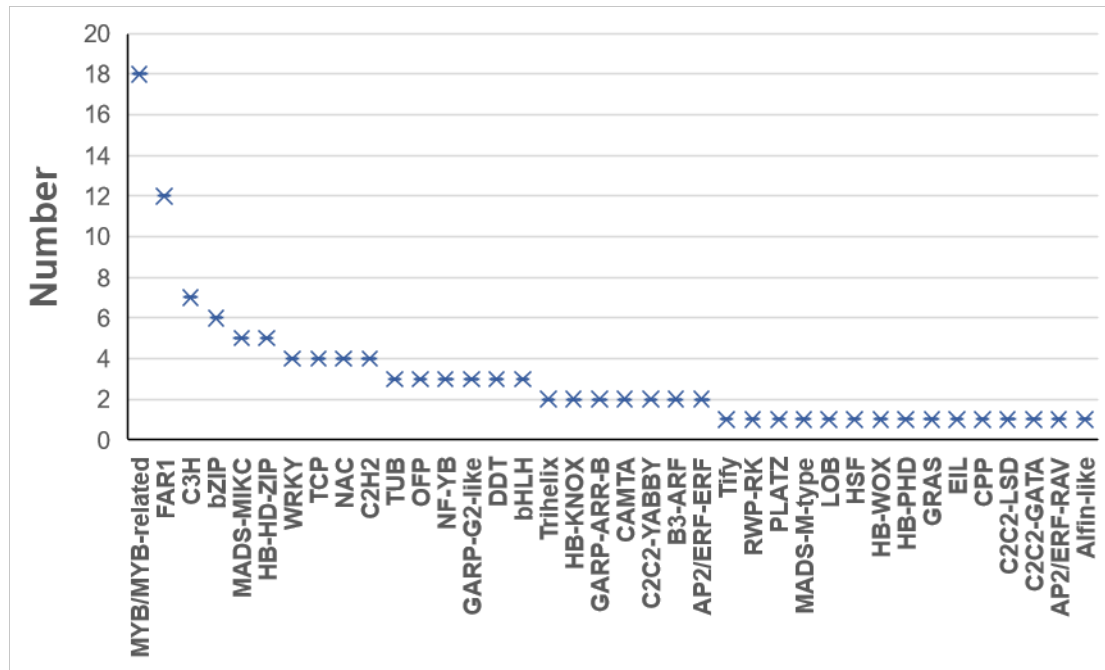


Figure S28. Transcription factors in all hotspots.

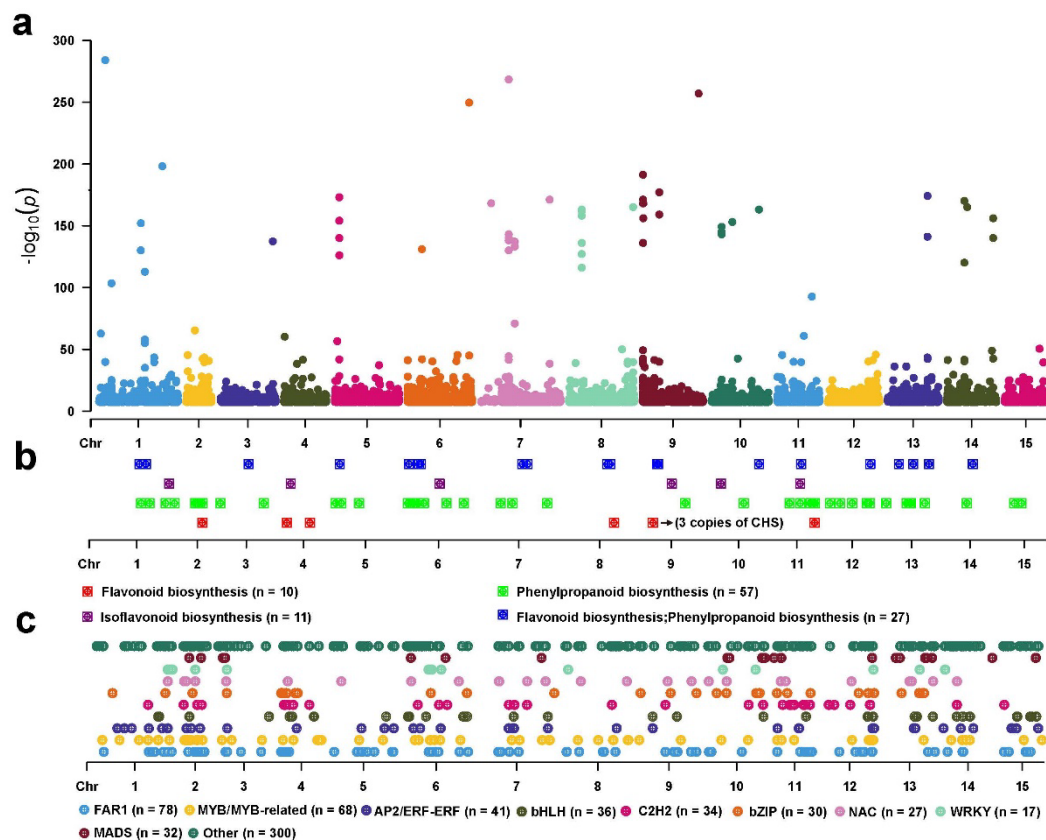


Figure S29. Toward genetic basis of flavonoid synthesis. **a**, Genome-wide significant SNPs (sSNPs) of 208 flavonoids above the significant threshold ($P < 5.24 \times 10^{-8}$, Bonferroni correction). **b** and **c**, Important pathway genes and transcription factors related to flavonoid synthesis.

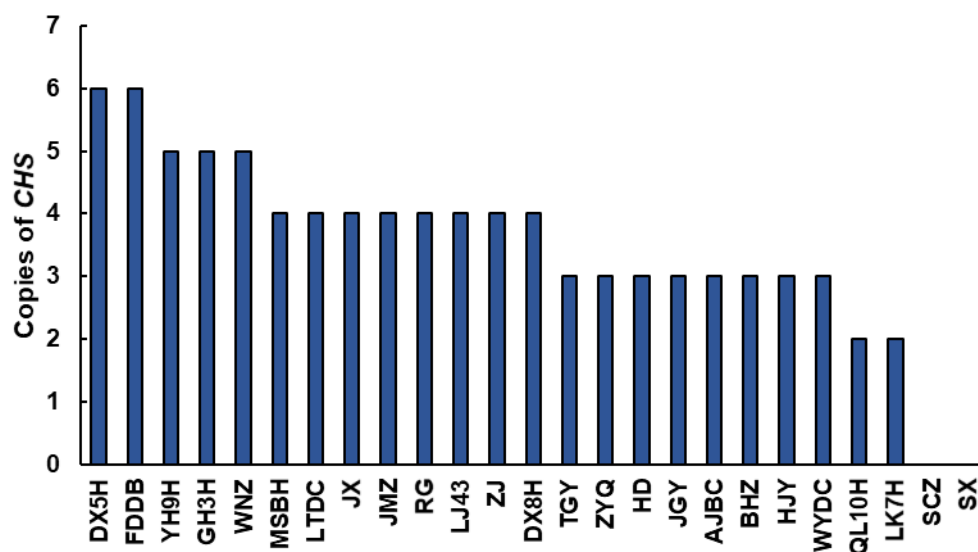


Figure S30. Pan-genomic copy number variation of *Chalcone synthase (CHS)* genes on *CHS* tandem region on chromosome 9.

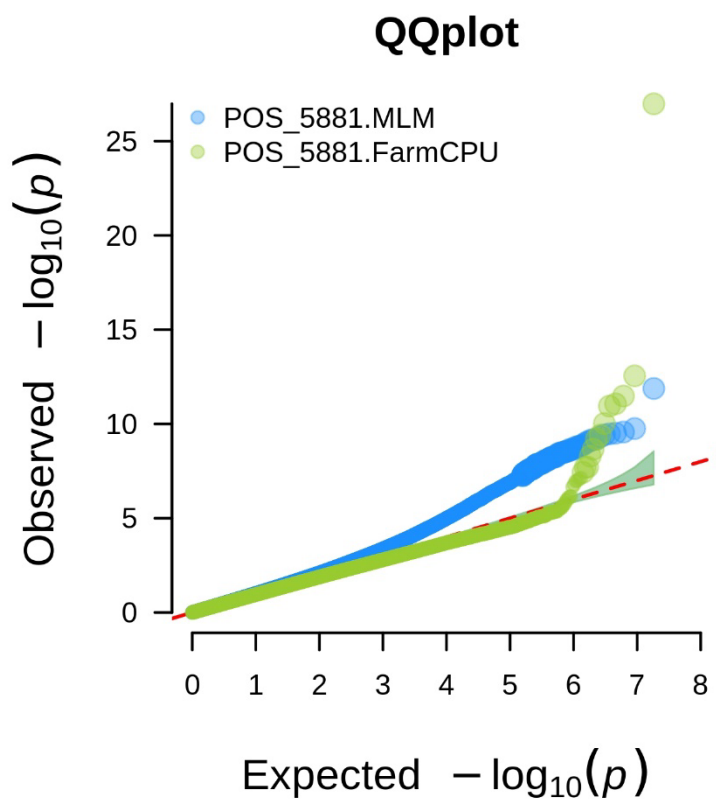


Figure S31. The QQ plot for GWAS signals of (-)-Epicatechin (POS_5881). The statistical analysis was performed with Bonferroni correction.