frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 July 2017
doi: 10.3389/fpls.2017.01205

Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing Approach for In-depth Understanding of Genes in Secondary Metabolism Pathways of *Camellia sinensis*

Qingshan Xu, Junyan Zhu, Shiqi Zhao, Yan Hou, Fangdong Li, Yuling Tai, Xiaochun Wan and ChaoLing Wei**

State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China

OPEN ACCESS

Edited by:
Luo Jie,
Huazhong Agricultural University, China

Reviewed by:
Alain Tissier,
Leibniz-Institute of Plant Biochemistry, Germany
Vinay Kumar,
Central University of Punjab, India

****Correspondence:***
ChaoLing Wei
weichl@ahau.edu.cn
Xiaochun Wan
xcwan@ahau.edu.cn

Specialty section:
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

Received: *19 December 2016*
Accepted: *26 June 2017*
Published: *11 July 2017*

Citation:
Xu Q, Zhu J, Zhao S, Hou Y, Li F, Tai Y, Wan X and Wei C (2017) Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing Approach for In-depth Understanding of Genes in Secondary Metabolism Pathways of Camellia sinensis. Front. Plant Sci. 8:1205. doi: 10.3389/fpls.2017.01205

Characteristic secondary metabolites, including flavonoids, theanine and caffeine, are important components of *Camellia sinensis*, and their biosynthesis has attracted widespread interest. Previous studies on the biosynthesis of these major secondary metabolites using next-generation sequencing technologies limited the accurately prediction of full-length (FL) splice isoforms. Herein, we applied single-molecule sequencing to pooled tea plant tissues, to provide a more complete transcriptome of *C. sinensis*. Moreover, we identified 94 FL transcripts and four alternative splicing events for enzyme-coding genes involved in the biosynthesis of flavonoids, theanine and caffeine. According to the comparison between long-read isoforms and assemble transcripts, we improved the quality and accuracy of genes sequenced by short-read next-generation sequencing technology. The resulting FL transcripts, together with the improved assembled transcripts and identified alternative splicing events, enhance our understanding of genes involved in the biosynthesis of characteristic secondary metabolites in *C. sinensis*.

Keywords: *Camellia sinensis*, single-molecule sequencing, full-length transcript, alternative splicing, characteristic secondary metabolite

INTRODUCTION

The tea plant (*Camellia sinensis*) is an important horticultural crop and source of one of the most popular natural non-alcoholic beverages consumed across the world (Chen et al., 2007; Zhang et al., 2015). The rich flavors of tea are mainly attributable to the characteristic secondary metabolites including flavonoids, theanine and caffeine (Liang et al., 2001; Mamati et al., 2006). These secondary compounds have been confirmed to be beneficial to human health (Hertog et al., 1993; Cabrera et al., 2006; Khan and Mukhtar, 2007) and contribute to the nutrient content and unique taste of tea (Chu and Juneja, 1997; Chen et al., 2008). Flavonoids such as flavanones, flavones, dihydroflavonols, flavonols, and flavin-3-ols (catechins) are derived from multiple branches of the phenylpropanoid pathway (Dixon and Pasinetti, 2010). Theanine is synthesized from glutamic acid and ethylamine by theanine synthetase (TS) in the roots of the tea plant (Deng et al., 2012).

Caffeine is a purine alkaloid that is abundant in the leaves of tea plant (Takeda, 1994; Ashihara et al., 1995). A thorough understanding of the genes underlying the biosynthesis of characteristic metabolites are essential for functional genomic studies.

Due to the large genome size (∼4.0 Gigabases) (Tanaka et al., 2006) of *C. sinensis* and genetic barriers in tea plant tissue culture and transformation, little genomic information is available currently. The genes encoding characteristic secondary metabolite biosynthetic enzymes were mostly discovered through Sanger sequencing (Singh et al., 2008, 2009b) or next-generation sequencing (Shi et al., 2011; Wu et al., 2013, 2014; Wang et al., 2014; Li et al., 2015). Sanger sequencing of full-length (FL) cDNA clones is the most reliable means of transcript discovery, but this method has fallen out of fashion somewhat following the advent of cheaper next-generation sequencing technologies (Wang et al., 2016). Using RNA-seq technology, most of the essential genes that regulate theanine, caffeine, and flavonoid biosynthesis were identified from whole tissues of tea (Shi et al., 2011). By studying transcription profiles of different tissues at different developmental stages, the gene network responsible for the regulation of the secondary metabolic pathways was also elucidated in tea plant (Li et al., 2015). However, the relatively short length of the reads generated from next-generation sequencing prevented to assemble the FL transcripts accurately (Minoche et al., 2014; Dong et al., 2015). Furthermore, in some cases, incorrect annotation can result from the low-quality transcripts generated by short-read RNA-seq sequencing (Au et al., 2012, 2013).

AS is an important post-transcriptional regulatory mechanism in multicellular eukaryotes that significantly enhances transcriptome diversity (Kalsotra and Cooper, 2011; Reddy et al., 2013). Next-generation sequencing revealed that over 60% of multi-exon genes are alternatively spliced in plant, such as *Oryza sativa* (Zhang et al., 2010), *Arabidopsis thaliana* (Marquez et al., 2012), and *Glycine max* (Shen et al., 2014). Up to now, very little was known about the alternative splicing in tea plant for the absence of genome information (Li et al., 2015). Additionally, short reads generated from next-generation sequencing require computational *de novo* assembly, therefore, identification of gene isoforms are not well supported by direct experimental evidence and may suffer from a high incidence of false positives (Au et al., 2012). More recently, single-molecule sequencing (SMS) technology eliminates the need for assembly with much longer reads (Sharon et al., 2013; Tilgner et al., 2014, 2015), providing direct evidence for transcript isoforms of each gene (Au et al., 2013; Chen et al., 2014; Abdelghany et al., 2016). These long-read transcripts can greatly increase the accuracy of transcriptome characterization compared with transcript tags assembled from short RNA-seq reads (Dong et al., 2015). Moreover, the higher error rate associated with SMS sequencing has been addressed by self-correction which involves the use of circular-consensus reads (Li Q. et al., 2014; Xu et al., 2015).

In this study, we employed an SMS approach to generate a more complete/FL transcriptome of *C. sinensis*. Based on long-read databases and genome sequences from bacterial artificial chromosome (BAC) libraries, we acquired FL transcripts and observed alternative splicing events for flavonoid, theanine and caffeine biosynthetic genes. The longer reads improved the quality and accuracy of transcripts generated from short-read assembly. This is the first study to use SMS technology to get the global overview of FL transcripts and alternative splicing events in tea. These results are necessary to deduce the nature of the encoded protein and in assessing a splice variant's role in gene regulation for *C. sinensis*.

MATERIALS AND METHODS

Plant Materials

Tea plants (*C. sinensis cv. Shuchazao*) were grown in the 916 Tea Plantation in Shucheng County, Anhui Province, China. Eight samples of different tissues were collected from exactly the same tea plant. The tissues sampled were as follows: apical bud, first leaf, mature leaf, old leaf, stems, flowers, fruits and roots. Apical bud, first leaf, mature leaf and stems were collected on June 15, 2015; Old leaf were collected on November 13, 2015; Flowers, fruits and roots were collected on October 12, 2015. All samples were immediately frozen in liquid nitrogen and stored at −80°C until further use.

RNA Isolation

Total RNA was extracted using the RNeasy Plus Mini kit (Qiagen, Valencia, CA, United States). Total RNA from each sample was quantified and the quality assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States). Equal amounts of total RNA from each sample were pooled to provide 90 μg of total RNA. PolyA RNAs was isolated from total RNA using Dynal oligo (dT) 25 beads (Invitrogen^{TM} Life Technologies, Carlsbad, CA, United States) according to the manufacturer's protocol. The isolated polyA RNAs was eluted with 20 μl of RNase-free water and subjected to RNA-seq library construction.

Library Preparation and Single-Molecule Sequencing

The library was prepared according to the PacBio ISO-Seq experimental workflow (Supplementary Figure S1). The first cDNA strand was synthesized from purified polyA RNAs using a Clontech SMARTer PCR cDNA Synthesis Kit (Clontech, Mountain View, CA, United States). After PCR optimization, large-scale PCR was performed to synthesize second strand cDNA for BluePippin size selection (Sage Science, Inc., Beverly, MA, United States) with size ranges of 0−1 kb, 1−2 kb, 2−3 kb, and 3−6 kb. After size selection, another amplification was performed, and amplified, size selected cDNA products were made into SMRTbell template libraries (0−1 kb, 1−2 kb, 2−3 kb, and 3−6 kb) according to the manufacture's instruction.

Libraries were prepared by annealing a sequencing primer (SMRTbell Template Prep Kit 1.0) and binding polymerase to the primer-annealed template. Sequencing was performed on a PacBio RS II platform. A total of seven SMRT cells were conducted in this study (Supplementary Table S1).

Data Analyses of Single-Molecule Sequencing Data

Raw data from four libraries produced by Pacific Biosciences RS II were processed following the BGI PacBio transcriptome analysis procedure (SMRT analysis 2.3.0) (**Figure 1**). In this pipeline, the 'Reads of Insert' that could either be a FL transcript (as defined by the presence of 5′ primer, 3′ primer, and the polyA tail if applicable) or a non-full-length transcript were generated using a minimum filtering requirement of 0 and a minimum read accuracy of 0.75. In the cluster panel, the options of "Predict Consensus Isoforms using the ICE Algorithm" and "Call Quiver to Polish Consensus Isoforms" were applied to get high quality, FL, and polished consensus transcripts. Finally, the high quality consensus transcripts of multiple libraries were merged together and redundancy removed based on CD-HIT-EST (-c 0.98 -T 6 -G 0 -aL 0.90 -AL 100 -aS 0.98 -AS 30) to obtain final FL isoforms.

Functional Annotation

Final FL isoforms were searched against NCBI non-redundant (NR), NCBI nucleotide sequence (NT), Swiss-Prot, Cluster of Orthologous Groups (COG) and Kyoto Encyclopedia of Genes and Genomes (KEGG, version 58) databases with a threshold *E*-value $\leq 10^{-5}$. Gene Ontology (GO) annotations were determined based on the best BLASTX hit from the NR database using the Blast2GO software version 2.3.5 (*E*-value $\leq 10^{-5}$). KEGG pathway analyses were performed using the KEGG Automatic Annotation Server (KAAS[1]).

Unigene and Isoform Prediction of Major Secondary Metabolites Biosynthetic Genes

To identify candidate genes, isoforms encoding enzymes from characteristic secondary metabolic pathways were clustered using CD-HIT software version 4.6.6 (cd-hit-est $c = 0.90$) (Li et al., 2001). The longest isoform of each cluster was defined as the candidate gene (Li and Godzik, 2006).

Alternative splicing isoforms were analyzed using BLAST[2] by employing transcripts from each cluster with genome sequences from the BAC library. Alternative splicing isoforms found by BLAST were viewed using the Gene Structure Display Server[3].

Validation of Alternative Splicing Isoforms by RT-PCR

For PCR validation of alternative splicing events, 1 μg of total RNA obtained from the eight different tissues was used for reverse transcription (RT) in 20 μl reactions with SuperScript III reverse transcriptase (Invitrogen) and N6 random hexamers (TaKaRa, Dalian, China). Gene-specific primers were designed with Primer Premier 6 to span the predicted splicing events (Supplementary Table S2). PCR was performed as follows: 3 min at 94°C, followed by 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for a time period proportional to the predicted product size. PCR amplification was monitored by 2.5% agarose gel electrophoresis.

PCR products were excised from the gel and purified using a gel extraction kit (Qiagen, Hilden, Germany). Purified products were cloned into the pGEM-T easy vector (Promega, United States) and plasmids were isolated using the Qiagen plasmid mini-isolation kit and confirmed by sequencing. Sequences were aligned with related isoforms to confirm the predicted alternative splicing isoforms.

Comparison with Short-Read Assemblies

Short-read sequences based on Illumina Hiseq2000 sequencing were selected for comparison with *C. sinensis* FL transcripts. Illumina data were obtained from same eight tea plant (*C. sinensis cv. Shuchazao*) tissues (buds, first leaf, mature leaf, old leaf, stems, flowers, fruits and roots) in our previous study (unpublished data). Clean reads for each tissue were assembled and annotated to generate unigenes, which were merged into the final dataset and redundancy removed by CD-HIT-EST (-c 0.98 -T 6 -G 0 -aL 0.90 -AL 100 -aS 0.98 -AS 30).

Candidate secondary metabolic pathway genes were identified using CD-HIT software (cd-hit-est $c = 0.90$) (Li and Godzik, 2006). Comparison of FL and Illumina-derived candidate secondary metabolic pathway genes was performed using local BLASTN ($1e^{-10}$ cut-off).

RESULTS

High Quality Reads Were Obtained from *Camellia sinensis* by Full-Length Sequencing

To identify as many isoforms as possible, eight different *C. sinensis* tissues were harvested for RNA isolation. Equal amounts of total RNA from each tissue were pooled together and reverse-transcribed. To minimize bias that favors sequencing of shorter transcripts, multiple size-fractionated libraries (<1, 1–2, 2–3 and 3–6 kb) were made using BluePippin. Four ISO-Seq libraries were constructed for one sample, and seven cells were sequenced using the Pacific Bioscience RS II platform, generating 361,947 reads. The mean read lengths of inserts from different libraries (<1, 1–2, 2–3, and 3–6 kb) produced by SMS sequencing were 768, 2160, 3023, and 3885 bases, respectively (Supplementary Table S1).

SMRT analyses (Reads of Insert, Classify and Cluster) were used to obtain high-quality consensus isoforms (**Figure 1**). Reads of Insert from different libraries (<1, 1–2, 2–3, and 3–6 kb) were classified into 38,131, 83,638, 64,244 and 24,669 FL non-chimeric transcripts, respectively, depending on whether 5′ and 3′ primer sequences or polyA tails were detected (Supplementary Table S3). ICE and Quiver were then used to cluster and polish the non-chimeric transcripts. After clustering and polishing, 21,093, 34,891, 26,633, and 9,021 high quality, FL, and polished consensus transcripts were generated for the four libraries, respectively (Supplementary Table S4). The quality distribution of

[1] www.genome.jp/tools/kaas/

[2] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[3] http://gsds.cbi.pku.edu.cn/

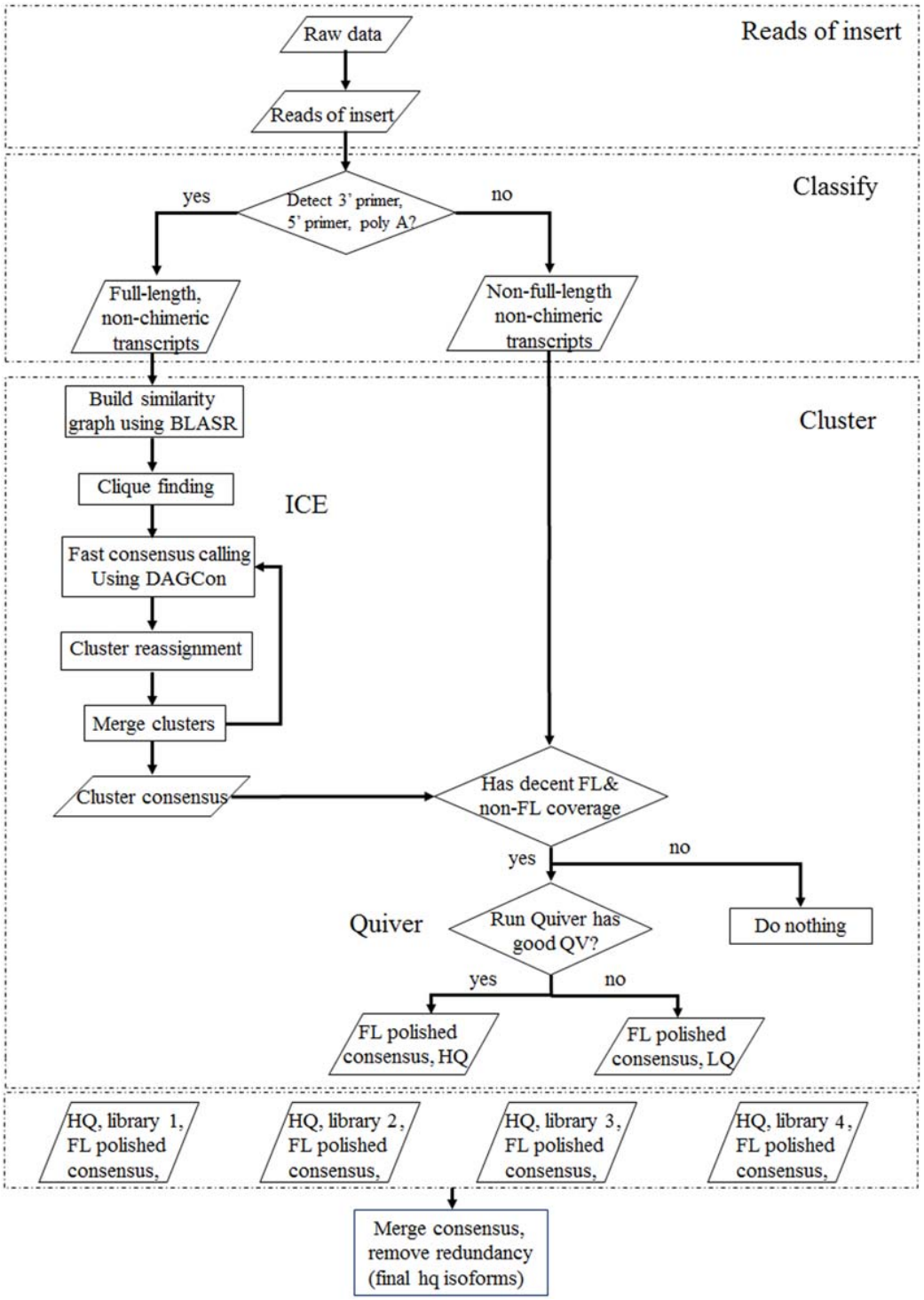

FIGURE 1 | SMRT analysis of each library to obtain high-quality consensus isoforms. Qualified sequencing data produced by Pacific Biosciences RS II were processed using SMRT analysis (Reads of Insert, Classify, Cluster) to obtain consensus full-length isoforms.

consensus isoforms were closed to 1 (Supplementary Figure S2). Finally, 91,638 high-quality consensus isoforms of the four libraries were merged into 80,217 isoforms with an average length of 1,781 bp and N50 of 2,459 bp (**Table 1**). In total, 68,360 isoforms (85.2%) were longer than 500bp, and 59900 isoforms (74.7%) were longer than 1 kb (**Figure 2**).

TABLE 1 | Summary of final *C. sinensis* consensus isoforms.

Sample	Total isoforms	Total bases (bp)	Mean length (bp)	N50 (bp)
Total	80,217	142,878,553	1,781	2,459

Functional Annotation and Categorization of the Isoforms

To predict and analyze the function of the 80,217 isoforms, we use BLAST (Altschul et al., 1990), BLAST2GO (Conesa et al., 2005), and InterProScan 5 (Quevillon et al., 2005) to perform functional annotation (using NR, NT, SwissProt, KEGG, COG, GO, and InterPro databases). A total of 72,877 isoforms were successfully matched to known proteins in at least one out the five databases, and 21,192 isoforms received high scores with proteins in all five databases (**Figure 3** and **Table 2**).

To functionally classify the *C. sinensis* transcripts, GO terms were assigned to each isoform using BLAST2GO based on the best BLASTx hit from the NR database. In total, 15,119 isoforms were assigned GO terms, which were classified into three major categories (molecular function, cellular component and biological process; Supplementary Figure S3). For molecular function classification, major categories were "catalytic activity" (GO: 0003824) and "binding" (GO: 0005488). In the cellular component category, isoforms involved in the "cell part" (5,977, 39.5% of the total), "cell" (5,977, 39.5%) and "organelle" (4,278, 28.3%) were highly represented. The major subgroups of biological processes were "cellular process" (GO: 0009987) and "metabolic process" (GO: 0008152).

Cluster of Orthologous Group contains protein sequences encoded in 21 prokaryotic and eukaryotic genomes, and this database was used to evaluate the completeness of the isoforms and the validity of the annotations. A total of 28,940 isoforms were assigned to 25 functional clusters (Supplementary Figure S4). "General function prediction only" (24.8%, 7,177), "replication, recombination and repair" (15.5%, 4,484), "transcription" (14.4%, 4,165), "post-translational modification, protein turnover, chaperones" (12.2%, 3,528), and "signal transduction mechanisms" (11.0%, 3,169) were the five largest categories. Secondary metabolites are very important for the taste and quality of tea, and approximately 4.1% (1,178) of isoforms were clustered into the secondary metabolism category (Supplementary Figure S4).

In order to explore the biological functions and interactions of genes in *C. sinensis*, isoforms were searched against the KEGG database. A total of 51,149 isoforms were annotated and assigned to 135 functional categories (**Table 2** and Supplementary Figure S5). Among these pathways, the "biosynthesis of secondary metabolites" pathway included 2176 isoforms, providing a valuable resource for further gene function research.

Unigenes and Isoforms in Flavonoid Pathway

Based on the KEGG database, a total of 301 isoforms were observed in flavonoid pathway (Supplementary Data S1) clustering into 90 candidate genes by CD-HIT-EST ($c = 0.90$) software. Flavonoids are synthesized via the phenylpropanoid pathway by the enzymes phenylalanine ammonia lyase (*PAL*),

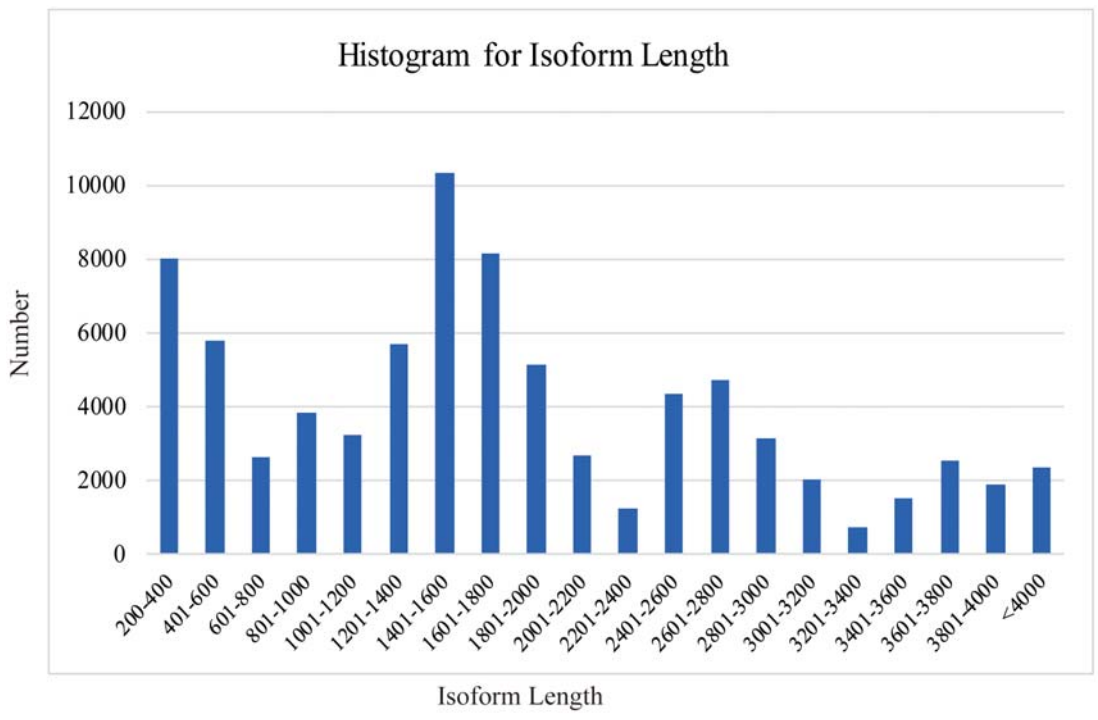

FIGURE 2 | Length distribution of *C. sinensis* transcripts.

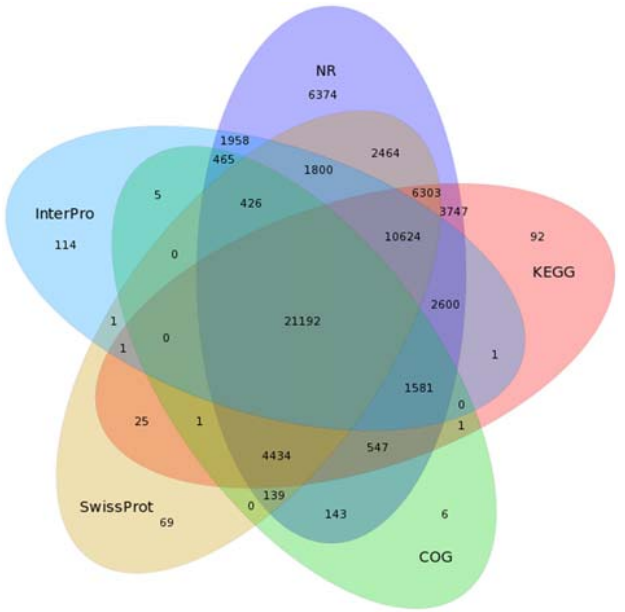

FIGURE 3 | Venn diagram of NR, COG, KEGG, SwissProt, and InterPro results for the *C. sinensis* transcriptome.

cinnamate 4-hydroxylase (*C4H*) and 4-coumarate CoA ligase (*4CL*). Fifteen, six and seven genes were annotated as *PAL*, *C4H* and *4CL*, respectively (**Figure 4A**). Among them, fifteen *PAL* were generated from 37 isoforms, and the isoforms "tea17336" and "tea3529" in different clusters may be transcribed from the same gene by alternative splicing (Supplementary Figure S6). Furthermore, *PAL* isoforms, "tea20264," "tea22666" and "tea19184" shared significant similarity with *CSPAL* (GenBank accession number: AY694188) which was associated with catechin accumulation (Singh et al., 2009a), and *PAL* isoform tea22927 showed 81.7% identity with *PtPAL2* (GenBank accession number: AF480620) which was expressed in heavily lignified structural cells of Quaking Aspen shoots (Kao et al., 2002).

Chalcone synthase (*CHS*) is the first enzyme of the general flavonoid pathway, and this enzyme mediates the influx of substrate from the phenylpropanoid pathway. By mapping isoforms to genome sequences in the BAC library, *CHS* (tea49771 and tea53048) were characterized as alternative 5′ splice sites (**Figure 5**). Subsequently, the stereo-specific cyclization of chalcones into naringenin is catalyzed by chalcone isomerase (*CHI*). Flavonoid 3′-hydroxylase (*F3′H*) and flavonoid 3′, 5′-hydroxylase (*F3′5′H*) catalyze the formation of eriodictyol and dihydrotricetin from naringenin (Wang et al., 2014). Five *F3′5′H* (tea27534, tea24671, tea22363, tea35097, and tea43773) genes were obtained from 18 isoforms in the present study. Of them, one gene (tea43773) with isoform (tea27827) from another gene (tea24671) cluster may undergo alternative splicing events (Supplementary Figure S6). A BLAST search of "tea27827" and "tea24671" revealed 98.5 and 98.4% identity with *CSF3′5′H'* (GenBank accession number: DQ194358) which played a critical role in the accumulation of tea catechins (Wang et al., 2014).

The formation of flavan-3-ols (e.g., catechin and gallocatechin) can be produced from leucoanthocyanidins by leucoanthocyanidin reductase (*LAR*) (Liu et al., 2016). Of the 13 *LAR* genes identified in this study, "tea53448" and "tea51087" were characterized as intron retention sites. Moreover, some *LAR* isoforms (tea51293, tea55264, tea51087, and tea51953) in our database shared significant similarity with *CSLAR* gene (GenBank accession no. GU992401) whose overexpression in tobacco leading to the accumulation of higher levels of epicatechin and its glucoside than of catechin (Pang et al., 2013). The generation of epi-flavan-3-ols (epicatechin and epigallocatechin) were achieved through a two-step reaction of leucoanthocyanidin catalyzed by leucoanthocyanidin oxidase (*ANS*) and anthocyanidin reductase (*ANR*) (Li et al., 2015). There were eight *ANS* genes and eight *ANR* genes from the long-read transcripts. Among them, tea52647 and tea52640 from different clusters were fell into the intron retention class (Supplementary Figure S6). Furthermore, by aligning long-read sequences with complete CDS data from NCBI, 50 genes were designated as FL transcripts (Supplementary Data S2).

Unigenes and Isoforms in Theanine Pathway

In total, 123 isoforms involved in theanine biosynthesis were annotated by the KEGG database (Supplementary Data S1). Based on clustering analysis with CD-HIT-EST ($c = 0.90$), 44 candidate genes were identified from these isoforms and 28 genes were considered to be FL transcripts followed by the alignment of long-read sequences with complete CDS data from NCBI (Supplementary Data S2).

Theanine is synthesized from glutamic acid and ethylamine via *TS*, alanine transaminase (*ALT*), arginine decarboxylase (*ADC*), glutamine synthetase (*GS*), glutamate synthase (*Fe-GOGAT*), and glutamate dehydrogenase (*GDH*) (Li et al., 2015). There were thirty-three *GSs/TSs*, four *GOGATs (NADPH)*, two *GOGATs (Fe)*, three *ALTs* and two *ADCs* in our database (**Figure 4B**). Of 33 *GSs/TS* genes, tea48459 and tea 11573 were

TABLE 2 | Summary of functional annotation results for *C. sinensis* transcripts.

Values	Total	Nr annotated	Nt annotated	SwissProt annotated	KEGG annotated	COG annotated	InterPro annotated	GO annotated	Overall
Number	80,217	64,797	69,456	47,479	51,149	28,940	40,768	15,119	72,887
Percentage	100%	80.78%	86.59%	59.19%	63.76%	36.08%	50.82%	18.85%	90.86%

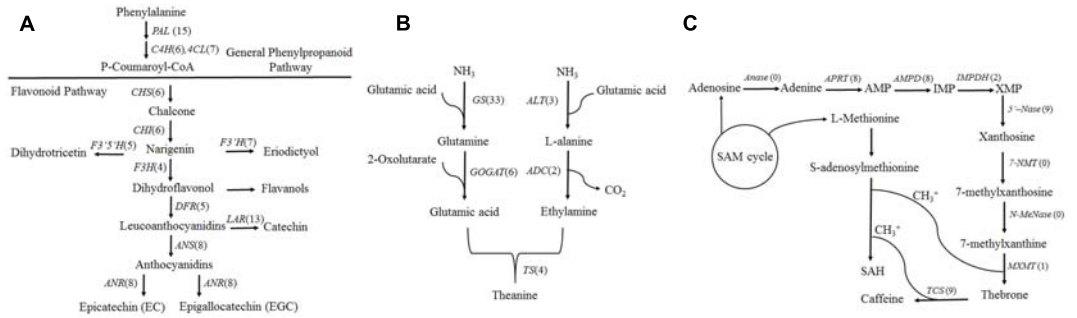

FIGURE 4 | *Camellia sinensis* transcripts involved in three secondary metabolite biosynthetic pathways. Numbers in brackets following each gene indicate the number of unigenes identified in C. *sinensis*. **(A)** Flavonoid biosynthesis pathway. **(B)** Theanine biosynthesis pathway. **(C)** Caffeine biosynthesis pathway.

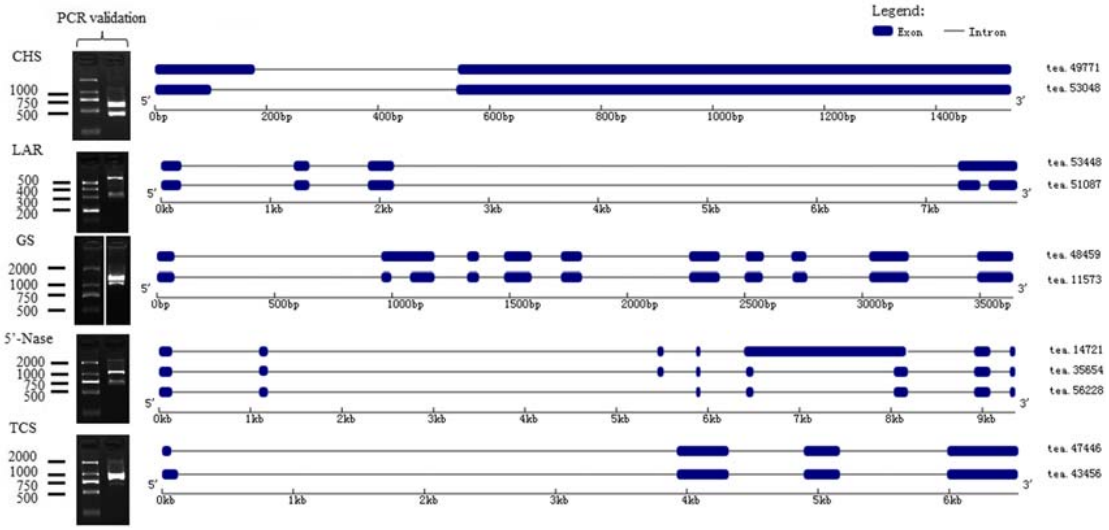

FIGURE 5 | Alternative splicing (AS) isoforms of characterized genes covered by PacBio long reads.

characterized as intron retention by mapping isoforms to genome sequences of the BAC library (**Figure 5**). Additionally, sequence analysis of *GS* isoforms tea47901, tea53333 and tea43896 revealed 99.5, 99.1, and 99.0% identity with *CsGS* (Genbank accession No. EF055882) whose expression was stimulated in response to abscisic acid, salicylic acid, and hydrogen peroxide in tea plant (Rana et al., 2008).

Unigenes and Isoforms in Caffeine Pathway

In our database, 105 isoforms annotated by KEGG database in caffeine pathway were clustered into 37 candidate genes by CD-HIT-EST ($c = 0.90$) (Supplementary Data S1). The caffeine biosynthesis pathway is part of purine metabolism and comprises purine biosynthesis and purine modification steps (Zrenner et al., 2005). Purine biosynthesis starts from adenosine, and involves adenosine nucleosidase (*Anase*), adenine phosphoribosyltransferase (*APRT*), AMP deaminase (*AMPDA*), IMP dehydrogenase (*IMPDH*), and 5′-nucleotidase (*5′-Nase*). Eight *APRTs*, eight *AMPDs*, two *IMPDHs* and nine *5′-Nases* were identified in the present study. Among them, nine *5′-Nase* were yielded from 15 isoforms. Of the 15 isoforms tea14721, tea35654 and tea56228 from the same cluster appeared to undergo exon skipping and intron retention (**Figure 5**). Moreover, tea28112 and tea31199 from different *5′-Nase* were characterized as exon skipping (Supplementary Figure S6).

Purine modification steps include one nucleosidase reaction and three methylations (Shi et al., 2011). Caffeine is derived from xanthosine (*XR*) via 7-methylxanthosine synthase (*7-NMT*), N-methylnucleotidase (*N-MeNase*), theobromine synthase (*MXMT*), and tea caffeine synthase (*TCS*). There were one *MXMT* and nine *TCSs* in our database (**Figure 4C**). Among nine *TCSs*, tea47446 and tea43456 were identified as alternative 5′ splicing event (**Figure 5**). Another *TCS* isoforms (tea39349 and tea26065) from different clusters appeared to undergo intron retention (Supplementary Figure S6). In addition, 16 genes were detected as FL transcripts by aligning long-read sequences with complete CDS data from NCBI (Supplementary Data S2).

PacBio Isoforms Improved the Quality of Transcripts from Short-Read Assembly

A total of 208 short-read transcripts annotated by KEGG database were obtained from our previous study (Supplementary Data S1), of which 143, 35, and 30 transcripts were involved in flavonoid, theanine and caffeine biosynthesis, respectively. These 208 transcripts were then clustered into 147 candidate genes by CD-HIT-EST ($c = 0.90$), including 105 genes in the flavonoid pathway, 18 genes in the theanine pathway, and 24 genes in the caffeine pathway (Supplementary Data S1).

We compared 147 candidate genes from Illumina sequencing with our 171 long-read genes using local BLASTN. The comparison revealed a good agreement between the short-read unigenes and the long-read database at the nucleotide level (**Figure 6** and Supplementary Data S3). For the flavonoid pathway, we identified 74 (70.5%) unigenes from Short-Seq with a high degree of consistency with our database. For the theanine pathway, 15 (83.3%) short-seq unigenes were mapped to 10 long-read genes with high homology. For the caffeine pathway, 23 (95.8%) short-seq unigenes shared significant homology with 17 long-read genes.

However, SMS-Seq genes were longer than Short-Seq unigenes. Approximately 25.9% of the assembled unigenes from short-seq reads were <500 bases, whereas only 2.3% of the isoforms from the PACBIO reads were <500 bases (**Figure 7**). Notably, many of these short-seq unigenes were completely mapped to the same genes of long-read sequences. For example, CL9445.Contig2 (ANS), CL9445.Contig3 (ANS), Unigene6902 (ANS), and Unigene22214 (ANS) were all mapped to tea48033 (ANS) with 100% homology (Supplementary Figure S7). These results suggest that most candidate genes assembled from the Short-Seq reads did not represent FL cDNAs. Our long-read data therefore improved the quality of transcripts assembled from Illumina short reads.

In addition, a few SMS genes shared significant homology with much longer Short-Seq unigenes. Interestingly, these Short-Seq unigenes included two complete CDS regions encoding two identical or different genes, whereas the related SMS genes were only homologous to part of the Short-Seq unigenes. For instance, tea5819 sequences were completely mapped to CL10428.Contig2, and the overlapping regions share the same complete CDS as GOGAT-Fe. However, the region of CL10428.Contig2 without SMS sequence coverage includes another complete CDS region encoding the ATP-dependent RNA helicase-like protein DB10. Tea9026 (GOGAT-Fe) sequences were mapped to two regions of CL10260.Contig7 that have the same CDS region encoding GOGAT-Fe (data not shown). This result suggests that transcripts generated from Short-Seq data may be susceptible to misassembly.

Validation of Alternative Splicing Events Identified by Multiple Alignment

To experimentally confirm the accuracy of the identified alternative splicing isoforms, three genes involved in flavonoid, theanine and caffeine biosynthesis annotated as a single transcript but present as two or more isoforms were selected for RT-PCR analysis. Primers were designed and synthesized (Supplementary Table S2) and used for RT-PCR using RNA from seven different tissues. The results showed that size of the fragments and the bands on the agarose gel were consistent with the alternative splicing isoforms (**Figure 5**). We cloned the DNA fragments corresponding to the predicted sizes and verified the isoforms by sequencing. Sequences and alignment of the verified isoforms are shown in Supplementary Data S4.

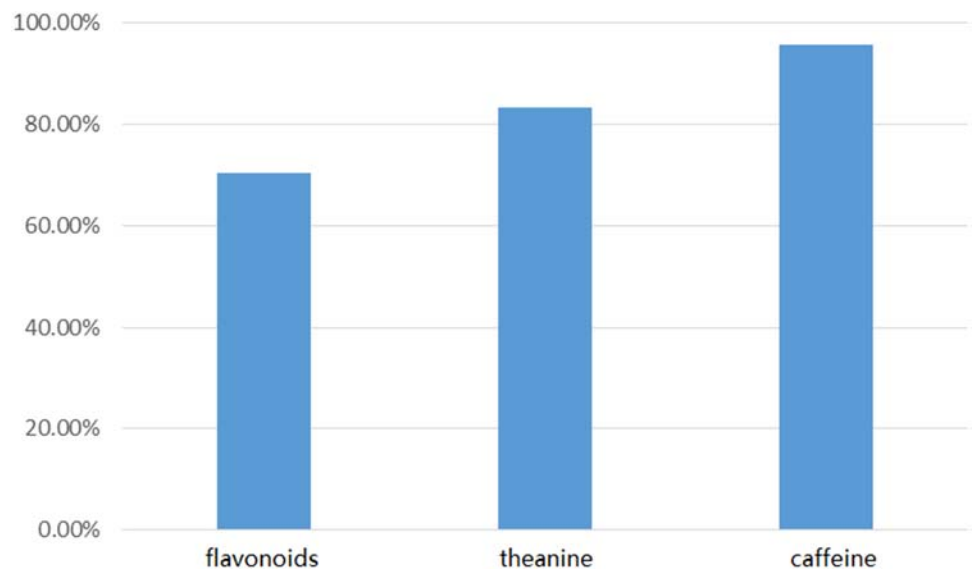

FIGURE 6 | BLAST comparison of Short-Seq sequences and long-read data.

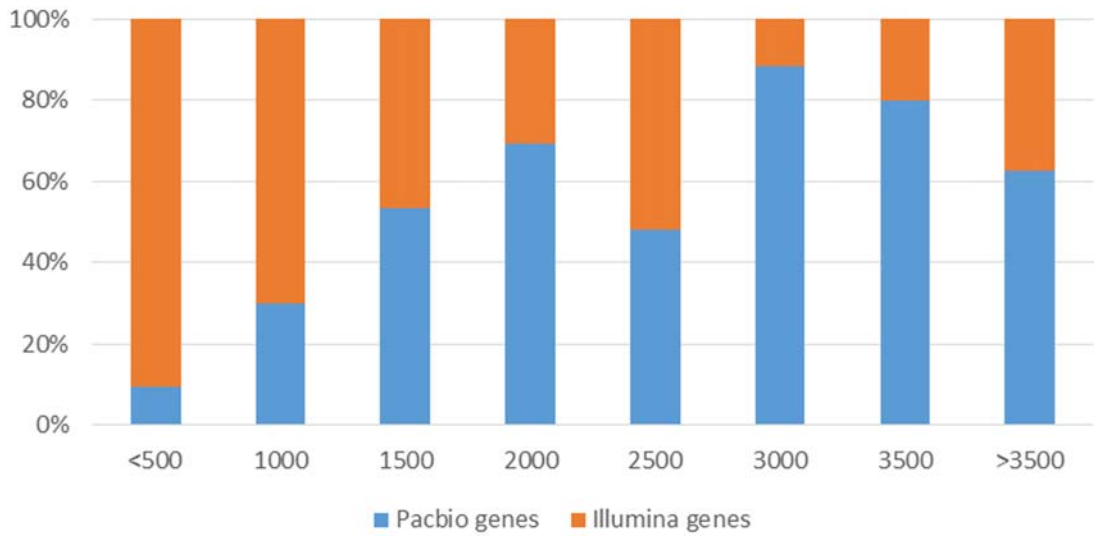

FIGURE 7 | Comparison of transcript length distribution from different sequencing platforms.

DISCUSSION

High-throughput mRNA sequencing studies using next-generation sequencing technologies have opened up a new era of transcriptome-wide research (Au et al., 2012; Mutz et al., 2013). Such approaches are particularly suitable for transcription profiling in non-model organisms that lack genomic sequences (Kawaharamiki et al., 2011; Shi et al., 2011). To date, most *C. sinensis* transcript studies have been based on next-generation sequencing (Wu et al., 2014; Zhang et al., 2015), and the short reads resulting from this approach have prevented the accurate assembly of FL transcripts in the absence of genomic sequence information (Au et al., 2013). In the present study, several Short-Seq reads from next-generation sequencing can be completely aligned to the same gene in our dataset (Supplementary Figure S7). Moreover, some misassembly transcripts were found by comparing with the long-read isoforms in our dataset. This result confirmed previous studies that transcripts generated from next-generation sequencing may suffer from misassembly (Schliesky et al., 2012; Li B. et al., 2014) and long reads produced by SMS sequencing technology can facilitate gene identification and annotation (Dong et al., 2015).

Previous studies have demonstrated the ability of SMS sequencing technology to generate continuous long reads (Abdelghany et al., 2016; Wang et al., 2016; Xu et al., 2016). Similar results were also observed in our study, resulting 80,217 isoforms with an average length of 1,781 bp were directly obtained by using SMS (Supplementary Table S5). By contrast, 55,088 transcripts were assembled and annotated from mixed tissue samples of *C. sinensis* based on next-generation sequencing, with an average unigene length of 355 bp (Shi et al., 2011). A total of 347,827 assembly transcripts were yielded from 13 different tea samples of various organs and developmental stages, with an average size of 791.2 bp (Li et al., 2015). On the other hand, we also identified 94 FL transcripts involved in the biosynthesis of flavonoids, theanine and caffeine by employing NCBI complete CDS. The above evidence indicated our results included a large number of longer transcripts specific to *C. sinensis* with known functions, which will be useful for improving the accuracy and quality of *C. sinensis* transcripts.

Due to its advantages, such as the highly accurate reads and the low costs, Illumina based RNA-seq is widely used for transcriptome analysis (Liu et al., 2012). However, for the alternative splicing events analysis, short reads require additional computational *de novo* assembly, therefore, it is difficult to infer the accuracy of gene model prediction (Steijger et al., 2013; Tilgner et al., 2013; Wang et al., 2016). These limitations were overcome with the emergence of SMS sequencing technology which can generate kilobase-sized sequencing reads in the absence of PCR amplification, where one read usually represents one FL transcript (Sharon et al., 2013; Gordon et al., 2015). In this research, alternative splicing events of some characteristic metabolic genes was predicted followed by the alignment with the BAC library, and confirmed by RT-PCR and sanger sequencing (**Figure 5**). Our results demonstrated that SMS sequencing is highly powerful in alternative splicing event discovery and provides a rich data resource for later functional studies of different isoforms in *C. sinensis*.

CONCLUSION

In summary, we identified numerous long-read isoforms specific to *C. sinensis* and characterized FL transcripts and alternative splicing events related to flavonoid, theanine and caffeine biosynthesis. The availability of FL isoforms can improve *C. sinensis* transcriptome characterization. The identification of alternative splicing events can deduce the nature of the encoded protein and in assessing a splice variant's role in gene regulation for *C. sinensis*. Furthermore, the FL transcripts generated in our

study provide a more accurate depiction of gene transcription and will greatly improve *C. sinensis* genome annotation in the future.

ACCESSION CODES

Raw data and 529 isoforms generated from SMRT sequencing and 208 transcripts produced by ILLUMINA HiSeq have been submitted to the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under accession numbers SRR5460108 and SRX2433645.

AUTHOR CONTRIBUTIONS

CW and XW conceived and designed the study. QX analyzed the data and wrote the manuscript. JZ performed PCR validation experiments. YH, FL, and SZ given the advice for data analyzing. YT provided bacterial artificial chromosome libraries. All authors have read and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [grant number 31171608], the Special Innovative Province Construction in Anhui Province in 2015 [grant number 15czs08032], the Vitalizing Plan of Tea Industry in Anhui Province [2012–2015], and the Program of Changjiang Scholars and Innovative Research Team in University [grant number IRT1101].

ACKNOWLEDGMENT

We would like to thank the 916 Tea Plantation in Shucheng, Anhui Province, China for providing samples of tea plants.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.01205/full#supplementary-material

REFERENCES

Abdelghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706. doi: 10.1038/ncomms11706

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Ashihara, H., Shimizu, H., Takeda, Y., Suzuki, T., Gillies, F. M., and Crozier, A. (1995). Caffeine metabolism in high and low caffeine containing cultivars of *Camellia sinensis*. *Z. Naturforsch. C* 50, 602–607.

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., et al. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4821–4830. doi: 10.1073/pnas.1320101110

Au, K. F., Underwood, J. G., Lee, L., and Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS ONE* 7:e46679. doi: 10.1371/journal.pone.0046679

Cabrera, C., Artacho, R., and Giménez, R. (2006). Beneficial effects of green tea—a review. *J. Am. Coll. Nutr.* 25, 79–99. doi: 10.1080/07315724.2006.10719518

Chen, D., Milacic, V., Chen, M. S., Wan, S. B., Lam, W. H., Huo, C., et al. (2008). Tea polyphenols, their biological effects and potential molecular targets. *Histol. Histopathol.* 23, 487–496. doi: 10.14670/HH-23.487

Chen, L., Kostadima, M., Martens, J. H. A., Canu, G., Garcia, S. P., Turro, E., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. *Science* 345, 1543–1549. doi: 10.1126/science.1251033

Chen, L., Zhou, Z. X., and Yang, Y. J. (2007). Genetic improvement and breeding of tea plant (*Camellia sinensis*) in China: from individual selection to hybridization and molecular breeding. *Euphytica* 154, 239–248. doi: 10.1007/s10681-006-9292-3

Chu, D. C., and Juneja, L. R. (1997). "General chemical composition of green tea and its infusion," in *Chemistry & Applications of Green Tea*, eds T. Yamamoto, L. R. Juneja, D. C. Chu, and M. Kim (Boca Raton, FL: CRC Press), 13–22.

Conesa, A., Götz, S., Garcíagómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation and visualization in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Deng, W. W., Wang, S., Qi, C., Zhang, Z. Z., and Hu, X. Y. (2012). Effect of salt treatment on theanine biosynthesis in *Camellia sinensis* seedlings. *Plant Physiol. Biochem.* 56, 35–40. doi: 10.1016/j.plaphy.2012.04.003

Dixon, R. A., and Pasinetti, G. M. (2010). Flavonoids and isoflavonoids: from plant biology to agriculture and neuroscience. *Plant Physiol.* 154, 453–457. doi: 10.1104/pp.110.161430

Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J. S. C., et al. (2015). Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genomics* 16:1039. doi: 10.1186/s12864-015-2257-y

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS ONE* 10:e0132628. doi: 10.1371/journal.pone.0132628

Hertog, M. G., Hollman, P. C., Katan, M. B., and Kromhout, D. (1993). Intake of potentially anticarcinogenic flavonoids and their determinants in adults in The Netherlands. *Nutr. Cancer* 20, 21–29. doi: 10.1080/01635589309514267

Kalsotra, A., and Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729. doi: 10.1038/nrg3052

Kao, Y. Y., Harding, S. A., and Tsai, C. J. (2002). Differential expression of two distinct phenylalanine ammonia-lyase genes in condensed tannin-accumulating and lignifying cells of quaking aspen. *Plant Physiol.* 130, 796–807. doi: 10.1104/pp.006262

Kawaharamiki, R., Wada, K., Azuma, N., and Chiba, S. (2011). Expression profiling without genome sequence information in a non-model species, pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. *PLoS ONE* 6:e26043. doi: 10.1371/journal.pone.0026043

Khan, N., and Mukhtar, H. (2007). Tea polyphenols for health promotion. *Life Sci.* 81, 519–533. doi: 10.1016/j.lfs.2007.06.011

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., et al. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.* 15, 553. doi: 10.1186/s13059-014-0553-5

Li, C. F., Yan, Z., Yao, Y., Zhao, Q. Y., Wang, S. J., Wang, X. C., et al. (2015). Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). *BMC Genomics* 16:560. doi: 10.1186/s12864-015-1773-0

Li, Q., Li, Y., Song, J., Xu, H., Xu, J., Zhu, Y., et al. (2014). High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytol.* 204, 1041–1049. doi: 10.1111/nph.12966

Li, W., and Godzik, A. (2006). Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282–283. doi: 10.1093/bioinformatics/17.3.282

Liang, Y., Ma, W., Lu, J., and Ying, W. (2001). Comparison of chemical compositions of *Ilex latifolia* Thumb and *Camellia sinensis* L. *Food Chem.* 75, 339–343. doi: 10.1016/S0308-8146(01)00209-6

Liu, C., Wang, X., Shulaev, V., and Dixon, R. A. (2016). A role for leucoanthocyanidin reductase in the extension of proanthocyanidins. *Nat. Plants* 2:16182. doi: 10.1038/nplants.2016.182

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Ray, P., et al. (2012). Comparison of next-generation sequencing systems. *Biomed Res. Int.* 2012:251364. doi: 10.1155/2012/251364

Mamati, G. E., Liang, Y., and Lu, J. (2006). Expression of basic genes involved in tea polyphenol synthesis in relation to accumulation of catechins and total tea polyphenols. *J. Sci. Food Agric.* 86, 459–464. doi: 10.1002/jsfa.2368

Marquez, Y., Brown, J. W., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* 22, 1184–1195. doi: 10.1101/gr.134106.111

Minoche, A. E., Dohm, J. C., Schneider, J., Holtgräwe, D., Viehöver, P., Montfort, M., et al. (2014). Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* 16, 184. doi: 10.1186/s13059-015-0729-7

Mutz, K. O., Heilkenbrinker, A., Lönne, M., Walter, J. G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.* 24, 22–30. doi: 10.1016/j.copbio.2012.09.004

Pang, Y., Abeysinghe, I. S. B., He, J., He, X. Z., Huhman, D., Mewan, K. M., et al. (2013). Functional characterization of proanthocyanidin pathway enzymes from tea and their application for metabolic engineering. *Plant Physiol.* 161, 1103–1116. doi: 10.1104/pp.112.212050

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., et al. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, 116–120. doi: 10.1093/nar/gki442

Rana, N. K., Mohanpuria, P., and Yadav, S. K. (2008). Cloning and characterization of a cytosolic glutamine synthetase from *Camellia sinensis* (L.) O. Kuntze that is Upregulated by ABA, SA, and H_2O_2. *Mol. Biotechnol.* 39, 49–56. doi: 10.1007/s12033-007-9027-2

Reddy, A. S. N., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* 25, 3657–3683. doi: 10.1105/tpc.113.117523

Schliesky, S., Gowik, U., Weber, A. P. M., and Bräutigam, A. (2012). RNA-seq assembly – are we there yet? *Front. Plant Sci.* 3:220. doi: 10.3389/fpls.2012.00220

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014. doi: 10.1038/nbt.2705

Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., et al. (2014). Global dissection of alternative splicing in paleopolyploid soybean. *Plant Cell* 26, 996–1008. doi: 10.1105/tpc.114.122739

Shi, C.-Y., Yang, H., Wei, C.-L., Yu, O., Zhang, Z.-Z., Jiang, C.-J., et al. (2011). Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC Genomics* 12:131. doi: 10.1186/1471-2164-12-131

Singh, K., Kumar, S., Rani, A., Gulati, A., and Ahuja, P. S. (2009a). Phenylalanine ammonia-lyase (PAL) and cinnamate 4-hydroxylase (C4H) and catechins (flavan-3-ols) accumulation in tea. *Funct. Integr. Genomics* 9, 125–134. doi: 10.1007/s10142-008-0092-9

Singh, K., Kumar, S., Yadav, S. K., and Ahuja, P. S. (2009b). Characterization of dihydroflavonol 4-reductase cDNA in tea [*Camellia sinensis* (L.) O. Kuntze]. *Plant Biotechnol. Rep.* 3, 95–101. doi: 10.1007/s11816-008-0079-y

Singh, K., Rani, A., Kumar, S., Sood, P., Mahajan, M., Yadav, S. K., et al. (2008). early gene of the flavonoid pathway, flavanone 3-hydroxylase, exhibits a positive relationship with the concentration of catechins in tea (*Camellia sinensis*). *Tree Physiol.* 28, 1349–1356. doi: 10.1093/treephys/28.9.1349

Steijger, T., Abril, J. F., Engstrom, P. G., Kokocinski, F., Hubbard, T. J., Guigo, R., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi: 10.1038/nmeth.2714

Takeda, Y. (1994). Differences in caffeine and tannin contents between tea [*Camellia sinensis*] cultivars, and application to tea breeding. *Japan Agric. Res. Q.* 28, 117–123.

Tanaka, J., Taniguchi, F., Hirai, N., and Yamaguchi, S. (2006). Estimation of the genome size of tea (*Camellia sinensis*), camellia (*C. japonica*), and their interspecific hybrids by flow cytometry. *Tea Res. J.* 101, 1–7. doi: 10.5979/cha.2006.1

Tilgner, H., Gubert, F., Sharon, D., and Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9869–9874. doi: 10.1073/pnas.1400447111

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., et al. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742. doi: 10.1038/nbt.3242

Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M., and Snyder, M. (2013). Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3*, 387–397. doi: 10.1534/g3.112.004812

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708. doi: 10.1038/ncomms11708

Wang, Y. S., Xu, Y. J., Gao, L. P., Yu, O., Wang, X. Z., He, X. J., et al. (2014). Functional analysis of flavonoid 3',5'-hydroxylase from tea plant (*Camellia sinensis*): critical role in the accumulation of catechins. *BMC Plant Biol.* 14:347. doi: 10.1186/s12870-014-0347-7

Wu, H., Chen, D., Li, J., Yu, B., Qiao, X., Huang, H., et al. (2013). De novo characterization of leaf transcriptome using 454 sequencing and development of EST-SSR markers in tea (*Camellia sinensis*). *Plant Mol. Biol. Rep.* 31, 524–538. doi: 10.1186/s12870-014-0347-7

Wu, Z. J., Li, X. H., Liu, Z. W., Xu, Z. S., and Zhuang, J. (2014). De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in *Camellia sinensis*. *BMC Plant Biol.* 14:277. doi: 10.1186/s12870-014-0277-4

Xu, Z., Luo, H., Ji, A., Zhang, X., Song, J., and Chen, S. (2016). Global identification of the full-length transcripts and alternative splicing related to phenolic acid biosynthetic genes in *Salvia miltiorrhiza*. *Front. Plant Sci.* 7:100. doi: 10.3389/fpls.2016.00100

Xu, Z., Peters, R. J., Weirather, J., Luo, H., Liao, B., Xin, Z., et al. (2015). Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. *Plant J.* 82, 951–961. doi: 10.1111/tpj.12865

Zhang, G., Guo, G., Hu, X., Yong, Z., Li, Q., Li, R., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* 20, 646–654. doi: 10.1101/gr.100677.109

Zhang, H. B., Xia, E. H., Huang, H., Jiang, J. J., Liu, B. Y., and Gao, L. Z. (2015). De novo transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. *BMC Genomics* 16:298. doi: 10.1186/s12864-015-1494-4

Zrenner, R., Stitt, M., Sonnewald, U., and Boldt, R. (2005). Pyrimidine and purine biosynthesis and degradation in plants. *Annu. Rev. Plant Biol.* 57, 805–836. doi: 10.1146/annurev.arplant.57.032905.105421

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.