https://doi.org/10.17221/116/2025-CJGPB

Whole genome sequencing analysis of two sinensis tea (*Camellia sinensis* var. *sinensis*) clones: Assessment of molecular variations to search for breeding markers

Dwinita Wikan Utami[1], Adhityo Wicaksono[2], M. Khais Prayoga[3], Heri Syahrian[3], Vitria P. Rahadi[3], Erdiansyah Rezamela[3], Budi Martono[4], Tri Joko Santoso[1], Nur Kholilatul Izzah[4], Haris Maulana[1], Ady Daryanto[1], Rerenstradika Tizar Terryana[5], Imas Rita Saadah[1], David Virya Chen[2]*

[1]*Research Center for Horticulture, Research Organization for Agriculture and Food, National Research and Innovation Agency (BRIN), Cibinong, Bogor, West Java Province, Indonesia*
[2]*Scientific Department, Genomik Solidaritas Indonesia (GSI Lab), Jakarta, Indonesia*
[3]*Research Institute for Tea and Cinchona, Bandung, Indonesia*
[4]*Research Center for Estate Crops, Research Organization for Agriculture and Food, National Research and Innovation Agency (BRIN), Cibinong, Bogor, West Java Province, Indonesia*
[5]*Research Center for Applied Botany, Research Organization for Life Sciences and Environment, National Research and Innovation Agency (BRIN), Cibinong, Bogor, West Java Province, Indonesia*

**Corresponding author: dwin011@brin.go.id*

Citation: Utami D.W., Wicaksono A., Prayoga M.K., Syahrian H., Rahadi V.P., Rezamela E., Martono B., Santoso T.J., Izzah N.K., Maulana H., Daryanto A., Terryana R.T., Saadah I.R., Chen D.V. (2026): Whole genome sequencing analysis of two sinensis tea (*Camellia sinensis* var. *sinensis*) clones: Assessment of molecular variations to search for breeding markers. Czech J. Genet. Plant. Breed., 62: 76–88.

Abstract: Tea (*Camellia sinensis* (L.) Kuntze) is a globally important crop valued for its flavour diversity and health benefits. Whole-genome sequencing (WGS) was performed to compare genomic variation and functional potential between clone Yabukita and locally adapted clone I.1.93. Using next-generation sequencing, approximately 10× genome coverage was achieved for both clones, with high mapping efficiency (98.24% for Yabukita and 97.88% for clone I.1.93), ensuring reliable downstream analyses. Single nucleotide polymorphism (SNP) analysis revealed distinct genomic patterns, with Yabukita showing a more uniform chromosomal SNP distribution, while clone I.1.93 exhibited higher SNP densities on specific chromosomes, particularly chromosomes 5 and 13. Silent mutations predominated in Yabukita (48.21%), whereas missense mutations were more frequent in clone I.1.93 (57.97%), suggesting greater functional divergence. Most SNPs occurred in non-coding regions, indicating potential regulatory roles. GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses revealed highly similar shared pathways, including photosynthesis and protein interactions, alongside clone-specific enrichment related to photosynthesis in Yabukita and stress responses in clone I.1.93. miRNA profiling identified distinct regulatory patterns, including the clone-specific miR530 in clone I.1.93. Biosynthetic gene cluster analysis further predicted secondary metabolite pathways associated with terpenoid, polyketide, and saccharide biosynthesis. These findings provide valuable genomic insights for tea improvement and breeding programs.

Keywords: comparative genomics; next-generation sequencing; single-nucleotide polymorphism; tea breeding

Supported through RIIM competition funding from the Indonesia Endowment Fund for Education (LPDP) under the Ministry of Finance of the Republic of Indonesia and National Research and Innovation Agency of Indonesia according to the contract number: 37/II.7/HK/2023 and the Visiting Researcher scheme (Contract number: 64/II/HK/2022).

https://doi.org/10.17221/116/2025-CJGPB

Tea (*Camellia sinensis* (L.) Kuntze) is widely recognised for its diverse flavour profiles and associated health benefits, making it one of the most important beverages worldwide. Two major tea varieties, Assamica tea (*C. sinensis* var. *assamica*) and Sinensis tea (*C. sinensis* var. *sinensis*), are extensively cultivated in Indonesia (Martono & Syafaruddin 2018). Tea flavour, a critical determinant of quality and global market value, is influenced by genetic variability, environmental conditions, and agricultural practices. Among these factors, genetic variation plays a fundamental role in determining flavour-related traits. To assess genetic variability among tea clones, modern genomic approaches such as whole-genome sequencing (WGS) provide a powerful tool for elucidating the molecular basis underlying distinct tea characteristics.

Genomics assessment through WGS enables a comprehensive understanding and precise profiling of distinct groups. However, tea possesses a relatively large genome (~3.0 Gb), which makes genome-wide variant discovery time-consuming and costly (An et al. 2022). With the advances in sequencing technologies, particularly long-read sequencing, have improved the characterisation and comparison of plant genetic profiles, including the identification of single-nucleotide polymorphisms (SNPs) through variant calling. Unlike short-read sequencing, long-read sequencing enables more accurate resolution of recently duplicated and highly repetitive genomic regions and reduces biases associated with extreme guanine and cytosine (GC) content, thereby improving variant discovery in complex genomes (Zverinova & Guryev 2022). Long-read technologies, including Oxford Nanopore Technology (ONT), have become increasingly powerful for comprehensive genome characterisation, structural variant detection, and haplotype phasing, advancing beyond the limitations of short reads (Gupta et al. 2024). When applied to plant genomes, long reads facilitate the identification of genome-wide SNPs within genic and intergenic regions following annotation, supporting detailed genetic profiling. These SNP variants represent valuable markers for crop breeding and the development of desirable agronomic traits (Liu et al. 2025).

The tea cultivar Yabukita (*Camellia sinensis* var. *sinensis* cv. Yabukita) has been known as a Japanese tea cultivar recognised for its distinct flavour and aroma (Yagi et al. 2010). Owing to its extensive use and well-established genetic background, Yabukita is commonly employed as a reference or control genotype in genetic and genomic studies of tea. Besides, the Indonesian tea clone I.1.93, developed by the Indonesian Research Institute for Tea and Cinchona (IRITIC), represents a high-performing sinensis variety known for its efficient nutrient utilisation (Prayoga et al. 2022). Given the distinct genetic backgrounds of the two clones, this study aimed to investigate genomic differentiation between Yabukita (used as the control) and clone I.1.93 through genome-wide SNP profiling and comprehensive genomic annotation, including gene ontology (GO) categories, metabolic pathways, secondary metabolite gene clusters, and predicted miRNAs. By integrating these datasets, this study seeks to uncover key markers for plant breeding. Such markers may be leveraged to improve future breeding efforts focused on enhancing tea quality and environmental resilience.

MATERIAL AND METHODS

Plant material. Two distinct tea clones, Yabukita and locally developed clone I.1.93, were selected for this study based on their contrasting flavour profiles and nitrogen use efficiency (Figure 1). Healthy, mature leaves were collected from plants grown under similar environmental conditions to minimise variability. According to Prayoga et al. (2022), Yabukita exhibits a lower annual production yield (2.92 kg/ha/year) compared with clone I.1.93 (3.42 kg/ha/year). Consistent with this difference, clone I.1.93 demonstrates considerably higher nitrogen use efficiency, exceeding that of Yabukita by more than twofold (64.10% versus 28.72%). In contrast, sensory quality assessments indicate that Yabukita possesses a superior flavour profile. When processed using both panning and steaming methods and evaluated by expert panellists, Yabukita achieved a higher sensory score (83.00) than clone I.1.93 (80.63).

Yabukita is a Japanese-origin tea clone introduced to Indonesia, where it has adapted well and is widely used as a benchmark for sinensis-type tea varieties, making it a suitable control in comparative studies. Meanwhile, clone I.1.93 is an Indonesian tea clone developed by the Research Institute for Tea and Cinchona, originating from germplasm exploration in the Pasirsarongge area of Cianjur Regency (Prayoga et al. 2022).

Whole genome sequencing (WGS). The whole genome sequencing of two tea clones was performed using the ONT sequencing platform. High-molecular-weight genomic DNA was extracted from approximately 6–8 g of fresh leaf tissue per sample using the

https://doi.org/10.17221/116/2025-CJGPB

CTAB protocol with minor modifications (Doyle & Doyle 1987). The purity, concentration, and integrity of the extracted DNA were assessed using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Inc., Wilmington, USA) to measure their A260/A230 and A260/A280 ratios, as well as 1% agarose gel electrophoresis. The extracted DNA was subsequently prepared for long-read sequencing using the protocol provided in the SQK-LSK110 genomic sequencing kit (Oxford Nanopore Technologies, Oxford, UK). Quality control and processing were conducted using MinKnow v23.11.7 (Oxford Nanopore Technologies) with basecaller Dorado v7.2.13 (Oxford Nanopore Technologies). This long-read sequencing approach provides valuable genetic insights that can help elucidate the underlying mechanisms driving the differences observed between the two tea varieties.

Genomic assembly and annotation. Following the sequencing process, the resulting FASTQ files of both clones' reads were processed on both stand-alone computers and Galaxy Europe (https://usegalaxy.eu) (Community 2024). The reads were first mapped to the tea reference genome *Camellia sinensis* genome assembly HZAU_G240_1.0 (GenBank ID: GCA_013676235.1) (sequenced by PacBio, assembled to chromosome level, submitted by Huazhong Agricultural University, China) using minimap2 (Li 2018) default alignment settings to generate the binary alignment map (BAM) files. The variant call format (VCF) files that list the SNP within the genome were called from the BAM files, and the consensus FASTA files that were built out of the reference genome were generated using BCFtools v1.15.1 (Li et al. 2009). The consensus FASTA genome files of both clones were processed to mask the repeat elements by RepeatMasker (Ver. 4.1.5, 2024) (with curated repeat dataset from Dfam) to reduce the potential annotation bias caused by the abundance of repeat elements in the genome. To structurally annotate the genomes, Augustus v3.4.0 (Stanke & Morgenstern 2005) was used. First, the reference genome FASTA file and the general feature format (GFF) file were used to train Augustus' algorithm. Then, both genomes were annotated to obtain the coding sequences (CDSs), predicted peptide sequences, and GFF files. To date, the tea genome (including the reference used here), despite being annotated, has most of the annotated proteins listed as hypothetical proteins. Hence, functional annotation of the genomes was conducted using UniProtKB-Swiss-Prot (update March 2023) database on *Arabidopsis thaliana* annotated proteins with Double Index Alignment of Next-Generation Sequencing Data (DIAMOND) Ver. 2.0.15 BLAST feature (Buchfink et al. 2014).

Downstream bioinformatics analysis. After the functional annotations of the genomes, the SNP files in VCF and the list of annotations were processed

Figure 1. Tea clones used in this study: Yabukita (A, B) and clone I.1.93 (C, D)

https://doi.org/10.17221/116/2025-CJGPB

for downstream analysis. The detected SNPs were then annotated by using SNP Eff v5.2 (Cingolani et al. 2012) to characterise the variants, including change rate per chromosome, percentage of effects by functional classes, and the variant-affected regions. The list of functionally annotated genes of both genomes with UniProtKB accession IDs were used as input to Database for Annotation, Visualization, and Integrated Discovery (DAVID) web server (https://david.ncifcrf.gov) (Huang et al. 2009) to obtain GO terms enrichment of the annotated genes, comprising biological process (BP), cellular component (CC), and metabolic function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment. The genome-wide miRNAs of both genomes were detected using INFERence of RNA ALignment (INFERNAL) v1.1.4 (Nawrocki & Eddy 2013), manually sorted only with miRNA on the list as a file of covariant model (CM) format and ran under CM-Search program using "trusted cutoff" (TC) setting, the lowest among other cutoffs, (e.g., noise cutoff or NC and gathering cutoff or GC) to generate false positive results from the covariant model dataset and the results were ensured to have a E-value lower than 10^{-5} (1E-15). The CM for miRNA prediction was obtained from the Rfam database v15 (https://rfam.org), and the miRNA dataset was manually curated by sorting it from other RNAs. Lastly, plant secondary metabolite gene clusters were detected using the PlantiSMASH web server (http://plantismash.secondarymetabolites.org) (Kautsar et al. 2017).

RESULTS

Whole genome sequencing data. The WGS was conducted using next-generation sequencing (NGS) technology, producing a substantial dataset for the tea Yabukita clone (control) and clone I.1.93. As the results, 10× coverage of genome data was produced. The sequencing process generated a significant amount of raw sequence data, which underwent rigorous quality control to ensure the reliability of downstream analyses (Table 1).

Table 1. The sequencing results of both tea samples

Sequencing results	Samples	
	Yabukita	I.193
No. of reads generated	5 868 053	8 255 326
No. of bases generated	31 466 603 166	30 677 811 256
Average read length	5 362.4	3 716.1
N50	12 130	12 635

N50 – minimum length of scaffolds in which (and longer) exactly half of the sequenced bases are assembled

Mapping statistics indicated high efficiency for both clones (Table 2), with 98.24% of Yabukita reads and 97.88% of clone I.1.93 reads mapped to the reference genome. Primary mapped reads constituted 93.18% and 92.30% of the total for Yabukita and clone I.1.93, respectively. Notably, duplicate reads were undetectable, ensuring high reliability for downstream analyses.

SNP variation analysis. SNP analysis revealed distinct genomic variations between the two clones. Chromosome-level SNP distributions highlighted that Yabukita showed a more uniform SNP distribution, while clone I.1.93 had denser SNP occurrences on specific chromosomes (e.g., Chr. 5 and 13) (Figure 2). Both clones exhibited shared SNP rates in chromosomes 3, 6, and 12.

Functionally, the SNPs were categorised into silent, missense, and nonsense mutations. Silent mutations dominated in Yabukita (48.21%), while missense mutations were prevalent in clone I.1.93 (57.97%), potentially influencing functional traits. The distribution of SNPs across genomic features

Table 2. Genome mapping profiles of the two tea clones

Metric	Assembly scoring values	
	Yabukita (control)	clone I.1.93
Total (QC-passed reads + QC-failed reads)	19 431 511 + 0	30 027 230 + 0
Primary	5 024 855 + 0	8 255 326 + 0
Secondary	10 880 121 + 0	17 587 965 + 0
Duplicates	0 + 0	0 + 0
Primary duplicates	0 + 0	0 + 0
Mapped	19 088 963 + 0 (98.24%: N/A)	29 391 327 + 0 (97.88%: N/A)
Primary mapped	4 682 307 + 0 (93.18%: N/A)	7 619 423 + 0 (92.30%: N/A)

https://doi.org/10.17221/116/2025-CJGPB

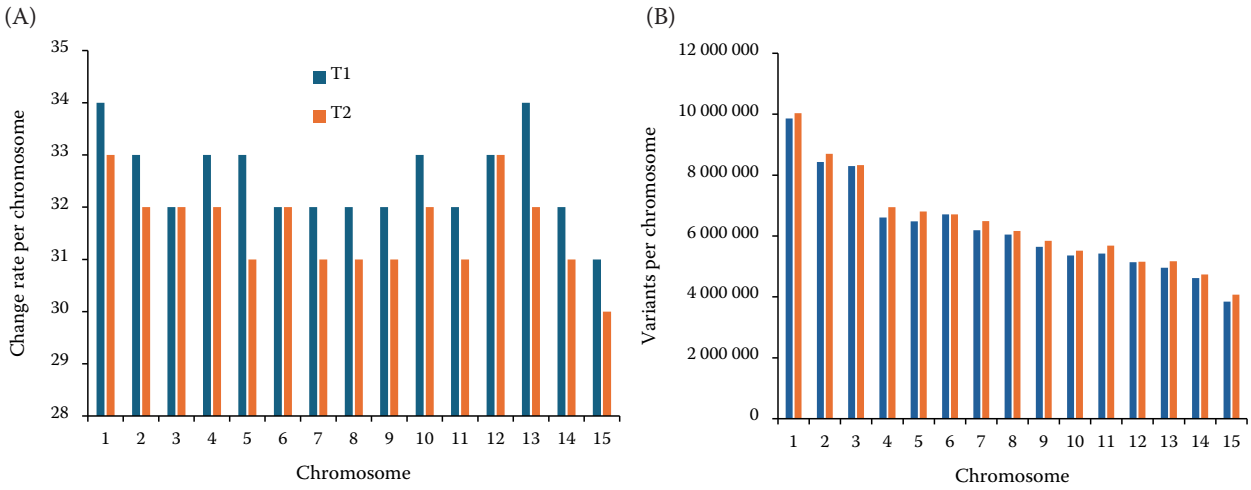
(upstream, introns, and downstream regions) was consistent between the clones, emphasising their presence in non-coding regions that might impact regulatory functions.

Functional annotation of SNPs: pathway and process analysis. GO and KEGG pathway analyses revealed biological processes, cellular components, and molecular functions associated with SNPs in both clones (Figure 3). Yabukita was enriched in photosynthesis-related processes, while clone I.1.93 demonstrated adaptations to environmental stresses, including salt stress and seed dormancy.

The comparative analysis of the BP, CC, MF, and KEGG pathways associated with SNPs in both tea clones reveals distinct differences and shared characteristics that could influence their phenotypic traits. The results showed that the shared features on both clones are highly similar to each other, as indicated by the very narrow differences on the bar plots (Figure 3).

From the perspective of non-shared/unique features of the genes on each genome, the GO of both clones are highly distinctive to each other, as shown in the unique genes GO BP (Table 3), CC (Table 4), MF (Table 5), and KEGG pathways (Table 6). Shared pathways included photosynthesis and protein interactions, with unique pathways such as carotenoid biosynthesis (Yabukita) and folate biosynthesis (clone I.1.93).

miRNA profiling. The miRNA profiling is a prediction-based, crucial procedure for understanding the genetic regulation underlying various biological processes in plants. In this analysis, genomic sequencing data for tea clones Yabukita and clone I.1.93 reveal significant differences in predicted miRNA patterns, which may impact stress resilience and overall plant productivity. miRNA sequencing revealed differences



(C)	T1	T2
Silent	40.23	40.42
Nonsense	1.66	1.60
Missense	58.11	57.99

(D)	T1	T2
Upstream gene variant	2.10	2.11
Intron variant	4.67	4.70
Intergenic region	93.24	93.19

Figure 2. Comparative single nucleotide polymorphism (SNP) profiles between two tea plant cultivars: distribution of mutational change rates across 15 chromosomes (A), amounts of SNP found across 15 chromosomes (B), relative proportion of SNP functional classes, including missense, nonsense, and silent mutations (C), distribution of SNPs within the top 3 genomic regions (intergenic, intron, and upstream) (D)
T1 – *Camellia sinensis* var. Yabukita (control); T2 – *C. sinensis* clone I.1.93

https://doi.org/10.17221/116/2025-CJGPB

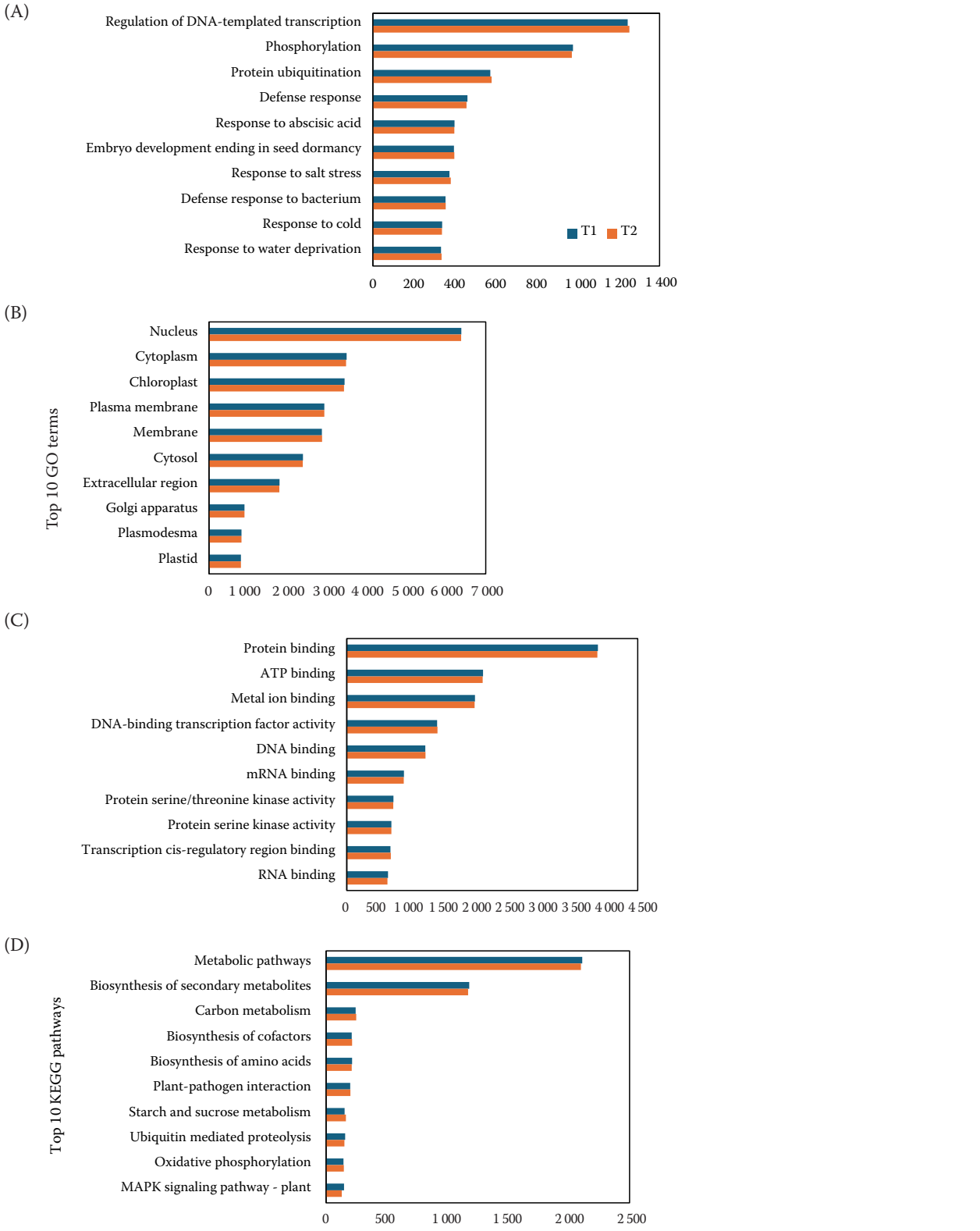


Figure 3. Functional enrichment and pathway analysis of shared genes; gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were annotated using the UniProt database based on *Arabidopsis thaliana* protein orthologs: biological processes (A), cellular components (B), molecular functions (C), and KEGG pathways (D)
T1 – *Camellia sinensis* var. Yabukita (control); T2 – *C. sinensis* clone I.1.93; the categories shown represent significant enrichments found in both tea variants

https://doi.org/10.17221/116/2025-CJGPB

Table 3. Gene ontology terms on biological processes generated by DAVID Bioinformatics (NCBI) using UniProt protein *Arabidopsis thaliana* annotation that are uniquely found in both tea genomes

Term (T1)	Count	*P*-value	Benjamini FDR	Term (T2)	Count	*P*-value	Benjamini FDR
Pollen tube guidance	11	9.30E-10	4.30E-07	embryo development ending in seed dormancy	13	4.20E-03	1.70E-01
Chloroplast organization	10	1.60E-04	1.00E-02	response to salt stress	11	1.20E-02	3.40E-01
Intracellular protein transport	10	5.70E-04	3.00E-02	response to abscisic acid	10	4.60E-02	5.70E-01
Response to oxidative tress	10	8.30E-03	2.00E-01	defense response to bacterium	8	9.70E-02	7.70E-01
Protein transport	10	2.20E-02	3.50E-01	chloroplast organization	7	8.20E-03	3.10E-01
Photosynthesis	9	1.00E-04	7.80E-03	response to heat	7	2.10E-02	3.40E-01
mRNA splicing, via spliceosome	9	1.10E-03	4.10E-02	seed development	6	1.20E-03	9.30E-02
Lipid catabolic process	8	4.80E-03	1.30E-01	plant-type hypersensitive response	6	3.50E-03	1.60E-01
DNA repair	7	3.50E-02	4.10E-01	mRNA processing	6	5.70E-02	6.40E-01
Cell differentiation	6	6.30E-02	5.00E-01	sucrose biosynthetic process	5	2.80E-05	6.30E-03

T1 – control (Yabukita); T2 – clone I.1.93; DAVID – database for annotation, visualization, and integrated discovery; FDR – false discovery rate

in expression levels and profiles between tea clones Yabukita and clone I.1.93 (Table 7).

The data indicate that several miRNAs are expressed at significantly higher levels in clone I.1.93 compared to Yabukita. Notably, miR530 stands out as a unique miRNA in clone I.1.93. This finding is particularly relevant, as miR530 is known to play a role in regulating genes associated with disease resistance and enhancing yield, making it vital for tea plants facing environmental challenges.

Gene clusters and secondary metabolites. The analysis of gene clusters in tea plants reveals a diverse

Table 4. Gene ontology terms on cellular components generated by DAVID Bioinformatics (NCBI) using UniProt protein *Arabidopsis thaliana* annotations that are uniquely found in both tea genomes

Term (T1)	Count	*P*-value	Benjamini FDR	Term (T2)	Count	*P*-value	Benjamini FDR
Chloroplast	106	5.80E-13	9.60E-11	chloroplast	90	3.10E-10	4.60E-08
Mitochondrion	65	9.30E-03	8.60E-02	mitochondrion	65	3.70E-04	1.40E-02
Cytosol	45	1.20E-02	1.00E-01	cytosol	47	3.20E-04	1.40E-02
Chloroplast stroma	26	5.00E-07	2.10E-05	chloroplast thylakoid membrane	19	6.70E-07	5.00E-05
Chloroplast thylakoid membrane	23	1.00E-08	8.60E-07	plastid	16	5.20E-02	3.20E-01
Plastid	21	5.70E-03	5.90E-02	chloroplast envelope	14	4.10E-03	6.00E-02
Chloroplast envelope	19	7.20E-05	1.50E-03	chloroplast stroma	13	4.90E-02	3.20E-01
Chloroplast thylakoid	13	1.70E-05	4.60E-04	thylakoid	10	5.30E-04	1.50E-02
Endomembrane system	9	3.70E-04	6.10E-03	chloroplast thylakoid	10	6.10E-04	1.50E-02
Pollen tube tip	8	2.60E-08	1.50E-06	chloroplast membrane	7	4.00E-03	6.00E-02

T1 – Control (Yabukita); T2 – clone I.1.93; DAVID – database for annotation, visualization, and integrated discovery; FDR – false discovery rate

https://doi.org/10.17221/116/2025-CJGPB

Table 5. Gene ontology terms on metabolic functions generated by DAVID Bioinformatics (NCBI) using UniProt protein *Arabidopsis thaliana* annotations that are uniquely found in both tea genomes

Term (T1)	Count	*P*-value	Benjamini FDR	Term (T2)	Count	*P*-value	Benjamini FDR
Protein binding	75	1.40E-05	1.10E-03	protein binding	64	1.70E-04	4.70E-03
Metal ion binding	33	9.70E-02	4.30E-01	mRNA binding	16	9.40E-02	6.00E-01
mRNA binding	19	4.70E-02	2.80E-01	zinc ion binding	12	8.60E-02	6.00E-01
RNA binding	18	1.20E-02	1.70E-01	rRNA binding	9	1.00E-05	6.20E-04
GTP binding	10	5.50E-03	1.10E-01	magnesium ion binding	8	4.80E-04	9.60E-03
rRNA binding	9	2.60E-05	1.20E-03	serine-type endopeptidase activity	6	6.80E-03	1.20E-01
Triglyceride lipase activity	8	5.90E-09	1.40E-06	sucrose-phosphate phosphatase activity	5	3.10E-08	6.90E-06
GTPase activity	8	1.00E-02	1.70E-01	protein heterodimerization activity	5	4.60E-02	4.20E-01
4 Iron, 4 sulfur cluster binding	5	7.90E-03	1.40E-01	thiamine binding	4	2.80E-06	3.10E-04
Lipoate synthase activity	4	1.00E-05	1.10E-03	thiamine diphosphokinase activity	4	1.40E-05	6.20E-04

T1 – Control (Yabukita); T2 – clone I.1.93; DAVID – database for annotation, visualization, and integrated discovery; FDR – false discovery rate

array of gene types distributed across different chromosomes, highlighting the complexity of metabolic pathways that contribute to the plant's unique characteristics (Table 8–9). The exploration of gene clusters associated with secondary metabolites in tea plants reveals a rich tapestry of biosynthetic pathways that contribute to the plant's unique biochemical properties. Secondary metabolites are crucial for the plant's interaction with its environment, providing defence against pests, diseases, and abiotic stresses, while also enhancing sensory qualities like flavour and aroma.

The data identifies nine distinct gene clusters, primarily located on chromosomes 3, 6, 7, and 12. These clusters encompass various types of secondary metabolite biosynthesis, including saccharides, lignans, terpenes, and putative compounds, each playing a significant role in the plant's overall health and product quality.

The PlantiSMASH analysis predicted several biosynthetic gene clusters (BGCs) associated with terpenoid, polyketide, and saccharide metabolism in both clones (Tables 7 and 8). These clusters are of significant interest, given the well-established roles of terpenoids and polyketides in the biosynthesis of plant aroma and flavour compounds. However, it is crucial to note that these annotations are derived from *in silico* pre-

Table 6. Kyoto Encyclopedia of Genes and Genomes terms generated by DAVID Bioinformatics (NCBI) using UniProt protein *Arabidopsis thaliana* annotations that are uniquely found in both tea genomes

Term (T1)	Count	*P*-value	Benjamini FDR	Term (T2)	Count	*P*-value	Benjamini FDR
Photosynthesis	7	0.0014	0.094	biosynthesis of cofactors	8	0.025	0.4
Homologous recombination	5	0.016	0.52	base excision repair	6	0.00034	0.016
Carotenoid biosynthesis	3	0.079	1	folate biosynthesis	4	0.0058	0.14
				photosynthesis	4	0.069	0.65
				mismatch repair	3	0.059	0.65

T1 – Control (Yabukita); T2 – clone I.1.93; DAVID – database for annotation, visualization, and integrated discovery; FDR – false discovery rate

https://doi.org/10.17221/116/2025-CJGPB

Table 7. The predicted miRNA profiles of both clones. The miRNA marked in bold letters indicates the unique feature

miRNA	Yabukita	Clone I.1.93	miRNA	Yabukita	Clone I.1.93
mir-156	4	4	MIR390	3	4
mir-166	9	11	MIR408	2	2
mir-124	1	1	MIR171_2	3	3
mir-160	6	6	MIR398	3	3
mir-399	8	8	MIR397	2	2
mir-395	27	27	**MIR530**	–	1
mir-172	4	7	MIR535	3	5
MIR159	14	18	MIR162_2	1	3
MIR167_1	3	8	MIR477	1	1
MIR171_1	24	32	mir-286	14	15
MIR169_2	15	19	MIR403	1	1
MIR164	3	4	MIR169_5	14	19
MIR396	4	4	MIR811	2	2
mir-190	1	1	MIR828	1	1
MIR168	9	9	mir-393	4	4
MIR394	5	6	MIR2275	6	6

Table 8. Secondary metabolites gene clusters found in Yabukita

Cluster	Chromosome	Type	From	To	Size (kb)
1	1	polyketide	304 042 233	304 075 951	33.72
2*	3	saccharide	34 964 928	35 022 997	58.07
3*	3	putative	224 389 443	224 482 504	93.06
4*	3	lignan	236 496 836	236 598 027	101.19
5*	6	terpene	176 324 438	176 408 280	83.84
6*	6	terpene	178 086 753	178 213 998	127.25
7*	7	putative	27 139 358	27 183 109	43.75
8	9	saccharide	46 031 047	46 095 351	64.3
9*	11	saccharide	142 878 219	142 925 777	47.56

*Also found in T2 (clone I.1.93)

Table 9. Secondary metabolites gene clusters found in clone I.1.93

Cluster	Chromosome	Type	From	To	Size (kb)
1*	3	saccharide	34 964 928	35 022 997	58.07
2*	3	putative	224 389 443	224 482 504	93.06
3*	3	lignan	236 496 836	236 598 027	101.19
4*	6	terpene	176 324 438	176 408 280	83.84
5*	6	terpene	178 086 753	178 213 998	127.25
6*	7	putative	27 139 358	27 183 109	43.75
7*	11	saccharide	142 878 219	142 925 777	47.56
8	12	putative	76 370 295	76 414 655	44.36
9	12	terpene	118 297 547	118 406 031	108.48

*Also found in T1 (Yabukita)

https://doi.org/10.17221/116/2025-CJGPB

dictions and have not been validated experimentally in this study. While the presence of these clusters may plausibly contribute to the distinct flavour and aroma profiles of Yabukita and Clone I.1.93, the direct functional link between these specific genomic regions and the observed sensory traits remains speculative. Future integrated omics approaches – combining transcriptomics, metabolomics, and sensory analysis – are required to definitively correlate cluster activity with metabolite production and the resulting sensory attributes.

DISCUSSION

Plant profiles of Yabukita and clone I.1.93. In this study, the use of Oxford Nanopore long-read sequencing facilitated the alignment and variant calling across complex genomic regions, contributing to the high mapping rates and detailed intergenic SNP profiles observed (Table 1, Figure 2). However, it is important to note that variant detection accuracy is influenced by multiple factors, including sequencing error profiles and bioinformatic parameters (Zverinova & Guryev 2022; Harvey et al. 2023). Both of the tea WGS results are mapped well to the reference genome of *Camellia sinensis* genome assembly HZAU_G240_1.0, indicated by higher than 90% of primary mapped reads for both of the clones (Table 1). Based on the previous study by Prayoga et al. (2022), the Yabukita clone is superior in aroma and flavour profile, while clone I.1.93 has a good NUE profile, making the plant grow with a bigger structural profile. At the molecular level, the majority of the shared functional enrichments, both GO and KEGG, indicated high similarities (Figure 2) with only a few differences. Comparable patterns were also observed for predicted miRNAs (Table 6), and secondary metabolites gene clusters (Table 7–8), but some unique data started to get revealed. The whole genome data profiles within the same species are nearly identical, with some distinctive features caused by SNP differences across the individual or cultivars, caused by mutations (Xu & Bai 2015; Lozada et al. 2022; Singh et al. 2023). However, random mutations are not always occurring within the gene exons, and could happen on the introns, untranslatable regions (UTR) that affect the gene promoters and terminators, or even extragenic/intergenic (Aziz & Masmoudi 2025). Long read whole genome sequencing allows the detection of the SNP profile differences in greater details than normal short read sequencing (Harvey et al. 2023).

The SNP profile differences, miRNA profiles, and gene clusters. The SNP annotations revealed the overall effect of SNP mutations within the genome and their general distribution. The annotation results (Figure 1) indicated that a few differences in the change rate per chromosome were detected between the two clones, and the majority of the mutations are both missense (top mutation) by 57–58% and silent mutation by 40%, with only very few amounts of nonsense mutations. Moreover, the mutations are nearly all (93%) mutations are intergenic. Despite the large number of missense mutations that occurred, the fact that the majority of the mutations are intergenic means all of these mutations may not have a direct effect on gene expression. Within the intergenic regions, there are transcriptional regulatory elements and also the size of the intergenic regions might affect the overall genomic architecture (Nelson et al. 2004). The SNP annotation revealed that the vast majority (> 93%) of variants occurred in intergenic regions (Figure 2C). While these intergenic SNPs may not directly alter protein sequences, they could potentially influence gene regulation by affecting non-coding regulatory elements (Nelson et al. 2004). In this study, the functional consequence of these intergenic variants remains unknown and represents a limitation of our variant-centric analysis without complementary regulatory data (e.g., ATAC-seq or ChIP-seq).

Within the intergenic regions, one of the known regulatory components in gene expression is microRNA or miRNA. The intergenic miRNAs could be found as a singular miRNA gene or a cluster of miRNA genes (Olena & Patton 2010). Although the position of miRNA in tea is not yet well studied, the miRNA in general could affect the overall profile of the tea. Some miRNAs affect the flavour compound biosynthesis in tea, e.g. mir169a and 169 h, mir171 and mir319 h induce nongallated catechin; mir156a, mir156c, and mir156q inhibit biosynthesis of gallated catechin; while mir166j, mir172b, and mir172 induce the biosynthesis of gallated catechin; mir58 and mir101 induce caffeine production; and mir171o, mir71a, mir71b, mir71c, and mir7d induce linalool, geraniol, and 2-phenylethanol production (Li et al. 2021). In a previous study that also involved *in silico* identification (by using BLAST) and 47 452 expressed sequence tags (EST) of the tea plant, 51 miRNAs were identified (Zhu & Luo 2013). There are 31 shared miRNAs that were identified from both clones, but there is a single miRNA (mir530) that was detected only in clone I.1.93 genome using the strictest miRNA

https://doi.org/10.17221/116/2025-CJGPB

detection cutoff method to avoid bias (Table 6). A notable finding was the *in silico* prediction of miR530 exclusively in Clone I.1.93 (Table 6). While miR530 has been associated with stress response and secondary metabolism regulation in other plant species (Srivastava & Singh 2018; Hossain et al. 2022), its functional role in tea remains uncharacterized. Therefore, we cannot conclude that miR530 explains the flavour differences between the clones. Its discovery merely identifies it as a candidate for future functional studies to investigate its potential influence on metabolism and stress resilience. The other previous studies showed that mir530 is a miRNA detected on leaf organ, associated with secondary metabolite regulation as it upregulates cycloartenol synthase (CAS1), sterol delta-7 reductase 1, CYP82G, zeatin o-glycosyl transferase (UGTs), and secoisolariciresinol dehydrogenase (ABA2) that are essential in withanolide (a steroid compound) biosynthesis in ashwagandha (*Withania somnifera* (L). Dunal) (Srivastava & Singh 2018; Hossain et al. 2022). As Yabukita is known to be more aromatic than clone I.1.93, the upregulation of steroid compounds that are less associated with tea aroma and flavour could be the result. However, more studies to characterise mir530 regulation on tea secondary metabolites production are needed.

Other than miRNA, gene clusters for secondary metabolites are distinct in both clones (Tables 7, 8). Of which, in Yabukita, cluster 1 is responsible for polyketide biosynthesis and cluster 8 for saccharide (with glycosyltransferase), whereas in clone I.1.93, cluster 8 is responsible for dioxygenase putative and cluster 9 for terpene. Poliketides are important in the biosynthesis of flavonoids, e.g. naringenin, as found in fungi (Zhang et al. 2022) and polyphenols (Valentic et al. 2016). Also, terpenoids (Itoh et al. 2010) and phenylpropanoids (Martín & Liras 2022) contribute to aromatic compounds. Saccharide cluster contains sets of 3 glucosyltransferases that are important for carbohydrate production of dioxygenase enzymes – wide functions, from hypoxia stress (Iacopino & Licausi 2020), growth metabolism (Wei et al. 2021) to pigment synthesis (Christinet et al. 2004). Both clones have shared terpene clusters, which are involved in aroma and flavour production (Pichersky & Raguso 2018). However, it is unknown if the extra terpene cluster in clone I.1.93 might actually improve or disrupt the tea flavour and aroma profile. Both clones also shared terpene clusters, which are often linked to aroma and flavour production (Pichersky & Raguso 2018). Nevertheless, these functional assignments are predictive and derived from PlantiSMASH annotations without direct biochemical validation. Future studies integrating RNA-seq and metabolite profiling are needed to confirm the activity and products of these clusters.

The implications for GO and KEGG differences. Overall, there are high similarities in the shared GO and KEGG of both clones (Figure 2). However, if both clones have their non-shared genes put to analysis to display each own GO and KEGG profiles, there are prominent differences in the GO (Tables 2–4) mainly in the stress response traits of BP (Table 2). For KEGG pathways (Table 5), there are 3 genes involved in carotenoid biosynthesis found in Yabukita, while none in the unique genes of clone I.1.93. On the other hand, clone I.1.93 has folate biosynthesis genes (4 genes). As carotenoid is involved in golden coloration of plants, its derivative might also be involved in distinct aroma profile (Winterhalter & Rouseff 2001). Meanwhile, folate or vitamin B9 is an important cofactor (Hanson & Gregory 2011) and a key in the synthesis of many important biomolecules, e.g. amino acids, nucleic acids, and vitamin B5 (Gorelova et al. 2017). Uniqueness in carotenoid genes might be involved in Yabukita clone's favourable aroma and flavour, and the folate biosynthesis might explain the higher NUE profile in clone I.1.93.

Future directions for molecular breeding marker selection for tea. These findings provide a foundation for breeding programs aimed at improving tea quality and resilience. Genomic regions and SNPs associated with key pathways can serve as targets for marker-assisted selection or genome editing, fostering the development of superior tea varieties. Further research, including experimental validation and exploration of gene-environment interactions, is essential to fully leverage these genetic insights for practical applications in tea cultivation. Furthermore, other studies also need more focus on validating the roles of unique miRNAs and clusters through functional assays, ensuring their applicability in improving agricultural traits.

CONCLUSION

This study provides a comparative whole-genome analysis of the tea clones Yabukita and I.1.93, offering insights into genomic variation, regulatory mechanisms, and metabolic potential. Genome-wide SNP profiling revealed distinct variation patterns,

https://doi.org/10.17221/116/2025-CJGPB

with Yabukita exhibiting a more uniform chromosomal SNP distribution, while clone I.1.93 showed higher SNP densities on specific chromosomes. Functional annotations indicated a predominance of silent mutations in Yabukita and missense mutations in clone I.1.93, suggesting greater potential functional divergence in the latter.

GO and KEGG pathways analyses showed that both clones shared highly similar core biological processes, including photosynthesis and protein interaction pathways. However, clone-specific enrichment patterns were evident: Yabukita was primarily associated with photosynthesis-related pathways, whereas clone I.1.93 was enriched in stress-related processes, such as salt stress response and seed dormancy. Differences in predicted miRNAs profiles further supported this distinction, with several miRNAs, including the clone-specific miR530, showing higher expression in clone I.1.93 and potentially contributing to stress resilience and productivity. In addition, the identification of secondary metabolite biosynthetic gene clusters associated with terpenoid, polyketide, and saccharide metabolism highlights the genetic basis of tea quality traits, including flavour and aroma. The predicted secondary metabolite gene clusters identified here offer promising targets for future research, though their roles in tea quality remain to be validated through integrated omics and functional studies.

Acknowledgements. The authors gratefully acknowledge the Head of Research Center for Horticulture, Research Organization for Agriculture and Food, National Research, and Innovation Agency (BRIN), who has supported the laboratory facilities. We thank the Scientific Department team, GSI Lab, for the valuable help with bioinformatics analysis.

REFERENCES

An Y., Zhang X., Jiang S., Zhao J., Zhang F. (2022): TeaPVs: A comprehensive genomic variation database for tea plant (*Camellia sinensis*). BMC Plant Biology, 22: 513.

Aziz M.A., Masmoudi K. (2025): Molecular breakthroughs in modern plant breeding techniques. Horticultural Plant Journal, 11: 15–41.

Buchfink B., Xie C., Huson D.H. (2014): Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12: 59–60.

Christinet L., Burdet F.X., Zaiko M., Hinz U., Zrÿd J.P. (2004): Characterization and functional identification of a novel plant 4,5-extradiol dioxygenase involved in betalain pigment biosynthesis in *Portulaca grandiflora*. Plant Physiology, 134: 265–274.

Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., Ruden D.M. (2012): A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly, 6: 80–92.

Doyle J., Doyle J. (1987): A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin, 19: 11–15.

Galaxy Community (2024): The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Research, 52: W83–W94.

Gorelova V., Ambach L., Rébeillé F., Stove C., Van Der Straeten D. (2017): Folates in plants: Research advances and progress in crop biofortification. Frontiers in Chemistry, 5: 1–20.

Gupta P., O'Neill H., Wolvetang E.J., Chatterjee A., Gupta I. (2024): Advances in single-cell long-read sequencing technologies. NAR Genomics and Bioinformatics, 6: 1–11.

Hanson A.D., Gregory J.F. (2011): Folate biosynthesis, turnover, and transport in plants. Annual Review of Plant Biology, 62: 105–125.

Harvey W.T., Ebert P., Ebler J., Audano P.A., Munson K.M., Hoekzema K., Porubsky D., Beck C.R., Marschall T., Garimella K., Eichler E.E. (2023): Whole-genome long-read sequencing down sampling and its effect on variant-calling precision and recall. Genome Research, 33: 2029–2040.

Hossain R., Quispe C., Saikat A.S.M., Jain D., Habib A., Janmeda P., Islam M.T., Radha, Daştan S.D., Kumar M., Butnariu M., Cho W.C., Sharifi-Rad J., Kipchakbayeva A., Calina D. (2022): Biosynthesis of secondary metabolites based on the regulation of microRNAs. BioMed Research International, 2022: 1–15.

Huang D.W., Sherman B.T., Lempicki R.A. (2009): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4: 44–57.

Iacopino S., Licausi F. (2020): The contribution of plant dioxygenases to hypoxia signaling. Frontiers in Plant Science, 11: 1–8.

Itoh T., Tokunaga K., Matsuda Y., Fujii I., Abe I., Ebizuka Y., Kushiro T. (2010): Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. Nature Chemistry, 2: 858–864.

Kautsar S.A., Suarez Duran H.G., Blin K., Osbourn A., Medema M.H. (2017): PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Research, 45: W55–W63.

https://doi.org/10.17221/116/2025-CJGPB

Li H. (2018): Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34: 3094–3100.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. (2009): The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25: 2078–2079.

Li H., Lin Q., Yan M., Wang M., Wang P., Zhao H., Wang Y., Ni D., Guo F. (2021): Relationship between secondary metabolism and miRNA for important flavor compounds in different tissues of tea plant (*Camellia sinensis*) as revealed by genome-wide miRNA analysis. Journal of Agricultural and Food Chemistry, 69: 2001–2012.

Liu D., Zhang C., Ye Y., Mei P., Gong Y., Liu Z., Sun C., Zhao X., Ding S., Chen J., Chen L., Ma C. (2025): TEA5K: A high-resolution and liquid-phase multiple-SNP array for molecular breeding in tea plant. Journal of Nanobiotechnology, 23: 481.

Lozada D.N., Bosland P.W., Barchenger D.W., Haghshenas-Jaryani M., Sanogo S., Walker S. (2022): Chile pepper (*Capsicum*) breeding and improvement in the multi-omics era. Frontiers in Plant Science, 13: 1–10.

Martín J.F., Liras P. (2022): Comparative molecular mechanisms of biosynthesis of naringenin and related chalcones in actinobacteria and plants. Antibiotics, 11: 1–21.

Martono B., Syafaruddin (2018): Genetic variability of 21 tea genotypes (*Camellia sinensis*) based on RAPD markers. Journal of Industrial and Beverage Crops, 5: 77–86.

Nawrocki E.P., Eddy S.R. (2013): Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 29: 2933–2935.

Nelson C.E., Hersh B.M., Carroll S.B. (2004): The regulatory content of intergenic DNA shapes genome architecture. Genome Biology, 5: R25.

Olena A.F., Patton J.G. (2010): Genomic organization of microRNAs. Journal of Cellular Physiology, 222: 540–545.

Pichersky E., Raguso R.A. (2018): Why do plants produce so many terpenoid compounds? New Phytologist, 220: 692–702.

Prayoga M.K., Syahrian H., Rahadi V.P., Atmaja M.I.P., Maulana H., Anas. (2022): Quality diversity of 35 tea clones (*Camellia sinensis* var. *sinensis*) processed for green tea. Biodiversitas, 23: 810–816.

Singh J., Kumar D., Chauhan S., Dhillon H.K., Kumar S., Kumar V., Kapoor R. (2023): Role of omics approaches in vegetable breeding for insect pest resistance. SN Applied Sciences, 5: 1–10.

Srivastava S., Singh R. (2018): Comparative study of withanolide biosynthesis-related miRNAs in root and leaf tissues of *Withania somnifera*. Applied Biochemistry and Biotechnology, 185: 1145–1159.

Stanke M., Morgenstern B. (2005): AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33: 465–467.

Valentic T.R., Jackson D.R., Brady S.F., Tsai S.C. (2016): Comprehensive analysis of a novel ketoreductase for pentangular polyphenol biosynthesis. ACS Chemical Biology, 11: 3421–3430.

Wei S., Zhang W., Fu R., Zhang Y. (2021): Genome-wide characterization of 2-oxoglutarate and Fe (II)-dependent dioxygenase family genes in tomato during growth cycle and their roles in metabolism. BMC Genomics, 22: 1–14.

Winterhalter P., Rouseff R. (2001): Carotenoid-derived aroma compounds: An introduction. ACS Symposium Series, 7: 1–17.

Xu X., Bai G. (2015): Whole-genome resequencing: Changing the paradigms of SNP detection, molecular mapping and gene discovery. Molecular Breeding, 35: 1–10.

Yagi C., Ikeda N., Sato D. (2010): Characteristics of eight Japanese tea cultivars. College of Tropical Agriculture and Human Resources, 1167–1168.

Zhang H., Li Z., Zhou S., Li S.M., Ran H., Song Z., Yu T., Yin W.B. (2022): A fungal NRPS-PKS enzyme catalyses the formation of the flavonoid naringenin. Nature Communications, 13: 1–11.

Zhu Q.W., Luo Y.P. (2013): Identification of miRNAs and their targets in tea (*Camellia sinensis*). Journal of Zhejiang University – Science B, 14: 916–923.

Zverinova S., Guryev V. (2022): Variant calling: Considerations, practices, and developments. Human Mutation, 43: 976–985.

Received: November 24, 2025
Accepted: February 6, 2026
Published online: February 23, 2026