Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transcriptomics in identifying tea plant breeding lines: a case study on the elite progenies of Camellia sinensis ‘Emei Wenchun’

Wei Chen
Sichuan Agricultural University
Jing Zhang
Sichuan Agricultural University
Guo Chen
Sichuan Agricultural University
DanDan Tang
Sichuan Agricultural University
Dong Cui
Leshan Academy of Agricultural Sciences
Yongxian Chen
Sichuan Agricultural University
Weiqi Wen
Mingshan Tea Plant Breeding and Reproduce Farm of Sichuan Province
Qian Tang
Sichuan Agricultural University
Liqiang Tan
tanliqiang@sicau.edu.cn

Sichuan Agricultural University

Research Article

Keywords: Camellia sinensis, transcriptomics, tea breeding, elite progenies, SNP markers, DNA fingerprinting

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10056892/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Tea plant breeding typically progresses through individual selection, line evaluation, regional adaptability testing, and cultivar registration. Despite its pivotal role in genetic improvement, line identification remains a major bottleneck. In this study, we integrated transcriptomic and biochemical profiling of one-bud-two-leaf samples from 15 tea accessions to improve the precision and molecular depth of elite clonal line evaluation. We identified the regulatory networks underlying key quality components. Elevated epigallocatechin gallate (EGCG) accumulation positively correlated with the up-regulation of core structural genes such as *CHS*, *F3'H*, *LAR*, *ANS* and specific transcription factors like *CsMYB199* and *CsMYB111*. Concurrently, enhanced caffeine biosynthesis was driven by the increased expression of *TCS1/2* and *CsMYB184*, coupled with the suppression of *CsMYB14-like*. Furthermore, to resolve the ambiguous parentage typical of open-pollinated elite lines, we mined 1,470 highly expressed, synonymous single nucleotide polymorphism (SNP) markers from the transcriptome. Using exclusion-based parentage analysis, 13 progeny lines exhibited zero triplet mismatches with their putative parents, confirming their true parentage. By contrast, unrelated lines showed mismatch counts ranging from 53 to 316. To further identify and protect these elite lines, we selected a core set of 150 SNP markers to develop DNA fingerprint profiles. The cumulative probability of identity (PI) for a core set of 150 SNPs reached 4.26×10^{-57}, providing unprecedented resolution for DNA fingerprinting and plant variety protection (PVP). Ultimately, this study establishes a robust transcriptomics-driven framework that bridges molecular trait mechanisms with precision tea breeding strategies.

Introduction

The tea plant (*Camellia sinensis*) is a globally vital economic crop, making its genetic improvement and quality enhancement central themes in agricultural science. An efficient breeding system serves as a crucial foundation for enabling sustained development of the tea industry. Over recent decades, the paradigm of tea breeding has progressively shifted from prioritizing basic agronomic traits, such as high yield, toward developing cultivars with "specialized and diversified" quality profiles. Consequently, the establishment of a highly efficient breeding evaluation system has become a crucial foundation for modern tea science [1,2].

The systematic selection of elite tea accessions generally progresses through individual plant screening, clonal line evaluation, regional adaptability testing, and cultivar registration. While initial individual screening and subsequent field trials are well-established, the intermediate stage of clonal line identification remains a critical bottleneck [3]. Traditional line identification relies predominantly on morphological descriptors and agronomic observations. Although intuitive, these phenotypic traits are vulnerable to environmental plasticity and fail to capture the profound genetic differentiation and molecular regulatory networks driving secondary metabolite biosynthesis [4]. This methodological lag necessitates a paradigm shift from conventional phenotype-based screening to robust molecular diagnostics.

To bridge this gap, multi-omics technologies particularly transcriptomics, via systematic analysis of gene expression patterns, can capture functional-level differences not directly captured by genomic sequence variations, thereby providing a more direct molecular insight into phenotypic variations. Early studies have demonstrated its potential in linking gene expression with agronomic traits in various crops, such as corn, oat, and ginkgo, often yielding comparable or even superior prediction accuracy relative to genomic models [5,6,7]. Within tea research, transcriptomics has revolutionized our understanding of secondary metabolism and stress physiology, yielding breakthroughs such as mapping catechin biosynthesis at single-cell resolution [8], identifying critical regulators in naturally caffeine-free mutants [9], and elucidating metabolite accumulation under multiple stresses or between wild and cultivated varieties [10,11]. Furthermore, integrating transcriptomics with metabolomics, genome-wide association studies (GWAS), and pangenome analyses has decoded the multi-layered regulatory mechanisms underlying complex agronomic traits, including bud size, branching angle, metabolic heterosis, and trait diversification driven by structural variations [12,13,14,15]. Beyond functional gene discovery, transcriptome-derived markers have effectively resolved intraspecific relationships and identified selection signals, complementing broader genomic selection strategies [16,17,18].

Despite these advancements, the application of transcriptomics for the precise discrimination of closely related clonal lines remains limited. This challenge is further exacerbated by the reproductive biology of tea plants. As a predominantly outcrossing species [19], many elite lines are selected from open-pollinated offspring, resulting in ambiguous paternal origins that severely hinder accurate pedigree tracing [20]. While traditional marker systems suffer from low throughput, high-depth RNA sequencing offers a dual advantage. It simultaneously profiles gene expression networks and provides a rich repository of high-quality, transcriptionally active single nucleotide polymorphisms (SNPs) [21,22]. Utilizing these functional SNPs enables rigorous parentage verification and facilitates the construction of highly specific DNA fingerprint profiles. These molecular 'ID cards' are statistically robust, characterized by an exceptionally low cumulative probability of identity (PI), thereby offering irrefutable evidence for plant variety protection (PVP) and the commercial regulation of elite clones [23].

To systematically resolve these bottlenecks, we investigated 13 elite progeny lines derived from open-pollinated *C. sinensis* 'Emei Wenchun', an elite landrace highly prized for its extremely early sprouting. By integrating targeted biochemical quantification with high-throughput RNA-seq of standardized shoots, we comprehensively profiled their secondary metabolic architectures and transcriptomic signatures. Subsequently, transcriptional regulatory network analysis and large-scale synonymous SNP mining were applied to elucidate the molecular mechanisms underlying elite quality traits, validate true parentage via exclusion-based analysis, and develop highly discriminatory DNA fingerprint profiles. Ultimately, this integrative approach bridges the gap between theoretical secondary metabolism research and practical molecular diagnostics, supplying high-fidelity targets for precision breeding and genetic improvement in tea plants.

Experimental Materials and Methods

Plant Materials

Thirteen elite clonal lines (EWCM129, EWCM136, EWCM146, EWCM241, EWCM244, EWCM338, EWCM339, EWCM411, EWCM432, EWCM631, EWCM641, EWCM804, and EWCM825) were selected from an established F_1 population derived from the open-pollinated progeny of 'Emei Wenchun' (maternal parent) and 'Chuanmu 217' (putative paternal parent) [24]. All clonal lines were cultivated and sampled from the tea breeding base in Mingshan County, Sichuan Province, China (N28°59', E103°53'). To minimize environmental variance, all selected progeny and their parental lines were vegetatively propagated via cuttings and cultivated under uniform agronomic practices in the same experimental field for 2–3 years. In early spring, standard tender shoots (one bud and two leaves) were systematically harvested (Figure S1). These samples were processed in two ways. For biochemical analysis, samples were subjected to microwave fixation, followed by drying at 80°C for 5 hours, then ground through a 40-mesh sieve and stored at 4°C for subsequent testing. For transcriptomic sequencing, fresh samples were rapidly frozen in liquid nitrogen and stored at -80°C in an ultra-low temperature freezer for subsequent sequencing.

Quantification of biochemical components

Biochemical characterization was conducted strictly adhering to standard analytical protocols. Briefly, water-soluble extracts were quantified following the gravimetric method [25]. Total polyphenols were determined using the Folin-Ciocalteu colorimetric assay, while individual catechin monomers were profiled using High-Performance Liquid Chromatography (HPLC) according to GB/T 8313 – 2018 [26]. Caffeine and total free amino acids were quantified using ultraviolet spectrophotometry and the ninhydrin colorimetric assay, according to GB/T 8312 – 2013 and GB/T 8314 – 2013 [27, 28], respectively.

Transcriptome Sequencing and Analysis

Total RNA was extracted from the sampled tissues using the CTAB-PBIOZOL method combined ethanol precipitation. RNA concentration and integrity were assessed using a Qubit fluorescence quantitative analyzer and a Qsep400 high-throughput biochip analyzer. Following quality control, cDNA libraries were constructed and sequenced on an Illumina platform by Wuhan Maiwei Metabolic Biotechnology Co., Ltd. (Wuhan, China).

Raw sequencing data were quality-filtered using Fastp to remove low-quality reads and obtain clean reads [29]. Reads containing adapters were removed. Additionally, reads were discarded if the proportion of N bases exceeded 10% or if more than 50% of bases had a quality score ≤ 20. These clean reads were used for all subsequent analyses. The resulting clean reads were aligned to the *Camellia sinensis* 'Tieguanyin' reference genome using HISAT [30]. Transcripts were assembled and quantified using StringTie, with gene expression levels quantified as FPKM values [31].

Differential expression and enrichment analysis

Differential expression between the parental lines and clonal progeny was assessed using DESeq2 package, with P-values adjusted by the Benjamini–Hochberg method. Genes were considered significantly differentially expressed if the adjusted P-value (FDR) < 0.05 and |log2Fold Change| ≥ 1. Enrichment analysis was performed via the hypergeometric test, with KEGG pathway and GO term enrichments assessed at the pathway and term levels. For multiple KEGG comparisons, the 15 pathways with the lowest adjusted P-values (q-values) were selected for display.

SNP molecular marker screening and analysis

Genomic variant calling was executed via the GATK pipeline, followed by functional annotation using ANNOVAR [32]. To isolate high-fidelity core SNP markers for rigorous parentage verification, a stringent multi-tiered filtering strategy was applied: (1) restriction to synonymous substitutions within coding sequences (retention of one representative SNP per gene); (2) robust transcript expression, defined as FPKM ≥ 20 in over 90% of samples, to guarantee genotyping accuracy and circumvent allelic dropout; and (3) confirmed polymorphism between the parental accessions. To evaluate the discriminatory power of the finalized 150-SNP panel, the probability of identity (*PI*) for each locus was computed using the following equation [33]:

$$PI = 2(\sum p_i^2)^2 - \sum p_i^4$$

where p_i denotes the frequency of the i^(th) allele.

Statistical analysis

Biochemical data were organized using Microsoft Excel 2019 and analyzed via IBM SPSS Statistics 22. Multiple group comparisons were performed via one-way ANOVA followed by LSD post-hoc tests in SPSS. Heat maps were generated using Weishengxin and the Maiwei Cloud Platform. The Pearson's correlation coefficients r was employed to assess the significance of associations between transcriptome gene expression levels and biochemical component concentrations.

Results

Phenotypic variation of primary biochemical components and catechins

To characterize the phenotypic variation in quality-related metabolites among the elite progenies of *C. sinensis* 'Emei Wenchun', key biochemical components were quantified across the 13 clonal lines and their parental accessions (Fig. 1A and Table S1). Water-soluble extract ranged from 40.28% to 46.99%, with a coefficient of variation (CV) of 4.16%. Notably, lines EWCM241, EWCM244, and EWCM641

exhibited significantly higher accumulation than the maternal parent 'Emei Wenchun' (EW) ($p < 0.05$). Tea polyphenols ranged from 16.33% to 22.13% (CV = 8.48%), generally surpassing the maternal levels except in four specific lines (EWCM129, EWCM146, EWCM631, and EWCM804). Free amino acid content displayed the highest variability (CV = 11.56%), ranging from 4.60% to 6.47%, with line 244 recording the highest content. Caffeine content varied from 3.25% to 4.12% (CV = 7.33%), with all lines except EWCM338 lower than the putative paternal parent 'Chuanmu 217' (CM). The phenol-to-amino acid ratio across all lines ranged from 2.75 to 4.81, indicating the optimal range for premium green tea.

Catechin components analysis (Fig. 1B and Table S2) showed total catechin levels between 12.92% and 19.35%. With the exception of EWCM146 and EWCM631, all progeny lines surpassed both parents in total catechin content. Epigallocatechin gallate (EGCG) was the most abundant catechin (5.78%-9.26%), with a higher mean concentration in the progenies relative to both parents. Epicatechin gallate (ECG) and gallocatechin gallate (GCG) displayed similar average concentrations but distinct variability, with GCG exhibiting a notably higher CV. The calculated catechin quality index (mean = 590.01) strongly underscores the superior processing potential of these elite lines.

Transcriptomic profiling and functional enrichment analysis

To elucidate the transcriptomic basis underlying the biochemical divergence, RNA-seq was performed on one-bud-two-leaf samples, generating 30 high-quality cDNA libraries. Comprehensive quality control confirmed high sequencing fidelity, with Q20 and Q30 values exceeding 98.49% and 95.33%, respectively, and an average mapping rate to the reference genome surpassing 90.09% (Table S3). A total of 17,037 differentially expressed genes (DEGs) were identified. Notably, lines EWCM825, EWCM241, and EWCM244 exhibited the most extensive transcriptomic reprogramming relative to the maternal parent, encompassing over 5,400 DEGs each (Fig. 2A).

Venn diagram and hierarchical clustering analyses isolated DEGs uniquely shared or differentiated between the progenies and their parents (Fig. 2B-C). Functional enrichment of these transcripts highlighted pathways critical to secondary metabolism and stress responses (Table S4). By contrasting extreme phenotypic groups for free amino acids, EGCG, and caffeine, Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses consistently mapped the associated DEGs to "plant-pathogen interaction," "plant hormone signal transduction," and specific secondary metabolite biosynthesis pathways, such as flavonoid and flavanol biosynthesis for the EGCG cohort (Fig. 2D-F). These findings imply a profound interconnectedness between plant defense signaling, hormonal regulation, and the differential accumulation of quality-related metabolites.

Co-expression networks linking target metabolites and transcripts

To bridge the gap between metabolite accumulation and transcriptional regulation, Pearson correlation networks were constructed ($|r| \geq 0.8$, $p < 0.05$). The analysis pinpointed specific genetic hubs strongly associated with biochemical traits. For instance, transcripts TGY015354 and TGY121180 exhibited robust positive correlations with total free amino acids, whereas TGY073101 was tightly linked to EGCG content. Furthermore, a module of six distinct genes including TGY012701 and TGY040259 showed highly significant correlations with caffeine biosynthesis (Fig. 3). These hub genes serve as prime candidates for downstream functional validation.

Transcriptional regulation of flavonoid and caffeine biosynthetic pathways

Flavonoid biosynthesis in higher plants is orchestrally governed by a hierarchical network comprising both structural genes and diverse transcription factors. Utilizing EGCG accumulation as the primary phenotypic index, we analyzed transcriptomic flux across four high-EGCG (641, 825, 338, 244) and four low-EGCG (136, 146, 432, 631) clonal lines. Correlation analyses revealed that the up-regulation of early and intermediate structural genes, namely *CHS* (TGY079101, TGY047870), *F3'H* (TGY125466), *LAR* (TGY051808), and *ANS1* (TGY116873), positively drove EGCG accumulation (Table S5). Conversely, *DFR* homeologs exhibited negative correlations, suggesting a complex competitive branching in the downstream flavonoid pathway (Fig. 4). At the regulatory level, transcription factors *CsWRKY70* and *CsMYB156* were positively associated with EGCG, whereas *CsMYB119* exerted a potential repressive effect.

A parallel investigation into caffeine biosynthesis evaluated key genes including *TCS*, *MXMT*, *XMT*, and *SAMA* across high- (338, 339, 129, 411) and low-caffeine (244, 146, 136, 641) lines. Enhanced caffeine accumulation strictly coincided with the elevated expression of *MXMT* (TGY127102) and *TCS1/2*, alongside the critical transcription factor *CsMYB184* ($r = 0.639$). In stark contrast, *SAMA* and *XMT* transcripts demonstrated robust negative correlations ($r \leq -0.65$) (Fig. 5 and Table S5). These integrated datasets delineate the precise molecular checkpoints dictating metabolic divergence in elite tea lineages.

Variant analysis and core SNP screening

Utilizing the GATK variant calling pipeline, a comprehensive set of 840,615 putative SNP loci was identified across the evaluated lines, yielding a high genomic density with an average inter-marker distance of 0.62 kb. Analysis of substitution patterns revealed 545,911 transitions (Ts) and 294,704 transversions (Tv), resulting in a Ts/Tv ratio of 1.85 (Table 1). As expected, transition events were heavily dominated by A/G (272,897) and C/T (273,014) substitutions. Among the functionally annotated transcripts across the 15 tea accessions, individual lines harbored between 60,700 and 68,004 synonymous variants, alongside 53,195 to 59,896 non-synonymous variants (Table S6).

Table 1 Summary of the 840,615 putative single-nucleotide polymorphisms (SNPs) identified from the unigenes of the 15 tea genetic resources

Number of unigenes	1,350,334,536
Total bases (G)	234.3
Number of SNPs	840,615
SNP frequency	0.62 SNP/Kb
Transition	545,911
A/G	272,897
C/T	273,014
Transversion	294,704
A/C	71,489
A/T	90,842
C/G	61,583
T/G	70,790

Parentage analysis and construction of line fingerprint profiles

To establish an unequivocal identification framework, we curated a stringent panel of 1,470 synonymous SNPs localized within highly expressed coding regions. These markers exhibited uniform genomic distribution and high allelic specificity (Fig. 6A). Employing an exclusion-based parentage analysis via triplet mismatch counting, we unambiguously validated the true pedigree. Crucially, the triplet mismatch count between the EW and CM accessions against the validated progenies was consistently zero. By contrast, pairwise comparisons among unrelated siblings or non-parental combinations exhibited high mismatch frequencies, ranging from 53 to 316 (Table 2).

For practical application in plant variety protection (PVP), a core subset of 150 elite SNPs (10 per chromosome) was selected to construct a highly discriminatory DNA fingerprinting profile (Fig. 6C). Analysis of minor allele frequencies (MAF) confirmed robust polymorphism, primarily concentrated in the 0.201 to 0.261 range (Fig. 6B). The locus-specific probability of identity (PI) averaged 0.423. Under the assumption of independent assortment, the cumulative PI for this 150-SNP multiplex reached an extraordinary threshold of 4.26×10^{-57}. This mathematical stringency effectively eliminates the probability of identical genotype sharing among non-clonal individuals, cementing this panel as a molecular tool for tea cultivar discrimination.

Table2 Parent Child Relationship Analysis (Statistics of Triple Mismatch Count)

ID	EW x CM	EW x others F1 (n=12)	CM x others F1 (n=12)	F1 x F1 (n=66)
EWCM129	0	99~207 (153.50)	115-227 (187.10)	91~273(175.06)
EWCM136	0	111~175(146.75)	95~218(172.08)	53~236(156.53)
EWCM146	0	146~255(182.58)	154~272(181.83)	96~316(196.98)
EWCM241	0	114~220(163.42)	85~287(192.58)	102~309(175.00)
EWCM244	0	120~232(165.00)	103~250(180.25)	93~302(184.26)
EWCM338	0	115~227(165.50)	118~245(175.17)	112~243(170.88)
EWCM339	0	111~180(156.75)	86~219(166.75)	79~270(151.85)
EWCM411	0	125~262(183.75)	133~255(171.30)	102~304(194.52)
EWCM432	0	101~186(147.08)	138~196(171.50)	85~237(152.55)
EWCM631	0	100~249(164.33)	128~231(189.92)	85~279(189.53)
EWCM641	0	93~213(149.50)	124~247(188.42)	97~256(174.89)
EWCM804	0	123~179(145.92)	83~287(182.75)	92~261(160.14)
EWCM825	0	98~236(154.92)	133~199(162.00)	68~243(157.15)

Discussion

Accurate variety identification and phenotypic evaluation remain foundational to tea plant breeding. While traditional morphological assessments provide intuitive baselines, their susceptibility to phenotypic plasticity limits their capacity to capture the true genetic divergence among elite clonal lines. Secondary metabolites, notably catechins and caffeine, are dynamically governed by multi-layered transcriptional cascades. Therefore, leveraging comparative transcriptomics to bridge transcript abundance with metabolic phenotypes offers a robust strategy for decoding the genetic basis of quality traits. Our study systematically applies this paradigm to an elite progeny population, providing novel insights into the transcriptional rewiring that drives metabolic diversity.

By integrating biochemical and transcriptomic profiles across the hybrid progenies and their parents, we identified substantial expression divergence (Fig. 2), suggesting that genetic recombination actively drives metabolic phenotypic variation through widespread transcriptional reprogramming. Focusing on EGCG, the predominant and most bioactive catechin monomer, our co-expression analysis revealed that the up-regulation of early and intermediate structural genes (*CHS*, *F3'H*, *LAR*, and *ANS*) alongside the

transcription factor CsWRKY70 acts as a primary positive driver for EGCG accumulation (Fig. 3 and TableS5), consistent with previous findings [34,35,36,37]. Intriguingly, several transcription factors classically associated with the broad up-regulation of catechin biosynthesis, such as CsTCP3, CsTT8, and CsMYB12, exhibited pronounced negative correlations specifically with EGCG levels in our elite lines (Fig. S2; [38,39,40]). We postulate that this apparent discrepancy arises from the resolution of the metabolic target, whereas previous studies predominantly evaluated the total catechin pool, our analysis isolates the highly specific regulatory signals governing the EGCG monomer. Furthermore, large regulatory gene families like MYB and TCP are characterized by extensive functional redundancy, sequence divergence, and complex phenomena such as competitive promoter binding. Alternative splicing events within these metabolic hubs may also fine-tune the efficiency of specific biosynthetic branches [41]. Future functional assays must dissect these competitive binding dynamics and their specific roles in galloylation and stereoisomeric conversion to fully resolve the EGCG regulatory network.

Regarding caffeine biosynthesis, the expression of core structural genes (*TCS1* and *TCS2*) and the transcription factor CsMYB184 showed significant positive correlations with caffeine accumulation (Fig. 3 and Table S5). This provides compelling transcriptomic validation of the TCS1-MYB184 regulatory module previously identified via GWAS [42]. Conversely, the negative correlation of CsMYB14-like implies a putative repressive function, potentially operating through direct transcriptional repression or the recruitment of corepressor complexes. Beyond autonomous transcriptional control, it is essential to recognize the profound influence of phytohormone signaling matrices. Recent evidence highlights that integrated hormone networks critically modulate the biosynthesis of EGCG, theanine, and caffeine under varying environmental stimuli [43]. The hormone-metabolite cross-talk observed in our dataset likely reflects highly coordinated, spatiotemporally dynamic regulatory feedback loops, presenting a fertile ground for future multi-omics investigations, particularly those integrating stress physiology frameworks.

Beyond metabolic regulation, accurate genetic mapping and parentage verification are fundamental for perennial woody crops like the tea plant. Here, we analyzed the transcriptomes of two parental lines and 13 progenies, utilizing high-density synonymous SNPs derived from highly expressed genes for parentage verification and cultivar fingerprinting (Fig. 6). Because highly expressed genes endure strong evolutionary constraints, their coding regions exhibit high sequence conservation, low synonymous mutation rates, and high stability [44,45]. Thus, transcriptome-derived SNPs accurately reflect true genetic divergence and offer a direct link between sequence variation, gene expression, and phenotype. However, it is important to acknowledge that this approach inherently excludes non-expressed intergenic variation, introducing an ascertainment bias toward transcribed regions that should be complemented by whole-genome strategies in future studies.

Given their high density, codominant inheritance, and technical reproducibility, SNP markers have become central to tea germplasm identification. Previous studies provide useful benchmarks. For example, 973 core SNPs were employed to differentiate genetic backgrounds among 349 tea resources [46], and a 24-SNP molecular ID were established to distinguish 103 varieties [47]. The genetic similarity

thresholds reported in these studies align broadly with the mismatch thresholds applied here. From spring one-bud-two-leaf transcriptomes of 13 ‘Emei Wenchun’ progeny lines, we identified 1,470 synonymous SNPs derived from highly expressed genes. Because these markers originate from transcribed regions, they preferentially capture variation in functional loci while inherently excluding non-expressed intergenic variation; this advantage facilitates linkage to phenotype but also introduces ascertainment bias that must be carefully considered. Our approach is conceptually similar to high-density marker strategies such as SLAF-seq in seeking dense genome coverage, although transcriptome-derived and SLAF-derived markers differ in genomic representation [48]. Employing an exclusion-based triplet mismatch analysis, we unambiguously validated parent–offspring pairs (zero mismatches), while non-parent–child combinations displayed 53 to 316 mismatches. This conclusively verified ‘Chuanmu 217’ as the paternal parent of these superior lines. By further distilling this dataset to 150 core markers (10 per chromosome), we achieved an exceptionally low cumulative probability of identity (PI) of 4.26×10^{-57}. Although this SNP panel exhibits high discriminatory power, its universal applicability requires further validation across diverse tea germplasm.

Conclusion

In summary, this study integrated biochemical profiling and transcriptomics across the hybrid progeny of *C. sinensis* ‘Emei Wenchun’ and ‘Chuanmu 217’ to elucidate the genetic and molecular basis of quality trait variation. By correlating differential gene expression with metabolite accumulation, we characterized the transcriptional networks driving elevated EGCG and caffeine biosynthesis at the clonal line level. Simultaneously, the strategic mining of transcriptome-derived synonymous SNPs yielded a highly discriminative DNA fingerprinting framework, enabling unambiguous parentage verification and providing a robust tool for plant variety protection. Future research should focus on functional validation of candidate regulators, the application of this marker panel across broader germplasm, and the utilization of multi-omics network modeling to fully resolve the regulatory dynamics of tea quality formation.

Abbreviations

ANOVA, analysis of variance;

ANS, anthocyanidin synthase;

cDNA, complementary DNA;

CHS, chalcone synthase;

CM, Chuanmu 217;

CTAB, cetyltrimethylammonium bromide;

CV, coefficient of variation;

DEGs, differentially expressed genes;

DFR, dihydroflavonol 4-reductase;

EC, epicatechin;

ECG, epicatechin gallate;

EGC, epigallocatechin;

EGCG, epigallocatechin gallate;

EW, Emei Wenchun;

F1, first filial generation;

F3'H, flavonoid 3'-hydroxylase;

FDR, false discovery rate;

FPKM, fragments per kilobase of transcript per million mapped reads;

GATK, Genome Analysis Toolkit;

GC, gallocatechin;

GCG, gallocatechin gallate;

GO, Gene Ontology;

GWAS, genome-wide association study;

HPLC, high-performance liquid chromatography;

KEGG, Kyoto Encyclopedia of Genes and Genomes;

LAR, leucoanthocyanidin reductase;

LSD, least significant difference;

MAF, minor allele frequency;

MXMT, 7-methylxanthine methyltransferase;

PI, probability of identity;

PVP, plant variety protection;

RNA-seq, RNA sequencing;

SD, standard deviation;

SLAF-seq, specific-locus amplified fragment sequencing;

SNP, single-nucleotide polymorphism;

TCS, tea caffeine synthase;

Ts/Tv, transition-to-transversion ratio;

XMT, xanthosine methyltransferase.

Declarations

Author contribution

Liqiang Tan: Supervision, Funding acquisition, Conceptualization. **Dandan Tang:** Supervision, Investigation. **Guo Chen:** Investigation. **Dong Cui:** Investigation. **Jing Zhang:** Writing – original draft, Investigation, Data curation. **Wei Chen**: Writing – review & editing, Writing – original draft, Investigation, Data curation, Conceptualization. **Qian Tang:** Supervision, Resources, Investigation. **Weiqi Wen:** Investigation. **Yongxian Chen**: Investigation.

Funding

This work was supported by the Sichuan Innovation Team of National Modern Agricultural Industry System (sccxtd-2024-10), National Natural Science Foundation of China (32372767, 32402633), and the Ya'an City's "Open Call for Applications" Project (kczx2023-2025-01).

Data availability

The datasets generated and analysed during the current study are available in the GSA repository (https://ngdc.cncb.ac.cn/gsa/browse/CRA045160).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Acknowledgements

Not applicable.

References

1. Apostolides 1CL, Chen Z. Global tea breeding: achievements, challenges and perspectives. In: Chen L, Apostolides Z, Chen ZM, editors. Global Tea Breeding: Achievements, Challenges and Perspectives. Berlin: Springer; 2012. pp. 13–68. https://doi.org/10.1007/978-3-642-31878-8.
2. Song 2LH, Zhang K, Wang X, Dong D, Liu S. Application of multi-perspectives in tea breeding and the main directions. Int J Mol Sci. 2023;24(16):12643. https://doi.org/10.3390/ijms241612643.
3. Shi 3LY. Advances in tea plant genetics and breeding. J Tea Sci. 2015;35(2):103–9. https://doi.org/10.13305/j.cnki.jts.2015.02.001.
4. Han 4KNK, Shangguan J, Wang L, Kayesh C, Zhang E. Plant variety and cultivar identification: advances and prospects. Crit Rev Biotechnol. 2013;33(2):111–25. https://doi.org/10.3109/07388551.2012.675314.
5. Pardo 5ACB, VanBuren J, de los Campos R, Shiu G. Transcriptome-based prediction of complex traits in maize. Plant Cell. 2020;32(1):139–51. https://doi.org/10.1105/tpc.19.00332.
6. Yoon 6KTH, Park YM, Park JC, Kim JH, Kim KH. De novo transcriptome assembly and SNP discovery for the development of dCAPS markers in oat. Agronomy. 2022;12(1):184. https://doi.org/10.3390/agronomy12010184.
7. Han 7LH, Ruan X, Xu J, He L. Transcriptome analysis of Ginkgo biloba L. leaves across late developmental stages based on RNA-seq and co-expression network. Forests. 2021;12(3):315. https://doi.org/10.3390/f12030315.
8. Wu 8WQ, Peng Y, Cui A, Zhao J, Pan M. Single-cell transcriptome atlas reveals developmental trajectories and a novel metabolic pathway of catechin esters in tea leaves. Plant Biotechnol J. 2022;20(11):2089–106. https://doi.org/10.1111/pbi.13891.
9. Yang 9MX, Qiao C, Tang D, Guo M, Liang Y. De novo full-length transcriptome analysis of a naturally caffeine-free tea plant reveals specificity in secondary metabolic regulation. Sci Rep. 2023;13(1):6015. https://doi.org/10.1038/s41598-023-32435-5.
10. Han 10LZ, Zhou Y, Wang Y, Lian T, Yuan S. Transcriptomic analysis of tea plant (Camellia sinensis) revealed the co-expression network of 4111 paralogous genes and biosynthesis of quality-related key metabolites under multiple stresses. Genomics. 2021;113(1):908–18. https://doi.org/10.1016/j.ygeno.2020.10.023.
11. Yang 11WF, Li X, Ning W, Huang P. Metabolomics and transcriptomics reveal the accumulation of key metabolites and flavor formation in the fresh leaves of wild and cultivated tea plants. Ind Crops Prod. 2025;235:121780. https://doi.org/10.1016/j.indcrop.2025.121780.

12. Duan 12LY, Huang J, Ding F, Kang D, Luo Y. Integrated transcriptomic and metabolomic analysis reveals metabolic heterosis in hybrid tea plants (Camellia sinensis). Genes. 2025;16(12):1457. https://doi.org/10.3390/genes16121457.
13. Zhu 13TL, Hu J, Xu J, Wu Q, Chen J. Pangenome analyses of tea plants reveal structural variations driving gene expression alterations and agronomic trait diversification. Nat Commun. 2026;17:372. https://doi.org/10.1038/s41467-025-67060-5.
14. Chen 14ZS, Fu S, Li Z, Chen F, Ma Q. Integration of digital phenotyping, GWAS, and transcriptomic analysis revealed a key gene for bud size in tea plant (Camellia sinensis). Hortic Res. 2025;12(6):uhaf051. https://doi.org/10.1093/hr/uhaf051.
15. Li 15ZJ, Huang X, Wang J, Zheng L, Li Q. Transcriptomics and plant hormone analysis reveal the mechanism of branching angle formation in tea plants (Camellia sinensis). Int J Mol Sci. 2025;26(2):604. https://doi.org/10.3390/ijms26020604.
16. Li 16CL, Wang M, Han Y, Hao Q, Qiao Y. Transcriptome-based variations effectively untangling the intraspecific relationships and selection signals in Xinyang Maojian tea population. Front Plant Sci. 2023;14:1114284. https://doi.org/10.3389/fpls.2023.1114284.
17. Zhang 17FK, Wang J, Qian M, Sun W, Shen L. Development and application of SNP-KASP markers based on genes related to nitrogen uptake, assimilation and allocation in tea plant (Camellia sinensis L). Agronomy. 2022;12(10):2534. https://doi.org/10.3390/agronomy12102534.
18. Massawe 18LN, Mayes F, Gorjanc S, Bančič G. Genomic selection strategies to increase genetic gain in tea breeding programs. Plant Genome. 2023;16:e20282. https://doi.org/10.1002/tpg2.20282.
19. Wambulwa 19MMK, Li MC, Thapa MM, Sun KK, Möller YS. Domestication origin and breeding history of the tea plant (Camellia sinensis) in China and India based on nuclear microsatellites and cpDNA sequence data. Front Plant Sci. 2018;8:2270. 10.3389/fpls.2017.02270.
20. Matsuda 20KN, Yanagida T, Hotta C, Mimura Y, Kanda Y. Parentage analysis of tea cultivars in Japan based on simple sequence repeat markers. Breed Sci. 2021;71(5):594–600. https://doi.org/10.1270/jsbbs.20156.
21. Meinhardt 21FWP, Tan LW, Zhou HW, Mischke L, Zhang S. Varietal identification of tea (Camellia sinensis) using nanofluidic array of single nucleotide polymorphism (SNP) markers. Hortic Res. 2014;1:14035. https://doi.org/10.1038/hortres.2014.35.
22. Tong 22XEH, Wu W, Wei Q, Zhao S, Zhang J. Tea plant genomics: achievements, challenges and perspectives. Hortic Res. 2020;7:7. https://doi.org/10.1038/s41438-019-0225-4.
23. Xun 23WL, Aktar H, Zhang S, Wu R, Ni L. Development of SNP markers for original analysis and germplasm identification in Camellia sinensis. Plants. 2023;12(1):162. https://doi.org/10.3390/plants12010162.
24. Wei 24ZD, Zhang X. Variation analysis and quantitative trait loci mapping of 16 free amino acid traits in the tea plant (Camellia sinensis). BMC Plant Biol. 2025;25(1):194. https://doi.org/10.1186/s12870-024-06038-9.

25. Standardization Administration of China. Tea—Determination of water extracts content (GB/T 8305 – 2013). Beijing: Standards Press of China; 2013.
26. Standardization Administration of China. Determination of tea polyphenols and catechins content in tea (GB/T 8313 – 2018). Beijing: Standards Press of China; 2018.
27. Standardization Administration of China. Tea—Determination of caffeine content (GB/T 8312 – 2013). Beijing: Standards Press of China; 2013.
28. Standardization Administration of China. Tea—Determination of free amino acids content (GB/T 8314 – 2013. Beijing: Standards Press of China; 2013.
29. Zhou 29CS, Chen Y, Gu Y. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
30. Langmead 30KD, Salzberg B. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
31. Pertea 31PM, Antonescu GM, Chang CM, Mendell TC, Salzberg JT. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
32. Smouse 32PR. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics. 2012;28(19):2537–9. https://doi.org/10.1093/bioinformatics/bts460.
33. Smouse 33PR. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes. 2006;6(1):288–95. https://doi.org/10.1111/j.1471-8286.2005.01155.x.
34. Lv 34YJH, Liu YQ, Jin SR, Wang J, Wei YF. Effects of light intensity and spectral composition on the transcriptome profiles of leaves in shade grown tea plants (Camellia sinensis L.) and regulatory network of flavonoid biosynthesis. Molecules. 2021;26(19):5836. https://doi.org/10.3390/molecules26195836.
35. Li 35YS, Zhao P, Tan X, Ahmad M, Xu MZ. CsTCPs regulate shoot tip development and catechin biosynthesis in tea plant (Camellia sinensis). Hortic Res. 2021;8:104. https://doi.org/10.1038/s41438-021-00538-7.
36. Huang 36SX, Li X, Lin Q, Bai H, Zhu S. The WRKY transcription factor CsWRKY70 regulates EGCG biosynthesis by affecting CsLAR and CsUGT84A expressions in tea leaves (Camellia sinensis). Horticulturae. 2023;9(1):120. https://doi.org/10.3390/horticulturae9010120.
37. Liu 37JX, Liu N, Gao Y, Xia L. Exploration of the biosynthesis of galloylated catechins in tea plants. Beverage Plant Res. 2025;5:e023. https://doi.org/10.48130/bpr-0025-0009.
38. Xia 38LP, Fu E, Xu J, Zhao Y, Tong X. Diverse roles of MYB transcription factors in regulating secondary metabolite biosynthesis, shoot development, and stress responses in tea plants (Camellia sinensis). Plant J. 2022;110(4):1144–65. https://doi.org/10.1111/tpj.15729.
39. Tang 39WL, Zhang X, Xie S, Li X, Liu M. Tea GOLDEN2-LIKE genes enhance catechin biosynthesis through activating R2R3-MYB transcription factor. Hortic Res. 2022;9:uhac117. https://doi.org/10.1093/hr/uhac117.

40. Zhu 40LY, Li C, Zeng J, Qian L. MYB transcription factors regulate O-methylated EGCG biosynthesis in different leaf positions of tea plants (Camellia sinensis). Ind Crops Prod. 2025;230:121079. https://doi.org/10.1016/j.indcrop.2025.121079.
41. Yang 41QD, Chen C, Guo J, Li Y, Niu Y. Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (Camellia sinensis). Sci Rep. 2019;9:2709. https://doi.org/10.1038/s41598-019-39250-2.
42. Xia 42FK, Li Z, Jiang H, Qin X, Wang D. Genome-wide association analysis identified molecular markers associated with important tea flavor-related metabolites. Hortic Res. 2021;8:42. https://doi.org/10.1038/s41438-021-00477-3.
43. Liu 43ZQ, Lu L, Du X, Xiang X, Cheng P. The biosynthesis of EGCG, theanine and caffeine in response to temperature is mediated by hormone signal transduction factors in tea plant (Camellia sinensis L.). Front. Plant Sci. 2023;14:1142637. https://doi.org/10.3389/fpls.2023.1142637.
44. Jiao 44DL, Chen Y, Gladman J, Chopra N, Burow R. Sorghum MSD3 encodes an ω-3 fatty acid desaturase that increases grain number by reducing jasmonic acid levels. Int J Mol Sci. 2019;20(21):5359. https://doi.org/10.3390/ijms20215359.
45. Schreiber 45JM, Stein M, Mascher N. Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res. 2021;28(1):dsaa030. https://doi.org/10.1093/dnares/dsaa030.
46. Li 46LL, Liu X, Zhao F, Zhang J, Zheng Y. Preliminary investigation of essentially derived variety of tea tree and development of SNP markers. Plants. 2023;12(8):1643. https://doi.org/10.3390/plants12081643.
47. Yu 47FXJ, Cai WT, Lin CP, Wang Y, Fang ZH. Construction of molecular ID for tea cultivars using single-nucleotide polymorphism (SNP) markers. Sci Agric Sin. 2021;54(8):1751–60. https://doi.org/10.3864/j.issn.0578-1752.2021.08.014.
48. Huang 48MJQ, Ma L, Jin CL, Li JQ, Wang CF. Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). PLoS ONE. 2015;10(6):e0128798. https://doi.org/10.1371/journal.pone.0128798.

Figures

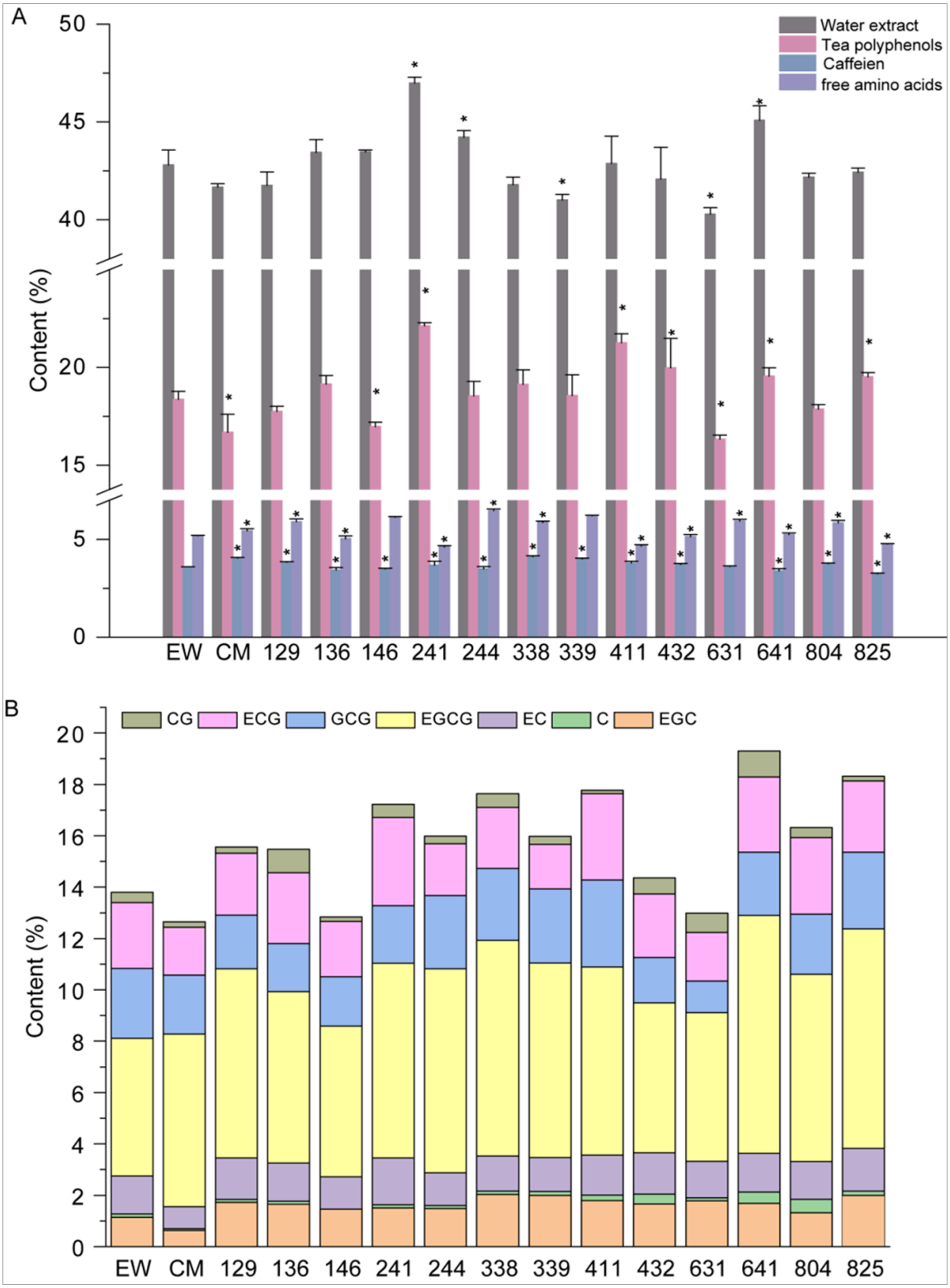


Figure 1

Phenotypic variation of primary biochemical components and catechins across elite tea progenies and parental lines

(A) Quantification of water-soluble extracts, total polyphenols, caffeine, and total free amino acids. (B) Content analysis of main catechin components. Data are presented as the mean±standard deviation

(SD) of three biological replicates (n=3). Asterisks indicate statistically significant differences between the progeny lines and the maternal parent 'Emei Wenchun' (EW) according to one-way ANOVA followed by LSD post-hoc test (*p < 0.05).

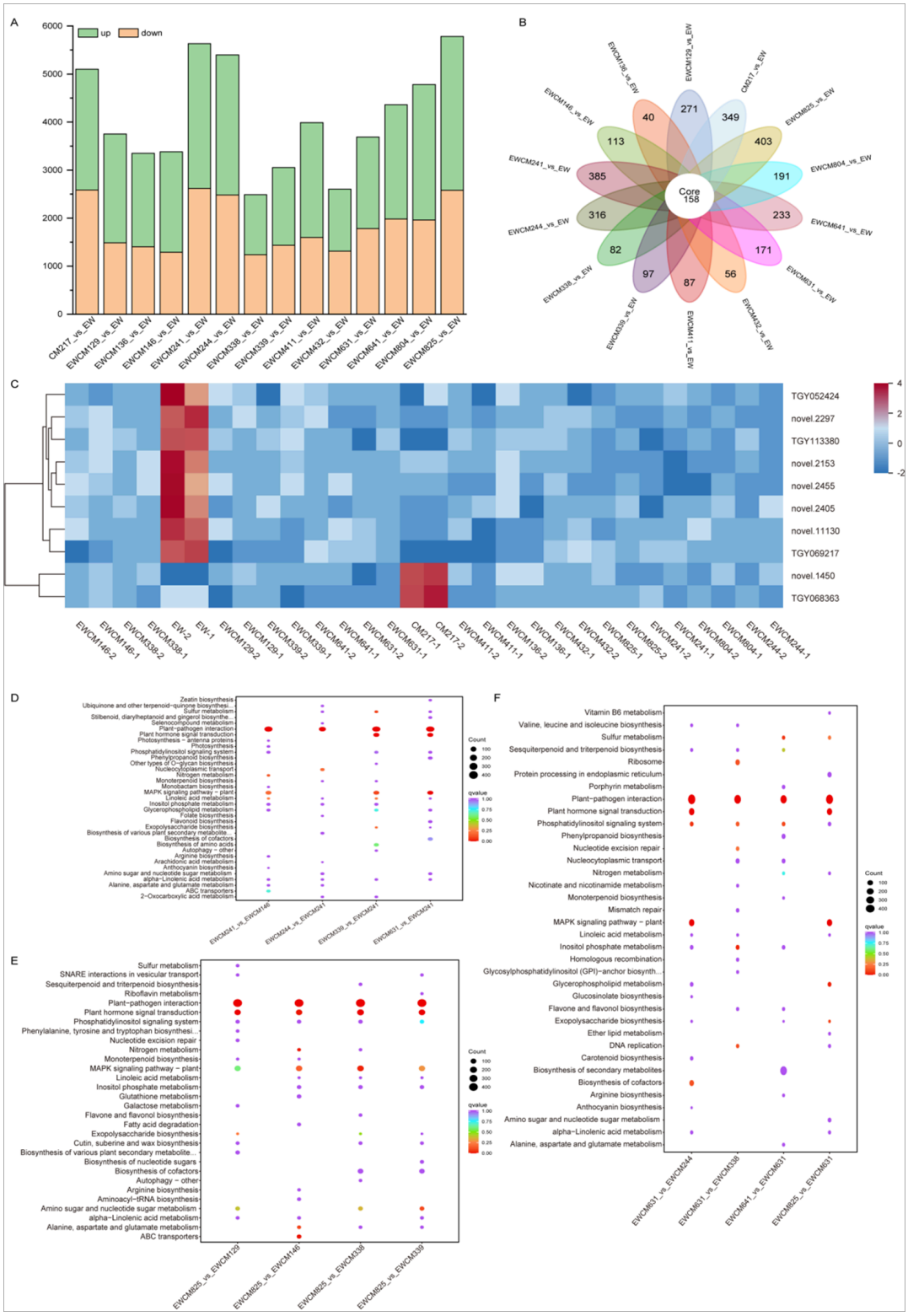


Figure 2

Global transcriptomic reprogramming and pathway enrichment in elite clonal lines

(A) Numbers of differentially expressed genes (DEGs) identified in pairwise comparisons against the maternal parent (EW). (B) Venn diagram illustrating the overlap of DEGs among the 13 progeny lines and CM217. (C) Hierarchical clustering heatmap of the shared core DEGs across all accessions. (D–F) KEGG pathway enrichment analysis of DEGs associated with contrasting phenotypic extremes for (D) total free amino acids, (E) caffeine, and (F) EGCG. Bubble size corresponds to the number of genes, and color gradients represent the adjusted enrichment significance (adjusted q-value).

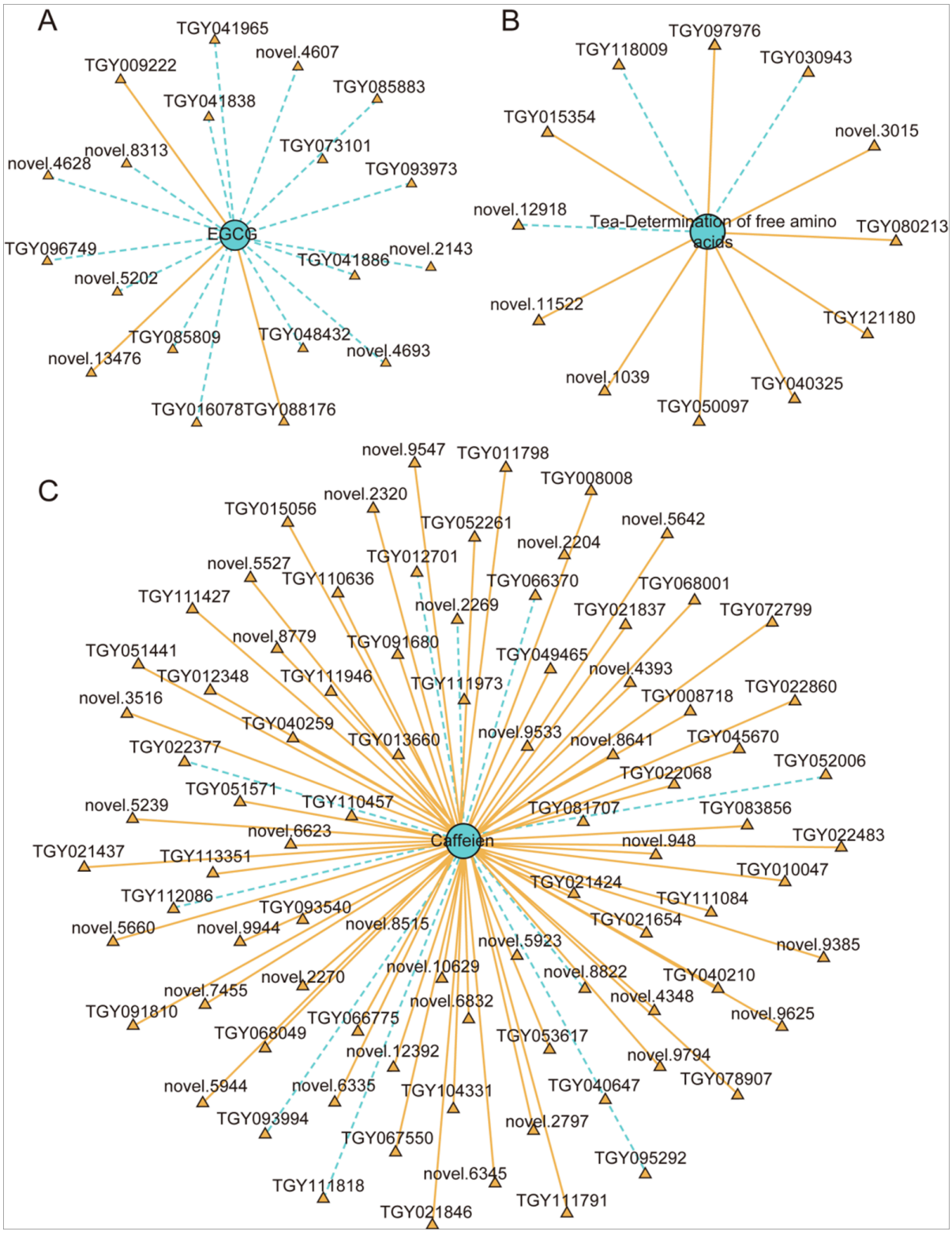


Figure 3

Co-expression networks connecting structural/regulatory transcripts to key quality metabolites

Pearson correlation networks were constructed for (A) EGCG, (B) total free amino acids, and (C) caffeine. Circular nodes denote the target biochemical components, while triangular nodes represent significantly correlated differentially expressed genes. Solid yellow lines indicate positive correlations, whereas dashed blue lines indicate negative correlations.

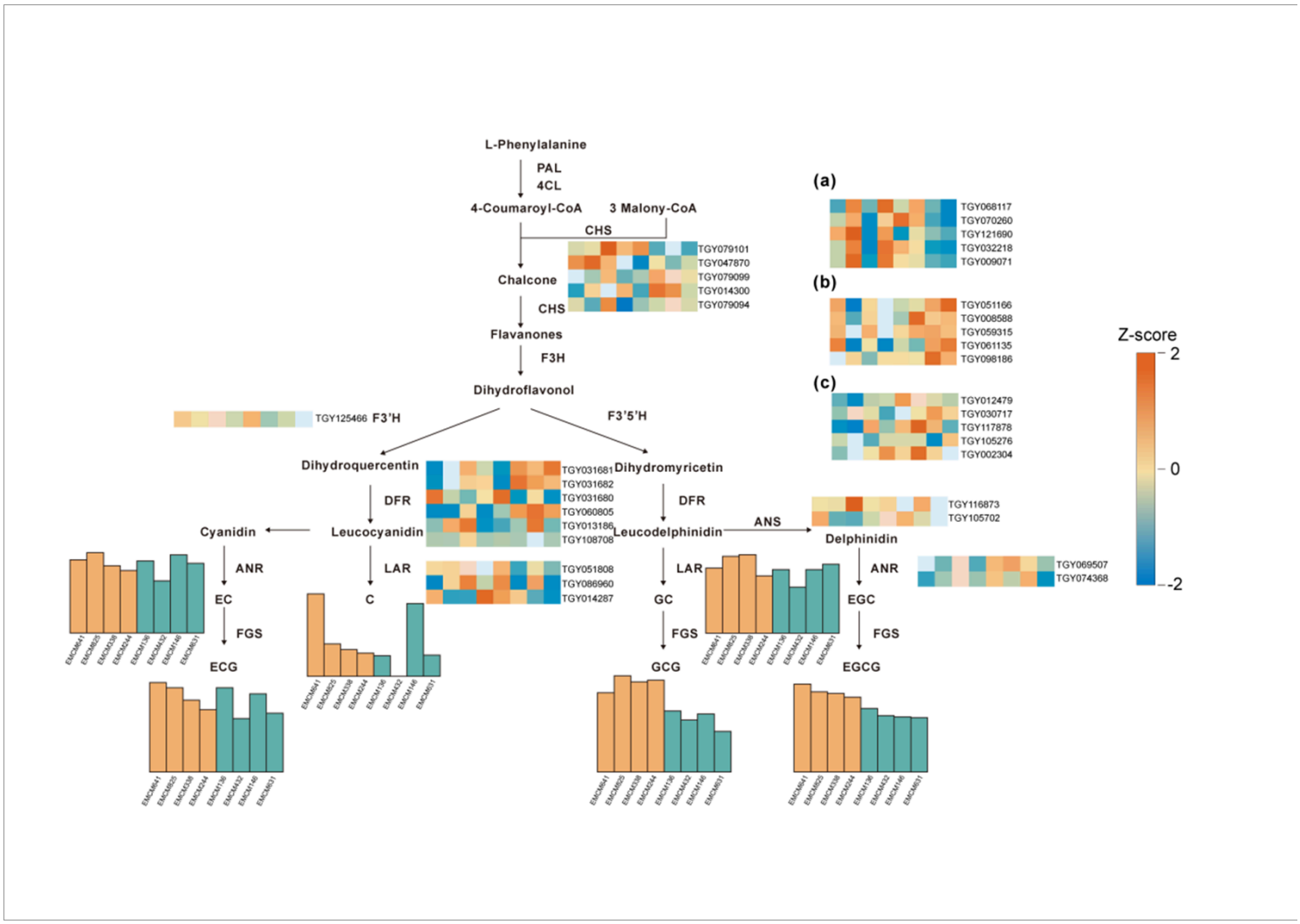


Figure 4

Integrated transcriptional and metabolic flux mapping of the flavonoid biosynthetic pathway

The heatmap visualizes the scaled expression profiles (row Z-score) of key structural genes and associated transcription factors across eight contrasting clonal lines. Color intensity ranges from blue (low expression) to red (high expression). The underlying bar chart presents the corresponding EGCG content (mg/g) for each line, strictly aligned with the respective heatmap columns. Annotations (a, positive), (b, negative), and (c, no correlation) denote the statistical association between specific transcription factors and downstream EGCG accumulation.

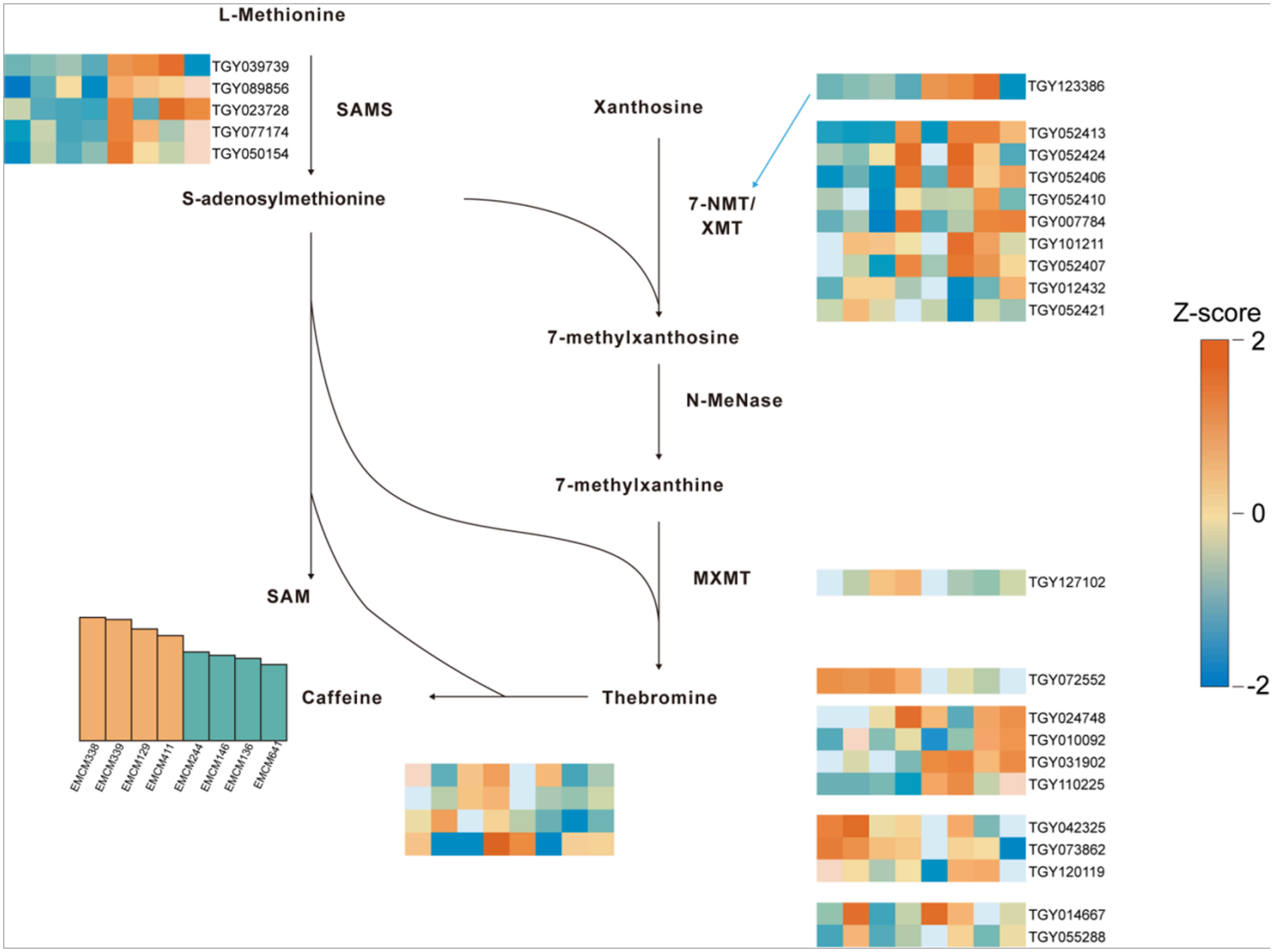


Figure 5

Integrated transcriptional mapping of the caffeine biosynthetic pathway

(A) Heatmap visualizing the expression profiles of core genes and transcription factors across eight clonal lines (blue = lower expression; red = higher expression). Columns are arranged in descending order of caffeine content. (B) Corresponding caffeine content bar chart, with blue arrows indicating pathway-associated transcription factors.

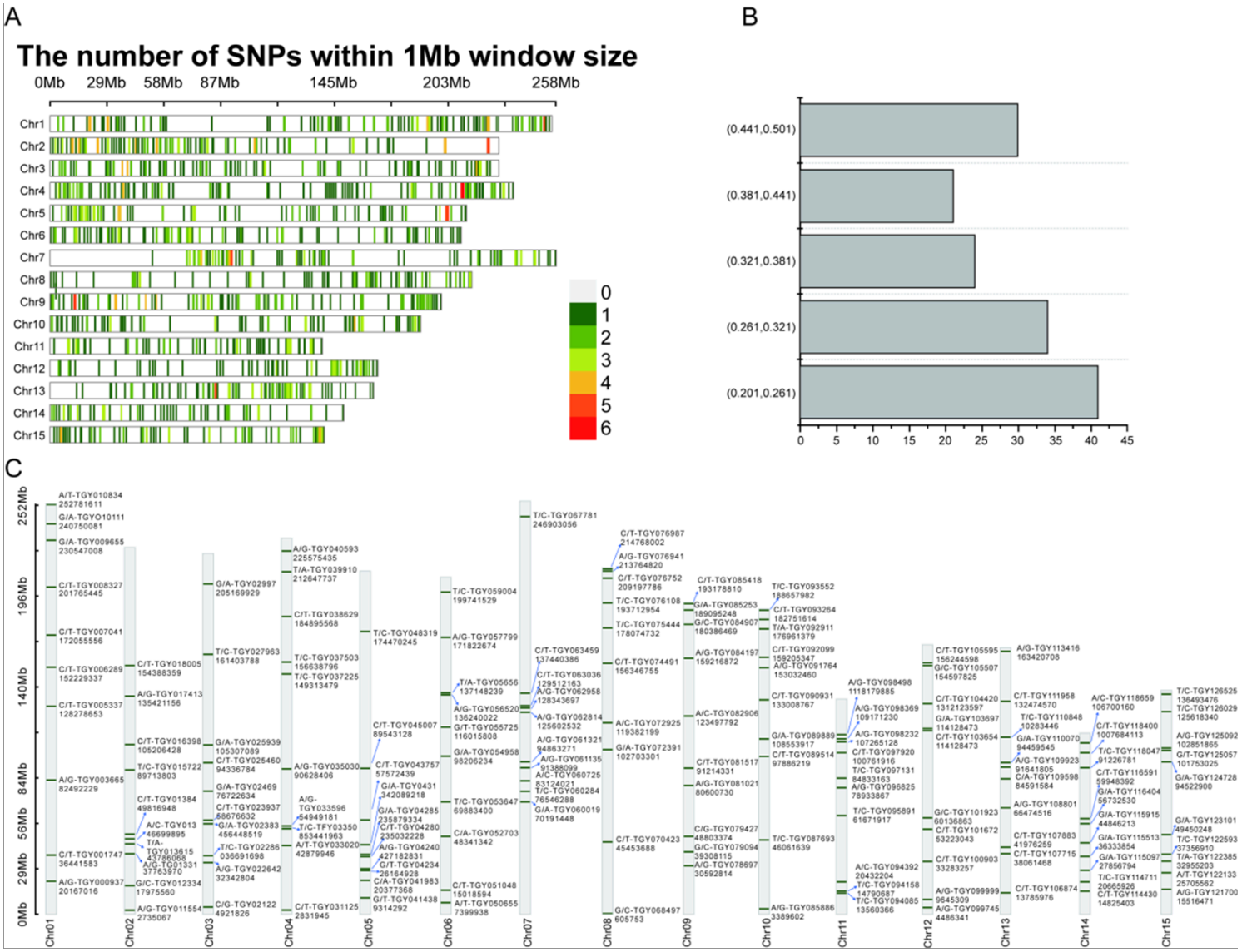


Figure 6

Overview of SNP distribution and genotypic fingerprints

(A) Chromosomal density mapping of the 1,470 core synonymous SNP markers curated from highly expressed transcripts, mapped against the reference genome. Color gradients indicate SNP density per 1-Mb sliding window. (B) Frequency distribution of minor alleles (MAF) for the selected 150 elite SNPs. (C) An example of the high-resolution DNA fingerprint barcode generated for the maternal line (EW), utilizing the 150-SNP multiplex panel to establish unique genotypic identification.

Supplementary Files

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

- SupportingTables.xlsx
- SuppFig.docx