Received: 27 March 2024 | Revised: 10 June 2024 | Accepted: 19 June 2024

DOI: 10.1111/ppl.14474

ORIGINAL RESEARCH

Physiologia Plantarum

Comparative transcriptome database for *Camellia sinensis* reveals genes related to the cold sensitivity and albino mechanism of 'Anji Baicha'

Xinghai Zheng[1,2] | Zahin Mohd Ali[2] | Peng Ken Lim[2] | Marek Mutwil[2] | Yuefei Wang[1]

[1]Tea Research Institute, Zhejiang University, Hangzhou, China

[2]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore, Singapore

Correspondence
Marek Mutwil,
Email: mutwil@ntu.edu.sg

Yuefei Wang,
Email: zdcy@zju.edu.cn

Funding information
"Pioneer" and "Leading Goose" R&D Program of Zhejiang, Grant/Award Number: 2023C02041

Edited by A.-J. van Dijk

Abstract

Tea, a globally popular beverage, contains various beneficial secondary metabolites. Tea plants (*Camellia sinensis*) exhibit diverse genetic traits across cultivars, impacting yield, adaptability, morphology, and secondary metabolite composition. Many tea cultivars have been the subject of much research interest, which have led to the accumulation of publicly available RNA-seq data. As such, it has become possible to systematically summarize the characteristics of different cultivars at the transcriptomic level, identify functional genes, and infer gene functions through co-expression analysis. Here, the transcriptomes of 9 tea cultivars were assembled, and comparative analysis was conducted on the coding sequences of 13 cultivars. To give access to this data, we present TeaNekT (https://teanekt.sbs.ntu.edu.sg/), a web resource that facilitates the prediction of gene functions of various tea cultivars. We used TeaNekT to perform a cross-cultivar comparison of co-expressed gene clusters and tissue-specific gene expression. We observed that 'Anji Baicha' possesses the highest number of cultivar-specific genes and the second-highest number of expanded genes. These genes in 'Anji Baicha' tend to be enriched in functions associated with cold stress response, chloroplast thylakoid structure, and nitrogen metabolism. Notably, we identified three significantly expanded homologous genes in 'Anji Baicha' encoding the ICE1, SIZ1, and MAPKK2, which are closely associated with the cold sensitivity of 'Anji Baicha'. Additionally, one significantly expanded homologous gene in 'Anji Baicha' encoding regulatory factor RIQ may play a crucial role in the abnormal chloroplast structure and absence of thylakoid membranes in 'Anji Baicha'.

1 | INTRODUCTION

As one of the most popular non-alcoholic beverages worldwide, tea contains a wide range of secondary metabolites beneficial to human health, such as polyphenols, alkaloids, and theanine (Wang et al., 2022). Meanwhile, the tea plant (*Camellia sinensis*) possesses a diverse range of germplasm resources (Chen et al., 2007). Different tea cultivars are each prized for certain desirable qualities in their own right and exhibit significant differences in plant morphology, leaf characteristics, growth habits, adaptability, and secondary metabolites (Chen et al., 2012; Zhao et al., 2023).

For instance, the temperature-sensitive albino cultivar 'Anji Baicha' (also known as 'Baiye 1') have attracted much attention due to its unique leaf color and secondary metabolite content (Zhang

et al., 2020). The new shoots of the 'Anji Baicha' cultivar are highly sensitive to cold temperatures and undergo a color transformation as the temperature gradually warms up in early spring (Cheng et al., 1999; Li, 2002). Albino (also known as albinism) is a common variation among the plant kingdom caused by one or more factors, including genotype, environment, hormone imbalance, nuclear-plastid genome incompatibility, plastid-DNA deletion, gene mutations related to chlorophyll biogenesis, and metabolic blockage of the chlorophyll biosynthesis pathway (Kumari et al., 2009). For most crops, albino often leads to adverse effects such as low yield and infertility, attributed to the impairment of photosynthetic organs resulting in an imbalance of carbon and nitrogen metabolism (Abbo et al., 2003; Parrott and Smith, 1986). However, during the albino period of 'Anji Baicha', the content of L-theanine in new shoot is 0.5 to 2 times that of regular tea cultivars, while the total catechin content is only half that of regular tea cultivars (Zhang et al., 2020). The taste of green tea made from new shoots of 'Anji Baicha' resulting from its unique ratio of phenols and amino acids (Zou et al., 2018), is highly sought-after, with its cost being tenfold that of ordinary cultivars (Ren et al., 2015). Therefore, it is essential to study the low-temperature sensitivity and albino mechanism of albino cultivars like 'Anji Baicha'. Previous studies have found that 'Anji Baicha' albino leaves exhibit reduced chlorophyll content, abnormal chloroplast structure, disappearance of grana, lack of evident lamellar membrane structure, and remaining in the proplastid stage (Li et al., 2011). Extensive research has been conducted on pigment metabolism, with a growing emphasis on chloroplast structural proteins (Ye et al., 2023). However, potential key genes directly linked to low-temperature response and albino phenotype have not been precisely identified (Ye et al., 2023).

With the unique qualities and abundant germplasm resources of tea plants attracting an increasing number of researchers to study various aspects of tea plants, such as secondary metabolites and phenotypic variations (Zhao et al., 2023; Wang et al., 2022; Liao et al., 2022), the transcriptomics dataset of tea plants has also become increasingly extensive. This has led to the use of systems biology approaches on sequencing data hosted on public databases (Tai et al., 2018; Xia et al., 2020; Zhao & Ma, 2021), such as gene co-expression analysis, becoming a trend in analyzing transcriptomics data of tea plants, providing tea researchers with a more macroscopic and comprehensive perspective. Gene co-expression analysis is built upon the theoretical foundation that functionally related genes typically exhibit similar gene expression patterns (Zhou et al., 2002). Therefore, by identifying gene clusters with highly similar expression patterns, we can pinpoint functionally related genes and infer the functions of unknown genes based on the functions of their neighboring genes (Usadel et al., 2009).

Given that significant differences among different tea cultivars may arise from variations in coding sequences, it is necessary to perform transcriptome assembly on several tea cultivars selected for comparison. Through comparative analysis, we can understand the genetic basis differences among cultivars, infer the biological functions of differentially expressed genes based on gene co-expression analysis, and further elucidate the mechanisms underlying the formation of cultivar differences. Meanwhile, the conservation of functionally related genes among different cultivars would confirm their importance (Hansen et al., 2014). In view of this, we are set to construct a comparative transcriptome analysis tool that can effectively compare transcriptomic data across multiple cultivars and extract predictive information on gene function and regulation from it. To this end, we utilize co-expression network toolkit (CoNekT) (Proost and Mutwil, 2018), a popular framework that has been used to construct comparative transcriptomic databases for plants and species from other kingdoms (Ng et al., 2020; Tan and Mutwil, 2020; Lim et al., 2020; Lim et al., 2022; Villanueva et al., 2022).

In this study, the transcriptomes of 9 tea cultivars were individually assembled using the reference genome of 'Shuchazao' as the guide genome. Then, the coding sequences of 13 tea cultivars were annotated and subjected to phylogenetic analysis. The expression profiles of these 13 tea cultivars were determined and used for constructing co-expression networks. All transcriptome data, including annotations and co-expression networks, of the 13 tea cultivars were imported into CoNekT, resulting in the creation of a co-expression network toolkit for tea plants (TeaNekT) (https://teanekt.sbs.ntu.edu.sg/). The phylogenetic analysis of the 13 cultivars effectively reconstructed their relationships. The analysis of orthologous groups (OGs) identified two types of genes showing differences among cultivars: cultivar-specific genes and expanded genes. Among them, cultivar-specific genes and expanded genes in 'Anji Baicha' are enriched in biological functions related to cold stress response, chloroplast thylakoid structure, and nitrogen metabolism. In further studies, we found 3 significantly expanded homologous genes in 'Anji Baicha' encoding the transcription factor ICE1, E3 SUMO-protein ligase SIZ1, and mitogen-activated protein kinase kinase 2, which may be closely associated with the low-temperature sensitivity of 'Anji Baicha'. Additionally, one significantly expanded homologous gene in 'Anji Baicha' encoding regulatory factor RIQ may play a crucial role in the abnormal chloroplast structure and absence of thylakoid membranes in 'Anji Baicha'.

2 | METHODS

2.1 | Data source and sample metadata annotation

By searching and filtering using the keyword "Camellia sinensis" in the NCBI SRA database, a total of 861 RNA-seq samples from 13 tea cultivars were obtained. Initial annotations of these RNA-seq raw data were performed using the metadata fields in the NCBI SRA database, including cultivar, plant tissue, sampling age, and experimental treatments. Subsequently, the corresponding original papers for each RNA-seq data were searched and retrieved to further supplement and correct the annotation information (Table S1).

For the convenience of calculating the specificity measure (SPM) in the TeaNekT database for all plant tissues, only six terms were retained: "leaf, bud", "root", "stem", "flower", "fruit" and "seed" (e.g., descriptions like "two leaves and a bud" were classified under "leaf, bud"). For the control group, samples in each experiment or samples were directly collected without any treatment and uniformly

labeled as "no treatment" in the experimental treatment column. The samples with missing annotations in the metadata fields of the NCBI SRA database and could not be found in the retrieved original papers were labeled as "missing".

2.2 | Genome-guided transcriptome assembly and completeness assessment

The coding sequences of the cultivar 'Shuchazao' were downloaded from TPIA (http://tpia.teaplant.org) (Xia et al., 2019), and the coding sequences of the cultivars 'Huangdan', 'Longjing 43' and 'Yunkang 10' were downloaded from TeaPGDB V1.0 (http://eplant.njau.edu.cn/tea) (Lei et al., 2021). The coding sequences of the other 9 tea cultivars were obtained using the following genome-guided transcriptome assembly method.

Firstly, several RNA-seq raw data containing different metadata entries were selected from each cultivar for transcriptome assembly. The RNA-seq raw data were processed using the fastp (Chen et al., 2018) to remove adapters and low-quality reads, resulting in high-quality clean data. Then, the HISAT2 sequence alignment software (Kim et al., 2015) was used to align the clean data reads to the reference genome of 'Shuchazao' downloaded from TPIA (http://tpia.teaplant.org) (Xia et al., 2019), generating BAM files. The BAM files were then assembled using the TRINITY with default parameters (Haas et al., 2013), resulting in the assembled transcripts. Next, the TransDecoder with default parameters (https://transdecoder.github.io/) was used to identify and predict open reading frames (ORFs) in the transcripts assembled by TRINITY, identifying potential protein-coding sequences (CDS). Finally, the CD-HIT with default parameters (Fu et al., 2012) was used to remove redundant coding sequences, resulting in the final CDS file, i.e., the coding sequences.

The completeness of the transcriptome assembly was evaluated using the BUSCO (Simão et al., 2015) based on the eudicots_odb10 dataset, which represents the dataset of eudicot plants for both the assembled transcripts generated by the TRINITY for the 9 tea cultivars and the coding sequences of the 13 tea cultivars.

2.3 | Conserved genes, cultivar-specific genes and expanded genes identification

The coding sequences of 13 tea cultivars were aligned using the OrthoFinder with default parameters (Emms & Kelly, 2019), and the orthologous groups (OGs) among these cultivars were constructed (Table S2).

The OG containing coding sequences from 13 cultivars is defined as the conserved OG, while the OG containing coding sequences from only a single cultivar is defined as the cultivar-specific OG. Coding sequences in the conserved OGs are referred to as conserved genes, and coding sequences in the cultivar-specific OGs are referred to as cultivar-specific genes (Table S3).

Additionally, for each cultivar, an expanded OG is defined by meeting the following conditions: It belongs to the conserved OG; the OG's coding sequences have more than two copies in this cultivar, while the copy number in the other 12 cultivars is less than two; the copy number in this cultivar is significantly higher than in the other cultivars, as determined by a one-sample t-test using the stats.ttest_1-samp method. The coding sequences in the expanded OGs are defined as expanded genes (Table S3).

2.4 | Gene annotation and functional enrichment analysis

The coding sequences of the 13 tea cultivars were annotated for biological functions using the Mercator v4 2.0 online tool (Lohse et al., 2014) (Table S4). The InterProScan (Quevillon et al., 2005) was used to obtain protein Pfam domains and gene ontology (GO) terms for each gene from the coding sequences of the 13 tea cultivars (Table S4). The KEGG annotation of the coding sequences of the 13 tea cultivars was performed using the BlastKOALA online tool (Kanehisa et al., 2016) available at https://www.kegg.jp/blastkoala/ (Table S4). iTAK with default parameters (Zheng et al., 2016) was used to predict and classify transcriptional factors from the coding sequences of the 13 tea cultivars (Table S4).

To perform functional annotation on the conserved genes, expanded genes and cultivar-specific genes, the hypergeometric test was conducted using the hypergeom tool from the scipy.stats package (Hahne et al., 2008). This test compared the genes in each gene set with the genes associated with each annotation term. Then, the p-values of all annotation terms corresponding to each conserved gene and cultivar-specific genes were corrected using the fdrcorrection tool from the statsmodels.stats.multitest package (Benjamini and Hochberg, 1995; Nazer et al., 2018) to obtain the false discovery rate (FDR) values (Table S5). Annotation terms with FDR values less than or equal to 0.05 were considered significant for functional annotation of the conserved, expanded, and cultivar-specific genes. Then, calculate the coefficient of variation (CV) for each annotation term across all cultivars. Sort the annotation terms in descending order based on their CV, and select the top 80 annotation terms for data visualization.

2.5 | Phylogenetic analysis of tea cultivars

A set of single-copy orthologous genes containing a single orthologous gene from each of the 13 tea cultivars was extracted from the conserved genes for phylogenetic analysis.

To construct the phylogenetic relationships among tea cultivars, the amino acid sequences of each single-copy orthologous gene pair were aligned using the MUSCLE with default parameters (Edgar, 2004). The TRIMAL (Capella-Gutierrez et al., 2009) was then used to remove poorly aligned regions. Based on the trimmed alignment files of each single-copy orthologous gene pair, phylogenetic trees were constructed using the RAXML (Stamatakis, 2014), with 'Yunkang 10' serving as the outgroup. Subsequently, all the phylogenetic tree files of the single-copy orthologous gene pairs were merged into a supermatrix file, and

the ASTRAL (Zhang et al., 2018) was used to construct the phylogenetic tree of the 13 tea cultivars.

2.6 | Gene expression levels and expression variation coefficients calculation

The clean data reads were pseudo-aligned to the coding sequences of the 13 tea cultivars using the Kallisto (Bray et al., 2016). This allowed for the quantification of the transcripts per million (TPM) values for all genes in each sample, serving as a measure of gene expression levels. Additionally, the percentage of sequencing reads aligned to the coding sequences was calculated as the pseudo-alignment read percentage, providing an assessment of the alignment efficiency of the transcriptomic sequencing data (Table S1). Subsequently, the coefficient of variation (CV) for the expression levels of each gene across all samples was calculated as an indicator of gene expression variability within each tea cultivar (Table S3).

2.7 | Co-expression networks and TeaNekT database construction

First, an expression profile was constructed for the 13 tea cultivars. Then, a co-expression network was built for the tea cultivars using the highest reciprocal rank (HRR) metric (Mutwil et al., 2010). The CoNekT database framework, with default settings, was used to populate the database with information such as gene expression levels, gene annotations, orthologous gene sets, and phylogenetic trees for the 13 tea cultivars (Proost and Mutwil, 2018). This resulted in the construction of TeaNekT (https://teanekt.sbs.ntu.edu.sg/), an online tea plants database equipped with co-expression clusters that can identify co-expressed genes, gene families and enrich specific biological functions. In the TeaNekT database, the heuristic cluster chiseling algorithm (HCCA) (Mutwil et al., 2010) was employed to generate co-expression clusters for the 13 tea cultivars, with a 100 genes per cluster limit.

3 | RESULTS

3.1 | Information on selected tea cultivars

This study selected a total of 13 tea cultivars, including 'Anji Baicha', 'Echa 1', 'Fuding Dabaicha', 'Huangdan', 'Huangjinya', 'Jinxuan', 'Longjing 43', 'Longjing Changye', 'Shuchazao', 'Tieguanyin', 'Zhongcha 108', 'Yunkang 10' and 'Zijuan' for transcriptome assembly and the TeaNekT database construction. These cultivars were selected based on the availability of a substantial amount of RNA-seq data in NCBI (at least 30 RNA-seq data), ensuring robust data support for this study and the database.

Of these 13 tea cultivars, 11 belong to *C. sinensis* var. *sinensis*, while two cultivars, 'Yunkang 10' and 'Zijuan', belong to *C. sinensis* var. *assamica* (Table 1). Only the coding sequences of 'Shuchazao' were found in the TPIA database, and the coding sequences of 'Yunkang 10', 'Longjing 43', and 'Huangdan' were found in the TeaPGDB V1.0 database, while this study assembled the coding sequences of the other 9 tea cultivars.

3.2 | Genome-guided transcriptome assembly yielded high completeness of coding sequences

Through downloading and genome-guided transcriptome assembly, coding sequences from 13 tea cultivars were obtained. The transcriptome assembly of the 9 tea cultivars was carried out based on the following pipeline (Figure 1A). Subsequently, the coding sequences of the 13 tea cultivars were annotated, subjected to phylogenetic analysis, expression quantification, co-expression network analysis, and construction of the TeaNekT database.

During the assembly process, the assembled transcripts generated by TRINITY yielded approximately 360,000 to 670,000 sequences, many of which could represent non-coding sequences or redundant transcripts (Table 2). To obtain coding transcripts, TransDecoder was used to filter and obtain approximately 110,000 to 180,000 coding sequences. Finally, CD-HIT removed duplicate sequences, resulting in approximately 49,000 to 67,000 coding sequences. The quality of the assembly for both the assembled transcripts of the 9 cultivars and the coding sequences of the 13 cultivars were assessed using BUSCO.

In the BUSCO results, we observed that the assembled transcripts assemblies of the 9 tea cultivars exhibit completeness values of over 90% (Figure 1B). Furthermore, the completeness values of the coding sequences are also above 85% (Figure 1C). These values are comparable to the quality of 'Shuchazao' and 'Huangdan', and significantly higher than that of 'Longjing 43' and 'Yunkang 10'. This indicates that the transcriptome assembly in this study is robust.

The number of transcripts of tea cultivars ranges from approximately 30,000 to 70,000 (Figure 1D). The assembled coding sequences generally contained a higher number of coding sequences compared to the coding sequences available in the TPIA and TeaPGDB V1.0 database. Specifically, the 'Yunkang 10' and 'Longjing 43' cultivars downloaded from the database had the least coding sequences, with 36,951 and 33,556 coding sequences, respectively. The GC content of all tea cultivars coding sequences remains relatively stable, fluctuating around 44% (Figure 1D).

In this study, conserved genes are those that exist in the coding sequences of homologous groups containing 13 tea cultivars, while cultivar-specific genes refer to those that exist in the coding sequences of homologous groups of only one tea cultivars. In coding sequences of all tea cultivars, the number of conserved genes remains stable, fluctuating around 20,000 (Figure 1D). The number of cultivar-specific genes does not

TABLE 1 Overview of 13 tea cultivars.

Abbreviation	Variety	Cultivar	Origin	Characteristics	Coding sequences source	Reference*
AJBC	*C. sinensis* var. *sinensis*	Anji Baicha	Discovered in the wild in 1980	Cold-sensitive, albino variety, amino acids-riched	Transcriptome assembly	Cheng et al., 1999
EC1	*C. sinensis* var. *sinensis*	Echa 1	Bred with Fuding Dabaicha as the female parent and Meizhan as the male parent	Strong vitality, high adaptability, high yield	Transcriptome assembly	Chen et al., 2010
FDDB	*C. sinensis* var. *sinensis*	Fuding Dabaicha	Obtained through individual plant selection method by local farmers in Fuding County, Fujian Province over a hundred years ago	Strong resistance, abundant trichomes, lush and vigorous buds/leaves	Transcriptome assembly	Qiu, 1965
HD	*C. sinensis* var. *sinensis*	Huangdan	Originally from Anxi County, Fujian Province, with a cultivation history of over a hundred years	Strong vitality, high adaptability, high yield	TeaPGDB V1.0	Chen, 1981
HJY	*C. sinensis* var. *sinensis*	Huangjinya	Selected from a natural variation branch of the tea tree in 1998	Light-sensitive, yellowing variety, amino acids-riched	Transcriptome assembly	Wang et al., 2008
JX	*C. sinensis* var. *sinensis*	Jinxuan	Bred with Yingzhi Hongxin as the male parent and Tainong 8 as the female parent	Deep green leaves, elliptical leaves, abundant trichomes, lush and vigorous buds/leaves	Transcriptome assembly	Mo, 2011
LJ43	*C. sinensis* var. *sinensis*	Longjing 43	Developed through systematic breeding method from the population of Longjing tea trees	Early germination, low tenderness	TeaPGDB V1.0	Chen, 1982
LJCY	*C. sinensis* var. *sinensis*	Longjing Changye	Developed through individual plant selection method from the population of Longjing tea trees	Elliptical leaves, high tenderness	Transcriptome assembly	Yang et al., 1995
SCZ	*C. sinensis* var. *sinensis*	Shuchazao	Developed through systematic breeding method from the population of Shucha tea trees	Early germination, high yield, tender leaves	TPIA database	Xia, 2000
TGY	*C. sinensis* var. *sinensis*	Tieguanyin	Discovered in the field in the early 1700s	Green leaves with hints of purple-red, late germination, few trichomes, lush and vigorous buds/leaves	Transcriptome assembly	Wang, 2004
ZC108	*C. sinensis* var. *sinensis*	Zhongcha 108	Bred from Longjing 43 through radiation mutagenesis	Early germination, strong vitality, high adaptability, high yield	Transcriptome assembly	Yang et al., 2003
YK10	*C. sinensis* var. *assamica*	Yunkang 10	Developed through individual plant selection method from the population of Nannuoshan tea trees	Strong drought and cold resistance, polyphenol-riched	TeaPGDB V1.0	Tian et al., 2011
ZJ	*C. sinensis* var. *assamica*	Zijuan	Developed through individual plant selection method from the population of Yunnan Daye tea trees	Purple variety, anthocyanins-riched	Transcriptome assembly	Yang et al., 2013

*The reference provides information in the table regarding the origin and characteristics.

exceed 5,000 in any cultivar. The number of conserved genes remains consistent across the cultivars, while the number of non-conserved genes (the collection of "Others" and "Cultivar-specific" genes) shows significant variation among the cultivars.

In coding sequences of all tea cultivars, the median transcript length ranges from 500 to 1000 (Figure 1E). The median transcript length values of the assembled coding sequences are lower than those of the coding sequences available in the TPIA and TeaPGDB V1.0 database.

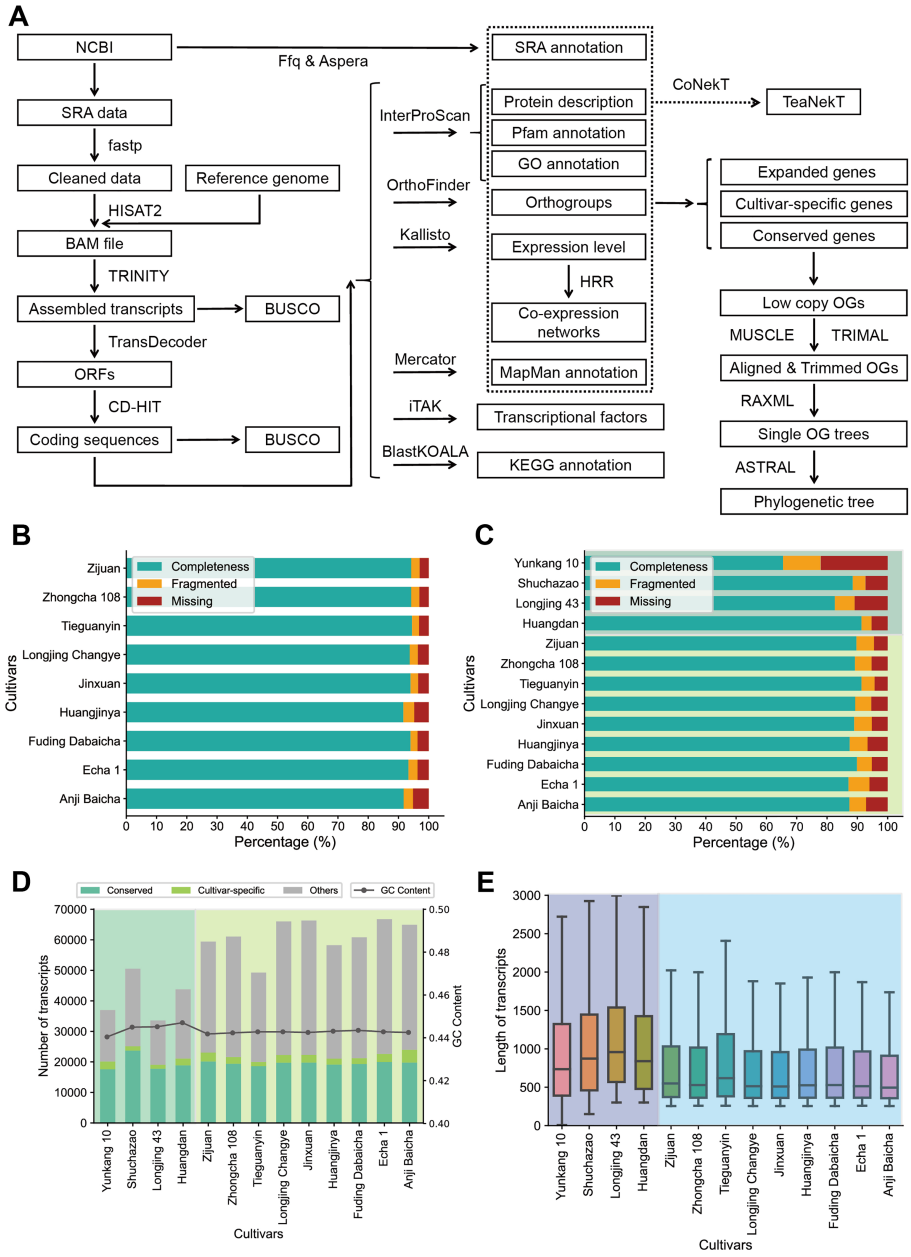

FIGURE 1 Pipeline of this study and the assembly quality of coding sequences of 13 tea cultivars. Cultivars with coding sequences downloaded from databases and cultivars with coding sequences self-assembled are distinguished by different colored background panels (Top/Left: Downloaded from the database, Bottom/Right: Self-assembled). (A) Pipeline for genome-guided transcriptome assembly, phylogenetic analysis, and TeaNekT database construction. (B) Completeness assessment of the assembled transcripts by BUSCO. (C) Completeness assessment of the coding sequences by BUSCO. (D) GC content of coding sequences and proportion of conserved genes and cultivar-specific genes. (E) Length distribution of coding sequences.

TABLE 2 Change in the number of sequences from assembled transcripts to the coding sequences.

Cultivars	AJBC	EC1	FDDB	HJY	JX	LJCY	TGY	ZC108	ZJ
Assembled transcripts	662791	587518	533846	508150	597634	591662	361842	522222	420751
ORFs	172926	176696	151144	146106	179914	174391	111671	160814	132113
Coding sequences	64925	66747	60805	58213	66303	65999	49257	61055	59405

3.3 | Analysis of orthologous groups (OGs) revealed the differences and phylogenetic relationships among tea cultivars

To compare and analyze the coding sequences of 13 tea cultivars based on sequence similarity, we utilized OrthoFinder to obtain orthologous groups (OGs). Analysis of these OGs involves comparing the coding sequences among different cultivars to unveil the phylogenetic relationships and functional conservation of these genes, aiding in exploring the differences and commonalities in the coding sequences of different cultivars.

Within all the OGs, two types of OGs are particularly noteworthy: Conserved OGs and cultivar-specific OGs. Conserved OGs contain coding sequences from homologous groups of all 13 tea cultivars,

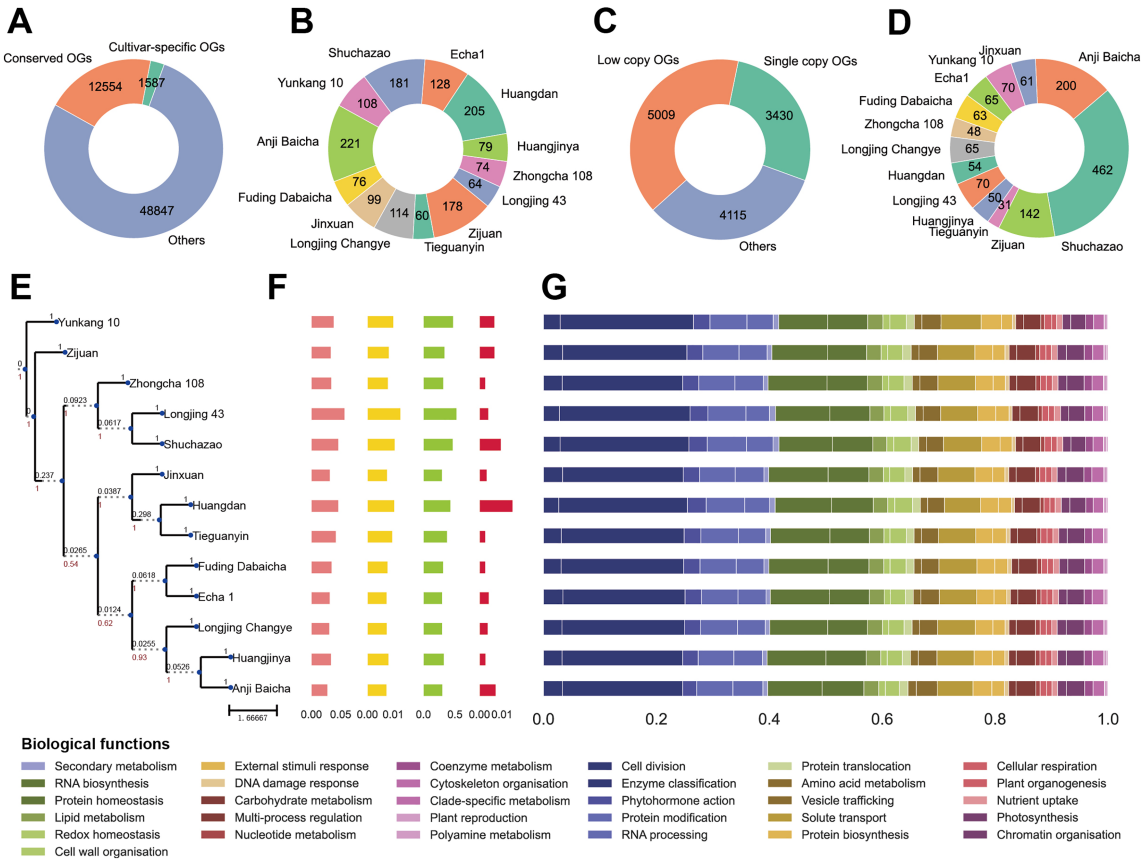

FIGURE 2 Phylogenetic analysis and comparative analysis of coding sequences of 13 tea cultivars. (A) Number of conserved OGs and cultivar-specific OGs. (B) Number of cultivar-specific OGs for each cultivar. (C) Number of single copy OGs and low copy OGs. (D) Number of expanded OGs for each cultivar. (E) Phylogenetic tree of 13 tea cultivars. Bootstrap values and posterior probabilities are colored and displayed on the branches of the phylogeny (Red for branch support values and black for branch length values). (F) Proportions of transcription factors (TFs), transcription regulators (TRs), conserved genes, and cultivar-specific genes (arranged in order from left to right) in the coding sequences of tea cultivars. (G) Proportion of various biological functional genes in the coding sequences of different cultivars.

whereas cultivar-specific OGs only include coding sequences from a single tea cultivar. We observed 12,554 conserved OGs and 1,587 cultivar-specific OGs among the OGs of the 13 tea cultivars (Figure 2A). Among cultivar-specific OGs, we further investigated the cultivar-specific OGs of each cultivar and found that the cultivar 'Anji Baicha' has the highest number of cultivar-specific OGs (221), followed by 'Huangdan' (205), 'Shuchazao' (181), and 'Zijuan' (178) (Figure 2B). This suggests that 'Anji Baicha' has more cultivar-specific genes than other cultivars.

Conserved OGs contain conserved homologous genes among the 13 tea cultivars, where single-copy OGs and low-copy OGs contain genes that are conserved and stable among the cultivars. Other OGs exhibit variability among the cultivars, such as amplifications specific to a particular cultivar. We observed 3,430 single-copy OGs and 5,009 low-copy OGs (Figure 2C). In other categories of OGs, we further investigated OGs significantly expanded in each cultivar and found that the cultivar 'Shuchazao' has the most expanded OGs (462), followed by 'Anji Baicha' (200) and 'Zijuan' (142) (Figure 2D).

To analyze the phylogenetic relationship and transcriptomic similarity among tea cultivars, a phylogenetic analysis was conducted on single-copy OGs of 13 tea cultivars (Figure 2E). ASTRAL utilizes maximum likelihood estimation and the bootstrap method to calculate branch support values. In this phylogenetic tree, branch support values are typically close to 1. A high branch support value (close to 1) indicates that the branch is well-supported across multiple datasets, allowing for confident inference of its presence and position. We found that the phylogenetic relationships and order of the 13 cultivars were consistent with the origin and characteristics of the cultivars listed in Table 1.

The phylogenetic order of the 13 tea cultivars, displayed from top to bottom (Figure 2E), is as follows: 'Yunkang 10' and 'Zijuan', as earlier diverging var. *assamica* cultivars, diverge from the other cultivars in the earliest branch. 'Zhongcha 108', a variant generated from radiation mutagenesis of 'Longjing 43', is grouped in a neighboring branch with 'Longjing 43'. 'Echa 1', a hybrid cultivar derived from the cross between 'Fuding Dabaicha' and 'Meizhan' as the maternal and paternal parent, respectively, is placed in the same branch as 'Fuding Dabaicha'. Finally, 'Anji Baicha' and 'Huangjinya', two unique colored tea cultivars that undergo leaf color changes, are sensitive to the environment and rich in amino acids, are grouped in a separate branch.

In the compositional analysis of the coding sequences from 13 tea cultivars, we found that 'Longjing 43', 'Shuchazao', 'Huangdan', and 'Tieguanyin' have the highest proportions of transcription factors, transcription regulators, and conserved genes (Figure 2F). Among them, 'Huangdan' has the highest proportion of cultivar-specific genes. In the distribution of biological functions of the coding sequences from the 13 cultivars, all cultivars exhibit very similar proportions across various biological functions, which is expected (Figure 2G). However, there are still subtle differences. For instance, 'Yunkang 10' has the highest proportion of genes related to the enzyme classification.

3.4 | Comparative analysis of cultivar-specific and expanded genes among tea cultivars

The analysis of orthologous groups (OGs) on coding sequences has identified two categories of genes that explain the differences among tea cultivars: cultivar-specific genes in cultivar-specific OGs and expanded genes in expanded OGs. We conducted functional enrichment analysis on these genes to further explore the functional differences between these two gene categories and understand the differences among cultivars at a biological level. In the functional enrichment analysis, we enriched all levels of terms in MapMan and then calculated the proportion of enriched sub-terms under each primary term.

In the functional enrichment analysis of cultivar-specific genes, we found that the cultivar 'Anji Baicha' exhibited the highest number of significantly enriched biological functions, followed by 'Shuchazao', 'Yunkang 10', and 'Zijuan' (Figure S1A). Compared to other cultivars, the cultivar-specific genes of 'Anji Baicha' were more inclined to be significantly enriched in categories such as protein homeostasis, amino acid metabolism, multi-process regulation, chromatin organization, and cell division. Additionally, in other cultivars, it was observed that the cultivar-specific genes of 'Longjing 43' were more inclined to be significantly enriched in redox homeostasis, while those of 'Huangjinya' were more inclined to be significantly enriched in polyamine metabolism.

In the functional enrichment analysis of expanded genes, we found that the cultivar 'Shuchazao' exhibited the most significantly enriched biological functions, followed by 'Anji Baicha' and 'Zijuan' (Figure S1B). Compared to other cultivars, the expanded genes of 'Anji Baicha' were more inclined to be significantly enriched in categories such as protein modification, chromatin organization, and cell division. Additionally, in other cultivars, it was observed that the expanded genes of 'Shuchazao' were more inclined to be significantly enriched in enzyme classification, while those of 'Echa 1' were more inclined to be significantly enriched in clade-specific metabolism.

Both artificial selection and adaptive evolution can lead to the extensive expansion of a gene family in a particular cultivar, and in this study, expanded genes in cultivars can reveal such processes (Emerson and Thomas, 2009). Therefore, we conducted enrichment analysis on expanded genes in the 13 cultivars across three dimensions: transcription factors/regulators, gene symbols, and biological functions.

At the transcription factor/transcription regulator level, we observed significant differences among cultivars in the transcription factors bHLH, bZIP, and AP2/ERF-ERF encoded by expanded genes (Figure 3A). Similarly, cultivars had notable differences in the transcription regulators SET, PHD, and GNAT encoded by expanded genes. At the gene symbol level, we found that 'Anji Baicha' and 'Huangjinya' had noticeably more pathway genes encoded by expanded genes (Figure 3B). At the biological function level, we observed that expanded genes in 'Zijuan' could be enriched into more biological functions (Figure 3C).

3.5 | Enrichment analysis revealed biological functions related to cultivar differences and the uniqueness of 'Anji Baicha'

In the orthologous groups (OGs) and functional enrichment analysis, we observed that the cultivar 'Anji Baicha' has a high number of cultivar-specific genes and expanded genes with diverse and complex biological functions, indicating further research value. Additionally, as a variant of albino tea plants known for their unique characteristics, such as cold sensitivity and high amino acid content, 'Anji Baicha' itself holds significant research value (Zhang et al., 2020). Therefore, we further explored the enrichment of conserved, cultivar-specific, and expanded genes in 'Anji Baicha' at a more detailed level of biological function terms (Figure 4).

Cultivar-specific genes and expanded genes in 'Anji Baicha' tend to be significantly enriched in biological functional terms related to cold stress, photosynthesis, and carbon-nitrogen metabolism balance. Here, we have selected to showcase more detailed descriptions of the relevant terms. We observed that expanded genes of 'Anji Baicha' are significantly enriched in biological functions such as ICE-CBF-COR cold acclimation signaling, photosystem II.thylakoid grana stacking, chloroplast.outer envelope guidance and insertion, and plastid-encoded RNA polymerase (PEP) complex. The expanded genes in these biological functions may be related to the unique cold sensitivity and leaf albino mechanism of 'Anji Baicha'.

Additionally, expanded genes of 'Anji Baicha' are significantly enriched in transcription factor families such as bHLH class-X transcription factor, bHLH class-IIIb transcription factor, BZR-type transcription factor, RAV-NGATHA transcription factor, AP2/ERF family, ARR-B-type transcription factor, and PLATZ transcription factor. The cultivar-specific genes of 'Anji Baicha' are significantly enriched in the bHLH class-Va transcription factor and ARF transcription factor families. These transcription factors may be related to the uniqueness of 'Anji Baicha' and warrant further investigation.

3.6 | Three OGs related to cold-stress sensing and signaling are associated with the low-temperature sensitivity of 'Anji Baicha'

As a cold-sensitive albino tea cultivar, we observed enrichment of expanded genes of 'Anji Baicha' in several biological functions related

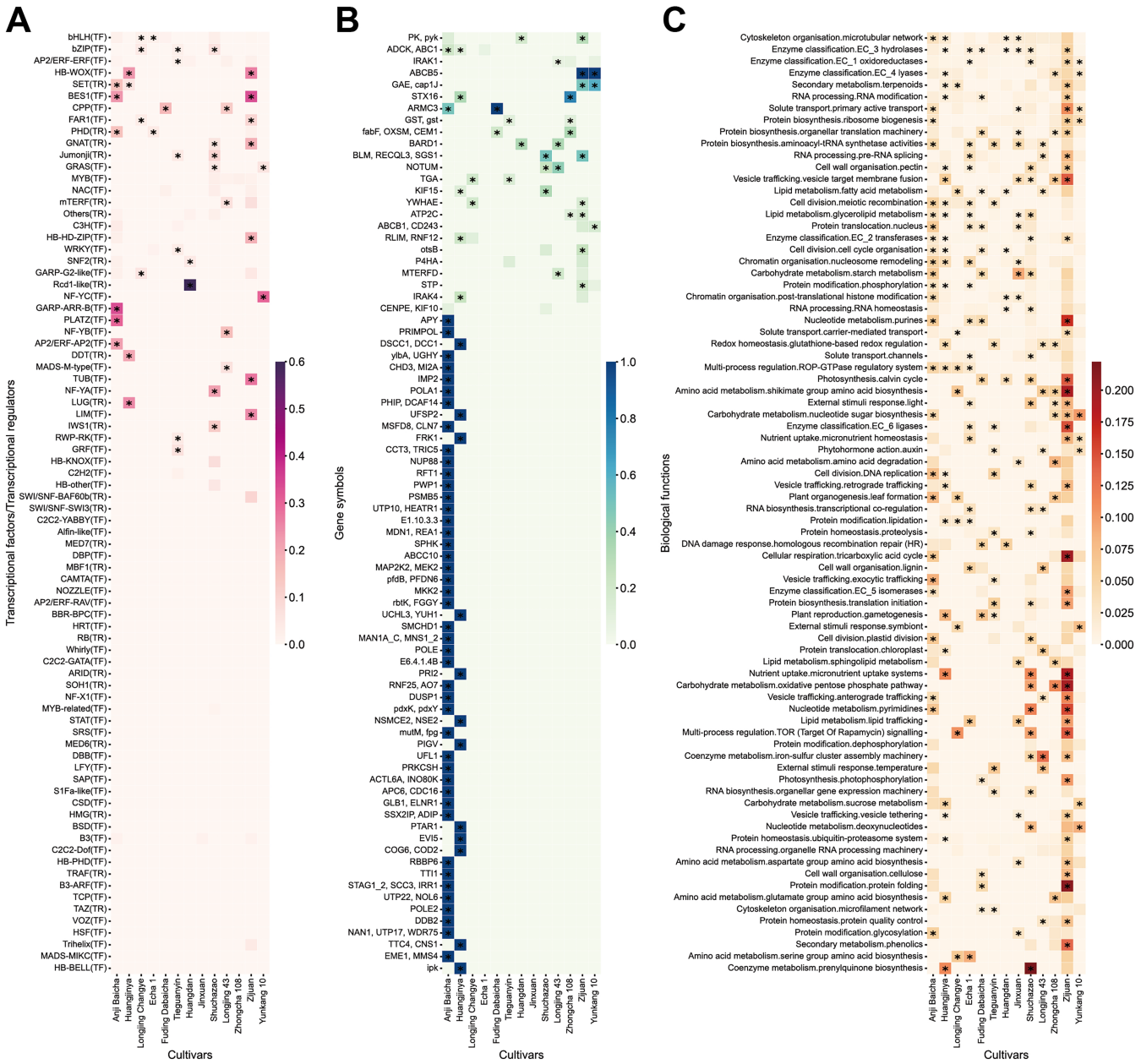

FIGURE 3 Enrichment analysis of expanded genes in the coding sequences of 13 tea cultivars. (A) Transcription factors (TFs) and transcription regulators (TRs). (B) Gene symbols. (C) Biological functions.

to cold-stress sensing and signaling, such as ICE-CBF-COR cold acclimation signalling, MAP kinase cascade signalling, and E3 SUMO ubiquitin ligase activities (Figure 4). This implies that 'Anji Baicha' expands a large number of genes in the cold-stress sensing and signaling pathways, making it more sensitive to low temperatures, which may also help enhance its resistance to cold stress during the albino phase.

To validate this hypothesis, we utilized the onboard functions of the TeaNekT database that we have constructed to assist in the study of key genes in the cold-stress sensing and signaling pathways. The homologous genes in OG0003915, encoding the transcription factor ICE1, were found to be significantly expanded in 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/tree/view/3915). Similarly, the homologous genes in OG0007208, encoding the E3 SUMO-protein ligase SIZ1, were also found to be significantly expanded in 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/tree/view/7208). We also observed that the homologous genes in OG0003915 and OG0007208 are predominantly expressed in leaves and buds, with lower expression in other tissues, implying tissue-specific expression of these two OGs in leaves and buds.

Additionally, we found that the homologous genes in OG0004279, encoding the mitogen-activated protein kinase kinase 2, were significantly expanded in 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/tree/view/4279). However, unlike OG0003915 and

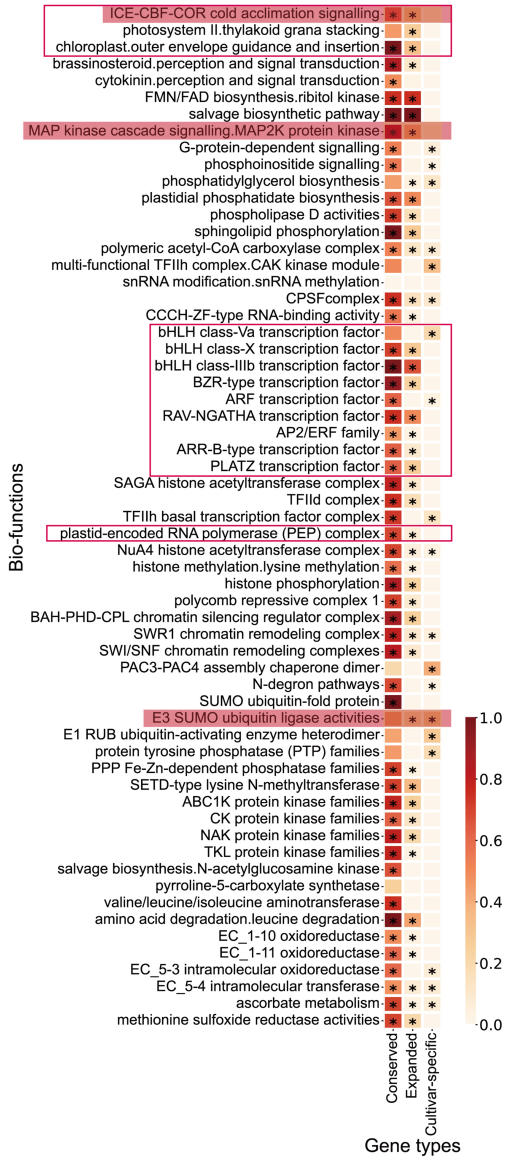

FIGURE 4 Enrichment analysis of the biological functions of conserved genes, cultivar-specific genes, and expanded genes in the cultivar 'Anji Baicha'. Detailed biological function information is presented, with the color intensity in the heatmap representing the recall value. "*" denotes significant/non-significant enrichment.

OG0007208, the homologous genes in OG0004279 are highly expressed in leaves and buds and exhibit high expression in other tissues, indicating a lack of tissue-specific expression.

3.7 | A OG related to photosystem II thylakoid grana stacking is associated with the albino mechanism of 'Anji Baicha'

In past research on the mechanism of 'Anji Baicha' albino, it was found that during the albino phase, the number of mature intact chloroplasts significantly decreased, chloroplast structure was abnormal, and there was no apparent lamellar structure (Li et al., 2011). In this study, we observed that expanded genes of 'Anji Baicha' are enriched in the biological function of photosystem II thylakoid grana stacking (Figure 4). In the biological function of photosystem II thylakoid grana stacking, we identified a gene family encoding the regulatory factor RIQ. RIQ stands for reduced induction of non-photochemical quenching, a regulatory factor associated with the organization of chloroplast grana structure and photosystem II complexes.

In addition to observing that the homologous genes in OG0010584, encoding the regulatory factor RIQ, were significantly expanded in 'Anji Baicha', we also found that these genes are predominantly expressed in leaves and buds, with lower expression in other tissues, implying tissue-specific expression of OG0010584 in leaves and buds (https://teanekt.sbs.ntu.edu.sg/tree/view/10584) (Figure 5A). The gene *ANJIBAICHA57615* from OG0010584 shows significantly lower expression at the bud-prealbinism stage (Figure 5B). *ANJIBAICHA45068* and *ANJIBAICHA57615*, which are part of the homologous genes in OG0010584, belong to cluster 72 and cluster 229, respectively. We observed that cluster 72 is mainly enriched in GO terms such as photosynthesis, photosystem II assembly, and photosystem II oxygen-evolving complex, while cluster 229 is mainly enriched in GO terms such as photosystem, photosystem I reaction center, and photosystem II oxygen-evolving complex, which illustrates the association between the OG0010584 and photosynthesis (Figure 5C). In a comparative study of cluster 229 with co-expression clusters from other cultivars, we observed a conserved co-expression relationship among cultivars involving the gene *ANJIBAICHA57615* and genes encoding LHCA, LHCB, CAB4, ATM4, PSAL, and PSBQ (Figure 5D and E).

4 | DISCUSSION

This study primarily investigates the differences and commonalities among cultivars of tea plants through two main aspects of work. Firstly, transcriptome assembly and comparative analysis were conducted on 13 cultivars, revealing the conservation and specificity of coding sequences among cultivars. Secondly, annotation, phylogenetic analysis, and co-expression network construction were performed on the coding sequences of 13 cultivars, which were integrated into the CoNekT framework to establish the TeaNekT database.

In previous studies, researchers tended to consider the *C. sinensis* species as a whole when conducting large-scale co-expression analysis of tea plant samples (Zhang et al., 2020). They generally used the genome of the tea cultivar 'Shuchazao' as the reference genome to calculate gene expression levels for constructing the co-expression network of tea plants. In this study, the transcriptome of different tea cultivars was separately assembled to obtain their respective coding sequences. Each cultivar possesses unique coding sequences, allowing us to explore gene conservation through sequence similarities and uncover cultivar-specific genes by analyzing these distinctive coding sequences (Figure 3). For instance, in our study of orthologous groups (OGs) across 13 cultivars, we found that 'Anji Baicha' cultivar has the highest number of cultivar-specific OGs and ranks second in

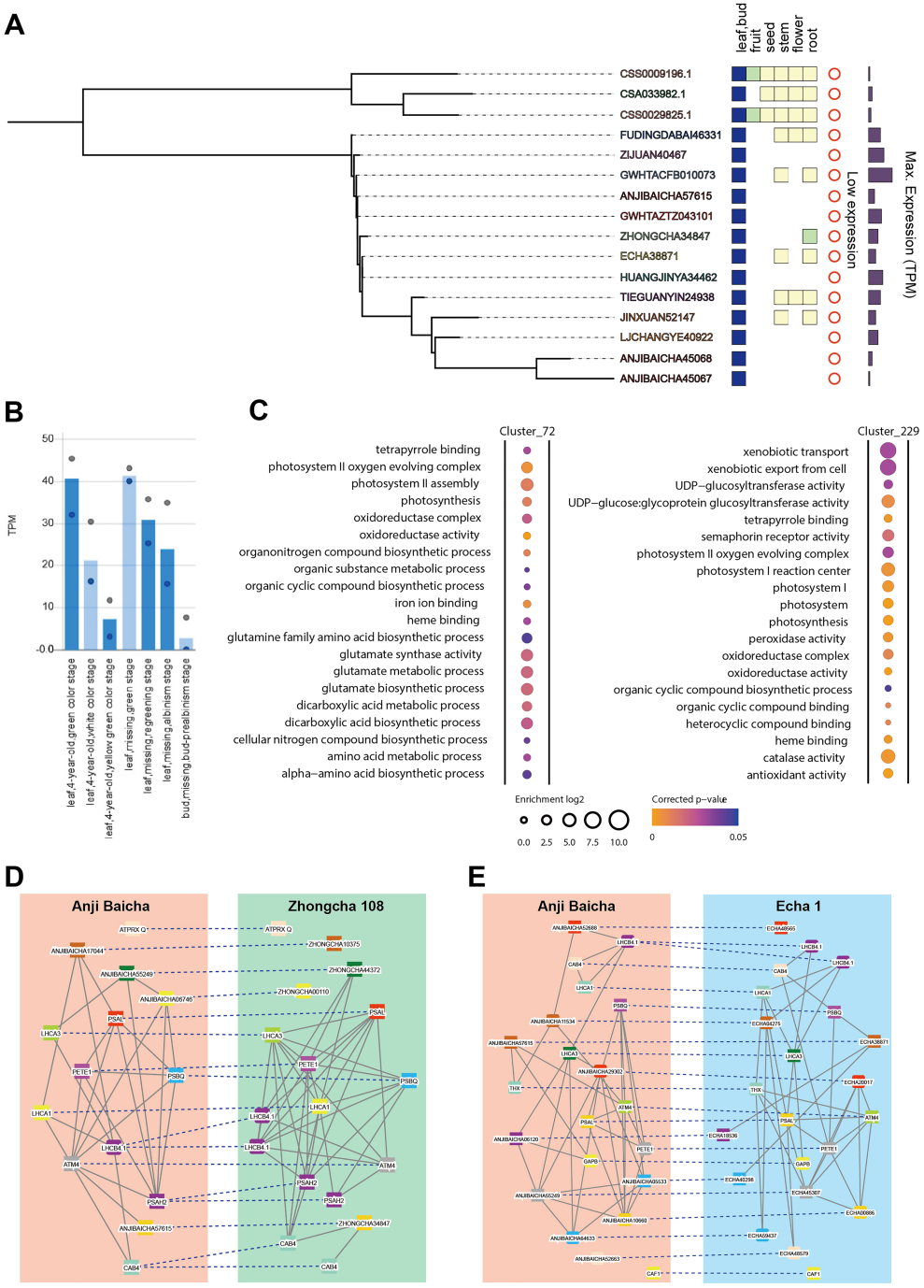

FIGURE 5 Biological function analysis of expanded genes in the orthogroup OG0010584 in the cultivar 'Anji Baicha'. (A) Phylogenetic tree and tissue-specific expression analysis of the orthogroup OG0010584. The heatmap shows the expression level in different tissues, full red dots show genes with low-expression, and the bar on the right indicates the maximum expression level (TPM). The color of a gene identifier indicates the cultivars. Note that OrthoFinder tree nodes do not contain bootstrap values, and should be interpreted with care. Missing data is indicated by an absent box. (B) Differential expression of *ANJIBAICHA57615* in the cultivar 'Anji Baicha' under different stages. (C) Gene Ontology (GO) enrichment analysis of clusters 72 and 229 in the cultivar 'Anji Baicha'. The size of the dots represents the Enrichment log2 value, while the color of the dots represents the corrected p-value. (D) Comparative analysis of the cluster 229 between 'Anji Baicha' and 'Zhongcha 108'. (E) Comparative analysis of the cluster 229 between 'Anji Baicha' and 'Echa 1'.

expanded OGs (Figure 2B and D). This gives us a direction to selectively study tea cultivars with significant differences.

Additionally, assembling the cultivar-specific transcriptomes provides us with an approach to study the phylogenetic relationships among tea cultivars. Such phylogenetic analysis can assist us in inferring the relationships between cultivars, enabling us to demonstrate the trends in differential formation during comparative analyses (Figure 2E, F and G).

To give access to the newly assembled transcripts and gene expression data of various tea cultivars, we constructed the TeaNekT database. The database provides a comparative analysis of gene expression profiles, enabling tea researchers to identify factors influencing the expression levels of a group of genes in tea plants. For example, we discovered that in the cultivar 'Jinxuan', most of the genes related to the flavonoid pathway are co-expressed in the HCCA cluster 126 (https://teanekt.sbs.ntu.edu.sg/cluster/view/7475). Therefore, we used TeaNekT to generate average expression profiles for the co-expression cluster 126 and the expression profiles of genes related to the flavonoid pathway in cluster 126 (https://teanekt.sbs.ntu.edu.sg/heatmap/cluster/7475) (Figure S2A and B). We observed that genes in cluster 126 exhibit similar expression patterns and are highly responsive to methyl jasmonate (MeJA) treatment, showing

significant upregulation in response to methyl jasmonate (MeJA) treatment.

The TeaNekT database is the first to employ the highest reciprocal rank (HRR) to measure co-expression relationships between genes and use the heuristic cluster chiseling algorithm (HCCA) to detect densely connected gene groups as co-expression clusters in tea research. This method can provide network partitions with relatively uniform cluster sizes and may reveal biological relationships that are not readily apparent from single-gene co-expression methods (Proost and Mutwil, 2018). Taking HCCA cluster 126 of 'Jinxuan' as an example, this co-expression cluster consists of 144 genes (https://teanekt.sbs.ntu.edu.sg/cluster/view/7475). The genes in cluster 126 are significantly enriched in gene ontology (GO) terms like oxidoreductase activity, UDP-glycosyltransferase activity, aromatic amino acid family biosynthetic process, organic cyclic compound biosynthetic process, and aromatic amino acid metabolic process, among others. With the assistance of TeaNekT, we were able to visualize the co-expression network for cluster 126, which provided us with a more detailed understanding of the functions of each gene within cluster 126 (https://teanekt.sbs.ntu.edu.sg/cluster/graph/7475) (Figure S2C).

TeaNekT allows for comparative studies of co-expression clusters within and between cultivars, providing insights into conserved co-expression relationships of clusters across different cultivars. For example, to investigate the conservation of co-expression relationships within the chalcone synthase (CHS) gene family, we used TeaNekT to compare cluster 126 of 'Jinxuan' with cluster 216 of 'Huangdan'. We observed that most of the gene families related to the flavonoid pathway in cluster 126 are conserved between two cultivars, including PAL, C4H, 4CL, CHS, F3H, and FLS gene families (https://teanekt.sbs.ntu.edu.sg/graph_comparison/cluster/2793/7475/1) (Figure S2D). A gene module refers to a subnetwork within a co-expression network composed of functionally related genes. TeaNekT uses expression context conservation (ECC) to identify gene modules that exhibit conservation under different cultivars. We observed that the gene *JINXUAN33459* in 'Jinxuan' has a homologous gene *GWHTAZTZ033250* with a high ECC score (0.18) in 'Huangdan'. By comparing the co-expression neighborhoods of two cinnamate 4-hydroxylase (C4H) genes, the conservation of the PAL, C4H, 4CL, CHS, F3H, ADT6, and TT7 gene families between two cultivars can be observed (Figure S2E).

By combining phylogenetic analysis of homologous genes with tissue expression profiles, the tool allows researchers to understand the phylogenetic relationships of homologous genes across different tea cultivars and the conservation of genes across tissues. The synthesis pathway of flavonoids can be divided into three steps: shikimic acid pathway, phenylpropanoid pathway, and flavonoid pathway (Figure S3A) (Shi et al., 2021). Among them, the latter two steps are particularly crucial for the synthesis of flavonoids. We investigated the encoding genes of key enzymes from the phenylpropanoid pathway to the flavonoid pathway and generated phylogenetic trees of these gene families as well as tissue expression profiles using TeaNekT (Figure S3B). We observed that, compared to the encoding genes of enzymes in the phenylpropanoid pathway, the encoding genes of enzymes in the flavonoid pathway exhibit specific expression in the leaves and buds of almost all cultivars. This may explain the abundant accumulation of flavonoids in the leaves and buds of tea plants (Zhang et al., 2017).

Past studies have shown that ICE1 is a bHLH transcription factor critical for plant responses to cold stress (Zhu, 2016). Under cold conditions, ICE1 is activated and its activity and stability are regulated through modification pathways such as sumoylation and polyubiquitylation (Chinnusamy et al., 2007). There is a regulatory relationship between ICE1 and C-repeat binding factors (CBFs), where ICE1 can regulate the expression of CBF genes, thereby affecting plant responses to cold stress and freezing tolerance (Zhu, 2016). SIZ1 is an E3 SUMO ligase that participates in ICE1 sumoylation (Chinnusamy et al., 2007). Under cold stress conditions, SIZ1 cooperates with ICE1 to regulate the expression of CBF genes, modulating plant responses to cold stress by influencing ICE1's stability and activity (Zhu, 2016). In this study, we found that homologous genes encoding the transcription factor ICE1 and E3 SUMO-protein ligase SIZ1, which are specifically expressed in the leaves and buds of tea plants, are significantly expanded in 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/tree/view/3915; https://teanekt.sbs.ntu.edu.sg/tree/view/7208). Moreover, we observed that these genes are highly expressed in the bud-prealbinism stage of 'Anji Baicha' leaves and buds in early spring, implying an association of the transcription factor ICE1 and E3 SUMO-protein ligase SIZ1 with the cold-sensitive characteristics of the low-temperature-sensitive albino tea plant cultivar 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/sequence/view/767800).

MKK2 is a component of the MAPK cascade pathway, located downstream of the MAP3K MEKK1 (Teige et al., 2004). It interacts with MAPK-related protein kinases such as MPK4 and MPK6, forming a MAPK cascade pathway that plays a crucial role in plant responses to cold stress (Teige et al., 2004). When subjected to cold or salt stress, MKK2 is activated and regulates the expression of COR genes, thereby modulating plant tolerance to freezing and salt stress (Zhu, 2016). In this study, we found that homologous genes encoding mitogen-activated protein kinase kinase 2 are significantly expanded in 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/tree/view/4279). Moreover, we observed that these genes are highly expressed throughout the entire process from the early spring green leaf stage to the end of the albino stage in 'Anji Baicha', which may also affect the cold sensitivity of 'Anji Baicha' (https://teanekt.sbs.ntu.edu.sg/sequence/view/775927).

Previous studies have found that 'Anji Baicha' albino leaves exhibit reduced chlorophyll content and impaired chloroplast and thylakoid development (Li et al., 2011). Extensive research has delved into pigment metabolism, with an increasing focus on chloroplast structural proteins, particularly LHCB complexes (Ye et al., 2023). However, in the tea plant, potential key genes directly associated with low-temperature response and albino phenotype have not been accurately identified (Ye et al., 2023). In this study, we identified a significantly expanded OG encoding regulatory factor RIQ in 'Anji Baicha', with genes in this OG showing significantly downregulated expression at the bud-prealbinism stage of 'Anji Baicha' (Figure 5A and B). The

regulatory factor RIQ is associated with the organization of chloroplast grana structure and photosystem II complexes, playing a crucial role in regulating granum structure and photosynthetic processes (Yokoyama et al., 2016). In plant cells, the RIQ protein plays a crucial role in connecting granum structures and organizing the photosystem II protein (LHCII), contributing significantly to granum formation and function (Yokoyama et al., 2016). In this study, the gene *ANJIBAICHA57615* encoding RIQ and genes encoding multiple LHCA and LHCB proteins exhibited a conserved co-expression relationship (Figure 5D and E).

Furthermore, in this study, we observed that the co-expression clusters 72 and 229 containing genes encoding RIQ were not only enriched in photosynthesis-related GO terms such as photosynthesis, photosystem II assembly, and photosystem II oxygen-evolving complex, but also enriched in nitrogen metabolism-related GO terms such as glutamine family amino acid biosynthetic process, amino acid metabolic process, and organonitrogen compound biosynthetic process (Figure 5C). This implies that the co-expression clusters 72 and 229 may link the albino mechanism of 'Anji Baicha' with its unique quality of high amino acid content.

5 | CONCLUSION

This study compiled the transcriptomes of 9 tea cultivars, yielding high-quality coding sequences. Through the annotation of coding sequences from 13 tea cultivars and the construction of co-expression networks using the CoNekT framework, TeaNekT was developed for comparative transcriptome studies among tea cultivars. The phylogenetic analysis of the 13 cultivars effectively reconstructed their relationships. Examination of orthologous groups (OGs) identified two categories of genes exhibiting variances among cultivars: cultivar-specific genes and expanded genes. Notably, cultivar-specific genes and expanded genes in 'Anji Baicha' are enriched in biological functions associated with responses to cold stress, chloroplast thylakoid structure, and nitrogen metabolism. Subsequent investigations unveiled three significantly expanded homologous genes in 'Anji Baicha' encoding the transcription factor ICE1, E3 SUMO-protein ligase SIZ1, and mitogen-activated protein kinase kinase 2, which may be closely linked to the low-temperature sensitivity of 'Anji Baicha'. Furthermore, a significantly expanded homologous gene in 'Anji Baicha' encoding the regulatory factor RIQ might play a critical role in the abnormal chloroplast structure and the absence of thylakoid membranes in 'Anji Baicha'.

AUTHOR CONTRIBUTIONS

X.Z. led the main work of this study, including project conception, data annotation, data analysis, and paper writing. P.K.L. provided X.Z. with suggestions on transcriptome assembly. Z.A. uploaded the transcriptome data to CoNekT to create TeaNekT. M.M. and Y.W. co-supervised X.Z. in completing this project. M.M. and P.K.L. both participated in the revision of the paper. The authors thank all members of the Mutwil Lab for their suggestions and assistance with this manuscript.

ACKNOWLEDGEMENTS

X.Z. is sponsored by a China Scholarship Council fellowship.

FUNDING INFORMATION

This project is funded by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C02041).

DATA AVAILABILITY STATEMENT

The raw RNA sequencing data used in this study are available from the NCBI Sequence Read Archive (SRA) under the accession numbers listed in Table S1. No sequencing data were generated during this study.

ORCID

Xinghai Zheng https://orcid.org/0009-0004-9016-4726
Marek Mutwil https://orcid.org/0000-0002-7848-0126

REFERENCES

Abbo, S., Berger, J., & Turner, N. C. (2003). Evolution of cultivated chickpea: four bottlenecks limit diversity and constrain adaptation. *Functional Plant Biology*, 30(10), 1081–1087.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), 525–527.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973.

Chen, L., Apostolides, Z., Chen, Z. M., Yao, M. Z., & Chen, L. (2012). *Tea Germplasm and Breeding in China*. Global tea breeding: Achievements, challenges and perspectives, 13–68.

Chen, L., Zhou, Z. X., & Yang, Y. J. (2007). Genetic improvement and breeding of tea plant (*Camellia sinensis*) in China: from individual selection to hybridization and molecular breeding. *Euphytica*, 154, 239–248.

Chen, R. (1981). The harvesting and processing techniques of "Huangjingui". *Collection*, (4).

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884-i890.

Chen, W. (1982). New tea cultivar - "Longjing 43". *Agricultural Science and Technology Communication*, 8.

Chen, X., Jia, S., Min, C., et al. (2010). Asexual reproduction technique of "Echa 1" tea tree. *Chinese Tea*, (8), 17–18.

Cheng, H., Li, S., Chen, M., et al. (1999). Physiological and biochemical nature of specific traits in "Anji Baicha". *Tea Science*, 19(2), 87–92.

Cheng, H., Li, S. F., Chen, M., Yu, F. L., Yan, J., Liu, Y., & Chen, L. G. (1999). Physiological and biochemical essence of the extraordinary characters of Anji Baicha. *Journal of Tea Science*, 19(2), 87–92.

Chinnusamy, V., Zhu, J., & Zhu, J. K. (2007). Cold stress regulation of gene expression in plants. *Trends in plant science*, 12(10), 444–451.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), 1792–1797.

Emerson, R. O., & Thomas, J. H. (2009). Adaptive evolution in zinc finger transcription factors. *PLoS genetics*, 5(1), e1000325.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20, 1–14.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … & Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8), 1494–1512.

Hahne, F., Huber, W., Gentleman, R., Falcon, S., Falcon, S., & Gentleman, R. (2008). Hypergeometric testing used for gene set enrichment analysis. *Bioconductor case studies*, 207–220.

Hansen, B. O., Vaid, N., Musialak-Lange, M., Janowski, M., & Mutwil, M. (2014). Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Frontiers in plant science*, 5, 93139.

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology*, 428(4), 726–731.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4), 357–360.

Kumari, M., Clarke, H. J., Small, I., & Siddique, K. H. (2009). Albinism in plants: a major bottleneck in wide hybridization, androgenesis and doubled haploid culture. *Critical Reviews in Plant Science*, 28(6), 393–409.

Lei, X., Wang, Y., Zhou, Y., Chen, Y., Chen, H., Zou, Z., … & Fang, W. (2021). TeaPGDB: tea plant genome database. *Beverage Plant Research*, 1(1), 1–12.

Liao, Y., Zhou, X., & Zeng, L. (2022). How does tea (*Camellia sinensis*) produce specialized metabolites which determine its unique quality and function: A review. *Critical Reviews in Food Science and Nutrition*, 62(14), 3751–3767.

Lim, J. J. J., Koh, J., Moo, J. R., Villanueva, E. M. F., Putri, D. A., Lim, Y. S., … & Mutwil, M. (2020). Fungi. guru: Comparative genomic and transcriptomic resource for the fungi kingdom. *Computational and Structural Biotechnology Journal*, 18, 3788–3795.

Lim, P. K., Davey, E. E., Wee, S., Seetoh, W. S., Goh, J. C., Zheng, X., … & Mutwil, M. (2022). Bacteria. guru: comparative transcriptomics and co-expression database for bacterial pathogens. *Journal of Molecular Biology*, 434(11), 167380.

Li, Q., Huang, J., Liu, S., Li, J., Yang, X., Liu, Y., & Liu, Z. (2011). Proteomic analysis of young leaves at three developmental stages in an albino tea cultivar. *Proteome science*, 9, 1–12.

Li, S. F. (2002). Studies on the mechanism of the leaf color change in Anjibaicha (*Camellia sinensis*). *Journal of China Institute of Metrology*, 13, 214–217.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., … & Usadel, B. (2014). M ercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment*, 37, 1250–1258.

Mo, X. (2011). Innovative utilization of "Jinxuan" tea in Vietnamese tea areas. *Chinese Tropical Agriculture*, (6), 28–30.

Mutwil, M., Usadel, B., Schuˇstte, M., Loraine, A., Ebenhoˇsh, O., & Persson, S. (2010). Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant physiology*, 152(1), 29–43.

Nazer, E., Dale, R. K., Palmer, C., & Lei, E. P. (2018). Argonaute2 attenuates active transcription by limiting RNA Polymerase II elongation in *Drosophila melanogaster*. *Scientific reports*, 8(1), 15685.

Ng, J. W. X., Tan, Q. W., Ferrari, C., & Mutwil, M. (2020). Diurnal. plant. tools: comparative transcriptomic and co-expression analyses of diurnal gene expression of the Archaeplastida kingdom. *Plant and Cell Physiology*, 61(1), 212–220.

Parrott, W. A., & Smith, R. R. (1986). Evidence for the existence of endosperm balance number in the true clovers (*Trifolium* spp.). *Canadian journal of genetics and cytology*, 28(4), 581–586.

Proost, S., & Mutwil, M. (2018). CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic acids research*, 46(W1), W133-W140.

Qiu, X. (1965). Economic value and regional adaptability of "Fuding Dabaicha". *Tea Science*, 2(04), 10–14.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic acids research*, 33(suppl_2), W116-W120.

Ren, F. J., Wang, W. L., & Yang, G. Q. (2015). The major problem facing the development of An'ji white tea industry and its countermeasures. *Journal of Huzhou University*, 37(9), 10–13.

Shi, Y., Jiang, X., Chen, L., Li, W. W., Lai, S., Fu, Z., … & Xia, T. (2021). Functional analyses of flavonol synthase genes from *Camellia sinensis* reveal their roles in anther development. *Frontiers in plant science*, 12, 753131.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.

Tai, Y., Liu, C., Yu, S., Yang, H., Sun, J., Guo, C., … & Wan, X. (2018). Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (*Camellia sinensis*). *Bmc Genomics*, 19, 1–13.

Tan, Q. W., & Mutwil, M. (2020). Malaria. tools—comparative genomic and transcriptomic database for *Plasmodium* species. *Nucleic acids research*, 48(D1), D768-D775.

Teige, M., Scheikl, E., Eulgem, T., Doczi, R., Ichimura, K., Shinozaki, K., … & Hirt, H. (2004). The MKK2 pathway mediates cold and salt stress signaling in *Arabidopsis*. *Molecular cell*, 15(1), 141–152.

Tian, Y., Xu, P., Zhu, X. (2011). Application and promotion of national excellent tea cultivar "Yunkang 10" in Yunnan Province. *Modern Agricultural Science and Technology*, (24), 118–119.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., … & Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, cell & environment*, 32(12), 1633–1651.

Villanueva, E. M. F., Lim, P. K., Lim, J. J. J., Lim, S. C., Lau, P. Y., Koh, K. T. S., … & Mutwil, M. (2022). Protist. guru: a comparative transcriptomics database for protists. *Journal of Molecular Biology*, 434(11), 167502.

Wang, C., Han, J., Pu, Y., & Wang, X. (2022). Tea (*Camellia sinensis*): a review of nutritional composition, potential applications, and Omics Research. *Applied Sciences*, 12(12), 5874.

Wang, K., Li, M., Liang, Y., et al. (2008). Research on the breeding of new tea cultivar "Huangjinya". *Chinese Tea*, (4), 21–23.

Wang, Z. (2004). Introduction to Anxi "Tieguanyin". *Tea Science and Technology*, (4), 28–30.

Xia, E. H., Li, F. D., Tong, W., Li, P. H., Wu, Q., Zhao, H. J., … & Wan, X. C. (2019). Tea plant information archive: a comprehensive genomics and bioinformatics platform for tea plant. *Plant biotechnology journal*, 17(10), 1938–1953.

Xia, E. H., Tong, W., Wu, Q., Wei, S., Zhao, J., Zhang, Z. Z., … & Wan, X. C. (2020). Tea plant genomics: achievements, challenges and perspectives. *Horticulture research*, 7.

Xia, X. (2000). Excellent tea cultivar - "Shuchazao". *Anhui Agriculture*, 3, 17.

Yang, S., Wang, Y., Yang, Y., et al. (1995). Breeding of early high-quality fragrant green tea cultivar "Longjing Changye". *Chinese Tea*, 17(6), 14–16.

Yang, X., Tian, Y., Huang, M., et al. (2013). Breeding and application of national plant protection cultivar "Zijuan" tea tree. *Hunan Agricultural Science*, (6), 3.

Yang, Y., Yang, S., Yang, Y., et al. (2003). Research on the breeding and application of a new high-quality early green tea cultivar - "Zhongcha 108". *Chinese Tea*, 25(2), 12–14.

Ye, J. J., Lin, X. Y., Yang, Z. X., Wang, Y. Q., Liang, Y. R., Wang, K. R., … & Zheng, X. Q. (2023). The light-harvesting chlorophyll a/b-binding proteins of photosystem II family members are responsible for temperature sensitivity and leaf color phenotype in albino tea plant. *Journal of Advanced Research*.

Yokoyama, R., Yamamoto, H., Kondo, M., Takeda, S., Ifuku, K., Fukao, Y., … & Shikanai, T. (2016). Grana-localized proteins, RIQ1 and RIQ2, affect the organization of light-harvesting complex II and grana stacking in *Arabidopsis*. *The Plant Cell*, 28(9), 2261–2275.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, 19, 15–30.

Zhang, C., Wang, M., Gao, X., Zhou, F., Shen, C., & Liu, Z. (2020). Multi-omics research in albino tea plants: past, present, and future. *Scientia Horticulturae*, 261, 108943.

Zhang, Q., Liu, M., & Ruan, J. (2017). Metabolomics analysis reveals the metabolic and functional roles of flavonoids in light-sensitive tea leaves. *BMC plant biology*, 17, 1–10.

Zhang, R., Ma, Y., Hu, X., Chen, Y., He, X., Wang, P., … & Zhang, S. (2020). TeaCoN: a database of gene co-expression network for tea plant (*Camellia sinensis*). *BMC genomics*, 21, 1–9.

Zhao, S., Cheng, H., Xu, P., & Wang, Y. (2023). Regulation of biosynthesis of the main flavor-contributing metabolites in tea plant (*Camellia sinensis*): A review. *Critical Reviews in Food Science and Nutrition*, 63(30), 10520–10535.

Zhao, Z., & Ma, D. (2021). Genome-wide identification, characterization and function analysis of lineage-specific genes in the tea plant *Camellia sinensis*. *Frontiers in Genetics*, 12, 770570.

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., … & Fei, Z. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular plant*, 9(12), 1667–1670.

Zhou, X., Kao, M. C. J., & Wong, W. H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20), 12783–12788.

Zhu, J. K. (2016). Abiotic stress signaling and responses in plants. *Cell*, 167(2), 313–324.

Zou, G., Xiao, Y., Wang, M., & Zhang, H. (2018). Detection of bitterness and astringency of green tea with different taste by electronic nose and tongue. *PLoS One*, 13(12), e0206517.

SUPPORTING INFORMATION

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

How to cite this article: Zheng, X., Ali, Z.M., Lim, P.K., Mutwil, M. & Wang, Y. (2024) Comparative transcriptome database for *Camellia sinensis* reveals genes related to the cold sensitivity and albino mechanism of 'Anji Baicha'. *Physiologia Plantarum*, 176(4), e14474. Available from: https://doi.org/10.1111/ppl.14474