the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) **119,** 2096–2115 doi: 10.1111/tpj.16874

RESOURCE

In-depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of *Camellia sinensis*

Arslan Tariq[1,†], Minghui Meng[2,†], Xiaohui Jiang[2], Anthony Bolger[3], Sebastian Beier[3], Jan P. Buchmann[1], Alisdair R. Fernie[4], Weiwei Wen[2,*] and Björn Usadel[1,3,*]

[1]*HHU Düsseldorf, Faculty of Mathematics and Natural Sciences, CEPLAS, Heinrich Heine University, Universitätsstrasse 1, Düsseldorf, Germany,*

[2]*National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Key Laboratory of Horticultural Plant Biology (MOE), College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan 430070, China,*

[3]*Institute of Bio- and Geosciences, IBG-4: Bioinformatics, CEPLAS, Forschungszentrum Jülich, Leo Brandt-Straße, Jülich 52425 Germany, and*

[4]*Max-Planck-Institute of Molecular Plant Physiology, Am Muehlenberg 1, Potsdam-Golm 14476 Germany*

Received 28 November 2023; revised 21 May 2024; accepted 25 May 2024; published online 14 June 2024.
*For correspondence (e-mail usadel@hhu.de; wwwen@mail.hzau.edu.cn).
†These authors contributed equally.

SUMMARY

Tea, one of the most widely consumed beverages globally, exhibits remarkable genomic diversity in its underlying flavour and health-related compounds. In this study, we present the construction and analysis of a tea pangenome comprising a total of 11 genomes, with a focus on three newly sequenced genomes comprising the purple-leaved *assamica* cultivar "Zijuan", the temperature-sensitive *sinensis* cultivar "Anjibaicha" and the wild accession "L618" whose assemblies exhibited excellent quality scores as they profited from latest sequencing technologies. Our analysis incorporates a detailed investigation of transposon complement across the tea pangenome, revealing shared patterns of transposon distribution among the studied genomes and improved transposon resolution with long read technologies, as shown by long terminal repeat (LTR) Assembly Index analysis. Furthermore, our study encompasses a gene-centric exploration of the pangenome, exploring the genomic landscape of the catechin pathway with our study, providing insights on copy number alterations and gene-centric variants, especially for Anthocyanidin synthases. We constructed a gene-centric pangenome by structurally and functionally annotating all available genomes using an identical pipeline, which both increased gene completeness and allowed for a high functional annotation rate. This improved and consistently annotated gene set will allow for a better comparison between tea genomes. We used this improved pangenome to capture the core and dispensable gene repertoire, elucidating the functional diversity present within the tea species. This pangenome resource might serve as a valuable resource for understanding the fundamental genetic basis of traits such as flavour, stress tolerance, and disease resistance, with implications for tea breeding programmes.

Keywords: Tea, Camellia sinensis, pangenome, ANS.

INTRODUCTION

Tea [*Camellia sinensis* (L.) O. Kuntze], a member of the genus *Camellia* (Theaceae), is one of the most popular beverages worldwide, with rich flavours and health benefits. Tea is usually subclassified into *assamica* and *sinensis* varieties, where *assamica* varieties feature larger leaves and are adapted to more humid and warmer climates, whereas *sinensis* varieties are more cold hardy, slower growing and have smaller leaves (Zhang et al., 2023). Booth varieties are flavourful, but they might differ in distinct flavonoid accumulation (Fang et al., 2021; Yu et al., 2020). Owing to the advances in sequencing

technologies and bioinformatic methodologies, significant progress in tea plant genomics has been achieved. Until now, several chromosome-level tea plant genomes have been assembled, and hundreds of diverse tea genotypes have been subjected to genome resequencing or RNA-sequencing. These data have greatly advanced our understanding of the evolution of the tea plant genome, including frequent intra- and interspecies introgressions, and have shed light on the genetic and molecular mechanisms of important tea traits (Chen, Wang, et al., 2023; Xia et al., 2020; Zhang et al., 2021; Zhang, Zhang, et al., 2020).

Many of the above-mentioned analyses were, however, based on a single reference genome species and short read resequencing of several accessions (An et al., 2020, 2021; Lei et al., 2022; Liu et al., 2023; Wang et al., 2020) or based on transcriptome-level sequencing only (Kong et al., 2022; Li et al., 2016; Shi et al., 2015; Xia et al., 2017; Zhang, Li, et al., 2020), not offering full genomic resolution except for the recent first generation Pacbio-based tea pangenome (Chen, Wang, et al., 2023). This is because studying a single individual's genome in depth in itself does not reveal the entire population's diversity, as has been shown for multiple crops using pangenomic approaches (Bayer et al., 2020; Wang et al., 2023). Indeed, traditional genetic analysis methods, such as quantitative trait loci mapping and genome-wide association studies (GWAS), relying on one reference only, are not able to explain as much observed phenotypic variance as when using an underlying pangenomic approach (Zhou et al., 2022). These pangenomic approaches are able to reveal the presence/absence variations (PAV) of genes, transposons, and/or regions. These studies have shown that besides the so-called core genome, which comprises of genes found in all accessions and represents the core set of genes usually having a conserved function, the pangenome comprises shell and cloud genes, which are found only in some or a few accessions. The latter two are especially interesting not only as these comprise many genes that are on the way to pseudogenization but also as this gene pool has been shown to often comprise genes involved in defence response, abiotic stress tolerance, and adaptation, which are particularly important for breeding (Hoopes et al., 2022; Kang et al., 2023). In addition, copy number variation (CNV) of genes is frequently observed and contributes to crop phenotypic diversity, but both PAV and CNVs are difficult to identify using a single reference genome. Hence, analysing and comparing multiple high-quality genome assemblies is considered a gold standard for pangenomic analyses. However, as generating such full genome-based pangenomes requires a significant sequencing effort for larger genomes such as tea, which was often simply not possible before the latest updates to third-generation sequencing (Gui et al., 2023), a stop-gap pan-transcriptome analysis was introduced to detect some genetic and additional expression variations at a lower cost. Indeed, a recent pan-transcriptome analysis of the tea genome revealed multiple molecular differences between *assamica* and *sinensis* varieties, as well as expression differences in flavonoid metabolism and regulation (Kong et al., 2022). These results were in line with earlier data (Zhang, Zhang, et al., 2020) that used a novel tea genome and high-throughput metabolomic and transcriptomic data analysis to show that diverse ANTHOCYANIDIN REDUCTASE (*Cs*ANR), FLAVONOID 3′5′-HYDROXYLASE (*Cs*F3′5′H), and *Cs*MYB5 alleles exist, likely explaining differences in catechin composition as two different *Cs*ANR genomic variants exhibited different enzymatic parameters. Given the rapid improvements in long-read sequencing technologies (Gui et al., 2023; van Rengs et al., 2022), nowadays even extremely large plant genomes, such as the one of the faba bean (Jayakodi et al., 2023), can be assembled. Hence, in the last few years, eight high-quality tea genomes comprising one wild variety, DASZ, one *assamica* variety, Yunkang10, as well as six *sinensis* varieties, have become available.

Given the above, we set about providing a comprehensive analysis of the tea pangenome in order to obtain insights into the genetic diversity of tea and its evolution, identify unique and useful genes for breeding and biotechnology. In the current study, we present a detailed analysis of the tea pangenome using eight published genome sequences, as well as three novel high-quality genomes. Our novel genomes comprise those of the purple-leaved *assamica* cultivar "Zijuan", and the temperature-sensitive *sinensis* cultivar "Anjibaicha" which have both been extensively studied using omics technologies (Li et al., 2015, 2016; Xu et al., 2018), as well as that of a wild accession "L618" facilitating analyses between plant lines.

RESULTS

Genome assembly

The three new *C. sinensis* cultivars, Anjibaicha (AJ hereafter, a temperature-sensitive albino cultivar), Zijuan (ZJ hereafter, an *assamica* variety from the Yunan province of China), and L618 (a wild accession), were selected, representing one member each for the two major varieties and a wild line. These three specimens were sequenced using PacBio high-fidelity (HiFi) sequencing, which produced 6.20, 6.35, and 6.15 million reads with an average read length of 16 575, 18 553, and 17 661 bases for AJ, ZJ, and L618, respectively. Assembling these genomes with Hifiasm resulted in very contiguous assemblies featuring N50 contig sizes of 62.73, 94.85, and 94.24 Mb for AJ, ZJ, and L618, respectively. The resulting assemblies were anchored to 15 chromosomes using Hi-C data with 69× read coverage for the AJ accession (Figure **S1**) and using reference-based scaffolding with RagTag against AJ for ZJ

Table 1 Genomic characteristics and quality of the three new and previously published tea assemblies

Genomes	Scaffold number	Total length (Gbp)	GC (%)	Scaffold N50 (Mbp)	# N's per 100 kbp	Compleasm (%)	LAI
Anjibaicha	1638	3.24	38.97	202	2.08	99.14	10.07
Zijuan	550	3.06	38.66	212	0.62	98.93	9.19
L618	212	3.01	38.58	198	0.27	99.05	8.93
Biyun	3495	2.92	38.24	196	135.46	96.73	8.79
DASZ	1231	3.11	38.98	204	52.45	98.50	9.79
DuyunMaojian	8936	3.13	38.67	211	0.02	94.75	4.32
Huangdan	913	2.95	38.63	213	63.46	98.32	8.9
LJ43	30 544	3.26	38.68	144	21.64	94.28	8.64
Shuchazao	1333	2.94	38.25	167	19.39	96.52	8.96
Tieguanyin	163	3.06	38.51	213	11.54	97.68	9.53
Yunkang10	11 573	3.02	39.62	187	14845.35	95.14	0

LAI values are chromosome anchored sequences only. Compleasm values were calculated with the BUSCO lineage "eudicots_odb10" and are comparable to BUSCO C-Scores.
LAI, LTR Assembly Index.

and L618. This resulted in high-quality genome assemblies with 1638, 550, and 212 scaffolds (Table 1), and an anchoring rate of 88.5, 97, and 98.5 to the 15 chromosomes for AJ, ZJ, and L618, respectively. The N50 scaffold sizes after genome-scale scaffolding reached values of 201.68, 211.85 Mb, and 197.99 Mb while the corresponding N90 was 14.54 Mb, 156.21 Mb, and 143.46 Mb for the three genomes of AJ, ZJ, and L618, respectively, thus representing genome-scale assemblies. Indeed, analysing the chromosome ends revealed that chromosomes 1, 4, and 8 of AJ showed telomeric repeats at both ends, while for chromosomes 2, 3, 7, 5, 9, 13, 14, and 15, telomeric repeats were identified for one end (Table S1). In case of the ZJ assembly, telomere repeats for chromosome 5 were found on both ends while for chromosomes 2, 3, 11, 12, 13, and 15, only telomeric repeats were found on one end. For L618, telomeric repeats for both ends were found only in chromosomes 5 and 8, and one end of telomeric repeats was found in chromosomes 2, 3, 7, 9, 10, 11, 13, and 15 (Table S1). In addition, the final total assembly size was 3.24, 3.06, and 3.01 Gb for AJ, ZJ, and L618, respectively (Table 1), which is close to the Genomescope 2.0 (Ranallo-Benavidez et al., 2020) sequence-based, estimated genome size, which also suggested a slightly larger AJ genome. In addition, Genomescope estimated heterozygosity rates of up to 2.03, 2.98, and 2.06 for AJ, ZJ and L618, respectively (Figure S2).

The genomes exhibited compleasm (Huang & Li, 2023)-based BUSCO (Simão et al., 2015) gene completeness values of 99.14, 98.93, and 99.05%, slightly surpassing previously released tea genomes (which range from 94.07 to 98.5%), where only the recently released pangenome of Chen, Wang, et al. (2023) approaches these values with a best value of 98.32%. This likely reflects improvements resulting from the latest high-quality PacBio Hifi technology. The Merqury k-mer-based quality analysis tool (Rhie et al., 2020) revealed impressive consensus base accuracy (QV) scores of 62.52, 64.42, and 67.40 for the genomes of AJ, ZJ, and L618, respectively, indicating very high-quality genome assemblies. Similarly, CRAQ quality analysis (Li et al., 2023) revealed exceptionally low average Clip-based Regional Errors (CRE = 0.14, 0.09, 0.06 for AJ, ZJ, and L618, respectively) and minimal Clip-based Structural Errors (CSE = 0.02, 0.03, and 0.02 for AJ, ZJ, and L618, respectively). Additionally, the assemblies exhibited outstanding average Regional and Structural Assembly Quality Indicators (R-AQI = 98.61, 99.07, 99.36, and S-AQI = 97.5, 97.04, 97.22 for AJ, ZJ, and L618, respectively). Notably, with an AQI surpassing 90, these results signify that our assemblies achieve reference-level assembly quality (Li et al., 2023). Consistent with these values, the long terminal repeat (LTR) Assembly Index (LAI) index, analysing the assembly completeness using long terminal repeats (LTRs) (Ou et al., 2018), was the highest for the AJ genome with a value of 10.07, whereas ZJ and L618 had LAI values of 9.19 and 8.93 (Table 1) within the typical range for the other tea genome assemblies that used long read technologies ranging from 8.64 (LJ43) to 9.79 (DASZ), with the exception of the nanopore-based assembly of DuyunMaojian, which fell off with a value of 4.32, only likely reflecting a more fragmented underlying contig N50 of <1 Mbp prior to scaffolding (Wang et al., 2022).

Genome annotation

The combination of *ab initio* gene prediction using the deep learning tool Helixer (Holst et al., 2023) and the RNASeq-based stringTie (Pertea et al., 2015) predicted transcripts were combined with a high-quality genome annotation with mikado (Venturini et al., 2018) to select the best transcript sets. This resulted in 41 429, 40 749, and

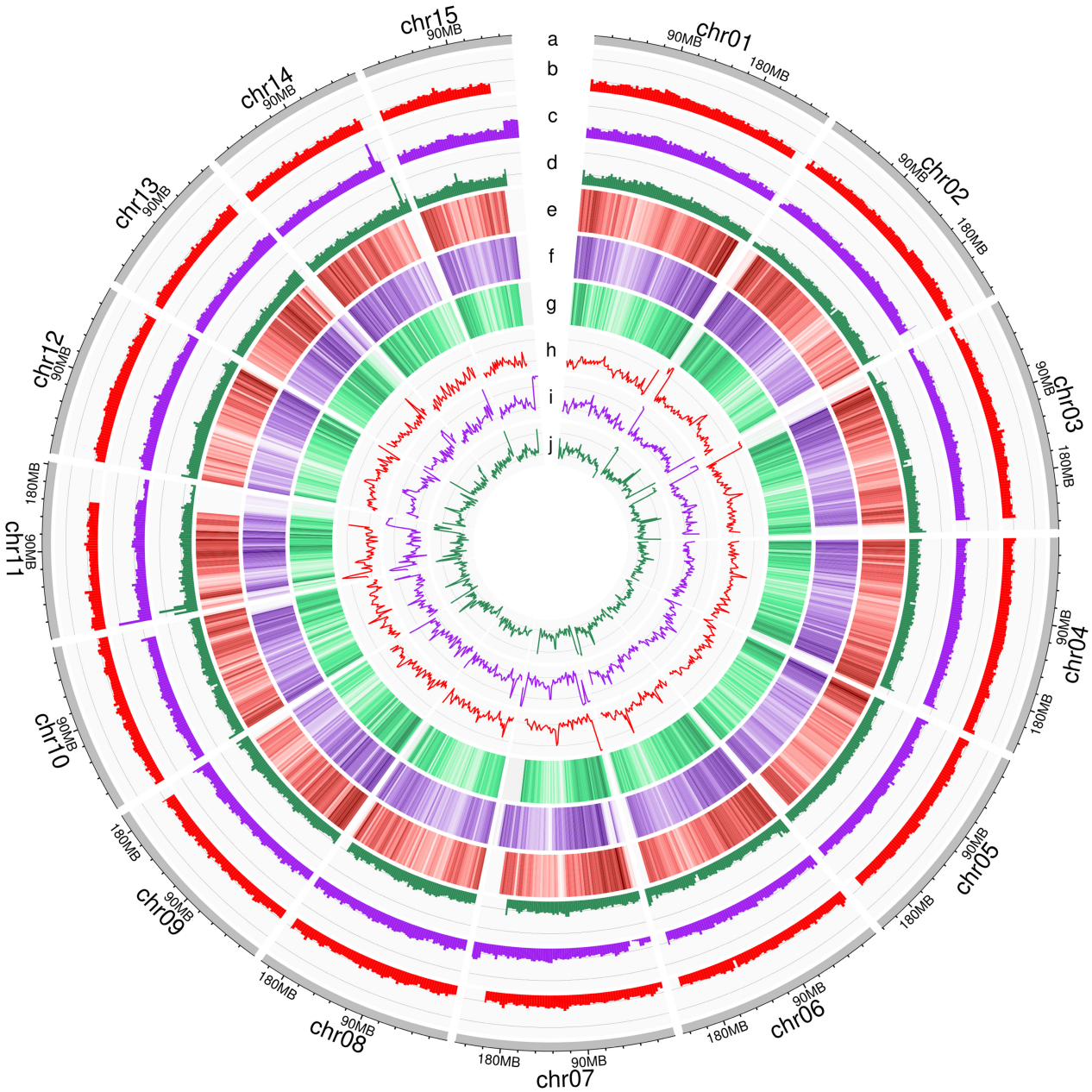

Figure 1. Comparative circular genome analysis of GC content, gene density, and transposable element (TE) distribution in AJ (in red), ZJ (in purple), and L618 (in sea green) using 3 Mbp Windows (Wang et al., 2023).
(a) Overall genome chromosome view.
(b–d) GC content for AJ, ZJ, and L618.
(e–g) Gene density for AJ, ZJ, and L618.
(h–j) TE Content for AJ, ZJ, and L618.

41 401 gene models from AJ, ZJ, and L618, respectively (Figure 1; Table S2), where for about 95% of these genes, a functional annotation could be derived using Mercator (Schwacke et al., 2019). In line with the almost complete genomic values, the resulting gene models exhibited completeness scores of 95.40, 95.25, and 94.97 for AJ, ZJ, and L618, respectively, when assessed by BUSCO (Table 2).

Pangenome annotation

In addition to these three genomes, we also obtained eight publicly available, high-quality genomes. These comprised Biyun, a commercial variety having relatively low heterozygosity (Zhang, Li, et al., 2020), DASZ, an ancient wild tea tree (Zhang, Zhang, et al., 2020), DuyunMaojian, a very old

Table 2 Reannotation of tea genomes

Genomes	Gene models	Published gene models	Annotated genes (%)	Published annotated genes (%)	BUSCO (%)	BUSCO published proteome (%)
Anjibaicha	41 429	NA	39 367 (95.02)	NA	95.40	NA
Zijuan	40 749	NA	38 697 (94.96)	NA	95.27	NA
L618	41 401	NA	39 338 (95.02)	NA	94.97	NA
Biyun	44 700	40 808	42 395 (94.84)	39 305 (96.32)	90.50	67.50
DASZ	41 562	33 021	39 457 (94.94)	30 863 (93.46)	88.39	81.20
DuyunMaojian	39 130	32 232	36 767 (93.96)	31 251 (96.96)	89.60	87.40
Huangdan	41 761	43 779	39 624 (94.88)	40 883 (93.38)	93.55	91.60
LJ43	39 365	33 556	37 104 (94.26)	31 778 (94.70)	88.78	82.80
Shuchazao	47 227	50 525	44 944 (95.17)	46 996 (93.02)	90.41	88.70
Tieguanyin	42 074	45 901	39 955 (94.96)	44 087 (96.05)	93.25	89.70
Yunkang10	37 995	36 951	35 688 (93.93)	34 251 (92.69)	90.15	66.10

The "eudiocots_odb10" lineage was used for BUSCO assessment.

cultivar most planted in Guizhou Province growing in harsh environments (Wang et al., 2022), Huangdan from Anxi County, Fujian Province, China and suitable for processing into oolong, green, and black teas (Wang et al., 2021), LJ43 which is highly cold resistant and early sprouting cultivar (Wang et al., 2020), Shuchazao planted in southeast China and having the most recent tandem duplication in the genomes (Xia et al., 2020), Tieguanyin (Oolong tea cultivar) (Zhang et al., 2021), and Yunkang10 (var. Assamica and widely grown in Southwestern China) (Xia et al., 2017). To provide a comparable and updated genome annotation, we reannotated these using the same annotation steps as for the newly sequenced genomes. As this reannotation resulted in a relatively low BUSCO completeness of 85.04%, we speculated that this might be due to the underlying assembly. Here, we observed several missing BUSCO genes that seemed to be due to single nucleotide polymorphisms, such as premature STOP codons. We wondered whether this was based on the assembly and polishing algorithms available at the time the genome was assembled and retrieved from the original assembly and raw read data. We polished the assembly using both short and long read data to obtain an improved genome. Indeed, after polishing multiple single nucleotide polymorphisms, even premature STOP codons could be removed, resulting in a new protein based BUSCO value of 89.60%. Hence, for these additional eight accessions, the number of predicted gene models ranged from 379 952 to 47 227, while exhibiting a BUSCO completeness ranging from 88.39 to 93.55% of which again almost all could also be functionally annotated (Table 2; Table S2). The new annotation, with our combination of the *ab initio* approach with Helixer and the extrinsic evidence from stringTie, improved the gene prediction results as compared to the published one between 2 (Huangdan from 91.6 to 93.55) and 23% points for the Biyun accession from 67.50 to 90.50% (Table 2). At the same time, gene number increased for almost all accessions; however, functional annotation rate stayed high at around 94% (Table 2), thus allowing for a more comprehensive genome analysis.

Transposable elements (TEs) are dynamic and sources of variation in a genome (Kidwell & Lisch, 1997). Relying on a pre-existing TE library for TE detection can introduce biases; *de novo* TE prediction proves superior for identifying species-specific TEs (Bell et al., 2022). Hence, TEs were identified with EDTA (Ou et al., 2019), which integrates a comprehensive suite of top-performing packages for the *de novo* detection of TE elements. This analysis revealed that 73.77, 72.58, and 69.26% of the assemblies of AJ, ZJ, and L618 were composed of repetitive elements, respectively (Tables S3 and S4). For the anchored genome only, 80.25, 80.75, and 80.64% of the genome are repetitive, representing relatively consistent numbers (Table 3). Here, Gypsy as a LTR element was most prevalent in the anchored parts of three genomes, where it accounted for 29.85, 29.76, and 29.32% of the total genome assembly of AJ, ZJ, and L618, respectively. This was followed by unknown LTR (21.57, 21.51, and 21.70% of AJ, ZJ, and L618) and then Copia LTRs, which accounted for approximately 7% of the genomes (7.22, 6.88, and 6.83% of AJ, ZJ, and L618). In the case of transposons with terminal inverted repeats (TIR), that is, DNA transposons, Mutator, and CACTA were most abundant and together made up ~11% of the genome. PIF Harbinger, TC1 Mariner, and hAT were also found in these genomes. LINE elements only contributed 0.15% of the genome. Helitron elements were predicted to account for 2.38, 2.35, and 2.27% of AJ, ZJ, and L618's genomes, respectively (Figures 1 and 2; Table 3).

For all 11 accessions, PanEDTA (Ou, Collins, et al., 2022) analysis revealed that the repetitive regions range from 68% to 81% in the scaffolded regions of these genomes, where DASZ contained the most repetitive regions while Yunkang10 contains fewer repetitive regions

Table 3 Transposable element (TE) analysis summary of the newly sequenced tea accessions Anjibaicha, Zijuan, and L618 in the chromosomes of the anchored genomes

		Genomes								
		Anjibaicha			Zijuan			L618		
Class	Subclass	Count	Mbp	%	Count	Mbp	%	Count	Mbp	%
LTR	Copia	256 421	207.19	7.22	252 904	204.31	6.88	252 528	202.62	6.83
	Gypsy	582 896	856.15	29.85	595 814	883.79	29.76	631 510	869.07	29.32
	Unknown	771 958	618.51	21.57	801 722	639.03	21.51	838 008	643.22	21.7
TIR	CACTA	176 238	85.51	2.98	180 455	86.33	2.91	187 382	87.40	2.95
	Mutator	536 449	210.79	7.35	585 429	258.44	8.7	640 092	260.09	8.77
	PIF Harbinger	165 112	49.91	1.74	165 018	48.73	1.64	164 483	48.19	1.63
	Tc1 Mariner	6592	2.52	0.09	6771	2.61	0.09	6539	2.52	0.08
	hAT	130 062	48.30	1.68	132 641	49.76	1.68	133 269	50.98	1.72
Low complexity		40	0.20	0.01	44	0.32	0.01	41	0.32	0.01
NonLTR	LINE element	7286	4.37	0.15	7519	4.40	0.15	7593	4.52	0.15
	Unknown	2663	1.07	0.04	2682	1.08	0.04	2671	1.07	0.04
NonTIR	Helitron	127 349	68.19	2.38	128 080	69.91	2.35	127 242	67.34	2.27
Repeat region		535 129	148.76	5.19	547 921	149.79	5.04	555 943	153.17	5.17
Total	Interspersed	3 298 195	2301.47	80.25	3 407 000	2398.49	80.75	3 547 301	2390.51	80.64

LTR, long terminal repeat; TIR, long terminal repeat.

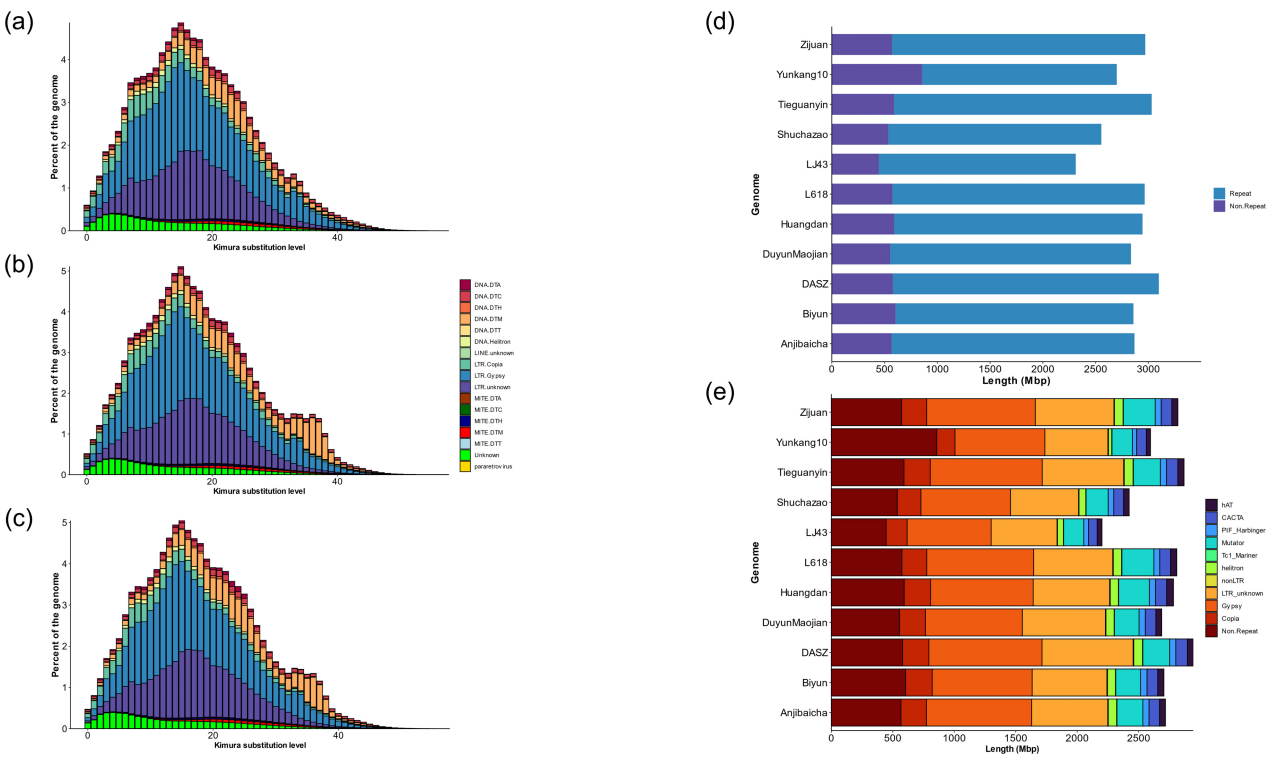

Figure 2. Transposable element landscape. (a–c) Transposable element (TE) landscape in the three newly assembled *Camellia sinensis* accessions AJ, ZJ, and L618, respectively, based on kimura distance-based copy divergence substitution analysis.
(d) TE length in 11 genomes (only anchored).
(e) TEs composition in 11 *C. sinensis* accessions (only anchored).

compared to other accessions. This observed reduction in repetitive elements in Yunkang10 may potentially be attributed to the presence of missing nucleotides ("N"s) in the sequences, likely stemming from the limitations of the underlying short-read sequencing and reflected in the LAI value of 0 (Table 1). In contrast, other accessions

(excluding the short-read sequenced Yunkang10) display a maximal variation across the total repeats of approximately less than 2% only (Table S3). The Gypsy element was found to have the highest genome coverage of all analysed genomes, with DASZ having the highest Gypsy coverage. The second most TE genome coverage among all analysed genomes was by "Unknown LTRs", followed by either Mutator or Copia elements. That said, these and other TE subclasses seem to be relatively consistent among all genomes when analysing the anchored genome regions (Figure 2b; Table S3; Figure S3). The Kimura distance-based copy divergence analysis revealed that the distribution of TE across all genomes was very similar; however, there were slightly more mutator (DNA-DTM) elements present in Huangdan, L618, and ZJ.

Comparative genomics and pangenomic gene family analysis

To shed light on genome dynamics, first all genomes were analysed using pairwise dot plots (Figure S4), revealing an overall high degree of genomic conservation. From the previously published genomes, both Huangdan and DASZ exhibited the highest genome-wide BUSCO gene completeness scores of 98.5 and 98.32% (Table 1), whereas Tieguanyin was haplotype resolved. Therefore, these genomes were earmarked as the potential most complete references to use for a comparison. Huangdan was used as a common reference for a more detailed analysis of the newly assembled genomes as it exhibited high similarities with the haplotype-resolved Tieguanyin. Hence, to compare the genome assemblies of AJ, ZJ, and L618, their chromosomes were mapped to the Huangdan genome, and detailed synteny analysis was performed using SyRI (Goel et al., 2019). The synteny analysis showed multiple translocations in all chromosomes of these assemblies compared to the reference genome Huangdan. There were some duplications in all chromosomes, but at the end of chromosome 12 of L618, large tandem duplication as well as inversion were present. There were also large inversions in ZJ, and AJ at chromosomes 4, 5, 6, 7, and 12. Although chromosome 15 has a large inversion at the end for AJ, it depicts the highest synteny among all chromosomes (Figure 3).

Orthofinder was used to construct the gene-family-based pangenome of the three newly assembled genomes with eight previously published genomes. This resulted in 40 600 gene families from 11 genomes, where 39, 15, 45, and only 1% were classified as core, soft-core, shell, and accession-specific gene families, respectively. The total gene set increased as more genomes were added and the core genes decreased (Figure 4a). There were more functionally annotated genes within the core (98.52%) and soft-core (97.20%) sets than in the shell (87.36%) and specific (85.97%) sets (Figure 4b). This trend aligns with the commonly observed pattern in previous pan-genome studies, where core genes tend to exhibit greater functional conservation (Jayakodi et al., 2020; Kang et al., 2023; Li et al., 2022; Liu et al., 2020; Wu et al., 2023). The results of the gene ontology (GO) enrichment analysis revealed that the core genes exhibited significant enrichment across a broad spectrum of 1364 generally conserved biological processes, encompassing RNA polymerase I preinitiation complex assembly, responses to oxidative stress, meiotic jasmonic acid biosynthetic processes, salt responses, regulation of photoperiodism, flowering, defence responses, transition metal ion transport, maltose catabolic processes, and DNA methylation, among others. Furthermore, these core genes demonstrated enrichment in 1014 resolved molecular functions, often representing small classes necessary for cell function, including rDNA binding, tRNA-intron endonuclease activity, C5 sterol desaturase activity, adenyl-nucleotide exchange factor activity, GTPase activator activity, myosin phosphatase activity, peptidase activity, and binding activities involving DNA, RNA, and proteins, as well as NADH kinase activity. Additionally, they were found to be enriched in 284 cellular components, such as the SAM complex, outer kinetochore, transcription factor TFIIA complex, nuclear inner membrane, transcription regulator complex, katanin complex, etioplast, plastid envelope, and chloroplast stroma thylakoid. Similarly, Mercator-based analysis at the pathway level identified almost all major pathways as enriched (Figures S5 and S6; Table S5).

Similarly, the softcore genes exhibited significant enrichment in 741 biological processes, ranging from spliceosomal complex assembly to xylulose and glycerol biosynthetic processes, DNA ligation, and peptide metabolic processes. Their 541 molecular functions included glycerol-1-phosphatase activity, biotin synthase activity, transaldolase activity, xylulokinase activity, xylose isomerase activity, telomerase inhibitor activity, and centromeric DNA binding activity. Additionally, the softcore genes were enriched in 130 cellular components, such as the plastid large ribosomal region, golgi lumen, chloroplast isoamylase complex, methylosome, and DNA-replication preinitiation complex (Figures S7 and S8; Table S6).

Meanwhile, the shell genes exhibited significant enrichment in 439 biological processes, ranging from lipid hydroxylation to the regulation of COPII vesicle coating, the IMP biosynthesis process, protein acetylation and deacetylation, and the glucose-mediated signalling pathway. Their 295 molecular functions included NADH dehydrogenase activity, ABC haeme transport activity, ion binding, mannose-binding, glutamate N-acetyltransferase activity, histone H3R2/R3 demethylase activity, glycolipid binding, and lipid kinase activity. Moreover, the shell genes were enriched in 65 cellular components, such as the primary cell wall, photosystem II antenna complex, plastid small

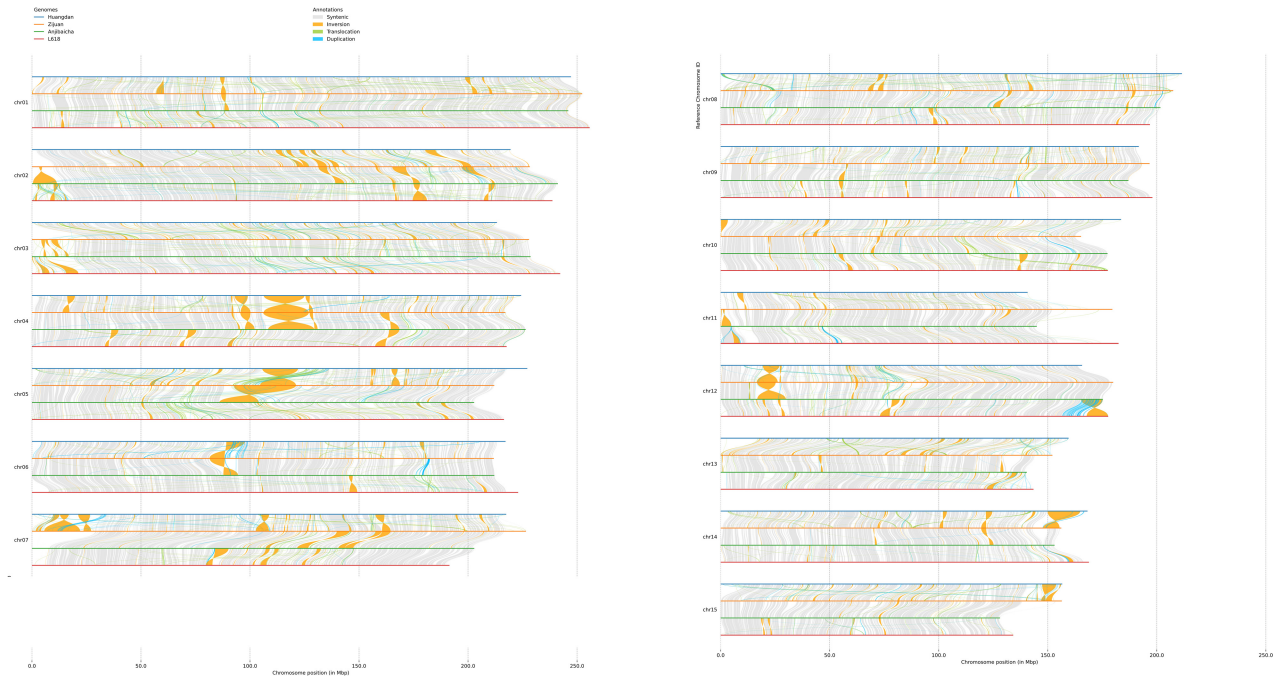

Figure 3. Detailed synteny analysis of the three new genomes AJ, ZJ, and L618 versus the published Huangdan reference.

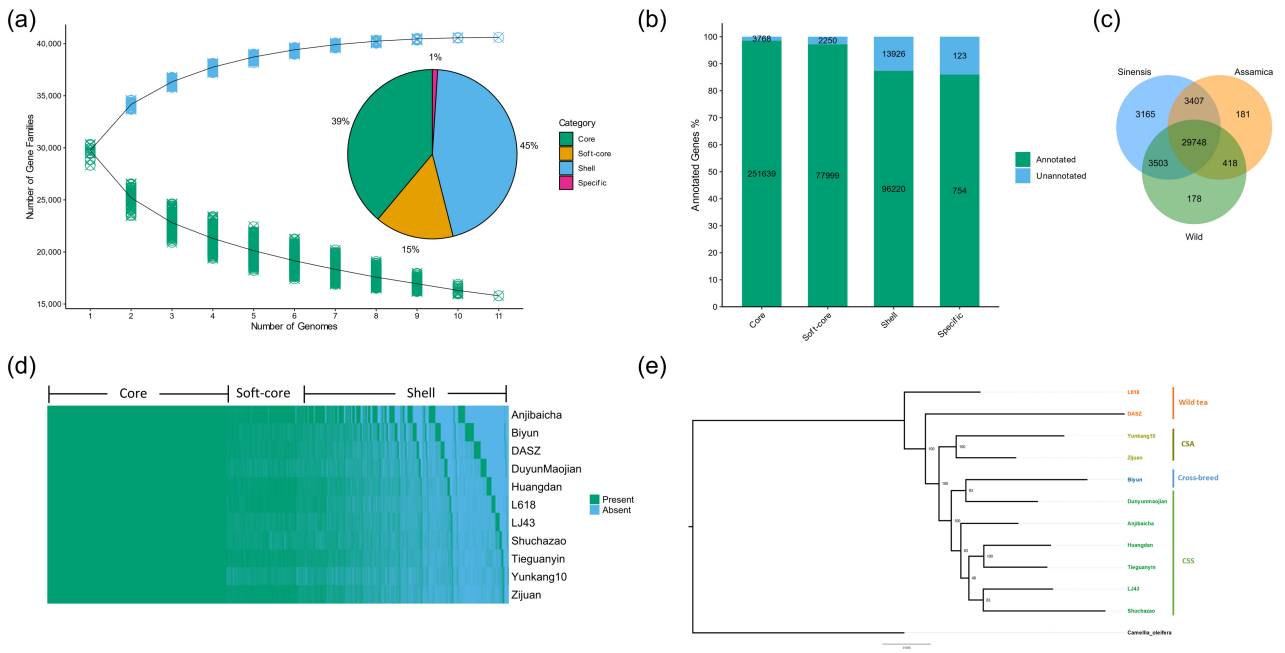

Figure 4. Gene centric pangenome analysis. (a) Pan and core genome size simulated using gene families clustering and gene family categories in terms of core, soft-core, shell, and accession specific.
(b) Proportion of gene annotation in each category.
(c) Venn diagram shown shared and unique gene families in *sinensis*, *assamica*, and wild varieties.
(d) Presence and absence of genes in 11 *Camellia sinensis* accessions where green indicated presence and blue absence.
(e) Phylogenetic tree depicting the evolutionary relationships among 11 genomes, constructed based on ortholog information and employing the GTR + F + R5 model.

ribosomal subunit, plasma membrane raft, storage vacuole, and aleurone grain membrane, whereas Mercator analysis showed some damage response and even secondary metabolism genes (Figures S9 and S10; Table S7).

Analysis of tea genome varieties

To analyse the different varieties for their genic content, we identified 29 745 gene families shared between *sinensis*, *assamica*, and wild tea, while these accessions exhibited 3165, 181, and 177 unique gene families, respectively. Wild and *sinensis* tea accessions demonstrated a greater sharing of unique gene families, with 3503 families in common, surpassing the *assamica* accessions, which shared 418 unique gene families with wild tea, while also sharing 3410 unique gene families with *sinensis* tea accessions (Figure 4c–e).

In the wild tea's GO enrichment analysis (i.e., DASZ and L618), 20 biological processes, such as response to maltose, histone H3-K27 demethylation, pyrimidine nucleotide biosynthetic process, pDNA replication proofreading, negative regulation of histone acetylation, base-excision repair, and gap-filling, were significantly enriched. Additionally, eight cellular components, predominantly related to ribosomes, and 17 molecular functions, primarily associated with kinase, phosphatase, and lipase activities, were significantly enriched (Figure S11a). Similarly, when using the Mercator MapMan ontology, a total of five mercator bins were significantly enriched, three related to cytoskeleton organization and others encompassing lipid metabolism (PLIP1) photosynthesis phosphoglycolate phosphatase (Figure S12a; Table S8).

In *assamica* varieties, 24 biological processes, such as regulation of the carotene biosynthetic process, positive regulation of developmental growth, homogalacturonan metabolic process, pigment biosynthetic process, and positive regulation of the cellular response to phosphate starvation, were identified. Only one cellular component, the SWI/SNF complex was significantly enriched, along with 15 molecular functions, including catechol oxidase activity, P-type transmembrane transporter activity, tryptophan synthase activity, and spermidine feruloyl/caffeoyl/sinapoyl CoA N acyltransferase activity (Figure S11b). Referring to the Mapman Mercator ontology, *assamica* varieties exhibited 14 enriched mercator bins, featuring functions like secondary metabolism, aureusidin synthase, aurones in secondary metabolism, solute transport, DNA damage response, protein homeostasis, and nutrient uptake (Figure S12b; Table S9).

In the *sinensis* tea accessions, the GO enrichment analysis revealed 69 biological processes, including response to low light intensity stimulus, fruit septum development, xyloglucan catabolic process, lipid hydroxylation, chloride ion homeostasis, and chlorophyll catabolic process. Twenty-two cellular components, such as the primary cell wall, transcription preinitiation complex, and pericentric heterochromatin, were identified, along with 58 molecular functions related to fatty acid in-chain hydroxylase activity, omega-3 fatty acid desaturase activity, sterol transporter activity, and phosphoric ester hydrolase activity (Figure S11c). Within the MapMan ontology, *sinensis* accessions demonstrated 24 significantly enriched pathways, encompassing diverse functions such as DEX2 involved in plant reproduction, external stimuli response elicitor peptide precursor (proPEP), signal transducer RHA2, RNA processing splicing factor MID1, the phytase PHY2, RNA processing CWC16, multi-process regulation of phytase activities, and RNA processing catalytic component – KAE1 (Figure S12c; Table S10).

A total of 2019 single-copy genes were identified in 11 tea plants and oil tea plants, and these were used to construct a phylogenetic tree (Fig 4e). The wild teas named L618 and DASZ were closely related to the outgroup species *C. oleifera*. Furthermore, two *Camellia sinensis* var. *assamica* (CSA) named Zijuan and Yunkang10 shared a common ancestor and showed obvious divergence compared with other *C. sinensis* var. *sinensis* (CSS) such as Huangdan, Tieguanyin, Shuchazao, LJ43, and Anjibaicha. Interestingly, Biyun has a relatively independent phylogenetic position compared with CSS, as it was a crossbreed of CSA and CSS in line with breeding records (National Inspection Variety of China, GS13044-1987).

Due to the importance of catechins, we analysed CNV in the genes involved in the catechin biosynthesis pathway (Figure 5). This revealed that the phenylalanine ammonia-lyase (PAL) gene is likely present in six copies in all accessions except DuyunMaojian and Yunkang10, where we found seven and four copies, respectively. Cinnamate 4-hydroxylase (C4H) exhibited two copies in all accessions except Shuchazao which has one extra copy in chromosome 11, which might be due to the high duplication in the genome, and Yunkang10 has only one copy. We detected three copies of 4-coumarate:CoA ligase (4CL) except Shuchazao, which has two copies in the anchored sequences while we identified a third copy on a non-chromosome anchored contig. Chalcone synthase (CHS) exhibited diverse copy numbers among all accessions, with six copies in DASZ, DuyunMaojian, L618, and Tieguanyin, while Biyun, Huangdan, and LJ43 have five copies. Anjibaicha has seven copies, while Yunkang10 has two copies, and Shuchazao has also exhibited two copies in anchored sequences and three copies in unanchored sequences. Chalcone isomerase (CHI); however, had consistently two copies in all accessions except LJ43 and Yunkang10, with only one copy and three copies in Biyun. Similarly, Flavanone 3-hydroxylase (F3H) was also present consistently in two copies except in Yunkang10 with one copy. For flavonoid 3′,5′-hydroxylase (F3′H), all accessions exhibited one gene on chromosome 15, whereas Biyun and Shuchazao

have two copies on chromosome 15, whereas no copy could be detected in Yunkang10. Dihydroflavonol 4-reductase (DFR) was also diverse, like CHS. We identified four copies of DFR in AJ, Biyun, DASZ, Huangdan, L618, ZJ, and Tieguanyin. Shuchazao has five copies, LJ43 and Yunkang10 have three copies, and DuyunMaojian has only one detectable copy. Leucoanthocyanidin reductase (LAR) has three copies in all accessions except Biyun and Huangdan, which have four and two copies, respectively. UDP-glucose: galloyl-1-*O*-β-D-glucosyltransferase (UGGT) has one copy in all genomes on chromosome 13 except Yunkang10 in which it is present in unanchored sequences (Figure 5; Table S11a,b). We detected three copies of Anthocyanidin synthase (ANS) in all accessions except Biyun, which has an extra copy on chromosome 14, and ZJ, AJ, and Tieguanyin, which has two functional ANS copies. Anthocyanidin reductase (ANR) showed consistent presence with two copies across all accessions, barring DASZ, Shuchazao, and Yunkang10, where a single copy is observed.

As ANS could be particularly important in case of the purple-leaved ZJ accessions, we deepened our analysis for ANS. Our pan-genome analysis identified ANS orthologs on chromosomes 12 and 14 in all accessions except DASZ, where we identified its orthologues on chromosome 4 instead of 12 (Figures S13 and S14). On chromosome 4 of DASZ and chromosome 12 of all other accessions, two ANS orthologues are present. Although we found ANS copies are conserved, they exhibited a somewhat varying degree of sequence similarity between the accessions in this locus. For instance, Huangdan and L618 exhibit high conservation compared to L618 and LJ43, where there is a significant variation in the ANS locus at chromosome 12. Indeed, for DASZ and ZJ, we even identified two and one gene, respectively, between the two ANS copies, which are present in a tandem configuration in the other accessions (Figure S13). The first ANS copy on chromosome 12 is partially present in ZJ, AJ, L618, and TGY. Sequence alignments of the ANS coding sequence from BY to ZJ revealed a premature stop codon in the first ZJ ANS copy (position 415) (Figures S13, S15 and S16). In ZJ, the second copy of the ANS gene on chromosome 12 has four unique substitutions in protein sequence, which can be observed in the amino acid alignment (position 47, 264, 299, and 366; Figure S15). We compared the ZJ ANS locus on chromosomes 12 and 14 with the recently published ZJ assembly (Chen, Wang, et al., 2023). Indeed, the copy numbers of ANS on chromosome 12 do not differ, and we identified a premature stop codon in one ANS copy (Figure S13). Interestingly we also observed three transposon insertions into this ANS copy, one DNA transposon and two LINE elements. In the syntenic ANS locus on TGY, only the first exon of the ANS copy and the two LINE elements were identified while the DNA transposon is missing. Except for ZJ and TGY, all other accessions harbour the DNA transposons in this syntenic ANS copy but are missing the LINE elements (Figure S13). Additionally, on chromosome 14, we identified one ANS, which was not annotated in Chen, Wang, et al. (2023). Given its consistent presence across all analysed accessions, except for Biyun which harbours two ANS orthologues separated by 214.78 kb (Figure S14), it appears that this region is missing in the Chen assembly (Figure S17).

Individual-specific presence variation and leaf colour-related genes

To identify individual specific presence variations (ISPV), all the available tea genomes were aligned to the genome of AJ, ZJ, and L618. In those three individuals, 759, 665, and 854 variations with average lengths of 7819, 5888, and 6543 bp were identified as ISPV and located within 234, 190, and 259 coding gene regions, respectively. For AJ and ZJ, a total of 14 671 and 13 998 genes showed differential expression levels between young leaves and mature leaves. Only 84 and 62 genes were located in the ISPV regions, and some of them had been annotated as leaf colour-related genes in the NR database (Table S12). CSAJ02G001820.1 was identified as Phosphoenolpyruvate/phosphate translocator (PPT) and CSAJ06G017860.1 was a MYB-related transcription factor named LHY in AJ. CSZJ01G004600.1 and CSZJ05G027170.1 were annotated as serine carboxypeptidase-like gene (SCPL) and CSZJ08G017510.1 was annotated as bHLH36 transcription factor in Zijuan. To detect potential candidate genes, we also checked differentially expressed genes (DEGs) located within 500 kb on both sides of the ISPV. A total of 649 showed differential expression patterns in young and mature leaves in AJ, those DEGs not only has been involved in KEGG pathways such as phosphatidylinositol signalling system, and brassinosteroid biosynthesis, which could regulate plant cell proliferation and differentiation, but also has been involved in some amino acid synthesis and photosynthetic related metabolic pathways such as porphyrin and chlorophyll metabolism, photosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, nicotinate, and nicotinamide metabolism. In ZJ, seven DEGs were defined as serine carboxypeptidase-like, which represents a large gene family involved in many biological processes, and the KEGG enrichment of the DEGs were focused on phagosome, nitrogen metabolism, galactose metabolism, and cutin, suberine, and wax biosynthesis (Table S13).

Comparative analysis of tea pangenomes

Finally, we compared our pangenome to the one of Chen, Wang, et al. (2023) that recently became available. Chen. Wang, et al. analysed a large set of genomes using first-generation PacBio sequencing. We speculated that this

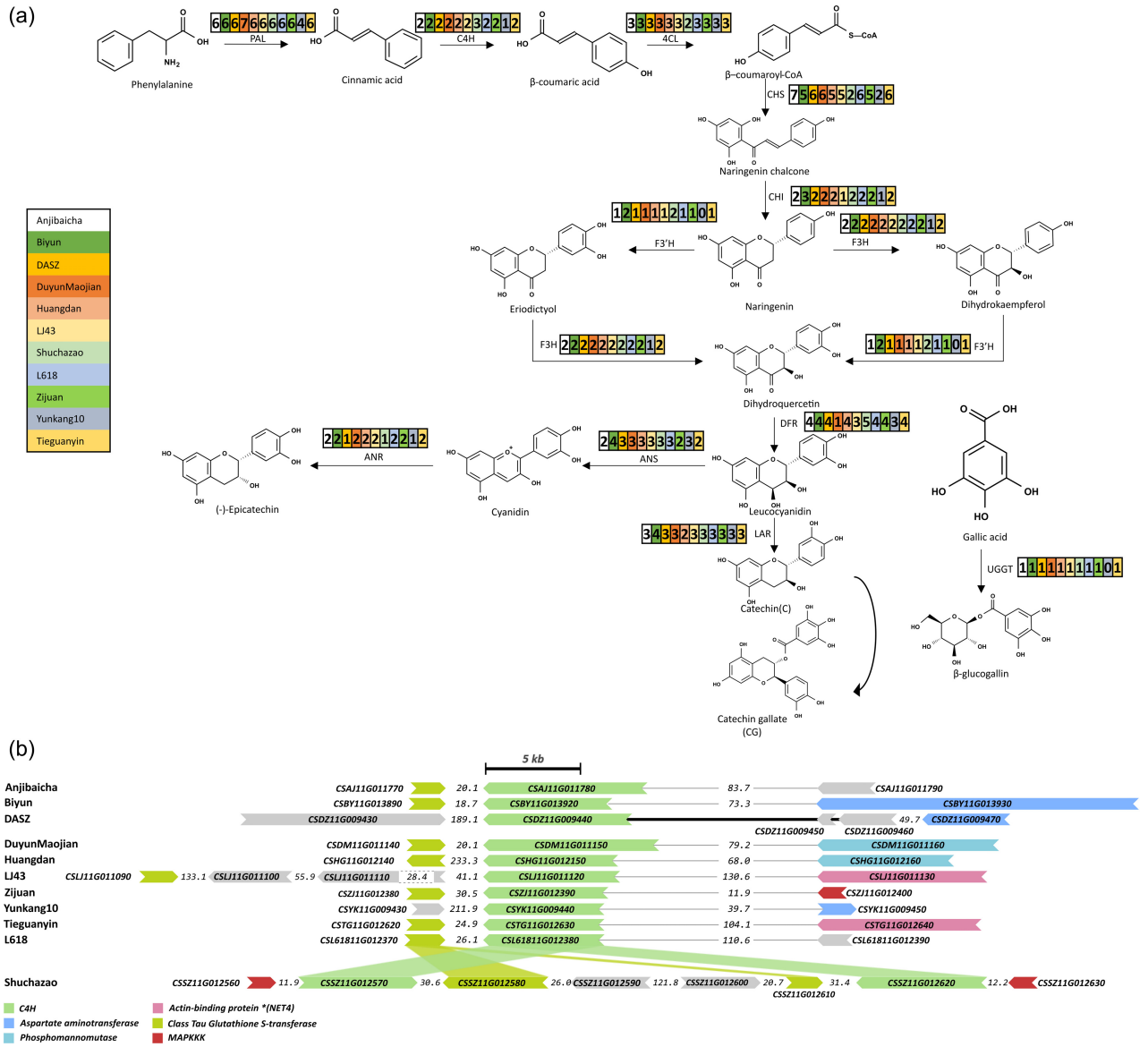

Figure 5. Analysis of the catechin pathway. (a) Catechin pathway in *Camellia sinensis* with protein copy number variation involved in the pathway that has been identified in anchored sequences of each accession. Phenylalanine ammonia-lyase (PAL), Cinnamate 4-hydroxylase (C4H), 4-Coumarate:CoA ligase (4CL), chalcone synthase gene (CHS), Chalcone isomerase (CHI), Flavanone 3-hydroxylase (F3H), flavonoid 3′,5′-hydroxylase (F3′H), Dihydroflavonol 4-reductase (DFR), Leucoanthocyanidin reductase (LAR), UDP-glucose: galloyl-1-*O*-β-D-glucosyltransferase (UGGT), anthocyanidin synthase (ANS), anthocyanidin reductase (ANR). (b) Schematic view of the genomic location of C4H in the Catechin pathway on chromosome 11 from all analysed accessions. The distances for genes being more than 10 kb apart are indicated by numbers in kb while genes are drawn to scale, except gene CSLJ11G011110 with an annotated length of 28.4 kb. Genes with the same functional annotations appearing more than once are colour coded as shown in the legend. Duplications on chromosome 11 in Shuchazao are indicated by coloured boxes linking the duplicated genes.

might lead to a somewhat lower quality in genome assemblies, which could be reflected in gene completeness as in the case of DuyunMaojian where new polishing algorithms improved the genome. Hence, we firstly assessed BUSCO gene completeness using the BUSCO gene sets with both the more sensitive compleasm (Huang & Li, 2023) and also using the original BUSCO tool (Simão et al., 2015) for our three new genomes as well as for the released genomes of Chen, Wang, et al. (2023). While compleasm yielded a near-perfect gene completeness for our three new genomes of at least 98.93% for three new genomes, the two best genomes from Chen reached 98.32%. Similarly, assessing the QV values, our genomes reached values of 62 and above, whereas the ones from Chen did not surpass a QV value of 43 when assessed with mercury (Table S14). The genome completeness values are

reflected in the annotation values. Here, we obtained BUSCO completeness values of 95% and above for our three genomes (Table S15) whereas the best genome "HD" from Chen reached a value of 91.9 when assessed with the eudicot conserved BUSCO gene set or 92.3% when assessed with the corresponding embryophyta BUSCO gene set. We obtained these high completeness values with less genes and transcripts than Chen, Wang, et al. (2023). As our data was not tuned by the BUSCO gene set, we wondered if this might be a better reflection of the true tea gene set and assessed these data using Mercator (Schwacke et al., 2019) which provides a consistent annotation and classification platform and indeed obtained annotation rates of 95% and classification rates of above 50% both in line with typical conserved plant genomes surpassing those of Chen, Wang, et al. (2023).

DISCUSSION

Here, we present three novel *de novo* tea genome assemblies that complement existing tea genome assemblies. Using the latest PacBio HIFI sequencing technologies, we were able to generate very contiguous genome assemblies that compared favourably to existing genome assemblies especially considering their BUSCO gene completeness, mercury and the new CRAQ metric, likely owing to the high precision of this technology. Interestingly for the LTR LAI indices, our genomes were similar to previous long-read technology tea assemblies when these were assembled to high contiguity, despite the AJ genome showing the best overall value. This likely indicates that either a dual long-read sequencing approach combining PacBio and Nanopore data might be necessary to further enhance LAI indices as done by van Rengs et al. (2022) and which is now directly available from the Hifiasm tool and/or additional sequencing and improvements could also further increase these values. Considering that LAI values of around 10 are already considered very good and Ou, Collins, et al. (2022) classified LAI indices above 10 as reference genomes and above 20 as gold standard genomes. A recent study comparing all available plant genomes (Mokhtar et al., 2023), showed that most genomes exhibited an LAI value of about eight, representing draft genomes of the current genome assemblies.

The genetically diverse *C. sinensis* species was thoroughly analysed in a recent pangenome study (Chen, Wang, et al., 2023). However, upon comparing these 22 genome assemblies and their annotations with our three assembled genomes and corresponding annotations, it became evident that our high-quality genome assemblies not only facilitate more complete genome annotations but also empower us to conduct a more exhaustive CNV analysis especially in case of the ANS gene (Tables S11a,b, 14 and 15). Thus, we could use the existing genomes to characterize the tea genomes in more detail. First, we analysed the phylogenetic relationship of 11 teas constructed by single copy genes that showed the obvious and expected differentiation of *sinensis* and *assamica* accessions and the independent evolution branch of the wild teas DASZ and L618, indicating a large genetic distance between ancient tea plants and the two cultivated varieties. The accessions we selected are representative enough for filling the gaps in the tea plant genome by providing a more continuous reference genome of *sinensis*, *assamica*, and wild tea plants.

In the pangenome analysis, the enrichment analysis showed that in the *assamica* tea varieties, pigment biosynthetic process, spermidine-related activities which are enriched in SWI/SNF complex, are directly involved in anthocyanidin synthase which contribute to leaf colour (Tang et al., 2020; Wang et al., 2017). Spermidine also plays a role to protect plants against high temperature stress by improving photosynthesis and regulating flavonoid biosynthesis (Sobolev et al., 2008; Sun et al., 2022). Mercator enrichment similarly showed that the *assamica* tea was enriched in secondary metabolism functions particularly aureusidin synthase and aurones (Figure S12b). These components play a crucial role in altering leaf colour, suggesting that this pathway plays a significant role in determining the coloration of ZJ leaves (Shakya et al., 2012). There are four extra copies of potential aureusidin synthases in ZJ on chromosome 3, which are completely absent in other accessions (Figure S18). DNA damage response, and protein homeostasis pathways might directly contribute to heat tolerance and anthocyanidin synthase (Ju et al., 1995; Nisa et al., 2019).

By performing analysis of DEGs in ISPV, we expected to find some genes associated with species-specific traits. Anjibaicha is a temperature sensitive albino mutant and rich in amino acids, PPT is a translocator located in the plastid inner envelope which could play an essential role in the development of chloroplasts and affect leaf colour (Tang et al., 2022), and the LHY was categorized as a morning gene of the circadian clock (Joo et al., 2017), which regulates the synthesis of photosensitive pigments which could change the colour of the tea plant (Yue et al., 2021).

The purple leaf of ZJ was mainly due to the large accumulation of anthocyanins, the family of SCPL had been presumed to be a regulator of anthocyanin acylation in carrot storage root (Curaba et al., 2019) and reported to be related to the accumulation of galloylated catechins in tea plants (Ahmad et al., 2020). Furthermore, many bHLH factors have been revealed to be associated with anthocyanin synthesis (Gonzalez et al., 2008; Hichri et al., 2010; Su et al., 2020), and bHLH36 was identified as a hub gene in flavonoid accumulation (Huang et al., 2018) as well as being involved in maize carotenoid metabolism (Ellison et al., 2017). Although the DEGs of ISPV were found in both this study and some previous reports, more

experiments are still needed to confirm their function in subsequent research. The cultivar ZJ is well-known for its characteristic colour, taste, and aroma, and its pharmacological potential for prevention of some chronic diseases due to its high accumulation of anthocyanidins (Khoo et al., 2017). Previous studies found that delphinidin, cyanidin, pelargonidin, and their glycoside derivatives were the major anthocyanins in ZJ and significantly contributed to its purple colour (Chen, Yang, et al., 2023; Jiang et al., 2013). The synthesis of anthocyanin downstream of the flavonoid biosynthetic pathway and flavonoids and procyanidins abundant in ZJ provide more substrates for anthocyanins accumulation. Increasing studies leveraging transcriptomic and metabolomic analyses have been conducted and a couple of genes related to the purple colour in tea plants were found (Huang et al., 2022; Sun et al., 2016). For example, transcriptomic analyses suggested the significant up-regulation of anthocyanidin synthase gene (*Cs*ANS1) and *Cs*AN1, a MYB family gene, positively regulated the expression of *Cs*ANS1, in anthocyanin-rich tea plants. Furthermore, a deletion mutation of *Cs*AN1 was found by cloning and alignment in anthocyanin-lacking cultivars (Huang et al., 2022). The function deficiency of *Cs*AN1 is predominantly responsible for the inability of anthocyanin accumulation, and this trait is heritable in progenies through hybridization (Huang et al., 2022). However, the genetic basis and molecular mechanisms of flavonoid and anthocyanin accumulation in tea plants have not been well elucidated.

To bridge this gap using our pangenome analyses, alongside published papers and our own data we first attempted comprehensive gene mining from the following representative publications with both candidate enzymes and transcription factors being identified but little functional evidence accrued to date (Huang et al., 2022; Sun et al., 2016). Secondly, comprehensive gene mining from our own mGWAS results of flavonoids, and our published parallel transcriptomic and metabolomic analyses were conducted (Qiu et al., 2020, 2023). Thirdly, while tea flavonoid and anthocyanin biosynthesis are largely unknown, glycosylated anthocyanidins are abundant and relatively stable in Zijuan (Chen, Yang, et al., 2023). Given that such modified anthocyanins are stored in the vacuole, one would assume a large number of UGTs would be identified in this accession. We therefore searched for such candidate genes here – finding that *Cs*ANS1 in Zijuan showed specific mutations in the coding region and led to mutations in several amino acids. In addition, we could identify a tandem repeat of three UGTs (somewhat similar to anthocyanidin 3-*O*-glucosyltransferase 5) where ZJ featured an insertion of about 180 bp in the 300 bp upstream region of the first UGT (CSZJ01G028560.1) As such, our newly constructed pan-genome facilitates in-depth exploration of the genomic diversity in tea varieties and would further aid in biological research and genetic improvements of *C. sinensis*.

Catechins are major polyphenol compounds that have a crucial role in determining the flavour of the tea and have numerous health benefits, including inhibition of growth of cancerous human colon (Uesato et al., 2001), reduction of inflammatory response (Fan et al., 2017), and neurodegenerative disease prevention (Pervin et al., 2018). In this study, we present a comprehensive exploration of the catechin biosynthesis pathway within tea genomes, leveraging multiple accessions to unravel the intricate genetic landscape governing this important secondary metabolite. The CNV among different tea genomes reveals the dynamic genomic architecture that underlies the diversity in catechin production among different accessions. This comprehensive analysis not only enhances our understanding of the genetic basis of catechin biosynthesis in tea plant but also provides a valuable resource for the future improvement of tea varieties through targeted breeding strategies. The identification of specific genes associated with catechin biosynthesis allows for the development of molecular markers that can be employed in marker-assisted breeding initiatives. Moreover, the CNVs among tea genomes provide breeders with a nuanced perspective, offering opportunities to harness genetic diversity for the development of novel tea cultivars with improved stress resistance, and quality.

Comparative analysis and outlook for future tea pangenomics

In comparison to the recently published tea pangenome (Chen, Wang, et al., 2023), we observed better QV values and better BUSCO completeness scores. Especially, the QV was more than 10 points larger, indicating 10-fold lower potential base error rates due to the logarithmic scaling of QV values. However, this is not too surprising and likely reflects the improved sequencing technology used for our three new genomes, which rely on highly precise HIFI PacBio reads, and they can be leveraged by the gene assembly algorithms. The improved base error precisions also make the assembly process less complex and more accurate due to more precise read alignments (Cheng et al., 2021). The improved assembly quality allowed us to analyse CNVs within biological pathways (e.g., the catechin pathway) and specific loci (e.g., the ANS locus) in more detail than previously reported. However, at the same time, it is impossible to exclude that these differences are based on underlying biology in the individual specimens that were explored.

Similarly, the genome precision was likely causal for the gene and protein annotation values where our HIFI read-based genomes obtained better completeness values. At the same time, we obtained more compact gene annotations potentially reflecting a more faithful representation of

the tea gene level as we obtained a high functional annotation and classification rate in line with other plant genomes. However, it cannot be excluded that tea does indeed contain several thousand additional genes whose function remains elusive and which are less well conserved in plants. In any case, as we obtained lower BUSCO gene and protein completeness values for our reannotation of legacy genomes, several important trends emerge. Firstly, with the advent of the more sensitive compleasm (Huang & Li, 2023) genome BUSCO values reach perfection when latest sequencing technology is used. However, this exacerbates the trend of gene structural annotation lagging slightly behind. Secondly, it seems that sequencing technology affects genome quality and has a pronounced effect on gene completeness which is why vertebrate genome projects stipulate extremely high underlying genome qualities mandating a QV of 60 and above (Rhie et al., 2021) which can be reached for tea using the latest HIFI reads. We expect that more HIFI based tea genomes will thus provide an even more precise analysis of the tea pangenome and our improvement of the DuyunMaojian shows that better genome assembly and analysis tools will aid as well. Further, additional deep long read genome and RNAseq data leveraging both PacBio and Nanopore platforms and more Hi-C data (Kong et al., 2023) will likely decrease the gap between the almost perfect genome BUSCO and annotated BUSCO sets. The tea pangenome era has thus just begun.

EXPERIMENTAL PROCEDURES

Plant material

The materials were planted in the tea germplasm resource nursery of Huazhong Agricultural University, Wuhan, China. The young leaves of three individuals named Anjibaicha, Zijuan, and L618 were sampled for whole genome sequencing, and the tender shoot of Anjibaicha was collected for the sequencing of Hi-C. The cetyltrimethyl ammonium bromide method was used to extract high-quality genomic DNAs from collected samples. Three SMRT libraries and a Hi-C library were constructed and sequenced on the PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA) and DNBSEQ-T7 platform (BGI, Shenzhen, China), respectively.

DNA extraction and assembly

The DNA from the tea plant samples was extracted and sequenced using highly accurate PacBio HiFi sequencing technology, a cutting-edge technique known for its accuracy and reliability. The resulting reads were assembled using hifiasm v0.19.5 (Cheng et al., 2021) with default parameters except -s 0.35. To further improve the assembly quality, Purge Haplotigs (Roach et al., 2018) was used to filter falsely duplicated contigs, ensuring a high-quality and accurate assembly. For the *Zijuan* assembly, we employed coverage cutoffs of -l 3 -m 27 -h 120 and utilized default parameters except for -a 80 to remove falsely duplicated contigs. Similarly, for the *Anjibaicha* assembly, we used coverage cutoffs of -l 2 -m 2 -h 95, while for the *L618* assembly, the coverage cutoffs were -l 3 -m 25 -h 100. To ensure the removal of falsely duplicated contigs, -a 80 with default parameters was used for all three accessions, while -l 8G, 2G, and 1G were used for *Anjibaicha*, *Zijuan*, and *L618*, respectively. For Hi-C scaffolding, *Anjibaicha* Hi-C Illumina reads were processed before YaHS scaffolding (Zhou et al., 2023) using the Arima pipeline. After YaHS scaffolding, the unanchored contigs with at least 80% similarity to the anchored sequences were removed using Purge Haplotigs (Roach et al., 2018), and chromosome orientation was harmonized using D-GENIES (Cabanettes & Klopp, 2018) dotplots against Huangdan (Wang et al., 2021). To further improve the assembly quality of *Zijuan* and *L618*, where no Hi-C data was available, we scaffolded the resulting clean assemblies against the high-quality reference genome of *C. sinensis* accession *Anjibaicha* using RagTag v2.1.0 (Alonge et al., 2022). We meticulously selected only the chromosomes of *Anjibaicha* to ensure the highest level of accuracy and reliability in our final assembly. To evaluate the completeness and accuracy of our final assembly, we performed a comprehensive assessment using compleasm (Huang & Li, 2023) with the eudicots_odb10 lineage. The *k-mer-based* quality assessment of the assembly was assessed using Merqury v1.3 (Rhie et al., 2020) with a *k-mer* size of 21 followed by Genomescope 2.0 analysis (Ranallo-Benavidez et al., 2020). The genome assembly (only anchored) was also assessed with LAI (beta 3.2) (Hufford et al., 2021; Ou et al., 2018) and with the reference-free tool Clipping Information for Revealing Assembly Quality (CRAQ) (Li et al., 2023) to identify structural and regional assembly error.

Genome annotation

Genomes were thoroughly annotated using a comprehensive pipeline that integrated both ab-initio and evidence-based approaches for gene prediction, ensuring the accuracy and completeness of the gene models. For ab-initio prediction, we utilized Helixer tool (Holst et al., 2023), which was first trained on a diverse set of tea genomes, including *Biyun* (Zhang, Li, et al., 2020; Zhang, Zhang, et al., 2020), *Huangdan* (Wang et al., 2021), *DASZ* (Zhang, Li, et al., 2020; Zhang, Zhang, et al., 2020), *LJ43* (Wang et al., 2020), and *Shuchazao* (Wei et al., 2018) accessions, resulting in highly accurate gene models. In the HelixerPost step, we used the parameters window size 100, edge thresh 0.10, peak thresh 0.90, and min coding length 60. For homology-based gene prediction, we employed a multi-step process, which involved mapping publicly available 78 RNASeq data (Table S16) from tea genomes using Hisat2 v2.2.1 (Kim et al., 2019) with default parameters except --dta, and subsequently merging all the mapped files using samtools v1.16.1 (Li et al., 2009) and sort with default parameters except -m 8G. We then utilized StringTie v2.2.1 (Pertea et al., 2015) to generate high-quality transcript assemblies with -c 50 and otherwise default parameters. Finally, we merged both the ab-initio and homology-based gene models using Mikado software v2.3.4 (Venturini et al., 2018) with weighting factors favouring extrinsic evidence (one for Helixer and three for StringTie) to obtain a consensus gene model with high accuracy and completeness. We utilized GffRead v0.12.4 (Pertea & Pertea, 2020) to extract protein and coding sequences (CDS) sequences from the final gene models, ensuring reliable functional annotation and classification. To get better protein-coding genes, we also reannotated other eight accessions DASZ (Zhang, Zhang, et al., 2020), DuyunMaojian (Wang et al., 2022), Biyun (Zhang, Li, et al., 2020), Huangdan (Wang et al., 2021), LJ43 (Wang et al., 2020), Shuchazao (Xia et al., 2020), Yunkang10 (Xia et al., 2017), and Tieguanyin (Zhang et al., 2021) with the above-mentioned method. For DuyunMaojian, we polished the existing genome with NextPolish v1.4.1 (Hu et al., 2020) using Oxford nanopore and Illumina reads from the same published study

(NCBI Bioproject: PRJNA841059) and subsequently ran our gene prediction pipeline to improve structural annotations.

To identify non-coding RNA genes, we employed RNAmmer (Lagesen et al., 2007) and tRNAscan-SE v2.0.12 (Chan et al., 2021) tools for rRNA and tRNA prediction, respectively. For miRNA identification, we utilized a homology-based approach by comparing each *C. sinensis* accession against the available plant miRNA database in miRBase using BLAST v2.2.26 (Altschul et al., 1990), with a stringent identity cutoff of >60% with parameters -W 6 -F -e 1e-5 -r 5 -q -4 -m 9 (Griffiths-Jones, 2010). For snoRNA identification, we employed the Infernal tool with the latest Rfam database v14.9 (Kalvari et al., 2021), ensuring reliable and comprehensive annotation of non-coding RNA genes. Furthermore, we utilized the advanced Extensive de novo TE Annotator (EDTA) v2.1.0 (Ou, Su, et al., 2022) to identify TEs within the tea genomes. The LTR, TIR, and Helitron modes were run separately, followed by a final EDTA run with -overwrite 0 and --cds options using Mikado's gene model CDS sequence, ensuring accurate and comprehensive annotation of TEs. For pan TE identification, the panEDTA module was used. The automated Mercator4 v6.0 pipeline with the ProtScriber extension (Schwacke et al., 2019) was used for functional annotation and classification, ensuring the reliable classification and interpretation of the annotated genes. BUSCO (Simão et al., 2015) with eudicot_odb10 lineage was used to assess the completeness of the annotated proteome.

Comparative genomics and gene family analysis

To compare the AJ, ZJ, and L618 genome assembly to the reference genome (Huangdan), all three genome assemblies (only chromosomes) were mapped to the reference genome using minimap2 (Li, 2021) and then the structural rearrangement and synteny analysis was performed using SyRI (Goel et al., 2019) and visualized by plotsr (Goel & Schneeberger, 2022). To analyse the gene families and shared orthology among all accessions, the gene clustering was performed using Orthofinder (Emms & Kelly, 2019), and only the longest protein was selected if there were multiple isoforms of a transcript. The gene families shared among all samples are defined as "core", gene families shared among 90% of samples are defined as "soft-core" and less than 90% are defined as "shell" gene families. If a gene family is only present in one accession, then it is considered as "specific". Then in the simulation to estimate the pan-genome size we used PanGP (Zhao et al., 2014) with the parameter algorithm: "totally random" algorithm, Sample Size: 500, and Sample Repeat 10.

To determine the phylogeny of available teas genomes, orthologous and paralogous gene families were clustered using coding sequences from 11 tea genomes, including Anjibaicha, Biyun, DASZ, DuyunMaojian, Huangdan, L618, LJ43, Shuchazao, Tieguanyin, Yunkang10, Zijuan, and an outgroup species *C. oleifera* by OrthoFinder v2.5.5 (Emms & Kelly, 2019) with default parameter. The multiple sequence alignment of single-copy genes was performed using MAFFT v7.459 (Katoh & Standley, 2013), and conserved blocks were selected by Gbocks v0.19b (Castresana, 2000). A supergene connected with the conserved blocks was used to construct the phylogenetic relationship by IQtree2 (Minh et al., 2020) based on the maximum likelihood method with 1000 times bootstrap.

To perform GO enrichment analysis, all proteins from the longest transcript present across the genomes were compared against SwissProt (release-2024_01) plant genes using BLAST with parameter "evalue 1e-3 -max_target_seqs 1". Subsequently, GO terms were transferred to the corresponding tea proteins. The enrichment analysis was executed in R using a Fisher exact test testing with the "greater" alternative, employing all genes from the Orthofinder analysis across the 11 genomes as background genes (i.e., 457 393 genes in total). For each enrichment analysis, genes within the core, soft-core, and shell categories were designated as genes of interest. The Benjamin-Hochberg method was applied to compute adjusted *P*-values, and only GO terms with an adjusted *P*-value <0.05 were deemed significantly enriched. Similarly, we employed the Mapman/Mercator ontology specific for plant terms (Bolger et al., 2021; Schwacke et al., 2019) relying on the automatic Mercator annotation. And for the Mercator term enrichment analysis, we used the online enrichment analysis tool, by selecting the parameter One-sided Fisher's exact test, over-representation analysis with FDR 0.05.

Catechin pathway CNV analysis

CDS sequences for the proteins involved in the Catechin pathway (PAL, C4H, 4CL, CHS, CHI, F3H, F3′H, DFR, LAR, UGGT) were selected from the corresponding Mercator4 v6.0 bins (PAL: 9.2.1.1.1, C4H:9.2.1.2, 4CL:9.2.1.3, CHS:9.2.2.2.1.1, CHI:9.2.2.3.1, F3H:9.2.2.5.1, F3′H:9.2.2.5.2, DFR:9.2.2.10.1, LAR:9.2.2.10.4, UGGT:19.1.1.1.2) for each analysed accession. The following CDS sequences from the accession Anjibaicha were used as baits in BLASTN searches against all analysed genomes: PAL: CSAJ01G004420.1, C4H:CSAJ11G011780.1, 4CL:CSAJ02G029730.1, CHS:CSAJ02G014430.1, CHI:CSAJ07G022710.1; DFR:CSAJ02G008660.1, FH3:CSAJ09G026400.1, F3′H:CSAJ15G013850.1, LAR: CSAJ06G006910.1, UGGT:CSAJ13G020450.1, ANS:CSAJ12G017530.1, ANR:CSAJ08G004240.1. BLASTN alignments (blastn -task dc-megablast) with a min. *E*-value of 10^{-01} on the same chromosome and clustering within a 60 kb window were extracted from the corresponding assembly and visually inspected using dotplots to identify full-length copies of the corresponding CDS. The molecular structures of the compounds in the catechin pathway were retrieved from KEGG in mol format and converted into SVG format using Open Babel (O'Boyle et al., 2011).

ANS synteny analysis

ANS loci were identified from the annotations on chromosome 12 (chromosome 4 on DASZ) and chromosome 14 and extracted from the corresponding assemblies. Syntenic regions were identified by dotplot analysis and the ANS genes manually re-annotated. In addition, we used KIPEs (Rempel et al., 2023) on protein sequences to identify ANS genes.

Individual specific presence variation

All the other tea genomes assembled based on the long reads of third-generation sequencing were aligned to the *de novo* assembly genomes of Anjibaicha, Zijuan, and L618 and the smartie-sv (Kronenberg et al., 2018) was employed to find pairwise structural variations. For each *de novo* reference genome, the sequence that could not be found in most other assembly genomes will be regarded as the individual-specific presence variation.

Expression analysis

The transcriptome data of young leaves and mature leaves of Anjibaicha and Zijuan were mapped to their genome respectively using HISAT2, and then StringTie v1.3.3b was employed for transcript assembly and FPKM computation. The R package DESeq2 v1.2 (Love et al., 2014) was used to identify the DEGs with *P*.adjust <0.05. ISPV overlapped DEGs were annotated with NR database (https://ftp.ncbi.nih.gov/blast/db/) using Blast+ (*E*-value >10^{-5}) (Camacho et al., 2009).

AUTHOR CONTRIBUTIONS

BU and WW jointly managed the project. AT, ARF, BU, WW, and MM wrote the manuscript. AT, BU, WW, MM, SB, and JPB performed the data analysis. XJ performed the experiments. AB contributed to initial assemblies and analyses.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (3211101118) NSFC-DFG collaborative project, National Natural Science Foundation of China (U23A20213), Fundamental Research Funds for the Central Universities (2662023PY011) to Weiwei Wen, Deutsche Forschungsgemeinschaft and the DFG-Project number 468870408 to Björn Usadel and Alisdair R. Fernie, and CEPLAS 390686111 to JPB. We thank Dr. Stella Eggels for data management. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Hi-C contact map of the Anjibaicha chromosome-length scaffold. The scaffolds are sorted according to their length, but chromosome scaffold numbering is given using the Huangdan order to facilitate comparison.

Figure S2. Genomescope 2.0 21k-mer profile for (a) AJ (b) ZJ (c) L618 generated from HiFi PacBio reads to estimate genome size, repetitiveness, and heterozygosity. The *x*-axis illustrates the diversity within the read set, while the y-axis portrays the cumulative frequency of these reads. The table below each graph is the genomescope summary output.

Figure S3. Transposable element dynamics on the anchored genomes for the analysed accessions.

Figure S4. Dotplot comparing all analysed accessions.

Figure S5. Core genes Gene Ontology (GO) enrichment analysis. The top 10 ontologies for each type based on enrichment ratio are shown.

Figure S6. Mercator-based core gene pathway enrichment analysis.

Figure S7. Soft-core genes Gene Ontology (GO) enrichment analysis. The top 10 ontologies for each type based on enrichment ratio are shown.

Figure S8. Mercator-based soft-core gene pathway enrichment analysis.

Figure S9. Shell Gene Ontology (GO) enrichment analysis. The top 10 ontologies for each type based on enrichment ratio are shown.

Figure S10. Mercator-based shell gene pathway enrichment analysis.

Figure S11. GO enrichment analysis for (a) wild accession, (b) cv *assamica*, and (c) cv *sinensis*. The top 10 ontologies for each type based on enrichment ratio are shown.

Figure S12. Mercator-based pathway enrichment analysis for (a) wild accession (b) cv *assamica* (c) cv *sinensis*.

Figure S13. Schematic diagram of the ANS region on chromosome 12 for all analysed accessions (chromosome 4 for DASZ). Homologous regions between the individual accessions are shown as grey and blue (inversion) blocks. The premature stop codon in the first ANS copy on Zijuan and the partial first ANS copy on Tieguanyin are indicated. Synthetic genes are indicated in green, while non-syntenic genes are indicated in red. Transposable elements identified in the first ANS copy on ZJ are indicated in yellow.

Figure S14. Schematic diagram of the ANS region of chromosome 14.

Figure S15. Multiple sequence alignments of all translated ANS coding sequences are present in all accessions. Red triangles indicate specific positions mentioned in the manuscript. The highlighted sequence CSZJ12G017560.1 contains the identified premature stop codon on position 139 identified on gene CSZJ12G017560.1. Sequence CSZJ12G017580 contains unique amino acid changes among all ANS orthologues.

Figure S16. Multiple sequence alignment of the first 417 bp from all ANS orthologue coding sequences identified in all accessions in this study. The identified premature stop codon on gene CSZJ12G017560.1 is emphasized in bold and indicated by the red arrow.

Figure S17. Exploring the ANS gene copy on chromosome 14 in Chen, Wang, et al. (2023) assembly through IGV visualization. (a) Representation of the ANS copy on chromosome 14 as assembled in this study (Zijuan). The top displays coordinates of the region, followed by PacBio read coverage and reads integrated into the assembly of this study. Subsequently, Chen, Wang, et al. (2023) read coverage and corresponding reads are depicted, along with genome annotation. (b) Corresponding ANS region in the assembly by Chen, Wang, et al. (2023), with the red rectangle highlighting the region showcased in panel (c). Similar to (a), the coordinates of the region are displayed at the top, followed by PacBio read coverage and reads integrated into the assembly by this study. Next, Chen, Wang, et al. (2023) read coverage and corresponding reads are presented, followed by genome annotation alongside the genome annotation generated in this study. (c) Zoomed-in view of the ANS region on chromosome 14, as indicated in panel (b).

Figure S18. Schematic diagram of the aureusidin synthase locus on chromosome 3 of Zijuan and its syntenic region in Anjibaicha. Syntenic genes are shown in green, while non-syntenic genes are shown in red. Grey regions indicate sequence homology between Anjibaicha and Zijuan.

Table S1. Telomere lengths for the newly sequenced tea accessions AJ, ZJ, and a wild tea L618. Unidentified telomeres in any accession are omitted, while telomeres not identified in individual accessions are indicated by "NA". The chromosome indicates the corresponding chromosome and telomere locations. The lengths are given in base pairs.

Table S2. Genes and additional genic features for all analysed assemblies in this study. All BUSCO analyses were performed using the eudicots_odb10 lineage. Avg len: average length; Pub: previously published data; Avg count/gene: Average count per gene.

Table S3. Coverage of identified transposable elements by panEDTA in Megabase Pairs for the anchored genomes of all analysed accessions in this study.

Table S4. Transposable element (TE) analysis summary for the whole assemblies (including unanchored sequences) of the newly assembled tea accessions Zijuan, Anjibaicha, and L618.

Table S5. Core genes Gene Ontology (GO) enrichment analysis.

Table S6. Soft-core genes Gene Ontology (GO) enrichment analysis.

Table S7. Shell genes Gene Ontology (GO) enrichment analysis.

Table S8. Wild accessions Gene Ontology (GO) enrichment analysis.

Table S9. Cultivar *assamica* Gene Ontology (GO) enrichment analysis.

Table S10. Cultivar *sinensis* genes Gene Ontology (GO) enrichment analysis.

Table S11. (a) Identified copy number variations of genes involved in the catechin pathway in *C. sinensis* and (b) Accessions for genes identified in the copy number variation analysis in the catechin pathway in *C. sinensis*.

Table S12. DEGs of AJ and ZJ found in ISPV.

Table S13. KEGG pathway analysis of DEGs found in ISPV.

Table S14. A comprehensive quality assessment of genome assemblies using BUSCO (v5.4.5) with eudicots_odb10 (Total orthologs groups = 2326) and embryophyta_odb10 (Total orthologs groups = 1614) orthologs databases, alongside Compleasm.

Table S15. BUSCO assessment of gene annotation. BUSCO (v5.4.5) was used with eudicots_odb10 (Total orthologs groups = 2326) and embryophyta_odb10 (Total orthologs groups = 1614), and BUSCO (v3.0.2) for embryophyta_odb9 (Total orthologs groups = 1440) as this database was used in the previously published study (Chen, Wang, et al., 2023).

Table S16. RNA-seq accessions used for annotation.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The genome and raw reads are available at https://www.cncb.ac.cn/ under the project accession of PRJCA006277, PRJCA006276 and PRJCA006159 as well as at ENA with accession PRJEB71966. The genome assembly of AJ, ZJ, and L618 is available under accession GCA_963931745, GCA_963931765, and GCA_963931755. Consistent genome annotation files, EDTA libraries and annotations etc are made available the DataHUB https://git.nfdi4plants.org/usadellab/Camellia_sinensis_genomics. It is available under DOI: https://doi.org/10.60534/30cpm-z1c67.

DATA AVAILABILITY STATEMENT

The genome and raw reads are available at China National Center for Bioinformation (https://www.cncb.ac.cn/) under the project accession of PRJCA006277, PRJCA006276, and PRJCA006159. Genome assemblies and raw reads are also available on the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/) under the project accession PRJEB71966. The genome assembly of AJ, ZJ, and L618 is available under accession GCA_963931745, GCA_963931765, and GCA_963931755. Consistent genome annotation files, EDTA libraries and annotations etc. are also being made available in the DataPLANT DataHUB (Weil et al., 2023) https://git.nfdi4plants.org/usadellab/Camellia_sinensis_genomics.
The genome browser for all the accessions included in our study is available at (https://www.plabipd.de/ceplas/?config=tea_pangenome.json).

REFERENCES

Ahmad, M.Z., Li, P., She, G., Xia, E., Benedito, V.A., Wan, X.C. *et al.* (2020) Genome-wide analysis of serine carboxypeptidase-like acyltransferase gene family for evolution and characterization of enzymes involved in the biosynthesis of galloylated catechins in the tea plant (*Camellia sinensis*). *Frontiers in Plant Science*, **11**, 848.

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S. *et al.* (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, **23**, 258.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

An, Y., Chen, L., Tao, L., Liu, S. & Wei, C. (2021) QTL mapping for leaf area of tea plants (*Camellia sinensis*) based on a high-quality genetic map constructed by whole genome resequencing. *Frontiers in Plant Science*, **12**, 705285.

An, Y., Mi, X., Zhao, S., Guo, R., Xia, X., Liu, S. *et al.* (2020) Revealing distinctions in genetic diversity and adaptive evolution between two varieties of by whole-genome resequencing. *Frontiers in Plant Science*, **11**, 603819.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J. & Edwards, D. (2020) Plant pan-genomes are the new reference. *Nature Plants*, **6**, 914–920.

Bell, E.A., Butler, C.L., Oliveira, C., Marburger, S., Yant, L. & Taylor, M.I. (2022) Transposable element annotation in non-model species: the benefits of species-specific repeat libraries using semi-automated EDTA and DeepTE de novo pipelines. *Molecular Ecology Resources*, **22**, 823–833.

Bolger, M., Schwacke, R. & Usadel, B. (2021) MapMan visualization of RNA-seq data using Mercator4 functional annotations. *Methods in Molecular Biology*, **2354**, 195–212.

Cabanettes, F. & Klopp, C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, **6**, e4958.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chan, P.P., Lin, B.Y., Mak, A.J. & Lowe, T.M. (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, **49**, 9077–9096.

Chen, S., Wang, P., Kong, W., Chai, K., Zhang, S., Yu, J. *et al.* (2023) Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nature Plants*, **9**, 1986–1999.

Chen, Y., Yang, J., Meng, Q. & Tong, H. (2023) Non-volatile metabolites profiling analysis reveals the tea flavor of "Zijuan" in different tea plantations. *Food Chemistry*, **412**, 135534.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, **18**, 170–175.

Curaba, J., Bostan, H., Cavagnaro, P.F., Senalik, D., Mengist, M.F., Zhao, Y. *et al.* (2019) Identification of an SCPL gene controlling anthocyanin acylation in carrot (*Daucus carota* L.) root. *Frontiers in Plant Science*, **10**, 1770.

Ellison, S., Senalik, D., Bostan, H., Iorizzo, M. & Simon, P. (2017) Fine mapping, transcriptome analysis, and marker development for Y2, the gene that conditions β-carotene accumulation in carrot (*Daucus carota* L.). *G3 (Bethesda)*, **7**, 2665–2675.

Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238.

Fan, F.-Y., Sang, L.-X. & Jiang, M. (2017) Catechins and their therapeutic benefits to inflammatory bowel disease. *Molecules*, **22**, 484.

Fang, Z.-T., Yang, W.-T., Li, C.-Y., Li, D., Dong, J.J., Zhao, D. *et al.* (2021) Accumulation pattern of catechins and flavonol glycosides in different varieties and cultivars of tea plant in China. *Journal of Food Composition and Analysis*, **97**, 103772.

Goel, M. & Schneeberger, K. (2022) Plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, **38**, 2922–2926.

Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, **20**, 277.

Gonzalez, A., Zhao, M., Leavitt, J.M. & Lloyd, A.M. (2008) Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. *The Plant Journal*, **53**, 814–827.

Griffiths-Jones, S. (2010) miRBase: microRNA sequences and annotation. *Current Protocols in Bioinformatics*, **12**, 12.9.1–12.9.10.

Gui, S., Martinez-Rivas, F.J., Wen, W., Meng, M., Yan, J., Usadel, B. *et al.* (2023) Going broad and deep: sequencing-driven insights into plant physiology, evolution, and crop domestication. *The Plant Journal*, **113**, 446–459.

Hichri, I., Heppel, S.C., Pillet, J., Léon, C., Czemmel, S., Delrot, S. *et al.* (2010) The basic helix-loop-helix transcription factor MYC1 is involved in the regulation of the flavonoid biosynthesis pathway in grapevine. *Molecular Plant*, **3**, 509–523.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F. *et al.* (2023) Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *bioRxiv*. 2023.02.06.527280. Available from: https://doi.org/10.1101/2023.02.06.527280

Hoopes, G., Meng, X., Hamilton, J.P., Achakkagari, S.R., de Alves Freitas Guesdes, F., Bolger, M.E. *et al.* (2022) Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular Plant*, **15**, 520–536.

Hu, J., Fan, J., Sun, Z. & Liu, S. (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, **36**, 2253–2255.

Huang, F., Duan, J., Lei, Y., Kang, Y., Luo, Y., Chen, Y. *et al.* (2022) Metabolomic and transcriptomic analyses reveal a MYB gene, CsAN1, involved in anthocyanins accumulation separation in F1 between 'Zijuan' (*Camellia sinensis* var. *assamica*) and 'Fudingdabaicha' (*C. sinensis* var. *sinensis*) tea plants. *Frontiers in Plant Science*, **13**, 1008588.

Huang, H., Yao, Q., Xia, E. & Gao, L. (2018) Metabolomics and transcriptomics analyses reveal nitrogen influences on the accumulation of flavonoids and amino acids in young shoots of tea plant (*Camellia sinensis* L.) associated with tea flavor. *Journal of Agricultural and Food Chemistry*, **66**, 9828–9838.

Huang, N. & Li, H. (2023) Compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics*, **39**, btad595.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J. *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.

Jayakodi, M., Golicz, A.A., Kreplak, J., Fechete, L.I., Angra, D., Bednář, P. *et al.* (2023) The giant diploid faba genome unlocks variation in a global protein crop. *Nature*, **615**, 652–659.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**(7837), 284–289.

Jiang, L., Shen, X., Shoji, T., Kanda, T., Zhou, J. & Zhao, L. (2013) Characterization and activity of anthocyanins in Zijuan tea (*Camellia sinensis* var. kitamura). *Journal of Agricultural and Food Chemistry*, **61**, 3306–3310.

Joo, Y., Fragoso, V., Yon, F., Baldwin, I.T. & Kim, S.-G. (2017) Circadian clock component, LHY, tells a plant when to respond photosynthetically to light in nature. *Journal of Integrative Plant Biology*, **59**, 572–587.

Ju, Z., Liu, C. & Yuan, Y. (1995) Activities of chalcone synthase and UDPGal: flavonoid-3-o-glycosyltransferase in relation to anthocyanin synthesis in apple. *Scientia Horticulturae*, **63**, 175–185.

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M. *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, **49**, D192–D200.

Kang, M., Wu, H., Liu, H., Liu, W., Zhu, M., Han, Y. *et al.* (2023) The pan-genome and local adaptation of Arabidopsis thaliana. *Nature Communications*, **14**, 6259.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Khoo, H.E., Azlan, A., Tang, S.T. & Lim, S.M. (2017) Anthocyanidins and anthocyanins: colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food & Nutrition Research*, **61**, 1361779.

Kidwell, M.G. & Lisch, D. (1997) Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 7704–7711.

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915.

Kong, W., Jiang, M., Wang, Y., Chen, S., Zhang, S., Lei, W. *et al.* (2022) Pan-transcriptome assembly combined with multiple association analysis provides new insights into the regulatory network of specialized metabolites in the tea plant *Camellia sinensis*. *Horticulture Research*, **9**, uhac100.

Kong, W., Yu, J., Yang, J., Zhang, Y. & Zhang, X. (2023) The high-resolution three-dimensional (3D) chromatin map of the tea plant (*Camellia sinensis*). *Horticulture Research*, **10**, uhad179.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S. *et al.* (2018) High-resolution comparative analysis of great ape genomes. *Science*, **360**, 6343. Available from: https://doi.org/10.1126/science.aar6343

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T. & Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35**, 3100–3108.

Lei, Y., Yang, L., Duan, S., Ning, S., Li, D., Wang, Z. *et al.* (2022) Whole-genome resequencing reveals the origin of tea in Lincang. *Frontiers in Plant Science*, **13**, 984422.

Li, C.-F., Xu, Y.-X., Ma, J.-Q., Jin, J.-Q., Huang, D.-J., Yao, M.-Z. *et al.* (2016) Biochemical and transcriptomic analyses reveal different metabolite biosynthesis profiles among three color and developmental stages in "Anji Baicha" (*Camellia sinensis*). *BMC Plant Biology*, **16**, 195.

Li, C.-F., Yao, M.-Z., Ma, C.-L., Ma, J.-Q., Jin, J.-Q. & Chen, L. (2015) Differential metabolic profiles during the albescent stages of "Anji Baicha" (*Camellia sinensis*). *PLoS One*, **10**, e0139996.

Li, H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, **37**, 4572–4574.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H. *et al.* (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature Communications*, **13**(1), 682.

Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. (2023) Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications*, **14**, 6556.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S. *et al.* (2020) Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.e13.

Liu, Z., Zhao, Y., Yang, P., Cheng, Y., Huang, F., Li, S. *et al.* (2023) Population whole-genome resequencing reveals the phylogenetic relationships and population structure of four Hunan typical tea landraces. *Beverage Plant Research*, **3**, 9.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. *et al.* (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, **37**, 1530–1534.

Mokhtar, M.M., Abd-Elhalim, H.M. & El Allali, A. (2023) A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index. *AoB Plants*, **15**, lad015.

Nisa, M.-U., Huang, Y., Benhamed, M. & Raynaud, C. (2019) The plant DNA damage response: signaling pathways leading to growth inhibition and putative role in response to stress conditions. *Frontiers in Plant Science*, **10**, 653.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. & Hutchison, G.R. (2011) Open babel: an open chemical toolbox. *Journal of Cheminformatics*, **3**, 33.

Ou, S., Chen, J. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, **46**, e126.

Ou, S., Collins, T., Qiu, Y., Seetharam, A.S., Menard, C., Manchanda, N. *et al.* (2022) Differences in activity and stability drive transposable element variation in tropical and temperate maize. *bioRxiv*. 2022.10.09.511471. Available from: https://doi.org/10.1101/2022.10.09.511471v1.abstract

Ou, S., **Su, W.**, **Liao, Y.**, **Chougule, K.**, **Agda, J.R.A.**, **Hellinga, A.J.** ***et al.*** (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, **20**, 275.

Ou, S., **Su, W.**, **Liao, Y.**, **Chougule, K.**, **Agda, J.R.A.**, **Hellinga, A.J.** ***et al.*** (2022) Author correction: benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, **23**, 76.

Pertea, G. & **Pertea, M.** (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, **9**, ISCB Comm J-304. Available from: https://doi.org/10.12688/f1000research.23297.2

Pertea, M., **Pertea, G.M.**, **Antonescu, C.M.**, **Chang, T.-C.**, **Mendell, J.T.** & **Salzberg, S.L.** (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

Pervin, M., **Unno, K.**, **Ohishi, T.**, **Tanabe, H.**, **Miyoshi, N.** & **Nakamura, Y.** (2018) Beneficial effects of green tea catechins on neurodegenerative diseases. *Molecules*, **23**, 1297.

Qiu, H., **Zhang, X.**, **Zhang, Y.**, **Jiang, X.**, **Ren, Y.**, **Gao, D.** ***et al.*** (2023) Depicting the genetic and metabolic panorama of chemical diversity in the tea plant. *Plant Biotechnology Journal*, **22**, 1001–1016. Available from: https://doi.org/10.1111/pbi.14241

Qiu, H., **Zhu, X.**, **Wan, H.**, **Xu, L.**, **Zhang, Q.**, **Hou, P.** ***et al.*** (2020) Parallel metabolomic and transcriptomic analysis reveals key factors for quality improvement of tea plants. *Journal of Agricultural and Food Chemistry*, **68**, 5483–5495.

Ranallo-Benavidez, T.R., **Jaron, K.S.** & **Schatz, M.C.** (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, **11**, 1–10.

Rempel, A., **Choudhary, N.** & **Pucker, B.** (2023) KIPEs3: automatic annotation of biosynthesis pathways. *PLoS One*, **18**, e0294342.

Rhie, A., **McCarthy, S.A.**, **Fedrigo, O.**, **Damas, J.**, **Formenti, G.**, **Koren, S.** ***et al.*** (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.

Rhie, A., **Walenz, B.P.**, **Koren, S.** & **Phillippy, A.M.** (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, **21**, 245.

Roach, M.J., **Schmidt, S.A.** & **Borneman, A.R.** (2018) Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 460.

Schwacke, R., **Ponce-Soto, G.Y.**, **Krause, K.**, **Bolger, A.M.**, **Arsova, B.**, **Hallab, A.** ***et al.*** (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, **12**, 879–892.

Shakya, R., **Ye, J.** & **Rommens, C.M.** (2012) Altered leaf colour is associated with increased superoxide-scavenging activity in aureusidin-producing transgenic plants. *Plant Biotechnology Journal*, **10**, 1046–1055.

Shi, J., **Ma, C.**, **Qi, D.**, **Lv, H.**, **Yang, T.**, **Peng, Q.** ***et al.*** (2015) Transcriptional responses and flavor volatiles biosynthesis in methyl jasmonate-treated tea leaves. *BMC Plant Biology*, **15**, 233.

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**, 3210–3212.

Sobolev, V.S., **Sy, A.A.** & **Gloer, J.B.** (2008) Spermidine and flavonoid conjugates from peanut (*Arachis hypogaea*) flowers. *Journal of Agricultural and Food Chemistry*, **56**, 2960–2969.

Su, W., **Tao, R.**, **Liu, W.**, **Yu, C.**, **Yue, Z.**, **He, S.** ***et al.*** (2020) Characterization of four polymorphic genes controlling red leaf colour in lettuce that have undergone disruptive selection since domestication. *Plant Biotechnology Journal*, **18**, 479–490.

Sun, B., **Zhu, Z.**, **Cao, P.**, **Chen, H.**, **Chen, C.**, **Zhou, X.** ***et al.*** (2016) Purple foliage coloration in tea (*Camellia sinensis* L.) arises from activation of the R2R3-MYB transcription factor CsAN1. *Scientific Reports*, **6**, 32534.

Sun, W., **Hao, J.**, **Fan, S.**, **Liu, C.** & **Han, Y.** (2022) Transcriptome and metabolome analysis revealed that exogenous spermidine-modulated flavone enhances the heat tolerance of lettuce. *Antioxidants (Basel)*, **11**, 2332.

Tang, S., **Peng, F.**, **Tang, Q.**, **Liu, Y.**, **Xia, H.**, **Yao, X.** ***et al.*** (2022) BnaPPT1 is essential for chloroplast development and seed oil accumulation in *Brassica napus*. *Journal of Advertising Research*, **42**, 29–40.

Tang, Y., **Fang, Z.**, **Liu, M.**, **Zhao, D.** & **Tao, J.** (2020) Color characteristics, pigment accumulation and biosynthetic analyses of leaf color variation in herbaceous peony (*Paeonia lactiflora* Pall.). *3 Biotech*, **10**, 76.

Uesato, S., **Kitagawa, Y.**, **Kamishimoto, M.**, **Kumagai, A.**, **Hori, H.** & **Nagasawa, H.** (2001) Inhibition of green tea catechins against the growth of cancerous human colon and hepatic epithelial cells. *Cancer Letters*, **170**, 41–44.

van Rengs, W.M.J., **Schmidt, M.H.-W.**, **Effgen, S.**, **Le, D.B.**, **Wang, Y.**, **Zaidan, M.W.A.M.** ***et al.*** (2022) A chromosome scale tomato genome built from complementary PacBio and nanopore sequences alone reveals extensive linkage drag during breeding. *The Plant Journal*, **110**, 572–588.

Venturini, L., **Caim, S.**, **Kaithakottil, G.G.**, **Mapleson, D.L.** & **Swarbreck, D.** (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, **7**, giy093.

Wang, F., **Zhang, B.**, **Wen, D.**, **Liu, R.**, **Yao, X.**, **Chen, Z.** ***et al.*** (2022) Chromosome-scale genome assembly of combined with multi-omics provides insights into its responses to infestation with green leafhoppers. *Frontiers in Plant Science*, **13**, 1004387.

Wang, L., **Pan, D.**, **Liang, M.**, **Abubakar, Y.S.**, **Li, J.**, **Lin, J.** ***et al.*** (2017) Regulation of anthocyanin biosynthesis in purple leaves of Zijuan tea (*Camellia sinensis* var. *kitamura*). *International Journal of Molecular Sciences*, **18**, 833. Available from: https://doi.org/10.3390/ijms18040833

Wang, P., **Yu, J.**, **Jin, S.**, **Chen, S.**, **Yue, C.**, **Wang, W.** ***et al.*** (2021) Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Horticulture Research*, **8**, 107.

Wang, S., **Qian, Y.-Q.**, **Zhao, R.-P.**, **Chen, L.-L.** & **Song, J.-M.** (2023) Graph-based pan-genomes: increased opportunities in plant genomics. *Journal of Experimental Botany*, **74**, 24–39.

Wang, X., **Feng, H.**, **Chang, Y.**, **Ma, C.**, **Wang, L.**, **Hao, X.** ***et al.*** (2020) Population sequencing enhances understanding of tea plant evolution. *Nature Communications*, **11**, 4447.

Wei, C., **Yang, H.**, **Wang, S.**, **Zhao, J.**, **Liu, C.**, **Gao, L.** ***et al.*** (2018) Draft genome sequence of var. provides insights into the evolution of the tea genome and tea quality. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, E4151–E4158.

Weil, H.L., **Schneider, K.**, **Tschöpe, M.**, **Bauer, J.**, **Maus, O.**, **Frey, K.** ***et al.*** (2023) PLANTdataHUB: a collaborative platform for continuous FAIR data sharing in plant research. *The Plant Journal*, **116**, 974–988.

Wu, D., **Xie, L.**, **Sun, Y.**, **Huang, Y.**, **Jia, L.**, **Dong, C.** ***et al.*** (2023) A syntelog-based pan-genome provides insights into rice domestication and de-domestication. *Genome Biology*, **24**(1), 179.

Xia, E., **Tong, W.**, **Hou, Y.**, **An, Y.**, **Chen, L.**, **Wu, Q.** ***et al.*** (2020) The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular Plant*, **13**, 1013–1026.

Xia, E.-H., **Zhang, H.-B.**, **Sheng, J.**, **Li, K.**, **Zhang, Q.J.**, **Kim, C.** ***et al.*** (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant*, **10**, 866–877.

Xu, Y.-X., **Chen, W.**, **Ma, C.-L.**, **Shen, S.-Y.**, **Zhou, Y.-Y.**, **Zhou, L.-Q.** ***et al.*** (2018) Corrigendum: proteome and acetyl-proteome profiling of cv. "Anji Baicha" during periodic albinism reveals alterations in photosynthetic and secondary metabolite biosynthetic pathways. *Frontiers in Plant Science*, **9**, 147.

Yu, X., **Xiao, J.**, **Chen, S.**, **Yu, Y.**, **Ma, J.**, **Lin, Y.** ***et al.*** (2020) Metabolite signatures of diverse *Camellia sinensis* tea populations. *Nature Communications*, **11**, 5586.

Yue, C., **Wang, Z.** & **Yang, P.** (2021) Review: the effect of light on the key pigment compounds of photosensitive etiolated tea plant. *Botanical Studies*, **62**, 21.

Zhang, Q.-J., **Li, W.**, **Li, K.**, **Nan, H.**, **Shi, C.**, **Zhang, Y.** ***et al.*** (2020) The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Molecular Plant*, **13**, 935–938.

Zhang, W., **Zhang, Y.**, **Qiu, H.**, **Guo, Y.**, **Wan, H.**, **Zhang, X.** ***et al.*** (2020) Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications*, **11**, 1–12.

Zhang, X., **Chen, S.**, **Shi, L.**, **Gong, D.**, **Zhang, S.**, **Zhao, Q.** ***et al.*** (2021) Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nature Genetics*, **53**, 1250–1259.

Zhang, Z.-B., **Xiong, T.**, **Chen, J.-H.**, **Ye, F.**, **Cao, J.-J.**, **Chen, Y.-R.** ***et al.*** (2023) Understanding the origin and evolution of tea (*Camellia sinensis* [L.]): genomic advances in tea. *Journal of Molecular Evolution*, **91**, 156–168.

Zhao, Y., **Jia, X.**, **Yang, J.**, **Ling, Y.**, **Zhang, Z.**, **Yu, J.** ***et al.*** (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, **30**, 1297–1299.

Zhou, C., **McCarthy, S.A.** & **Durbin, R.** (2023) YaHS: yet another hi-C scaffolding tool. *Bioinformatics*, **39**, 808. Available from: https://doi.org/10.1093/bioinformatics/btac808

Zhou, Y., **Zhang, Z.**, **Bao, Z.**, **Li, H.**, **Lyu, Y.**, **Zan, Y.** ***et al.*** (2022) Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, **606**, 527–534.