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.

Multi-omics analysis of the co-expression features of specific neighboring gene pairs suggests an association with catechin regulation in Camellia sinensis

Shuaibin Lian
Xinyang Normal University
Feixiang Ren
Xinyang Normal University
Shuanghui Cai
Xinyang Normal University
Zhong Wang
Xinyang Normal University
Youchao Tu
Xinyang Normal University
Ke Gong
Xinyang Normal University
Wei Zhang
chawenhua2009@163.com
Xinyang Normal University

Research Article

Keywords: Neighboring gene pairs, epigenetic regulation, intergenic distance, Hi-C, transcription factors, metabolic regulation

Posted Date: August 19th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7249588/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

Background:

The arrangement and positioning of genes on chromosomes are non-random in plant genomes. Adjacent gene pairs often exhibit similar co-expression patterns and regulatory mechanisms. However, the genomic and epigenetic features influencing such co-expression, particularly in perennial crops like tea (*Camellia sinensis*), remain largely uncharacterized.

Results:

Firstly, we identified 771 specific neighboring gene pairs (SNGs) in *Camellia sinensis* (YK10) and investigated the contributions of intergenic distance and gene length to SNGs co-expression. Results indicated that intergenic distance was significantly negatively correlated with co-expression strength, while gene length showed a positive correlation. Furthermore, these two features exerted synergistic effects with threshold characteristics. Secondly, we integrated multi-omics data including transcriptome, ATAC-seq, Hi-C and histone modification data to explore the factors influencing their co-expression and functional significance and found that SNGs marked by either ATAC-seq or H3K27ac peaks displayed significantly higher expression levels, suggesting that epigenetic regulation promotes co-expression. Thirdly, we employed logistic regression models to individually assess the contributions of nine factors—ATAC-seq, H3K27ac, Hi-C, GO, distance, length, promoter, enhancer, and expression level—to the co-expression of SNGs. Finally, by integrating co-expression networks with metabolic profiles, several transcription factors potentially involved in the regulation of catechin metabolic pathways were identified.

Conclusions:

Collectively, this study reveals a multilayered regulatory framework governing SNG co-expression and provides theoretical insights and candidate regulators for understanding metabolic regulation in tea plants.

Background

Mounting evidence suggests that gene organization in eukaryotic chromosomes follows non-random patterns[1]. Genes with similar expression profiles often cluster within the same genomic regions[2, 3], a pattern observed across numerous plant and animal species[4, 5]. For instance, in the mouse genome, genes associated with immune responses and essential survival functions are frequently organized into clusters[6]. Several mechanisms have been proposed to explain the co-expression of adjacent genes, including the presence of shared promoter elements[7], and coordinated histone modifications[8]. Notably, co-expression can persist even after gene pairs become physically separated during

evolution[9], likely due to their continued spatial proximity in the three-dimensional nuclear environment[10]. The higher-order folding of chromosomes facilitates the physical closeness of distant chromatin regions, enabling potential regulatory interactions among the genes located within them[11]. Recent studies have demonstrated that such spatial colocalization is functionally associated with transcriptional regulation[10, 12]. Based on these findings, gene pair colocalization can be broadly classified into three types: spatially colocalized, physically neighboring, and evolutionarily neighboring[13].

The leaves of *Camellia sinensis* serve as the primary raw material for various tea beverages, which represent the most widely consumed non-alcoholic drinks globally[14]. In China, the use of tea dates back approximately 3,700 years[15]. Tea leaves are rich in a variety of secondary metabolites, among which catechins—belonging to the group of tea polyphenols—are particularly prominent, typically accounting for 18–36% of the dry weight of tea leaves. Catechins not only contribute to the distinctive and pleasant flavor of tea but also exhibit multiple health-promoting properties, including antioxidant, anti-inflammatory, and free radical-scavenging activities[16, 17]. Additionally, tea is abundant in other functional compounds such as caffeine and theanine, which impart bitterness and umami sweetness, respectively[18, 19]. To date, nearly 4,000 bioactive compounds have been identified in tea leaves. The astringency of tea is primarily attributed to flavonoids, while its bitterness is mainly due to caffeine[20, 21].

In recent years, high-quality chromosomal assemblies have been achieved for several tea cultivars, including *Camellia sinensis* var. *assamica* (CSA, YK10), *Camellia sinensis* var. *sinensis* (CSS, Shuchazao, Biyun, Longjing 43, Tieguanyin, Huangdan), and ancient tea trees (DASZ) [22, 23]. These genomic resources have facilitated the exploration of 3D chromatin architecture in tea[24]. The three-dimensional organization of chromatin is not only essential for DNA replication and chromosome recombination, but also profoundly affects gene expression through modulating chromatin accessibility and regulatory element distribution, ultimately contributing to phenotypic variation[25]. Epigenetic regulation of gene expression is widespread across animals, plants, and fungi, playing crucial roles in processes such as development, disease, and environmental response[26, 27]. These epigenetic modifications remodel DNA–protein interactions within chromatin, resulting in transcriptional states that are active, poised, or silenced. Consequently, they influence DNA accessibility and nuclear positioning[28]. For instance, acetylation of histone H3 at lysine 27 (H3K27ac) is frequently associated with transcriptional activation[29]. Precise temporal and spatial regulation of transcription is vital for complex biological processes such as cell differentiation and response to environmental cues. This regulation is mediated by interactions between transcription factors (TFs) and cis-regulatory elements (CREs)[30, 31]. Deciphering CREs is fundamental to understanding transcriptional networks underlying tissue specificity and phenotypic diversity. Active CRE regions often exhibit open chromatin conformations, allowing access to regulatory proteins.

In this study, we identified 771 species-specific neighboring gene pairs in tea (*Camellia sinensis*), which are located adjacently in the tea genome but not in the genomes of other species. To explore the

biological significance of these tea-specific gene arrangements, we integrated multiple omics-based features that potentially influence gene expression and co-expression. We addressed the following key scientific questions: (1) What are the transcriptional expression patterns of these species-specific neighboring gene pairs? Are they associated with the biosynthesis of tea flavor-related metabolites? (2) Why do specific neighboring gene pairs tend to exhibit co-expression patterns? Which factors contribute to their co-expression features? To address these questions, we focused on the *Camellia sinensis* cultivar Yunkang-10 (YK10) as the primary reference genome, and conducted a comprehensive genomic analysis by integrating nine factors, including ATAC, H3K27ac, Hi-C, GO, distance, length, promoter, enhancer, and expression level, to investigate the co-expression features of specific neighboring gene pairs. The information of 11 other plant species and their phylogenetic relationships were shown in Fig. 1a.

Results

Identification and expression patterns of specific neighboring gene pairs (SNGs)

First, based on the phylogenetic relationships with *Camellia sinensis*, we selected 11 representative plant species to construct a phylogenetic tree (Fig. 1a) (see Methods for details). In this study, we identified a specific category of gene pairs with unusual genomic positioning—those that are not adjacent in other species but appear as neighboring genes in the tea genome (Fig. 1b). We define such gene pairs as specific neighboring gene pairs (SNGs), referring to genes which were separated in the evolutionary past but are now neighbors. To identify these SNGs, we first determined the sets of orthologous gene families between C. sinensis and each of the 11 selected species. For each orthologous group, we first examined whether the gene pair from *Camellia sinensis* is located adjacent to one another on the genome. If the two tea genes are adjacent, we then assessed whether their corresponding orthologs in the other species are also neighbors. If not, the pair is classified as an SNG. Subsequently, we intersected the SNGs identified across all 11 species and ultimately obtained a total of 771 SNGs. These represent gene pairs that are consistently adjacent only in *C. sinensis* but not in any of the other 11 species (Fig. 1b,1c) (see Methods for details).

To investigate the expression patterns and functional characteristics of the 771 identified SNGs during natural growth and development, we first performed expression clustering analysis. Based on their transcriptional profiles, SNGs were grouped into four distinct clusters (Fig. 1d). Notably, the majority of SNGs (272 pairs) exhibited coordinated high expression within the same cluster, indicating potential co-regulation. To further explore the biological roles of SNGs, we conducted Gene Ontology (GO) enrichment analysis for each cluster. Genes in cluster 1 showed upregulation of various metabolism-related processes, particularly enriched in amide metabolic process, peptide metabolic process, and carbohydrate derivative metabolic process. These genes were also significantly associated with photosynthesis, a core process of carbon metabolism. Functional annotations indicated that most genes

in cluster 1 encode proteins involved in nitrogen compound metabolism and photosynthetic activity, with additional enrichment for ion binding and mRNA binding, suggesting dual roles in basic metabolism and post-transcriptional regulation. Cluster 2 was enriched in genes involved in the organic hydroxy compound metabolic process and those related to metal cluster binding. These genes are potentially involved in redox-related metabolic pathways and play critical roles in cell development and differentiation, reflecting the plant's strategy of coordinating growth and stress responses through metabolic reprogramming. Genes in cluster 3 were significantly enriched in secondary metabolism-related pathways, including heterocycle metabolic process, aromatic compound metabolic process, and organophosphate metabolic process. This pattern suggests that genes in this module are likely involved in defense responses and the biosynthesis of quality-related secondary metabolites. In contrast, cluster 4 exhibited the most diverse metabolic annotation. Genes in this group were involved in a wide range of pathways, including cellular metabolic process, primary metabolic process, nitrogen compound metabolic process, and fatty acid metabolic process. This cluster also contained a substantial number of genes encoding catalytic proteins such as hydrolases acting on ester bonds and phosphatases, indicating a central role in macromolecule degradation and membrane-bound organelle function.

Taken together, the four clusters displayed a trend of functional complementarity in metabolic regulation. Specifically, cluster 4 may represent a key regulatory module for primary metabolism, while cluster 3 is more likely to be involved in the regulation of specialized secondary metabolism. This functional divergence underscores the complexity and diversity of metabolic strategies employed by tea plants in response to environmental cues and further highlights the role of SNG expression patterns in enriching gene functional diversity.

Genomic distance is negatively correlated with SNGs co-expression

In this study, we first calculated the Pearson correlation coefficient and intergenic distance for each pair of specific neighbor genes (SNGs). The study found a negative correlation between the Pearson correlation coefficients of SNGs and their intergenic distances; specifically, the Pearson correlation between SNGs decreases as the distance between them increases **(Additional file 1:Fig. S1a).** Next, gene pairs with Pearson correlation coefficients greater than 0.5 were defined as co-expressed SNGs (co-SNGs), while the remaining pairs were classified as not co-expressed SNGs (not co-SNGs). Comparing the distance distributions of these two groups showed that co-SNGs have significantly shorter intergenic distances than not co-SNGs **(**Fig. 2a**)**, suggesting that shorter distance may promote co-expression.

To further investigate the effect of distance, all SNGs were divided into three groups based on intergenic distance: short-distance group (distance < 8 kb, n = 212), medium-distance group (distance between 8 kb and 50 kb, n = 264), and long-distance group (distance > 50 kb, n = 295). The co-expression proportions were highest in the short-distance group (35.38%), followed by the medium-distance group (23.11%), and lowest in the long-distance group (22.03%) **(**Fig. 2b**).** The co-expression proportion in the short-distance group was approximately 53.09% and 60.50% higher than that in the medium- and long-distance groups, respectively, with the differences being statistically significant (Mann-Whitney U test, $P < 0.01$). This

indicates that shorter intergenic distances significantly strengthen co-expression levels. To ensure the observed phenomenon was not due to chance, we conducted randomization experiments by randomly selecting the same number of gene pairs as the SNGs and calculating their co-expression proportions, repeating the sampling 10,000 times. The results showed that the co-expression proportion of real SNGs was significantly higher than that of the randomized groups (Mann-Whitney U test, P < 2.2e-16, Fig. 2c), further validating the negative correlation between distance and co-expression.

In summary, our results clearly demonstrate that physical distance between special neighbor genes is significantly negatively correlated with their co-expression level: the shorter the distance, the higher the likelihood of co-expression.

Gene length is positively correlated with SNGs co-expression

To further explore the factors influencing the co-expression of SNGs, we evaluated whether the average gene length of each SNG pair is associated with their expression correlation. Specifically, we calculated the Pearson correlation coefficient between the expression levels of each gene pair and examined its relationship with their average gene length. The analysis revealed a positive correlation between average gene length and Pearson coefficient, indicating that SNGs with longer average lengths tend to exhibit stronger expression correlation **(Additional file 1:Fig. S1b).**

We subsequently observed that co-expression SNGs (co-SNGs) tended to have longer average gene lengths, with their density distribution curve clearly shifted toward longer lengths compared to non-co-expression SNGs (not co-SNGs) **(**Fig. 3a**).** To further investigate the influence of average gene length on SNG co-expression, all SNGs were grouped into three categories based on their average gene length: short-length group (length < 5.5 kb, n = 216), medium-length group (length between 5.5kb and 9 kb, n = 263), and long-length group (length > 9 kb, n = 292). The analysis revealed a gradual increase in the co-expression ratio with increasing gene length. The long-length group exhibited a co-expression ratio of 29.11%, significantly higher than that of the medium-length group (24.71%) and the short-length group (23.61%) Specifically, the co-expression ratio in the long-length group increased by approximately 17.81% and 23.30% compared to the medium- and short-length groups, respectively. These differences were statistically significant (Mann-Whitney U test, P < 0.05, Fig. 3b), further supporting the notion that longer average gene length promotes co-expression. In addition, to determine whether this observation was due to random chance, we performed a randomization test. We randomly sampled the same number of gene pairs and calculated the co-expression ratio, repeating this process 10,000 times. The results showed that the frequency of co-expression ratios in the randomized groups was significantly lower than that observed in the actual SNGs (Mann-Whitney U test, P < 2.2e-16, Fig. 3c**).**

Gene length and genomic distance synergistically strengthen SNGs co-expression.

We also investigated whether the combination of intergenic distance and average gene length could synergistically strengthen the co-expression level of SNGs. We first selected SNGs that simultaneously

met the criteria of distance < 8 kb and average length > 9 kb. The results showed that the co-expression ratio under this combined condition reached 40.00%, which was significantly higher than that of SNGs with only short distance (distance < 8 kb, 35.38%) or only long length (length > 9 kb, 29.11%), corresponding to increases of 13.14% and 37.41%, respectively. This suggests a stronger synergistic effect of the combined factors, and that distance may have a greater impact on co-expression than length, with an enhancement of 21.54% (Mann-Whitney U test, $P < 0.05$, Fig. 3d). To further analyze the effect of different combinations of distance and length on SNG co-expression, we evaluated two additional combination groups. Besides the “short distance and long length” group which exhibited the highest co-expression, the other two groups—short distance and short length (distance < 8 kb & length < 5.5 kb) and medium distance and medium length (distance with between 8kb and 50 kb & length with between 5.5kb and 9 kb)—showed markedly lower co-expression ratios of 20.48% and 15.29%, respectively. These values were more than halved compared to the most optimal group (Mann–Whitney U test, $P < 0.05$, Fig. 3e**)**. We further compared the combinatorial effects with those of the corresponding individual factors. In the medium distance and medium length group, the co-expression ratio dropped to 20.48%, while the ratios for medium distance and medium length alone were 23.11% and 24.71%, respectively, representing reductions of 11.38% and 17.12%. Similarly, in the group with long distance (distance > 50 kb) and short length (length < 5.5 kb), the co-expression ratio was only 15.29%, significantly lower than that of long distance (22.03%) or short length (23.61%) alone, showing decreases of 31.82% and 35.24%, respectively **(Additional file 1:Fig. S1c, S1d).**

In summary, our analysis demonstrates that average gene length contributes to the enhancement of SNG co-expression, and that short intergenic distance combined with long gene length exhibits a synergistic effect. However, once either factor exceeds its effective range, the combined effect may diminish or even suppress co-expression. These findings highlight a complex interplay between intergenic distance and gene length in the regulation of SNG co-expression.

ATAC-seq and H3K27ac peaks strengthen the transcriptional expression of SNGs.

Given that non-coding intergenic regions are enriched in various transcriptional regulatory elements, a genome-wide analysis of these elements is essential for understanding their influence on transcriptional activity[32]. To this end, we processed publicly available raw datasets of ATAC-seq and H3K27ac ChIP-seq from recent studies (see Methods for details). Our analysis revealed a strong enrichment of both ATAC-seq and H3K27ac ChIP-seq signals around transcription start sites (TSSs) **(**Fig. 4a**)**, suggesting a potential role in transcriptional activation. We then identified peaks from both datasets using the MACS2 algorithm and retained those consistently detected in biological replicates **(**Fig. 4b**).** Since ATAC-seq peaks, which represent accessible chromatin regions (ACRs), often encompass promoters and enhancers[32].We further overlapped H3K27ac peaks located in intergenic regions with ATAC-seq peaks. Regions showing co-occurrence were defined as active enhancers (n = 1706) **(Additional file 2:Fig. S2,Additional file 3:Table S1).**

To investigate whether the presence of ATAC-seq or H3K27ac peaks contributes to elevated transcriptional expression of SNGs, we annotated SNGs with both ATAC-seq and H3K27ac signals (see Methods for details). We then compared expression levels between SNGs with and without peak annotations. The results showed that SNGs annotated with either ATAC-seq or H3K27ac peaks exhibited significantly higher expression, with median values of 4.57 and 4.55, respectively, compared to 3.69 and 3.45 in the non-annotated groups. This corresponds to an expression increase of 23.85% and 31.89%, respectively (Mann–Whitney U test, $P < 0.001$; Fig. 5c, 5d). We further examined whether co-occurrence of both ATAC-seq and H3K27ac peaks exerts a synergistic effect on gene expression. SNGs carrying both peaks displayed the highest expression levels, with a median of 4.62, representing a 34.30% increase compared to the non-annotated group. To exclude potential confounding effects between the two peak types, we analyzed SNGs annotated with either ATAC-seq or H3K27ac peaks alone. Both subsets showed significantly higher expression than the non-annotated group, with no significant difference between them, indicating comparable regulatory strength. Notably, SNGs annotated solely with ATAC-seq peaks showed a more concentrated expression distribution (Mann–Whitney U test, $P < 0.001$, Fig. 5e).

Collectively, these results demonstrate that annotation with ATAC-seq and/or H3K27ac peaks is associated with elevated transcriptional expression of SNGs.

Logistic regression model indicates the molecular features of SNGs co-expression

To systematically evaluate the influence of multiple molecular features on SNG co-expression, we trained a logistic regression model by integrating nine factors, including ATAC, H3K27ac, Hi-C, GO, distance, length, promoter, enhancer, and expression level, to investigate the co-expression features of specific neighboring gene pairs. We used 80% of the SNG pairs for training and evaluated its performance on the remaining 20%.

Firstly, we evaluated the relationships between these features and SNG expression correlation (measured by Pearson correlation coefficient), and further assessed their discriminative power in a logistic regression model (Fig. 5a). Preliminary analyses showed that most features exhibited weak associations with expression correlation, and pairwise correlations between features were generally insignificant. Notably, only the Hi-C spatial interaction frequency showed a stable negative correlation with intergenic distance ($r = -0.286$), indicating that gene pairs with shorter distances tend to have higher spatial interaction frequencies (**Additional file 4**:**Fig. S3**). Given that spatial proximity may influence co-expression, and that gene pairs with closer TSSs usually have stronger spatial interactions, we further examined how varying Hi-C strength affects SNG co-expression. The results showed a negative correlation between Hi-C interaction frequency and co-expression levels among SNGs ($r = -0.453$). Specifically, co-expression decreased when Hi-C strength was below 10, but significantly increased when the strength exceeded 10 (**Additional file 5**:**Fig. S4c**). Combined with the negative correlation between Hi-

C and intergenic distance, this pattern likely reflects that SNGs with high Hi-C strength tend to have shorter intergenic distances, indirectly enhancing their co-expression. In the logistic regression model, however, the average AUC for the Hi-C feature was only 0.52, indicating limited discriminative power.

To further explore the connection between spatial regulation and co-expression, we examined whether SNGs share active enhancers identified by overlapping ATAC-seq and H3K27ac ChIP-seq peaks (see Methods for details). A total of 43 SNG pairs were annotated as sharing active enhancers, but their co-expression levels were not significantly higher than those of other SNGs (**Additional file 5**:**Fig. S4d**), and their average AUC in logistic regression was also modest (0.52; Fig. 5a). In terms of functional annotation, we quantified the number of shared promoter elements and shared GO biological process (BP) terms among SNGs. We found 763 SNG pairs sharing promoter elements and 219 sharing BP terms. Both features showed a strong positive correlation with co-expression proportion, with Pearson r values of 0.546 and 0.766, respectively (**Additional file 5**:**Fig. S4a, S4b**). Their corresponding average AUCs in logistic regression were 0.58 and 0.53, suggesting moderate predictive power (Fig. 5a).

Next, we further assessed whether ATAC-seq and H3K27ac annotations also contribute to SNG co-expression. Despite their known role in boosting gene expression, the presence of either ATAC-seq or H3K27ac peaks did not significantly strengthen co-expression levels (**Additional file 5**:**Fig. S4d**). In logistic regression models, the average AUCs for SNGs annotated with ATAC-seq or H3K27ac peaks alone were only 0.47 and 0.51, respectively, indicating weak predictive power (Fig. 5a). We also examined whether expression divergence between SNGs correlates with co-expression levels. By standardizing expression differences across tissues (see Methods for details), we divided SNGs into two groups based on whether their expression difference was greater than or less than 1. SNGs with smaller expression differences (expression < 1) showed slightly higher co-expression levels, and their average AUC in the logistic regression model reached 0.59, outperforming most single features (Fig. 5a).

Finally, among all individual features, intergenic distance showed the highest average AUC in the logistic regression model (AUC = 0.73), making it the most effective predictor of SNG co-expression. When all features were combined in a single integrated model, the resulting AUC exceeded that of any individual feature, highlighting the complementary roles of different molecular characteristics in explaining local gene co-expression (Fig. 5a). Moreover, logistic regression analysis further confirmed that gene distance is positively associated with co-expression, whereas gene length shows a negative association, consistent with our previous statistical observations (**Additional file 6**:**Fig. S5a).**

To further evaluate the relative importance of various molecular features in predicting the co-expression of SNGs, we used the "*varImp*" function from the R package "*caret*" to assess each variable's contribution within the logistic regression model. The results showed that genomic distance remained the most informative feature, exhibiting the highest variable importance score. Although Hi-C alone yielded a relatively low AUC when modeled individually, its importance ranked second in the integrated model, suggesting that Hi-C-mediated spatial interactions still play a role in regulating SNG co-expression (Fig. 5b).

Interaction effects between genomic distance and other molecular features on SNG Co-expression

Building on the Logistic regression model, we further investigated whether genomic distance interacts with other molecular features to influence SNGs co-expression. If such interactions exist, it would suggest that the effect of molecular features on co-expression is modulated by the distance between gene pairs, thereby revealing a more nuanced regulatory mechanism.

First, we incorporated interaction terms between distance and other molecular features into the logistic regression model to investigate how their combined effects influence the probability of SNG co-expression. The interaction effects were visualized using the "*effects*" package in R. In the interaction between distance and gene length, we observed that at shorter distances, SNGs with greater gene length exhibited the highest predicted co-expression probability (76.77%), which gradually decreased as the distance increased—consistent with previous observations (Fig. 6a). In the interaction between distance and Hi-C contact frequency, the highest co-expression probability (78.19%) was associated with low Hi-C intensity at shorter distances. As distance increased, the predicted probabilities for low and medium Hi-C frequencies decreased, while those for high Hi-C frequency increased, reaching 53.90% at the longest distances (Fig. 6b). These results suggest that although Hi-C interactions may negatively correlate with co-expression at short distances, they may exert a positive regulatory effect when the distance between SNGs exceeds a certain threshold.

A similar pattern was observed in the interaction between distance and shared promoter elements. At very short distances, SNGs sharing the fewest promoter elements had the highest predicted co-expression probability (72.74%) (Fig. 6c). While overall co-expression probability declined with increasing distance, SNGs sharing the greatest number of promoter elements eventually surpassed the other groups at certain distance ranges. This indicates that excessive promoter sharing between closely located genes may compromise their co-expression coordination. Finally, in the interaction between distance and shared Gene Ontology biological process (GO BP) terms, SNGs located in close proximity and annotated with the same GO BP term exhibited the highest predicted co-expression probability (up to 97.89%), which gradually declined with increasing distance (Fig. 6d). Similarly, in the interaction between expression differences and distance, SNGs with minimal expression differences exhibited the highest predicted co-expression frequency at short distances, which gradually declined as the distance increased. Notably, at longer distances, SNGs marked with H3K27ac peaks showed higher predicted co-expression frequencies compared to those without H3K27ac annotation. In contrast, the interaction effects between distance and either enhancer or ATAC-seq peaks were not pronounced (**Additional file 6**:**Fig. S5b)**. Collectively, these findings highlight the complexity and diversity of regulatory mechanisms underlying SNG co-expression and demonstrate that changes in genomic distance can significantly modulate the regulatory influence of other molecular features.

Co-expression network analysis

Weighted Gene Co-expression Network Analysis (WGCNA) is a highly robust method for classifying genes via hierarchical clustering of gene co-expression networks. To explore the metabolic significance of gene modules, we correlated gene expression profiles from the shoot tissue of the YK10 cultivar with the accumulation levels of eight metabolites (**Additional file 7**:**Fig. S6a)** (see Methods for details). A total of 1,449 specific neighboring gene pairs were included in the WGCNA analysis. After merging similar modules, seven modules were identified, each comprising between 98 and 2,503 genes (**Additional file 7**:**Fig. S6b**). Among these, five modules showed significant correlations with metabolites (correlation coefficient $r \geq 0.9$, $P < 0.05$, Fig. 7a). Specifically, the MEbrown module was significantly associated with caffeine, MEgreen with theobromine, and MEturquoise, MEred, and MEblack with catechins.

Hub genes within modules are typically regarded as representative of the module's biological function. Therefore, we constructed co-expression networks based on the top 30 genes with the highest module membership (KME) values from these five modules, selecting transcription factors (TFs) as key hub genes (Fig. 7b). In the Brown module, a bHLH TF was identified, which may play a critical role in caffeine biosynthesis. In the Black, Turquoise, and Red modules, eight TFs belonging to MYB, EIL, C2H2, LBD, NAC, WRKY, and Trihelix families were identified, potentially involved in catechin biosynthesis. Collectively, these results suggest that these transcription factors may contribute to the regulation of metabolite accumulation across different tea plant cultivars.

Conclusions

In this study, we systematically identified specific neighboring gene pairs (SNGs) in tea plant and comprehensively analyzed their expression patterns and the regulatory factors underlying their co-expression. Initially, we performed clustering of SNGs based on expression levels and found that genes within distinct expression clusters were enriched in metabolism-related biological processes, indicating that SNGs play important roles in metabolic regulation in tea plants.

Furthermore, we revealed the influence of intergenic distance and average gene length on SNG co-expression patterns: (1) intergenic distance exhibited a significant negative correlation with co-expression strength, where shorter distances corresponded to stronger co-expression, and co-expression decreased markedly as distance increased; (2) average gene length was positively correlated with co-expression, with longer gene pairs generally exhibiting higher co-expression levels, while shorter lengths were associated with reduced co-expression. Based on these observations, we explored the combined effect of distance and length, discovering that they synergistically strengthen SNG co-expression within certain thresholds, but this enhancement weakens or disappears once distance or length surpass critical values, suggesting complex threshold effects and nonlinear relationships in co-expression regulation.

In addition, chromatin accessibility and histone modifications significantly promote SNG transcriptional expression. We observed that SNGs annotated with both ATAC-seq and H3K27ac peaks displayed markedly elevated transcription levels, with a stronger combined effect, implying a coordinated role of

these epigenetic marks in regulating adjacent gene expression. Logistic regression modeling incorporating multiple molecular features demonstrated that: (1) intergenic distance is the most influential factor affecting SNG co-expression; (2) SNG co-expression is regulated by a combination of molecular mechanisms; (3) Hi-C spatial interaction intensity is negatively correlated with intergenic distance, and its effect on co-expression dynamically changes with distance, further reflecting the regulatory potential of three-dimensional genome architecture on gene expression. Finally, through integration of co-expression networks and metabolite accumulation data, we identified several key transcription factors potentially involved in metabolic synthesis regulation, revealing intrinsic links between SNG co-expression and metabolic control in tea plants.

Despite uncovering multilayered regulatory mechanisms of SNG expression and co-expression, this study has limitations. Firstly, we have yet to precisely determine the critical threshold of intergenic distance at which its enhancing effect on co-expression begins to diminish, and to fully elucidate how other factors intervene in this process. Secondly, the biological basis and evolutionary significance underlying the formation of SNGs—why certain genes remain adjacent and exhibit specific expression patterns following speciation or gene duplication—require further investigation. In summary, our study advances the understanding of the regulatory mechanisms governing neighboring gene co-expression in tea plants and provides valuable insights and theoretical foundations for future functional genomics research and metabolic regulation related to tea quality traits.

Methods

Identification of special neighboring gene pairs

In this study, we used the tea plant *Camellia sinensis* (YK10) as the focal species, along with 11 additional reference species, including eight eudicots, two monocots, and the basal angiosperm *Amborella trichopoda*. A phylogenetic tree was constructed using the Tree2GD software based on these species[33]. We then used the InParanoid tool to identify orthologous gene pairs between tea and the other 11 species based on homology analysis[34]. The parameters for orthology inference were set as follows: E-value < 1e-05 and confidence score ≥ 0.3, with all other parameters kept at default settings. By analyzing the locations and neighborhood relationships of orthologous gene pairs, we identified special neighboring gene pairs **(Additional file 8:Table S2)**. In particular, neighboring gene pairs refer to gene pairs which are linear neighbors in chromosome, separated gene pairs refer to gene pairs are not linear neighbors in chromosome.

Random experiments

We used the R programming language and conducted simulations based on the complete genome of YK10. In each iteration, two genes were randomly selected and their TPM values were used to calculate the Pearson correlation coefficient. For each simulation, the number of gene pairs sampled corresponded to the number of SNGs within various distance-defined ranges. The proportion of co-expressed pairs (defined as Pearson's $r > 0.5$) was recorded for each run. This process was repeated

100,000 times, and the frequency distribution of co-expression proportions across all iterations was subsequently analyzed.

Molecular feature metrics and datasets used

The following metrics were assessed for their potential to regulate local gene co-expression.

- Distance: The genomic distance between gene pairs was calculated as the absolute difference between the start site coordinates of gene1 and gene2, based on the YK10 reference annotation (gff3).
- Length: The average length of the transcribed region (TSS to TES) of each gene in the pairs.
- Hi-C: ICE-normalized Hi-C contact frequency between the 5-kb genomic bins containing the transcription start sites (TSSs) of each gene in the pair.
- ATAC-seq and H3K27ac ChIP-seq: These two features were included in the model as binary variables: a value of 1 indicates the presence of either an ATAC-seq or H3K27ac ChIP-seq peak in the gene; otherwise, 0.
- GO term sharing: The total number of shared Gene Ontology (GO) terms in the Biological Process (BP) category between the two genes in the pair. GO term matching was based on exact GO ID correspondence. GO annotations were obtained using the eggNOG-mapper datebase (http://eggnog-mapper.embl.de/) [35].
- Expression level difference: The relative difference in expression between the two genes in a pair, calculated as the absolute difference between their average expression levels, divided by the mean expression level of the pair.
- Enhancer sharing: This variable indicates whether the two genes in the pair share the same enhancer. In the model, a value of 1 is assigned if an enhancer is shared; otherwise, the value is 0.
- Promoter: Number of shared cis-regulatory elements in the promoter regions of SNGs

Logistic regression models

In this study, logistic regression models were constructed using custom scripts in the R programming language (v3.4.4), utilizing the "*glm*" function to assess multiple molecular features potentially associated with co-expression of specific neighboring gene pairs (SNGs). Co-expression status was encoded in a binomial format: SNGs exhibiting co-expression were labeled as positive cases (value = 1), while non-co-expression SNGs were treated as negative cases (value = 0). To ensure class balance in the dataset, an equal number of negative samples were randomly selected from the non-co-expressed group to match the number of positive cases.

Each model was trained on 80% of the gene pairs, randomly sampled without replacement, maintaining an equal proportion of positive and negative examples. The remaining 20% of gene pairs were used as the test set. Model performance was evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUC), which reflects the ability of the trained model to correctly classify the

co-expression status. This entire sampling and modeling process was repeated 50 times, and the mean AUC across all iterations was reported as a robust measure of predictive accuracy.

Construction of weighted gene co-expression network

Weighted gene co-expression network analysis (WGCNA) was conducted using the WGCNA R package to explore the potential associations between the expression profiles of specific neighboring genes (SNGs) and metabolite accumulation. A total of 1449 SNGs and 8 metabolites were included to construct a signed co-expression network based on Pearson correlation coefficients. The soft-thresholding power for network construction was set to 14, the minimum module size was defined as 30, and the threshold for module merging was set at 0.25; all other parameters were maintained at their default settings. The resulting module networks were visualized using Cytoscape software (version 3.9.0).

Statistical methods

Statistical significance between two independent samples was assessed using the Mann–Whitney U test (function '*wilcox.test*' in software R). The default significance threshold was set at 0.05. As a non-parametric test, the Mann–Whitney U method is designed to compare the medians of two independent groups and does not assume a normal distribution of the data, making it particularly suitable for skewed or non-normally distributed datasets.

To robustly estimate the error bars of proportions presented in bar plots, we employed a bootstrap resampling approach. Specifically, for each group of binary observations, we performed 1000 iterations of resampling with replacement, each time generating a bootstrap sample of the same size as the original dataset. For each resampled dataset, the proportion of positive cases was calculated, and the standard deviation of these 1000 proportions was taken as the standard error. This standard error was subsequently used to construct the error bars in the plots.

Declarations

Acknowledgments

Authors thank anonymous reviewers for their comments on the manuscript.

Funding

This work was supported by Henan Province Central Leading Local Science and Technology Development Fund Project Funding (Z20231811160).

Availability of data and materials

In this study, all data used in this study were obtained from public databases. We employed the high-quality reference genome of the tea cultivar *Camellia sinensis* (YK10), obtained from the latest genome

assembly available in the TeaBase database(http://teabase.ynau.edu.cn/) [36]. The RNA-seq datasets used in this study were obtained from public repositories with Bio-Project Accession No.PRJNA381277(https://www.ncbi.nlm.nih.gov/bioproject/PRJNA381277) [37],No.PRJNA524304(https://www.ncbi.nlm.nih.gov/bioproject/PRJNA524304) [38],No.PRJCA009753(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA009753)[39]. Detailed sample information is provided in **Additional file 9**:**Table S3**. Initially, raw sequencing reads were subjected to quality assessment using FastQC, followed by filtering and trimming using fastp with default parameters[40]. The cleaned RNA-seq reads were aligned to the YK10 reference genome using HISAT2[41], and gene-level quantification was performed with featureCounts[42]. Gene expression levels were normalized using the TPM (Transcripts Per Million) method.

The ATAC-seq and H3K27ac ChIP-seq datasets used in this study were obtained from public repositories with Bio-Project Accession No.PRJCA017759 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017759)[24]. Detailed sample information is provided in **Additional file 9**:**Table S3**. Quality control for these datasets was also performed using FastQC and fastp. Cleaned reads were mapped to the YK10 genome using Bowtie2 with default parameters[43], followed by filtering of low-quality alignments (Q10) using SAMtools[44]. Duplicate reads were removed using Picard (default settings), and MACS2 [45] was employed to call peaks from both ATAC-seq and ChIP-seq data. To evaluate peak reproducibility across biological replicates, we used DeepTools[46], and the identified peaks were annotated to genes in the tea genome using BEDTools[47].

In addition, Hi-C sequencing data used in this study were from public repositories with Bio-Project Accession PRJCA017759 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017759) [24].Detailed sample information is provided in **Additional file 9**:**Table S3**.We used FastQC and fastp to quality control and trimming . The processed reads were aligned to the YK10 genome and processed with HiC-Pro [48] to generate genome-wide chromatin interaction matrices at resolutions of 5 kb, 10 kb, 20 kb, 40 kb, and 100 kb. The Hi-C contact matrices were then normalized using the ICE (Iterative Correction and Eigenvector decomposition) method[49].

Moreover, we integrated metabolite accumulation data from the shoot tissue (one bud with two leaves) of the YK10 cultivar[50]. Promoter regions were defined as the 2 kb upstream of each gene's transcription start site (TSS) and were submitted to the PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) for the identification of cis-regulatory elements[51]. Transcription factors (TFs) were predicted and annotated using the PlantTFDB v5.0 database (https://planttfdb.gao-lab.org) [52]**(Additional file 10:Table S4)**.

The genetic data of the 11 species listed in Fig. 1a, including gene annotation data, CDS sequences, and protein sequences, were downloaded from two databases, EnsemblPlants (http://plants.ensembl.org/index.html) and Phytozomev14.0 (https://phytozome-next.jgi.doe.gov).

Competing interests

The authors declare that no competing interests exist.

Ethics approval and consent to participate

Not applicable

Authors' contributions

SBL and FXR implemented the algorithms and carried out the experiments. SBL and FXR drafted the manuscript. SBL and WZ designed the study and analyzed the results. SBL, FXR, SHC, WZ contributed to data collection and analysis. YCT, KG and WZ participated in discussion. SBL and SHC contributed equally. All authors read and approved the final manuscript.

Author details

1 College of Physics and Electronic Engineering, Xinyang Normal University, Xinyang, Henan, China

2 College of Life Sciences, Xinyang Normal University, Xinyang, Henan, China

References

1. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, Myers RM. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res. 2009;19(12):2163–71.
2. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91(3):243–8.
3. Woo YH, Li WH. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. Proc Natl Acad Sci U S A. 2011;108(8):3306–11.
4. Lercher MJ, Blumenthal T, Hurst LD. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 2003;13(2):238–43.
5. Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster. Nucleic Acids Res. 2005;33(5):1435–44.
6. Hentges KE, Pollock DD, Liu B, Justice MJ. Regional variation in the density of essential genes in mice. PLoS Genet. 2007;3(5):e72.
7. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462–70.
8. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128(4):707–19.
9. Wang GZ, Chen WH, Lercher MJ. Coexpression of linked gene pairs persists long after their separation. Genome Biol Evol. 2011;3:565–70.
10. Dai Z, Xiong Y, Dai X. Neighboring genes show interchromosomal colocalization after their separation. Mol Biol Evol. 2014;31(5):1166–72.

11. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4.
12. Xie T, Yang QY, Wang XT, McLysaght A, Zhang HY. Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance. Mol Biol Evol. 2016;33(9):2368–75.
13. Lian S, Liu T, Jing S, Yuan H, Zhang Z, Cheng L. Intrachromosomal colocalization strengthens co-expression, co-modification and evolutionary conservation of neighboring genes. BMC Genomics. 2018;19(1):455.
14. Yang CS, Hong J. Prevention of chronic diseases by tea: possible mechanisms and human relevance. Annu Rev Nutr. 2013;33:161–81.
15. Meegahakumbura MK, Wambulwa MC, Li MM, Thapa KK, Sun YS, Möller M, Xu JC, Yang JB, Liu J, Liu BY, Li DZ, Gao LM. 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.
16. Chen D, Chen G, Sun Y, Zeng X, Ye H. Physiological genetics, chemical composition, health benefits and toxicology of tea (Camellia sinensis L.) flower: A review. Food Res Int. 2020;137:109584.
17. Qu Z, Liu A, Li P, Liu C, Xiao W, Huang J, Liu Z, Zhang S. Advances in physiological functions and mechanisms of (-)-epicatechin. Crit Rev Food Sci Nutr. 2021;61(2):211–33.
18. Chacko SM, Thambi PT, Kuttan R, Nishigaki I. Beneficial effects of green tea: a literature review. Chin Med. 2010;5:13.
19. Rogers PJ, Smith JE, Heatherley SV, Pleydell-Pearce CW. Time for tea: mood, blood pressure and cognitive performance effects of caffeine and theanine administered alone and together. Psychopharmacology. 2008;195(4):569–77.
20. Lin CH, Shen YA, Hung PH, Yu YB, Chen YJ. Epigallocathechin gallate, polyphenol present in green tea, inhibits stem-like characteristics and epithelial-mesenchymal transition in nasopharyngeal cancer cell lines. BMC Complement Altern Med. 2012;12:201.
21. Ravichandran R, Parthiban R. Lipid occurrence, distribution and degradation to flavour volatiles during tea processing. Food Chem. 2000;68(1):7–13.
22. Xia E, Tong W, Hou Y, An Y, Chen L, Wu Q, et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. Mol Plant. 2020;13(7):1013–26.
23. Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, et al. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci U S A. 2018;115(18):E4151–8.
24. Kong W, Yu J, Yang J, Zhang Y, Zhang X. The high-resolution three-dimensional (3D) chromatin map of the tea plant (*Camellia sinensis*). Hortic Res. 2023;10(10):uhad179.
25. Zhang X, Pandey MK, Wang J, Zhao K, Ma X, Li Z, Zhao K, Gong F, Guo B, Varshney RK, Yin D. Chromatin spatial organization of wild type and mutant peanuts reveals high-resolution genomic

architecture and interaction alterations. Genome Biol. 2021;22(1):315.
26. Friedman N, Rando OJ. Epigenomics and the structure of the living genome. Genome Res. 2015;25(10):1482–90.
27. Freitag M. Histone Methylation by SET Domain Proteins in Fungi. Annu Rev Microbiol. 2017;71:413–39.
28. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20(4):207–20.
29. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33(5):510–7.
30. Lim DA. Transcriptional and epigenetic insights from stem cells and developing tissues. Development. 2015;142(15):2549–53.
31. Ohama N, Sato H, Shinozaki K, Yamaguchi-Shinozaki K. Transcriptional Regulatory Network of Plant Heat Stress Response. Trends Plant Sci. 2017;22(1):53–65.
32. Liao Y, Wang J, Zhu Z, Liu Y, Chen J, Zhou Y, Liu F, Lei J, Gaut BS, Cao B, Emerson JJ, Chen C. The 3D architecture of the pepper genome and its relationship to function and evolution. Nat Commun. 2022;13(1):3479.
33. Chen D, Zhang T, Chen Y, Ma H, Qi J. Tree2GD: a phylogenomic method to detect large-scale gene duplication events. Bioinformatics. 2022;38(23):5317–21.
34. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 2005;33:D476–80.
35. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021;38(12):5825–9.
36. Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, et al. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Mol Plant. 2017;10(6):866–77.
37. Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, et al. The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. Mol Plant. 2017;10(6):866–77.
38. Liu ZW, Shi XY, Duan SM, Nian B, Chen LJ, Zhang GH, Lv CY, Ma Y, Zhao M. Multiomics analysis of the mechanisms behind flavonoid differences between purple and green tender shoots of Camellia sinensis var. assamica. G3. (Bethesda). 2023;13(2):jkac297.
39. Ting GA, Zheng Y, Shan J. Differential metabolites and their transcriptional regulation in seven major tea cultivars (Camellia sinensis) in China. J Integr Agric. 2023;22(11):3346–63.
40. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.

42. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.
43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
44. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
45. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
46. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
48. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.
49. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999–1003.
50. Wu W, Lu M, Peng J, Lv H, Shi J, Zhang S, Liu Z, Duan J, Chen D, Dai W, Lin Z. Nontargeted and targeted metabolomics analysis provides novel insight into nonvolatile metabolites in Jianghua Kucha tea germplasm (*Camellia sinensis* var. *Assamica* cv. Jianghua). Food Chem X. 2022;13:100270.
51. Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002;30(1):325–7.
52. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45(D1):D1040–5.

Figures

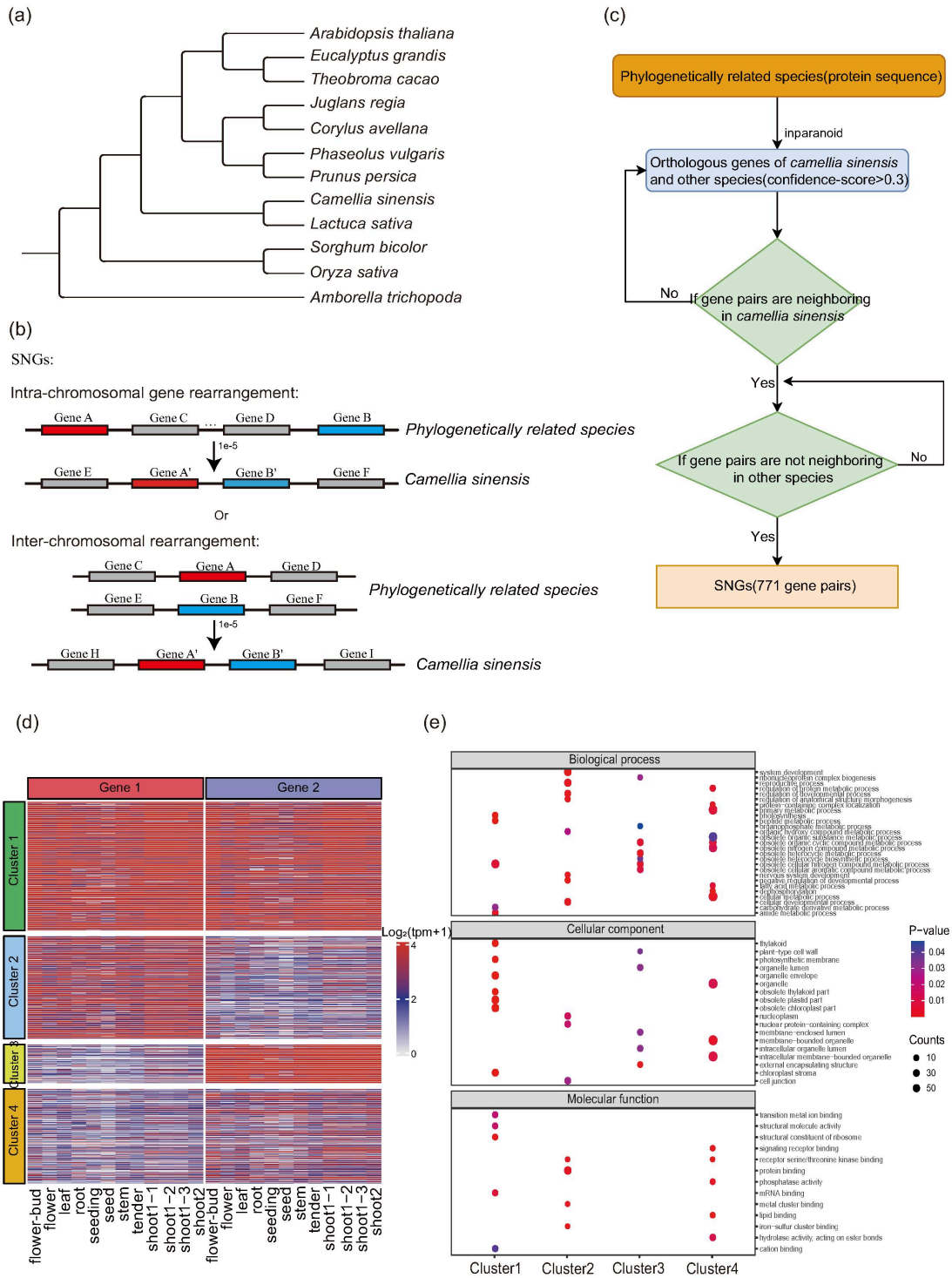

Figure 1

Identification and expression patterns of specific neighboring gene pairs (SNGs). **(a)** A phylogeny of species included in this study. **(b)** Schematic representation of specific neighboring gene pairs (SNGs) **(c)** Workflow of SNG identification **(d)**Heat map of differential expression patterns of SNGs across different developmental stages and tissues **(d)** Functional enrichment analysis of each cluster.

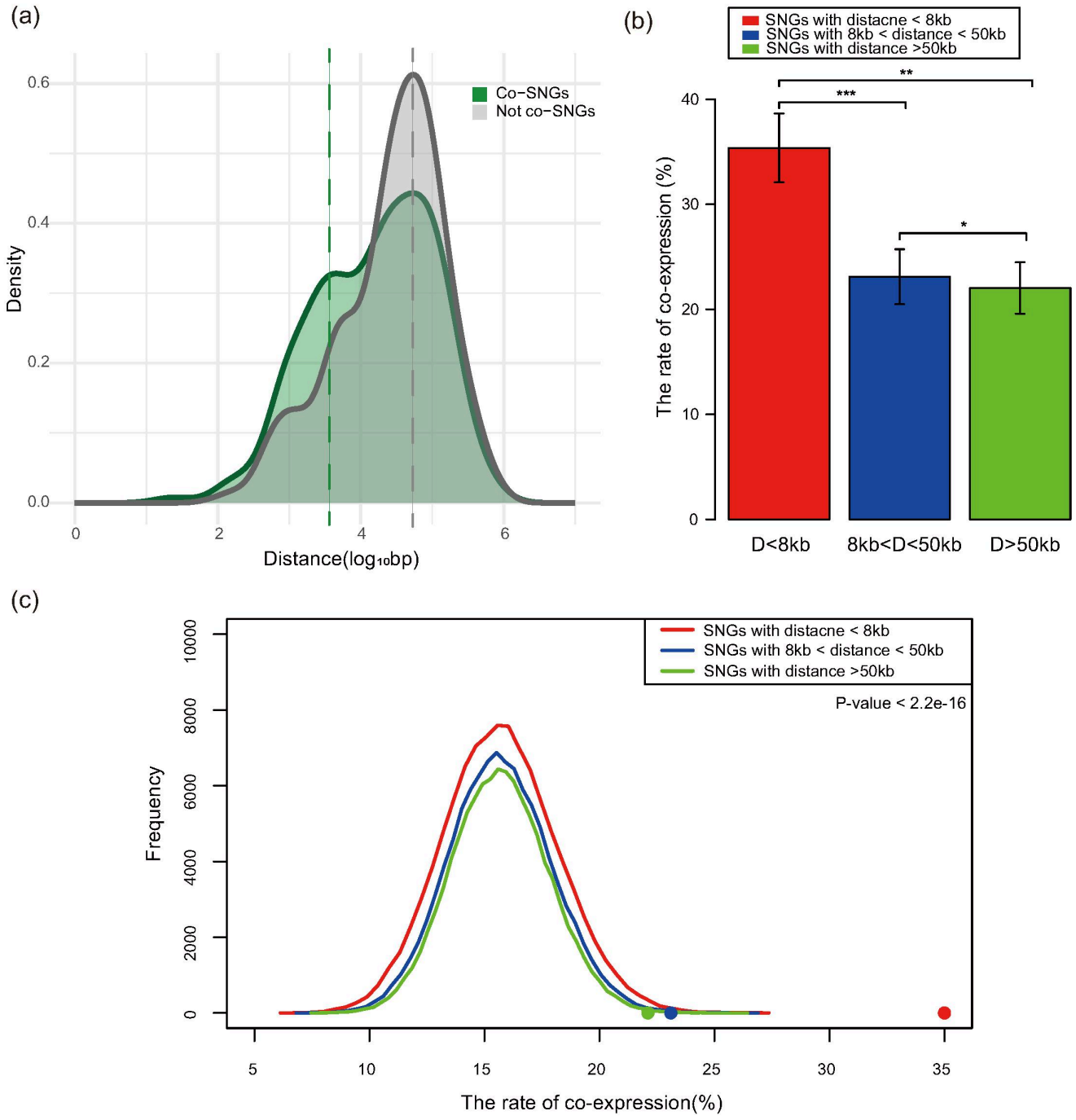

Figure 2

Genomic distance is negatively correlated with SNGs co-expression. **(a)** Distribution of the absolute distance between Co-SNGs and non- co-SNGs. **(b)** Co-expression ratios of SNGs in three defined genomic distance ranges. **(c)** The frequency distributions of co-expression ratios from 10,000 randomized experiments across three distance ranges. Curves of different colors represent the distributions for different genomic distance groups, with each randomization sampling the same number of gene pairs as in the real data. Error bars were calculated by bootstrapping. Significance values calculated from the Mann–Whitney U test are shown.

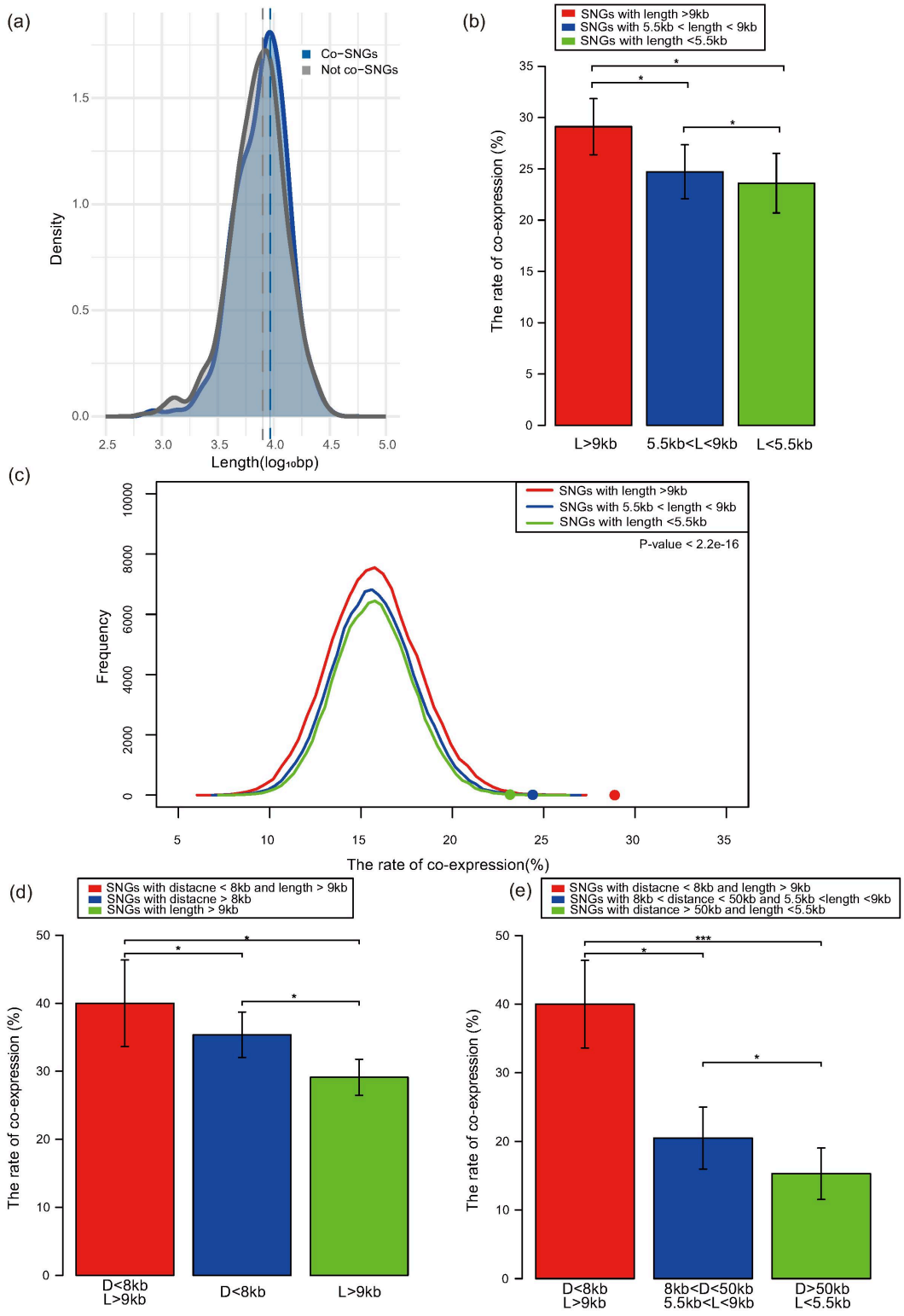

Figure 3

Average gene length is positively correlated with SNGs co-expression and acts synergistically with genomic distance **(a)** Distribution of the average length between Co-SNGs and non- co-SNGs. **(b)** Co-expression ratios of SNGs in three defined average gene length ranges **(c)** The frequency distributions of co-expression ratios from 10,000 randomized experiments across three length ranges. **(d)** Co-expression ratio under the shortest distance and longest gene length condition. **(e)** Co-expression ratios under

combinations of genomic distance and gene length. Error bars were calculated by bootstrapping. Significance values calculated from the Mann–Whitney U test are shown.

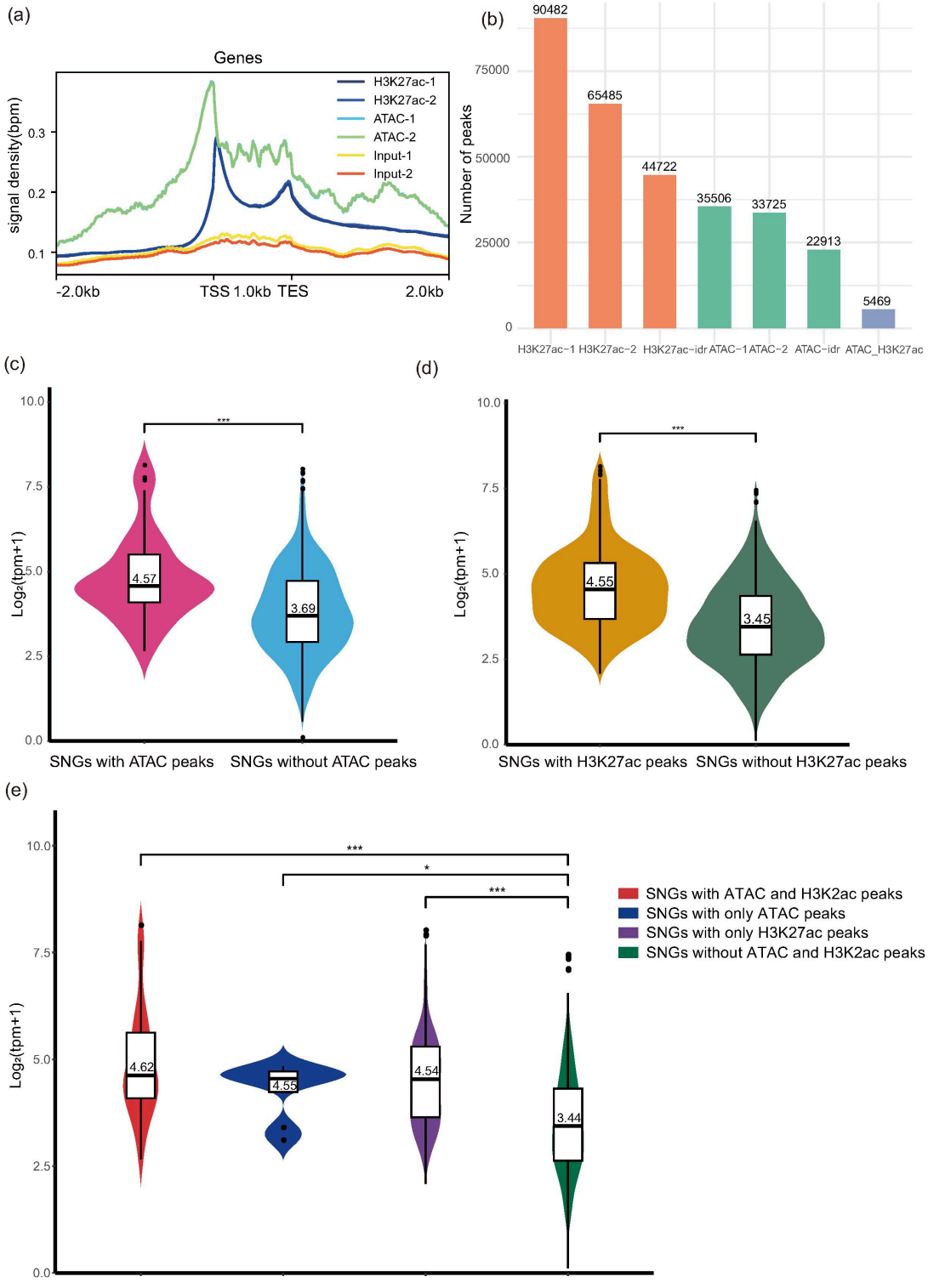

Figure 4

SNGs annotated with ATAC-seq or H3K27ac peaks exhibit higher transcriptional expression levels **(a)** Read depth in the 2 kb upstream and downstream of the transcription start sites (TSSs) and transcription end sites (TESs) of genes. **(b)** Number of ATAC-seq and H3K27ac peaks. **(c)** Expression

levels of SNGs annotated with ATAC-seq peaks. **(d)** Expression levels of SNGs annotated with H3K27ac peaks. **(e)** Expression levels of SNGs annotated with H3K27ac peaks and ATAC-seq peaks. Significance values calculated from the Mann–Whitney U test are shown.

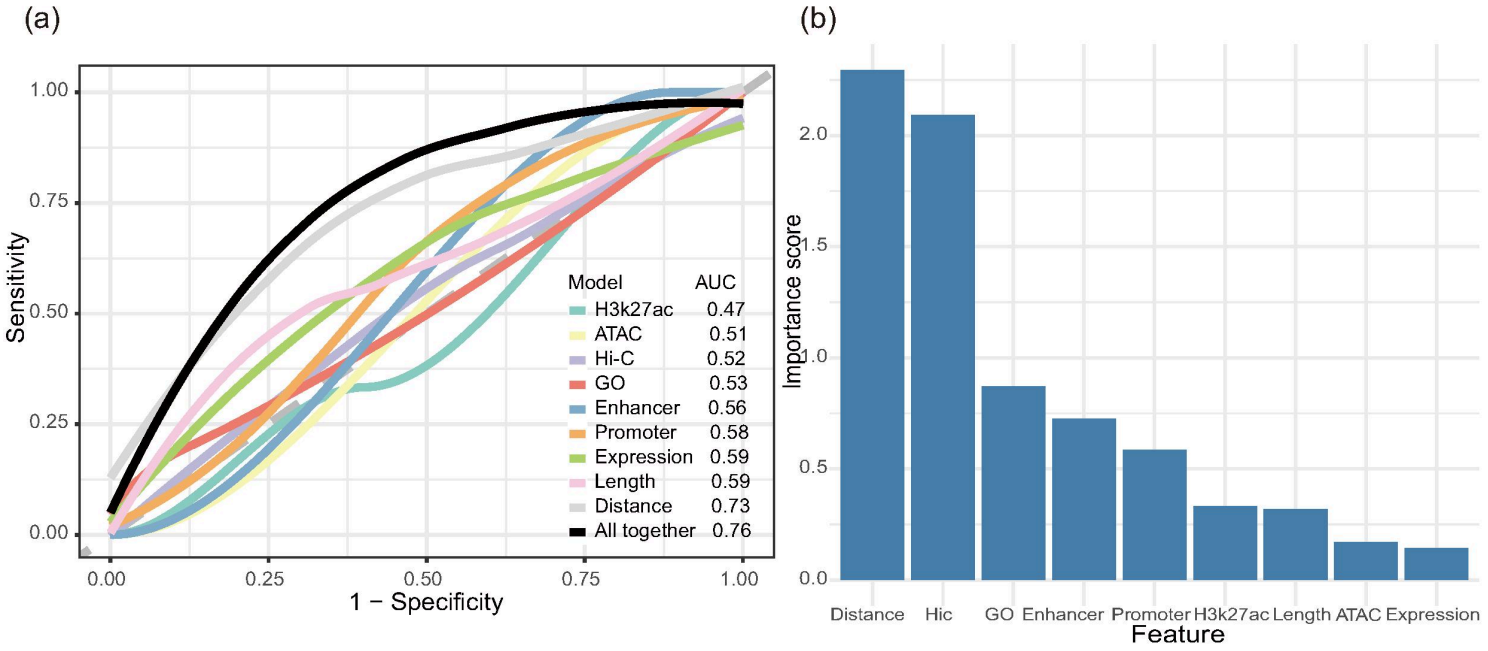

Figure 5

Molecular features associated with SNGs co-expression. **(a)** Receiver operating characteristic (ROC) curve of predicting Co-expression for several molecular features (logistic regression; N=771 for Co-SNGs and non Co-SNGs; see Methods for details). **(b)** Relative importance of each molecular feature in predicting co-expression.

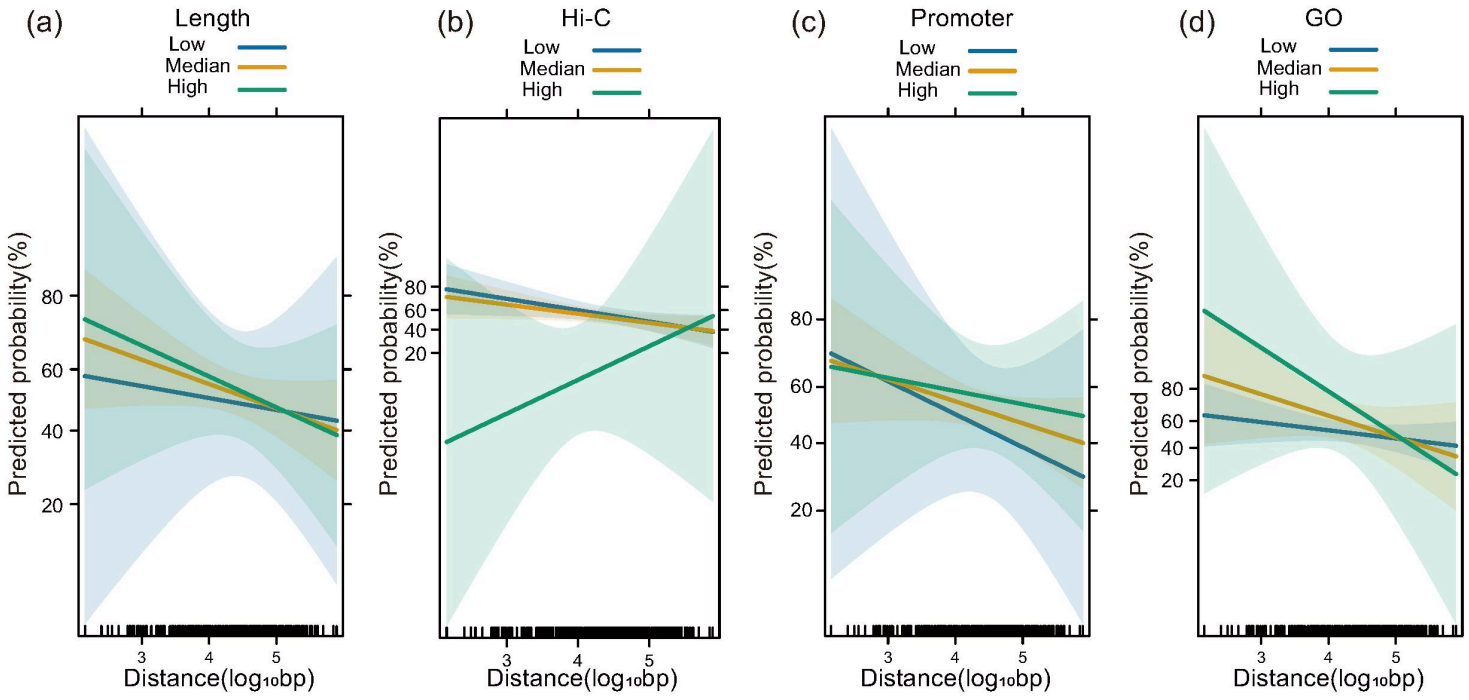

Figure 6

(a) Interaction effects between genomic distance and length on prediction frequency. **(b)** Interaction effects between genomic distance and Hi-C on prediction frequency. **(c)** Interaction effects between genomic distance and promoter on prediction frequency.**(d)** Interaction effects between genomic distance and GO on prediction frequency.

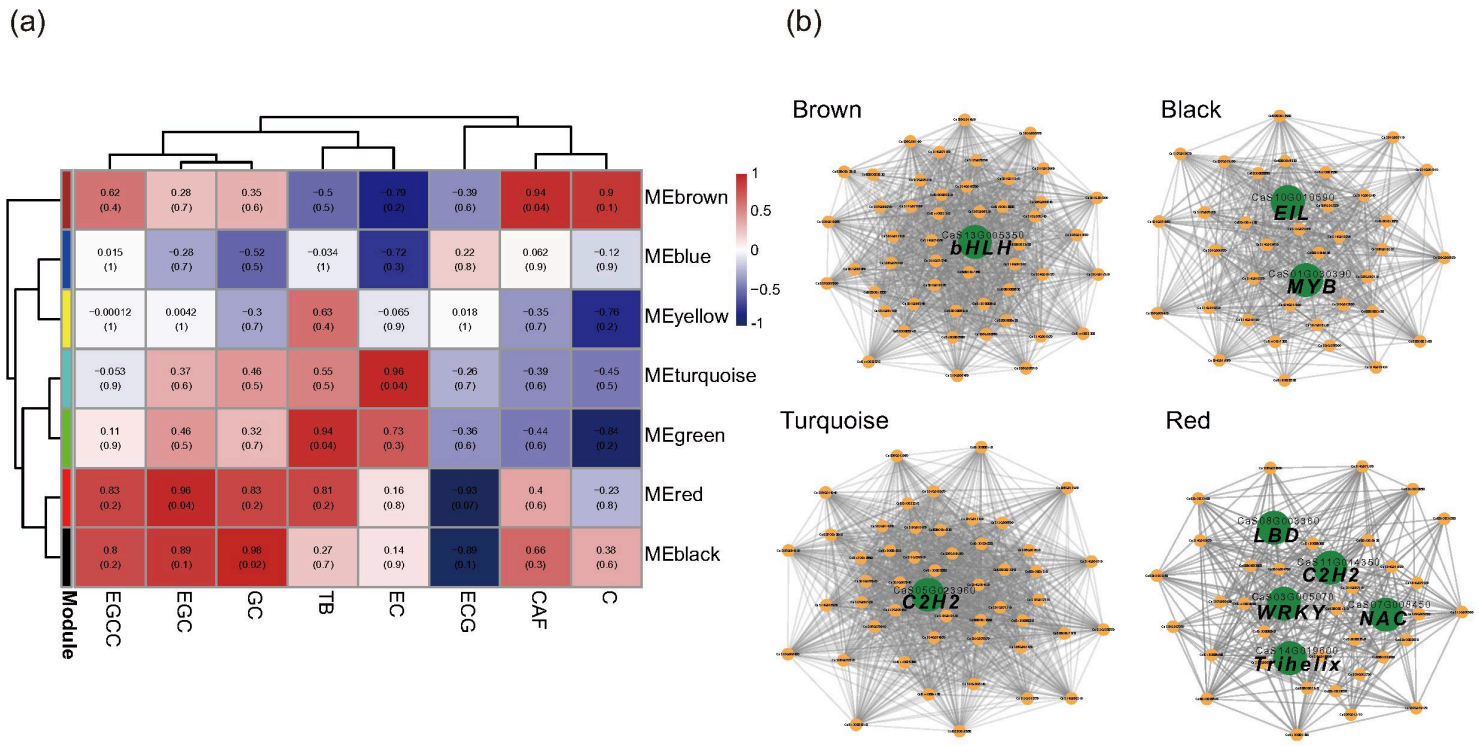

Figure 7

Co-expression network analysis **(a)** Matrix of module-metabolite associations. The data of gene expression profiles for the SNGs and themetabolite accumulationwere combined to perform the WGCNA. Correlation coefficients and P-values between modules and metabolites are shown at the row-column intersections. **(b)** Network visualization of hub genes in four modules, highlighting transcription factors (TFs) potentially involved in metabolite regulation.

Supplementary Files

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

- FigureS1.pdf
- FigureS2.pdf
- TableS1.xlsx

- FigureS3.pdf
- FigureS4.pdf
- FigureS5.pdf
- FigureS6.pdf
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx