# PacBio single molecule real-time sequencing-based full-length transcriptome of tree tomato (*Solanum betaceum* Cav.) and mining of simple sequence repeat (SSR) markers

**Honghong Deng**
Sichuan Agricultural University
**Ming'an Liao**
Sichuan Agricultural University
**Xiulan Lv**
Sichuan Agricultural University
**Lijin Lin** ( lij800924@qq.com )
Sichuan Agricultural University
**Qunxian Deng**
Sichuan Agricultural University
**Zhihui Wang**
Sichuan Agricultural University
**Jin Wang**
Sichuan Agricultural University
**Dong Liang**
Sichuan Agricultural University
**Xia Hui**
Sichuan Agricultural University
**Xun Wang**
Sichuan Agricultural University

**Research article**

**Keywords:** Tree tomato, full-length transcriptome, PacBio single-molecule long-read sequencing, simple sequence repeat, molecular marker

**Posted Date:** November 9th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-99778/v1

# Abstract

**Background:** Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) is a neglected, fast-growing, promising small fruit crop which provides a rich source of nutrition for human consumption. However, the transcriptome atlas of this important species is still lacking.

**Results:** In this study, RNA samples from a broad diversity of tissues (roots, leaves, stems, flowers and fruits) of tree tomato were sequenced using Pacific Biosciences' long-read single-molecule real-time sequencing technology. A total of 308699 full-length non-chimeric sequences with a mean length of 1005 bp and an N50 length of 1974 bp were obtained from our multi-tissue normalized cDNA libraries. A total of 140327, 104294, 135138, 78300, 53520, 152310 and 53520 transcripts were functionally annotated using Nr, Swiss-Prot, KEGG, KOG, GO, Nt and Pfam databases, respectively. Gene structural characteristics of the full-length transcriptome of tree tomato was subsequently investigated, including the predication of coding sequence and the identification of transcription factor families, long non-coding RNA and simple sequence repeat (SSR) marker. Thirty primers were randomly selected to evaluate the application of SSR markers, 23 of which obtained successfully amplification.

**Conclusions:** This is the first confident characterization of FL transcriptome profiling of tree tomato. The large-scale and high-quality transcriptome atlas and SSR molecular markers provided in the present study will facilitate further genetic studies of this important species.


# Background

Tree tomato (*Solanum betaceum* Cav., syn. *Cyphomandra betacea* (Cav.) Sendtn.), also familiarly known as tamarillo, is a neglected, fast-growing, promising small fruit crop native to the Andean region [1] and widely cultivated in the tropics and subtropics of South America, New Zealand, Australia and India, etc. [2–4]. As fresh fruit, it is an important and nutrient dense food source of human diets containing plenty of sugars, organic acids, minerals, ascorbic acid, provitamin A, carotenoids, vitamin $B_6$ and phenolics [5–7]; as processed product, it represents an important export commodity and stimulates both the local and overseas demand in fruit markets [3].

Previous studies mainly focused on its biochemical property [5, 6], phenology [8], and reproductive biology, including flower and pollen morphology, physiology, fruit characteristics, intraspecific hybridization and genetic diversity [4]. Despite the importance and recent progress, reference genome and transcriptome of tree tomato are not available, which severely impeded in-depth functional genomics, molecular genetics and genetic-assisted breeding of tree tomato. Additionally, *de novo* assembly of transcriptome sequence by the old-fashioned secondary-generation short-read sequencing, without a well-annotated reference genome, has been challenging [9]. The advent of Pacific Biosciences' (PacBio) long-read single-molecule real-time (SMRT) sequencing approach, also called third-generation sequencing technology, addressed these challenges and provided opportunities to obtain reliable genome-wide full-length (FL) transcripts directly [9].

The third-generation sequencing technology could generate an average read length more than 10 kb ('P6-C4' chemistry), thereby saving the need of further assembly and covering the size distribution of most transcripts in eukaryotes [10, 11]. Under such circumstances, PacBio sequencing has become an ideal tool to effectively and accurately capture FL or nearly FL transcripts of model or non-model species [9]. To overcome the drawback of high sequencing error rate in PacBio sequencing, a tailored analysis pipeline, Isoform Sequencing (Iso-Seq) pipeline, has been developed to calculate the circular consensus sequence (CCS) from more than two subreads [12].

To date, the Iso-Seq platform has been successfully applied to generate a well-characterized transcriptome for several plant species, even those species that lack reference genome including maize (*Zea mays*) [13], rice (*Oryza sativa* L.) [14], white myoga ginger (*Zingiber striolatum* Diels) [15], bermudagrass (*Cynodon dactylon* L.) [16], winged prickly ash (*Zanthoxylum planispinum*) [17], tea plant (*Camellia sinensis*) [18], strawberry (*Fragaria ananassa*) [19] and litchi (*Litchi chinensis* Sonn.) [20]. Use of Iso-Seq pipeline allows efficient characterization of exon-intron structure and accurate identification of FL alternative splicing and alternative polyadenylation sites [14, 21, 22], thus providing new light on the complexity and diversity of transcriptome [10, 13, 19, 20, 23].

In addition, transcriptome profiling has proved an effective approach for genome-wide development of simple sequence repeat (SSR) markers in many non-model plants at a large scale and low cost [15, 24–26]. SSRs are good DNA fingerprinting markers to assess genetic diversity and population structure and to distinguish closely-related cultivars because of the advantages of single locus, multiple allele variations and abundant polymorphism [27]. To date, only AFLP markers was used to measure the genetic diversity of different tree tomato varieties [28]. SSR markers identified and developed at the genome-wide scale of tree tomato are therefore highly desirable.

Herein, the PacBio SMRT sequencing technology was adopted to construct FL cDNA libraries from several tissues of tree tomato. The structural characteristics of transcripts was then investigated including the predication of coding sequence (CDS) and the identification of transcription factor (TF) families, long non-coding RNAs (lncRNA) and SSR markers. Distribution of SSR motifs was also investigated and SSR analysis was performed. This is the first confident characterization of FL transcriptome profiling of tree tomato. Molecular breeding of tree tomato will be accelerated by developing SSR markers associated with FL transcriptome. The results of this study have already opened exciting avenues in transcriptome-based studies for this important and promising fruit crop.

# Results

### Tree tomato full-length transcriptome sequencing with SMRT

To capture a representative FL transcriptome of tree tomato, RNA samples of ten different tissues were collected and equally pooled together for library preparation and sequencing. Using SMRT sequencing technology, a total of 9.92G subreads base were obtained, comprising 9,877,631 subreads, with an average subreads length of 1005 bp and an N50 length of 1974 bp. Some of these subreads were

extremely long (>5000 bp). Approximately 70.41% of the subreads fell in the size range of 200 to 1000 bp. Of the 416144 CCS isoforms, 308699 were identified as consensus FLNC reads with a mean length of 2099 bp (Table 1) based on the clustering algorithm of ICE. The length distribution of these subreads and FLNC are shown in Fig. 1A and 1B, respectively.

## Functional annotation of transcripts of tree tomato with Nr, Swiss-Prot, KEGG, KOG, GO, Nt and Pfam databases

To acquire a comprehensive reference FL transcriptome of tree tomato, transcripts were functionally annotated by sequence similarity search against seven different databases. A total of 140327, 104294, 135138, 78300, 53520, 152310 and 53520 transcripts were functionally annotated using Nr, Swiss-Prot, KEGG, KOG, GO, Nt and Pfam databases, respectively (Fig. 2; Additional file 1: Table S1). The annotation of Nr homologous species distribution showed the best blast hit with tree tomato were *Solanum tuberosum* (52712 isoforms), *Solanum pennellii* (21171 isoforms), *Solanum lycopersicum* (16666 isoforms) and *Capsicum annuum* (15851 isoforms) (Fig. 3; Additional file 2: Table S2).

Transcripts were successfully annotated with GO terms and enriched in three categories, including biological process, cellular component and molecular function (Fig. 4; Additional file 3: Table S3). In the biological process category, the share of the genes under metabolic process (27699 matched genes, 51.75%), cellular process (27089, 50.61%), single-organism process (20063, 37.49%), localization (7706, 14.40%), biological regulation (6786, 12.68%), regulation of biological process (6648, 12.42%) and response to stimulus (5893, 11.01%) were highly represented. The most abundant subcategory of cellular component was cell (12693 matched genes, 23.72%) and cell part (12693, 23.72%), followed by organelle (8699, 16.25%), macromolecular complex (7507, 14.03%), membrane (7344, 13.72%), membrane part (7019, 13.11%) and organelle part (3961, 7.40%). In the category of molecular function, binding (30712 matched genes, 57.38%), catalytic activity (26279, 49.10%), transporter activity (3491, 6.52%), molecular function regulator (2574, 4.81%), structural molecule activity (1939, 3.62%), nucleic acid binding transcription factor activity (1199, 2.24%) and molecular transducer activity (958, 1.79%) were the most prominently represented (Fig. 4; Additional file 3: Table S3).

The transcripts obtained were also compared with the KOG database. KOG analysis showed that tree tomato transcripts were assigned to a total of 26 categories (Fig. 5; Additional file 4: Table S4). The largest group belonged to general function prediction only (15323 matched genes, 19.57%), followed by post translational modification, protein turnover, chaperones (9750, 2.45%) and signal transduction mechanism (8614, 11.00%) (Fig. 5; Additional file 4: Table S4). A total of 5895 out of the 135138 transcripts were assigned to the signal transduction, thus making it the largest group (4.36%) among the major categories of KEGG functional classification. The three major categories assigned in KEGG pathways were translation (5233, 3.87%), folding, sorting and degradation (4989, 3.69%), and carbohydrate metabolism (4745, 3.51%) (Fig. 6; Additional file 5: Table S5). In addition, alignment results against Pfam, Swiss-Prot and Nt databases are summarized in Additional files 6, 7, and 8: Table S6, S7, and S8, respectively.

### Structure analysis of the full-length transcriptome of tree tomato

CDS from the full-length transcriptome of tree tomato were predicted using ANGEL software. The frequency for each length of CDS was evaluated. The most prevalent length of CDS ranges from 400 to 2000 bp (Fig. 7). A detailed breakup for each of such CDS categories is listed in Additional file 9: Table S9.

By predicting non-redundant transcripts using iTAK software, a total of 5114 genes were predicted to be TFs (Additional file 10: Table S10). These TFs belonged to different TF families, among which the most abundant observed was SNF2 (338 matched genes, 6.61%), followed by C3H (336, 6.57%), others (309, 6.04%), GRAS (213, 4.17%), MYB-related (188, 3.68%), bHLH (167, 3.27%), WRKY (163, 3.19%) and SET (161, 3.15%) (Fig. 8). A total of 43227, 42872, and 110333 noncoding RNAs candidates were predicted by CPC, CNCI and Pfam databases, respectively. Among them, 29453 transcripts were simultaneously identified by the three computational approaches (Fig. 9).

### SSR identification and validation of tree tomato

A screen of the 79549 genes using MicroSatete yielded diverse SSR types including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, hexanucleotide and some complex nucleotides. Among these, the mononucleotide repeats (63.97%) exhibited the highest frequency of occurrence, followed by dinucleotide (8.54%) and trinucleotide repeats (7.79%) (Fig. 10; Additional file 11: Table S11). For validation purposes, 30 primer pairs were randomly selected to evaluate the application of SSR markers, 23 of which were successfully amplified in the genomic DNA of tree tomato, resulting in clear PCR amplicons and expected product sizes. These 23 primer pairs showed reproducible bands and had stable repetition can be selected for further analysis (Fig. 11).

## Discussion

Transcriptome has become a powerful technique for investigating global gene expression profiles and has shaped our understanding of multiple biochemical pathways associated with physiological processes in the past few years [39]. Previously, due to the limitation of short sequencing reads, transcriptome analysis in species that lacks reference genome sequences often encounters complicated problem [9, 39]. Recent advances in PacBio SMRT sequencing technique enable the simultaneous and accurate interrogation of genome-wide gene expression [9]. The availability of PacBio SMRT sequencing technique has thus generated interest in understanding the complex transcriptome qualitatively and quantitatively [13–20].

Tree tomato has been identified as a promising small fruit crop high in antioxidants and nutritional value [2–7]. A high-confidence transcriptome atlas of this important species is still lacking. In the current study, through PacBio SMRT sequencing without assembly, we successfully obtained the first high-quality functionally annotated reference transcriptome for tree tomato. Moreover, transcriptome-derived SSR

markers were developed in this study. The large-scale and high-quality transcriptome atlas and molecular markers provided in the present study will facilitate further genetic studies of this important species.

In order to capture as many transcribed genes as possible, a broad diversity of tissues from major plant organs (roots, leaves, stems, flowers, fruits etc.) at different developmental stages (juvenile and adult) of tree tomato were collected for the RNA-Seq analysis in the current study. Collection encompassed the juvenile, vegetative and reproductive phases, representing a variety of transcriptional stages. Thus, the transcriptome atlas will be useful for the future study of tissue and development states.

PacBio generated 308699 FLNC sequences with a mean length of 1005 bp and an N50 length of 1974 bp (Table 1). N50 value is a weighted median describing half of the sum of the lengths of all contigs [40]. In previous studies an N50 of 3356 bp, 2459 bp and 3179 bp in white myoga ginger (*Z. striolatum* Diels) [15], tea plant (*C. sinensis*) [18] and litchi (*L. chinensis* Sonn.) [20] was reported, respectively. Comparatively speaking, larger N50 values represent more accurate and effective transcriptome assembly [40]. The differences among may be related with species. Because of a lack of NGS sequencing transcriptome, the full-length transcriptome obtained by PacBio SMRT sequencing here was not compared to its own previous version transcriptome.

Since no other reference transcriptome or draft genome data is available for tree tomato, it is imperative to assign transcripts to different biological functions and metabolic pathways. In this study, sequence-based alignments were therefore performed against multiple databases, resulting in significant BLAST hits in Fig. 2–6. For example, we used the GO annotations to assign each transcript to a set of GO slims including biological process, cellular component and molecular function categories (Fig. 4). The GO annotations results illustrate that the transcripts of tree tomato involved in diverse molecular functions and biological pathways [30]. The largest KOG group belonging to “general function prediction n only” (Fig. 5) generally denotes biochemical activity [32]. KEGG in Fig. 6 integrated the molecular interaction networks and metabolic pathways in tree tomato [31].

Another important aspect of our study was to analyze the structure of full-length transcriptome of tree tomato as shown in Fig. 7–9. CDS (Fig. 7), TFs (Fig. 8) and lncRNA (Fig. 9) were thoroughly analyzed. LncRNA represents a novel class of nonprotein coding transcripts and exerts a regulatory effect on numerous biological process [41]. In this study, a rigorous screening criterion combined with CPC, CNCI and Pfam databases lead to the identification of lncRNA (Fig. 9), which is useful for further investigating functional roles or evolution of lncRNA in tree tomato.

Full-length transcriptome that contains an enormous quantity of sequence information is a potentially rich source for SSR discovery [15]. Moreover, transcriptome-based SSR mining and development increased the likelihood of detecting SSR markers associated with functional genes due to the close linkage to expressed genes of transcriptome [42]. SSR markers have proved to be the most favored genetic marker for estimation of genetic diversity, phylogenetic analyses, genotype identification, marker-trait association, comparative mapping and genetic map construction [43]. Previously, genetic diversity study of tree tomato germplasm has mostly relied on AFLP marker [28]. To the best of our knowledge, no

SSR markers are available in tree tomato until now. The current study presents the first mining and development of SSR markers in tree tomato (Fig. 10; Additional file 11: Table S11). We believe that the SSR markers generated here would suffice the gap to some extent if not completely. Moreover, in view of lack of genome sequences for tree tomato, the SSR markers identified here contributes a valuable resource for marker-assisted breeding in tree tomato.

# Conclusion

Recent advances in PacBio long-read single-molecule real-time (SMRT) sequence approach enable to decipher the complex transcriptome in plant species even without reference genome sequences. In this study, we successfully obtained a high-quality full-length transcriptome of an important and promising small fruit crop, tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.), by using the PacBio SMRT sequencing technology. This is the first long-read transcriptome for tree tomato, which will be important for multiple gene discovery in tree tomato and for future delineation of gene function and annotations of the tree tomato genome sequence. In addition, the newly discovered SSR markers from transcriptome data will facilitate future molecular breeding of tree tomato.

# Materials And Methods

### Plant materials

Five-year old bearing tree tomato plants used in this study were grown at the experimental base of the College of Horticulture, Sichuan Agricultural University, Chengdu, China (latitude 30.71˚N, longitude 103.87˚E). Seven tissues including root tips, shoot tips, mature leaves, flower buds, flowers in full bloom, young fruit and mature fruit of three independent mature trees, and three tissues of root tips, shoot tips and leaves of three tree tomato seedlings were sampled and mixed afterward. Tree tomato seedlings were obtained by incubation of seeds at 22˚C and 95% relative humidity, which were randomly collected from the above mature trees.

### RNA extraction

Total RNA was extracted using the a PureLink RNA Mini Kit (Invitrogen Inc., Carlsbad, CA, USA), followed by DNase digestion and RNA purification using an on-column PureLink DNase Kit (Invitrogen Inc.) according to the manufacturer's instructions. 1% agarose gel was used to monitor whether there existed RNA degradation and potential contamination. The purity of RNA samples was determined by using a NanoPhotometer Spectrophotometer (Implen, Westlake Village, CA, USA). RNA concentration was measured using a Qubit 2.0 Fluorometer (Invitrogen Inc.). RNA integrity was checked using an RNA Nano 6000 Assay Kit on a BioAnalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA) before sequencing library preparation.

### Construction of Iso-Seq complementary DNA (cDNA) library and PacBio sequencing

Construction of Iso-Seq cDNA library and PacBio Sequencing were performed at Novogene Co., Ltd (Beijing, China). The mRNA was enriched using oligo-dT magnetic beads from 4.0 μg total RNA and reverse transcribed into cDNA using the SMARTer PCR cDNA Synthesis Kit (Clontech, now Takara, http://www.takarabio.com). The size-selected cDNA library was constructed according to the BluePippin Size Selection System protocol as described by PacBio (PN 100-092-800-03) and sequenced on the PacBio Sequel platform.

### Reads processing and error collection

Row data acquired after SMRT sequencing were processed using SMRTlink v5.0 software. CCS reads were yield from subread BAM files, and the full-length non-chimeric (FLNC) reads and non-full-length reads were determined by the simultaneous presence of the poly-A tail signal and the 5’ and 3’ cDNA primers from reads of insert (ROIs). The short reads were discarded. Subsequently, the FLNC sequences were isoform-level clustered with iterative clustering and error correction (ICE) software and herein generated one consensus isoform [29]. The non-full-length CCSs were polished with the Quiver algorithm. Finally, isoform with a minimum Quiver accuracy of 0.99 was considered high quality isoform and used for further analysis.

### Gene functional annotation

All isoforms were subjected to functional annotation using multiple protein and nucleotide databases, including the National Center for Biotechnology Information (NCBI) non-redundant protein (Nr, cutoff E-value $\leq 1^{e-5}$), NCBI non-redundant nucleotide (Nt, E-value $\leq 1^{e-5}$) (http://www.ncbi.nlm.nih.gov), gene ontology (GO) (http://www.geneontology.org/, E-value $\leq 1^{e-10}$) [30], kyoto encyclopedia of genes and genomes (KEGG, E-value $\leq 1^{e-3}$) [31], eukaryotic orthologous groups (KOG, E-value $\leq 1^{e-3}$) [32], Swissprot protein (http://www.expasy.ch/sprot, E-value $\leq 1^{e-5}$) and protein family (Pfam) database (E-value $\leq 0.01$) [33].

### Transcript analysis

Potential CDS regions within transcripts were predicted by ANGEL software, a long read implementation of ANGLE [34]. TFs were predicted with iTAK software from putative protein sequences [35]. TFs were downloaded from Plant Transcription Factor Database (v4.0) and blastp with default cutoff (E-value<0.05) parameters [36]. LncRNA was firstly screened via coding-non-coding-index (CNCI) with default parameters [37] and Coding Potential Calculator with NCBI eukaryotes' protein database (E-value $<1^{e-10}$) [38]. Then, each transcript was translated in three possible frames, and Pfam Scan with default parameters of -E 0.001 –domE 0.001 was utilized to determine whether there exists a domain of known protein family. SSRs within the transcriptome were identified by MIcroSAtellite (MISA) program (http://pgrc.ipk-gate rsleben.de/misa/), which allows the identification and localization of both the perfect and compound microsatellites. For PCR amplification, genomic DNA was extracted from fresh leaves of tree tomato using DNeasy Plant Mini Kit (Qiagen; Valencia, CA, USA) according to

manufacturer's protocol. 30 primer pairs used for PCR amplification are listed in (Additional file 12: Table S12).

# Declarations

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors declare that they have no competing interests.


### Funding

Not applicable


### Availability of data and materials

The data generated or analysed during this study are included in this published article and its supplementary information files.

### Authors' contributions

LL designed this study and contributed to the concept of this paper. HD performed the bioinformatic analysis and wrote the paper. ML, XL, and QD performed the transcriptome analysis. ZW, JW, DL, XH, and XW performed the SSR experiment. All authors have read and approved the manuscript.


### Acknowledgements

Not applicable


# References


1. Prohens J, Nuez F. The tamarillo (*Cyphomandra betacea*): A review of a promising small fruit crop. Small Fruits Rev. 2000;:43–68.
2. Samuels J. Biodiversity of food species of the solanaceae family: A preliminary taxonomic inventory of subfamily Solanoideae. Resources. 2015;4:277–322.
3. Carrillo-Perdomo E, Aller A, Cruz-Quintana SM, Giampieri F, Alvarez-Suarez JM. Andean berries from Ecuador: A review on botany, agronomy, chemistry and health potential. J Berry Res. 2015;5:49–69.

4. Ramírez F, Kallarackal J. Tree tomato (*Solanum betaceum* Cav.) reproductive physiology: A review. Sci Hortic (Amsterdam). 2019;248:206–15.
5. Acosta-Quezada PG, Raigón MD, Riofrío-Cuenca T, García-Martínez MD, Plazas M, Burneo JI, et al. Diversity for chemical composition in a collection of different varietal types of tree tomato (*Solanum betaceum* Cav.), an Andean exotic fruit. Food Chem. 2015;169:327–35.
6. Espin S, Gonzalez-Manzano S, Taco V, Poveda C, Ayuda-Durán B, Gonzalez-Paramas AM, et al. Phenolic composition and antioxidant capacity of yellow and purple-red Ecuadorian cultivars of tree tomato (*Solanum betaceum* Cav.). Food Chem. 2016;194:1073–80.
7. Lin L, Sun J, Cui T, Zhou X, Liao M, Huan Y, et al. Selenium accumulation characteristics of *Cyphomandra betacea* (*Solanum betaceum*) seedlings. Physiol Mol Biol Plants. 2020;26:1375–83.
8. Acosta-Quezada PG, Riofrío-Cuenca T, Rojas J, Vilanova S, Plazas M, Prohens J. Phenological growth stages of tree tomato (*Solanum betaceum* Cav.), an emerging fruit crop, according to the basic and extended BBCH scales. Sci Hortic (Amsterdam). 2016;199:216–23.
9. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21:1–16.
10. Marquez Y, Brown JWS, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res. 2012;22:1184–95.
11. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci U S A. 2014;111:9869–74.
12. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;16:1–13.
13. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016;7:11708.
14. Zhang G, Sun M, Wang J, Lei M, Li C, Zhao D, et al. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 2019;97:296–305.
15. Deng K, Deng R, Fan J, Chen E. Transcriptome analysis and development of simple sequence repeat (SSR) markers in *Zingiber striolatum* Diels. Physiol Mol Biol Plants. 2018;24:125–34.
16. Zhang B, Liu J, Wang X, Wei Z. Full-length RNA sequencing reveals unique transcriptome composition in bermudagrass. Plant Physiol Biochem. 2018;132:95–103.
17. Kim JA, Roy NS, Lee I hye, Choi AY, Choi BS, Yu YS, et al. Genome-wide transcriptome profiling of the medicinal plant *Zanthoxylum planispinum* using a single-molecule direct RNA sequencing approach. Genomics. 2019;111:973–9.
18. Xu Q, Zhu J, Zhao S, Hou Y, Li F, Tai Y, et al. 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. 2017;8:1–11.

19. Yuan H, Yu H, Huang T, Shen X, Xia J, Pang F, et al. The complexity of the *Fragariaananassa* (octoploid) transcriptome by single-molecule long-read sequencing. Hortic Res. 2019;6.
20. Zhou Y, Chen Z, He M, Gao H, Zhu H, Yun Z, et al. Unveiling the complexity of the litchi transcriptome and pericarp browning by single-molecule long-read sequencing. Postharvest Biol Technol. 2020;168:111252.
21. Hu H, Yang W, Zheng Z, Niu Z, Yang Y, Wan D, et al. Analysis of alternative splicing and alternative polyadenylation in *Populus alba* var. pyramidalis by Single-molecular long-read sequencing. Front Genet. 2020;11:1–10.
22. Xie L, Teng K, Tan P, Chao Y, Li Y, Guo W, et al. PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. Mol Genet Genomics. 2020;295:475–89.
23. Teng K, Teng W, Wen H, Yue Y, Guo W, Wu J, et al. PacBio single-molecule long-read sequencing shed new light on the complexity of the *Carex breviculmis* transcriptome. BMC Genomics. 2019;20:1–15.
24. Jia X, Tang L, Mei X, Liu H, Luo H, Deng Y, et al. Single-molecule long-read sequencing of the full-length transcriptome of *Rhododendron lapponicum* L. Sci Rep. 2020;10:1–11.
25. Wang S, Wang X, He Q, Liu X, Xu W, Li L, et al. Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. Plant Cell Rep. 2012;31:1437–47.
26. Yagi M, Yamamoto T, Isobe S, Hirakawa H, Tabata S, Tanase K, et al. Construction of a reference genetic linkage map for carnation (*Dianthus caryophyllus* L.). BMC Genomics. 2013;14.
27. Liu Q, Song Y, Liu L, Zhang M, Sun J, Zhang S, et al. Genetic diversity and population structure of pear (*Pyrus* spp.) collections revealed by a set of core genome-wide SSR markers. Tree Genet Genomes. 2015;11:1–22.
28. Acosta-Quezada PG, Vilanova S, Martínez-Laborde JB, Prohens J. Genetic diversity and relationships in accessions from different cultivar groups and origins in the tree tomato (*Solanum betaceum* Cav.). Euphytica. 2012;187:87–97.
29. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One. 2015;10:1–15.
30. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35.
31. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277-D280.
32. Koonin E V., Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 2004;5:R7.
33. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt Y, Eddy SR, et al. Pfam : the protein families database. 2014;42:D222–D230.

34. Shimizu K, Adachi J, Muraoka Y. Angle: A sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. J Bioinform Comput Biol. 2006;4:649–64.
35. Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant. 2016;9:1667–70.
36. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45:D1040–D1045.
37. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35:W345–W349.
38. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. 2013;41:e166.
39. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56.
40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
41. Wu L, Liu S, Qi H, Cai H, Xu M. Research progress on plant long non-coding RNA. Plants. 2020;9:1–10.
42. Taheri S, Abdullah TL, Yusop MR, Hanafi MM, Sahebi M, Azizi P, et al. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules. 2018;23:399.
43. Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: An overview of the recent progress in plants. Euphytica. 2011;177:309–34.

# Supplementary Information

Additional file 1: Table S1. Functional annotation of tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcriptome

Additional file 2: Table S2. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the non-redundant database

Additional file 3: Table S3. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the gene ontology (GO) database

Additional file 4: Table S4. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the eukaryotic orthologous groups (KOG) database

Additional file 5: Table S5. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the Kyoto encyclopedia of genes and genomes (KEGG) database

Additional file 6: Table S6. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the Pfam database

Additional file 7: Table S7. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the Swiss-Prot database

Additional file 8: Table S8. Tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) transcripts annotated in the Nt database

Additional file 9: Table S9. Coding sequences (CDS) of tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.)

Additional file 10: Table S10. Transcription factors (TFs) of tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.)

Additional file 11: Table S11. Simple sequence repeat (SSR) of tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.)

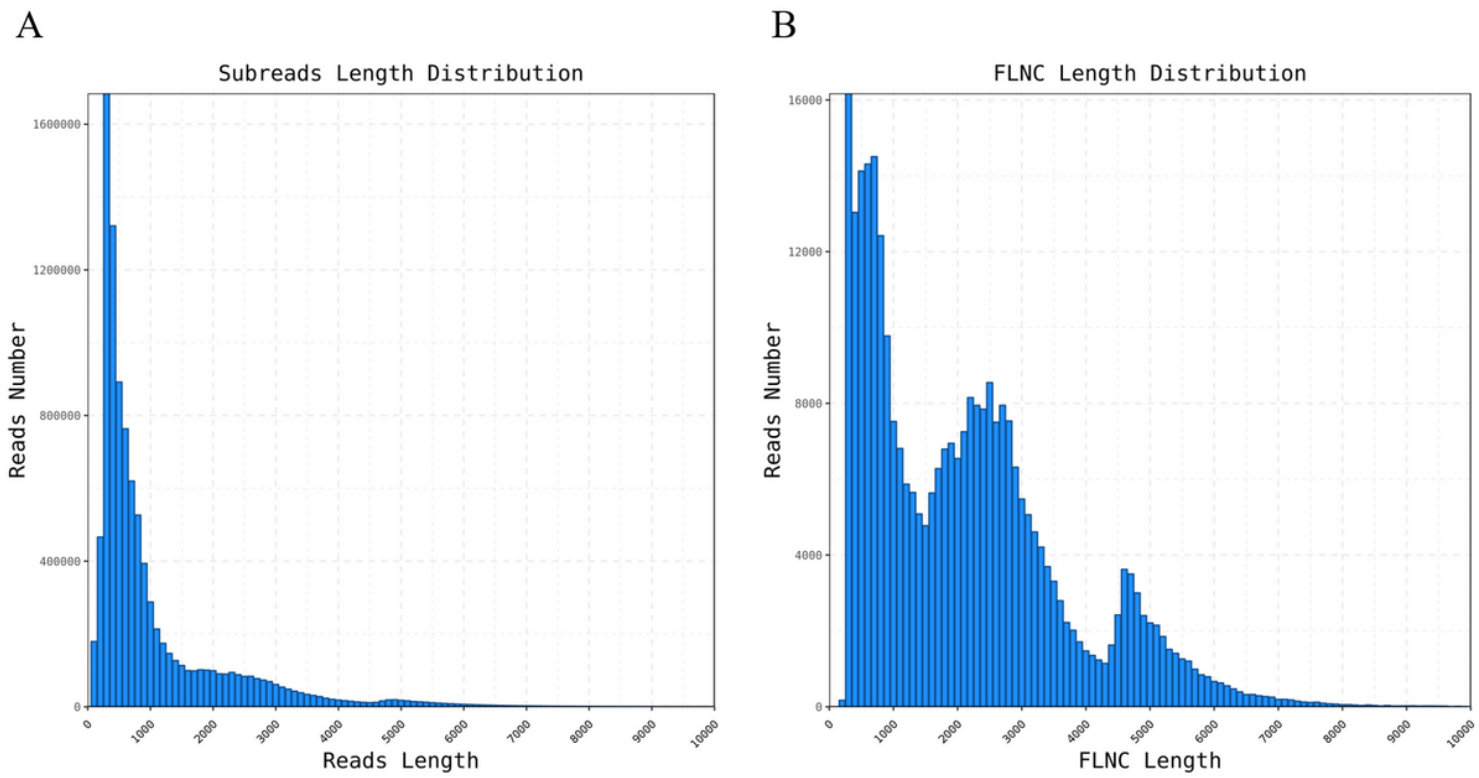
Additional file 12: Table S12. The primers used in this study

# Tables

Table 1. Summary of circular consensus sequence of tree tomato (*Cyphomandra betacea* (Cav.) Sendtn.) generated by SMRT sequencing technology

| Sample | CCS | 5'-primer | 3'-primer | Poly-A | Full length | FLNC | Average FLNC read length | Consensus reads |
|---|---|---|---|---|---|---|---|---|
| *betacea* | 416144 | 372441 | 381814 | 378908 | 322600 | 308699 | 2099 | 167191 |

# Figures




## Figure 1

The length distribution of subreads (A) and full-length non-chimeric (FLNC) reads (B) of tree tomato (Cyphomandra betacea (Cav.) Sendtn.). The horizontal axis represents the reads length.

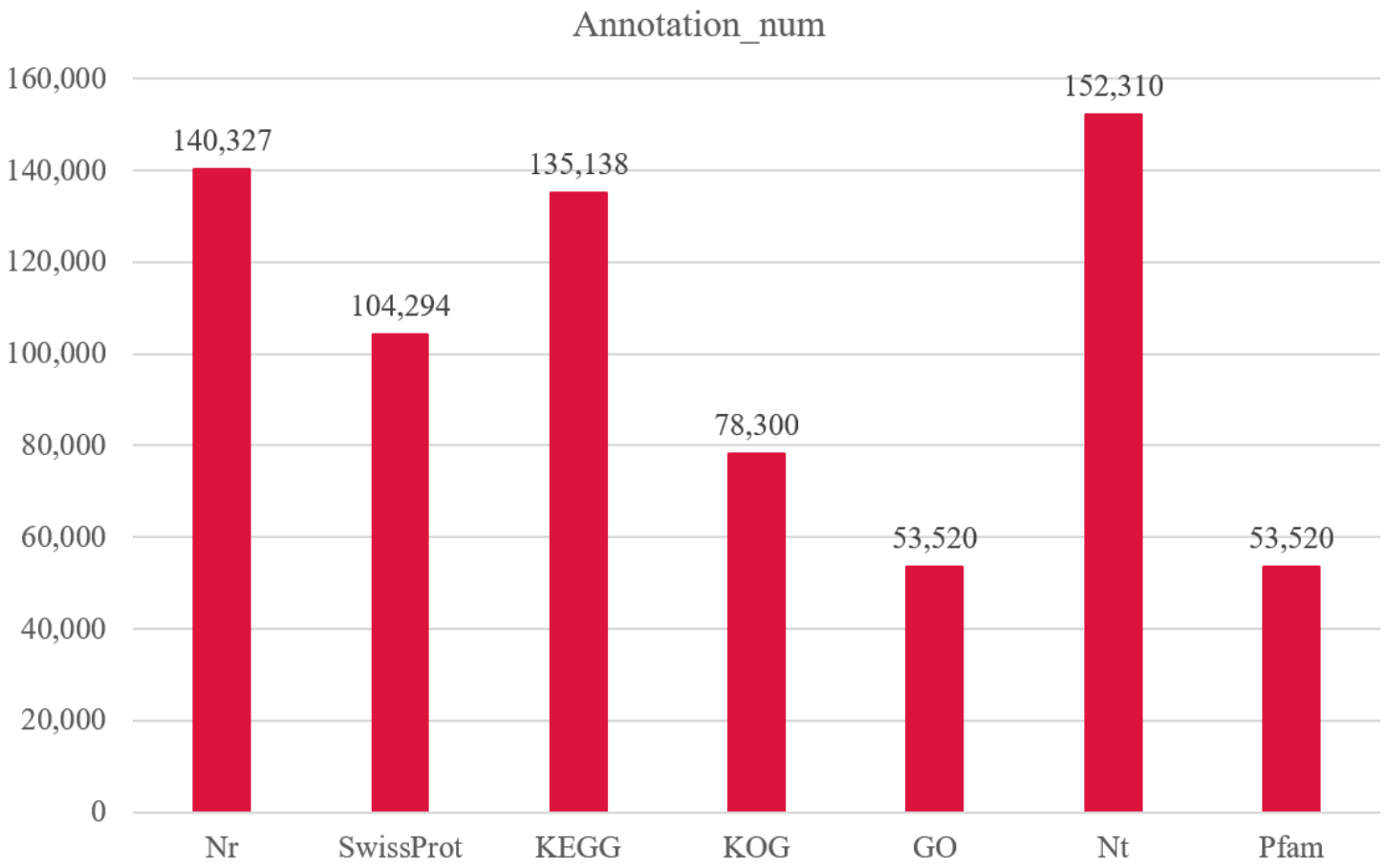



**Figure 2**

Summary of the functional annotation of tree tomato (Cyphomandra betacea (Cav.) Sendtn.) transcriptome using different databases.

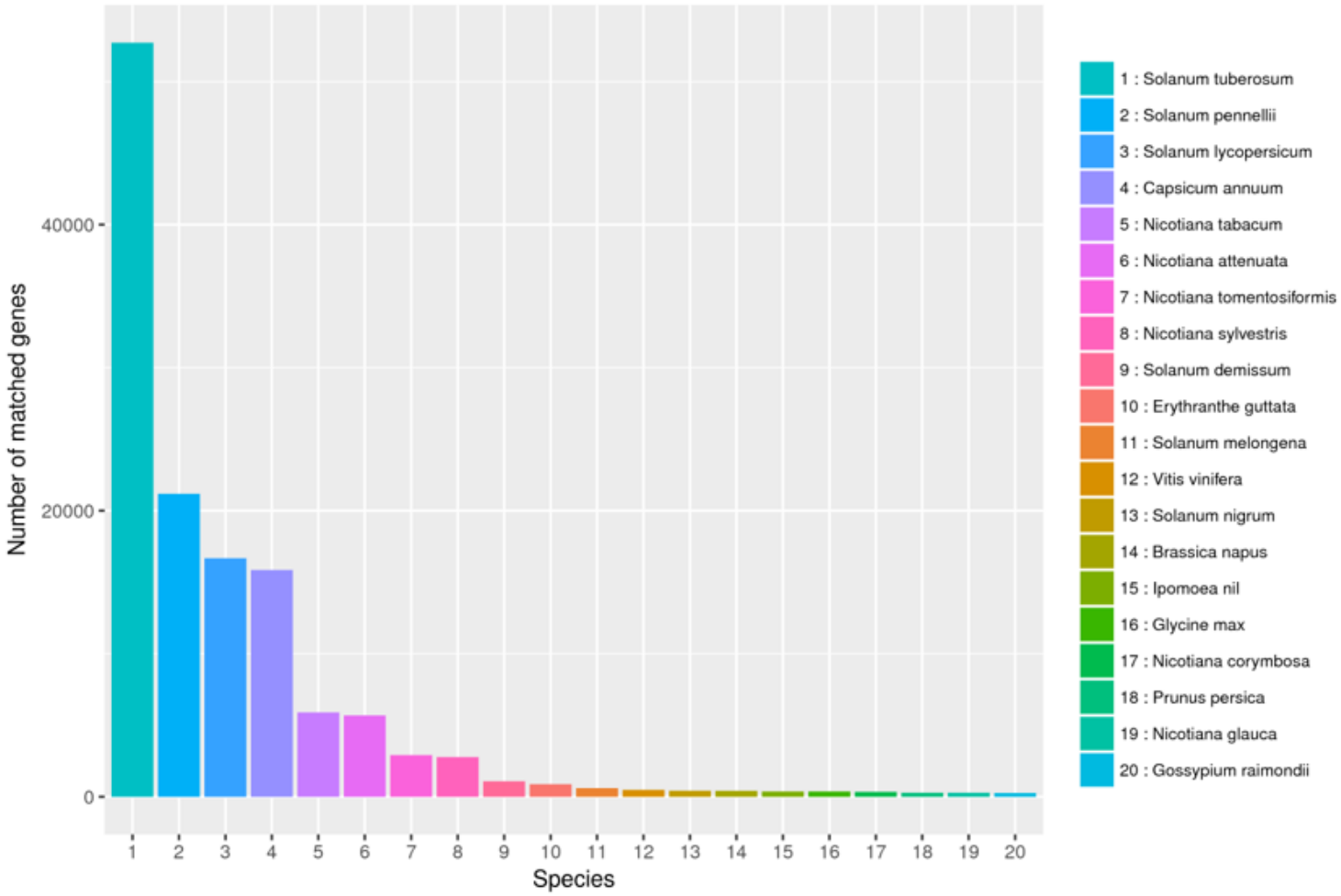
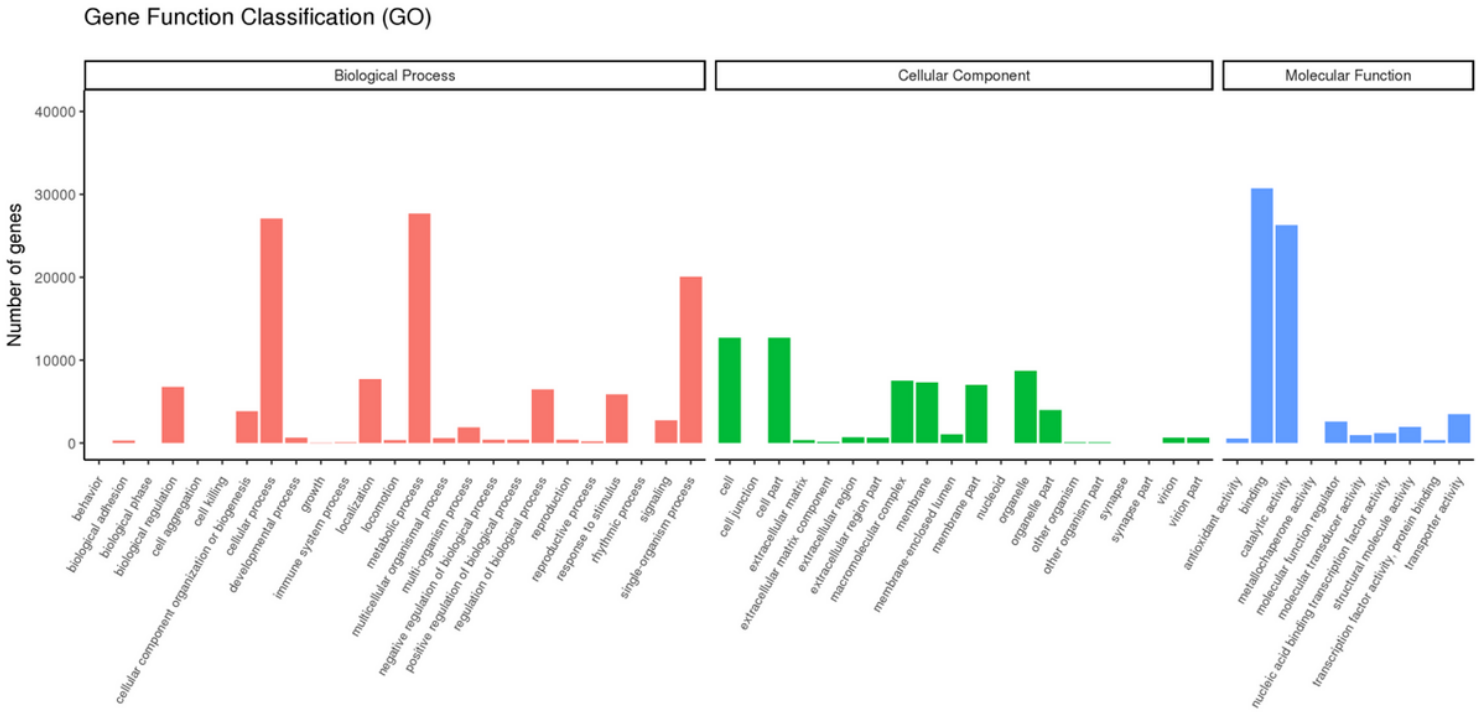


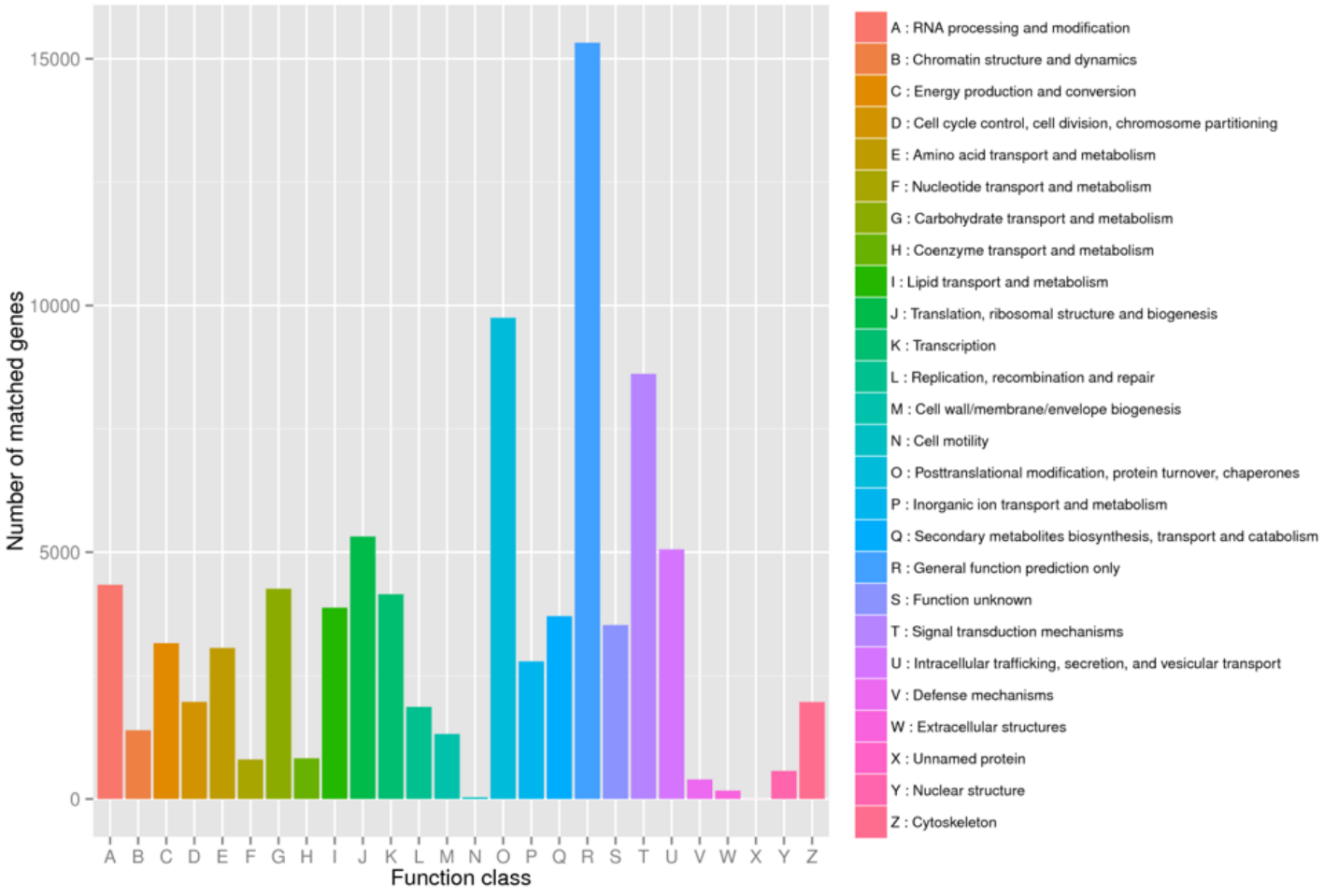

**Figure 3**

Homologous species distribution of tree tomato (Cyphomandra betacea (Cav.) Sendtn.) transcripts annotated in the non-redundant (Nr) database.



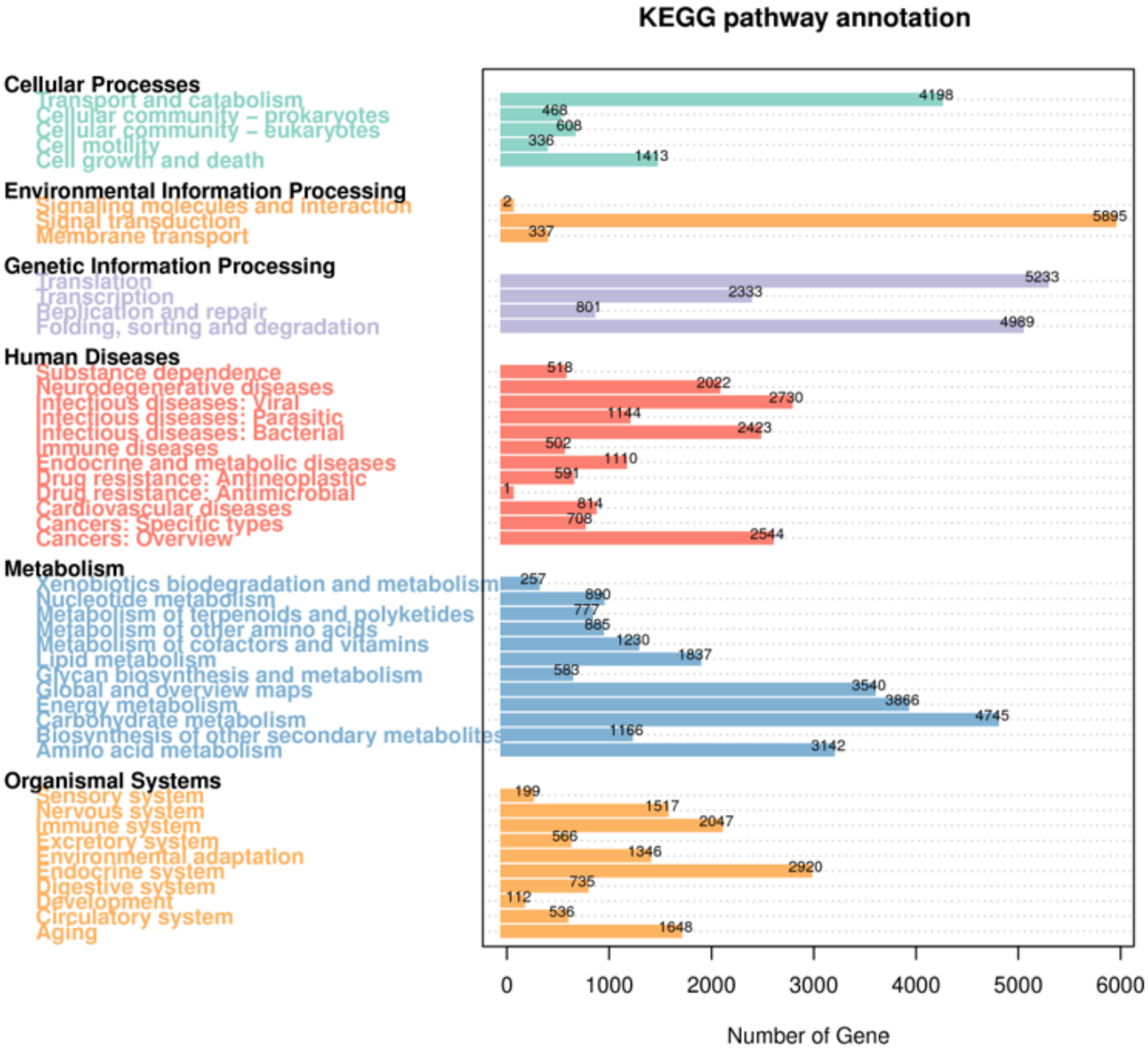

## Figure 4

Distribution of gene ontology (GO) terms for all annotated transcripts of tree tomato (Cyphomandra betacea (Cav.) Sendtn.) in biological process, cellular component and molecular function.




## Figure 5

Eukaryotic orthologous groups (KOG) classification of tree tomato (Cyphomandra betacea (Cav.) Sendtn.) transcripts. The x-axis represents the subcategories, and the y-axis indicates the number of transcripts in a specific functional cluster.




## Figure 6

Kyoto encyclopedia of genes and genomes (KEGG) enrichment of transcripts of tree tomato (Cyphomandra betacea (Cav.) Sendtn.).

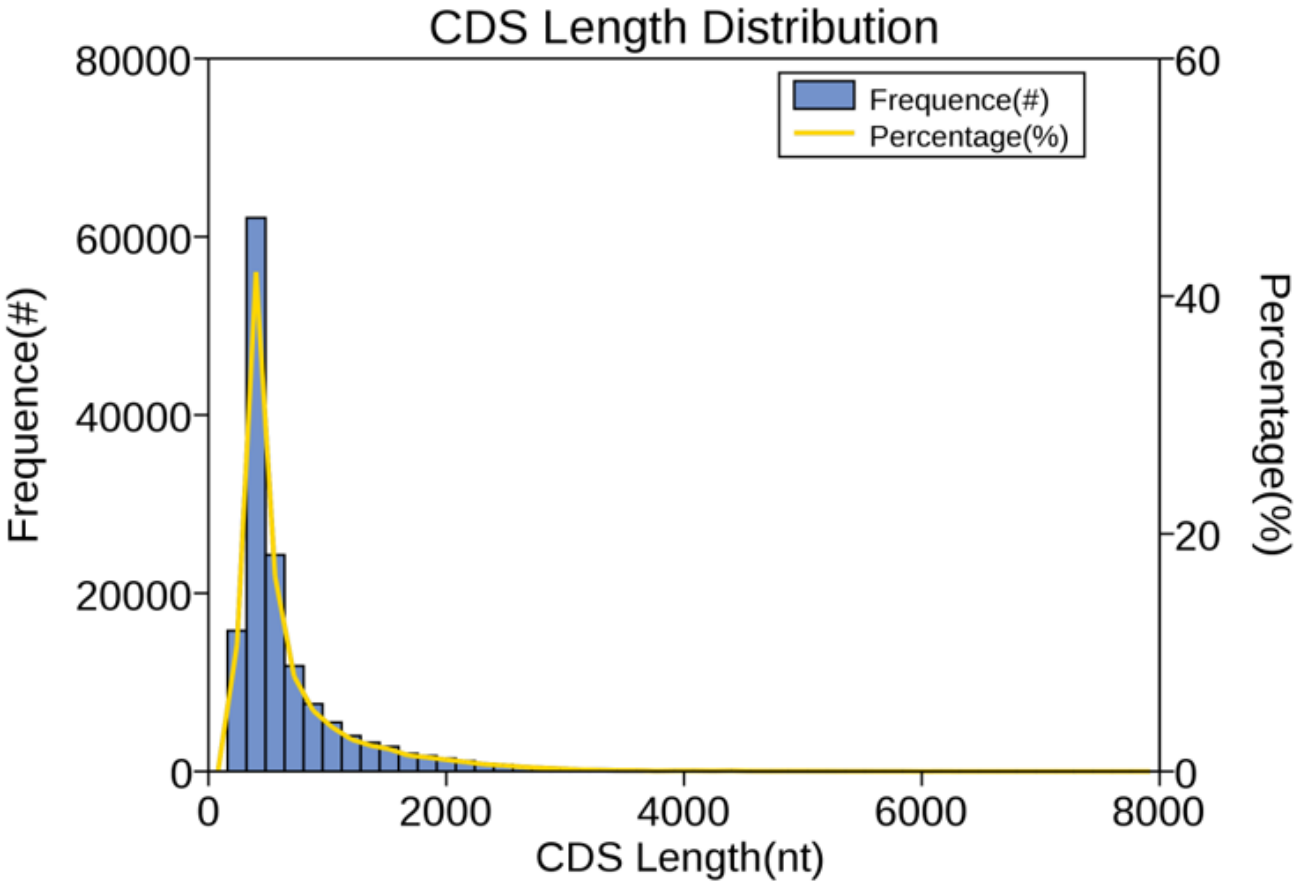
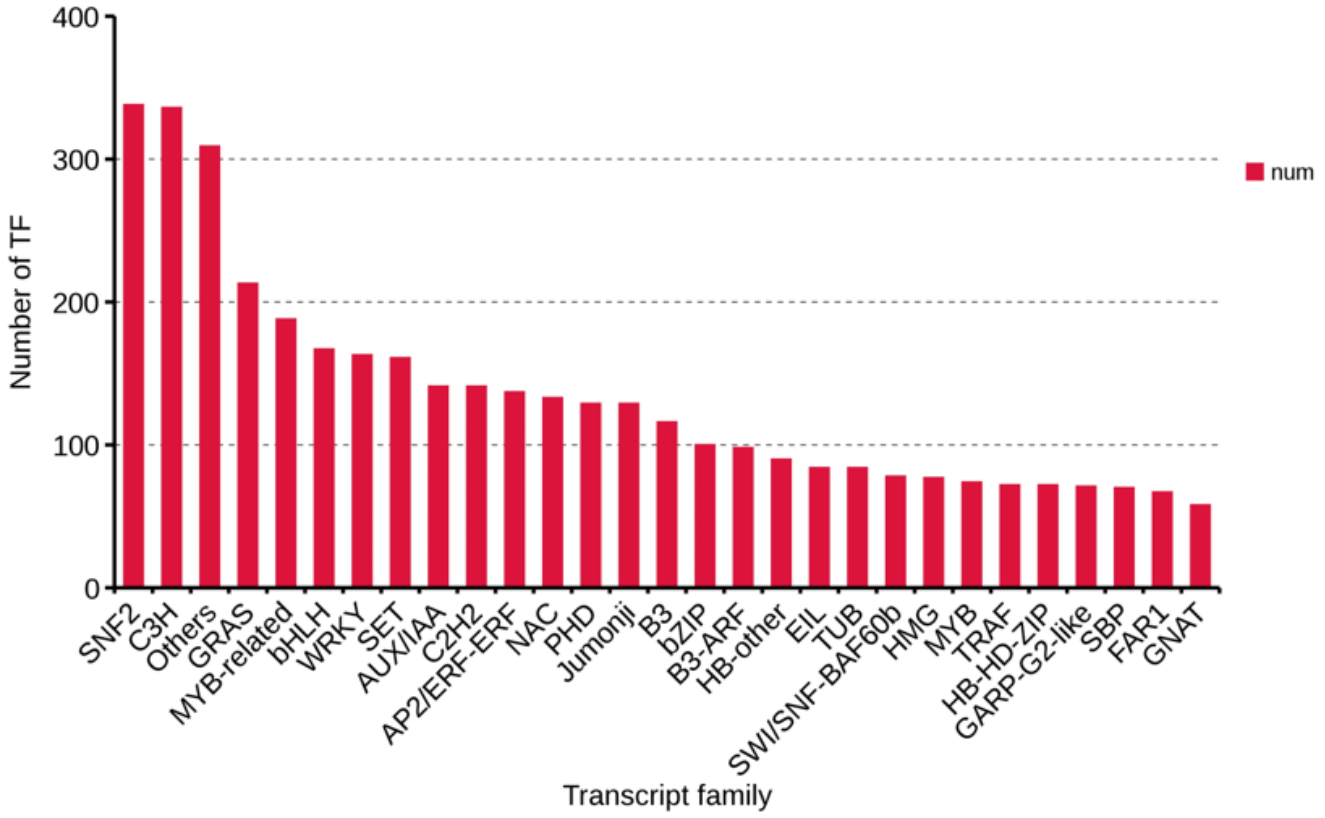



**Figure 7**

Frequency histogram depicting the length distribution of coding sequence (CDS) from full-length transcriptome of tree tomato (Cyphomandra betacea (Cav.) Sendtn.).




## Figure 8

Distribution of transcription factors (TFs) from full-length transcriptome of tree tomato (Cyphomandra betacea (Cav.) Sendtn.).

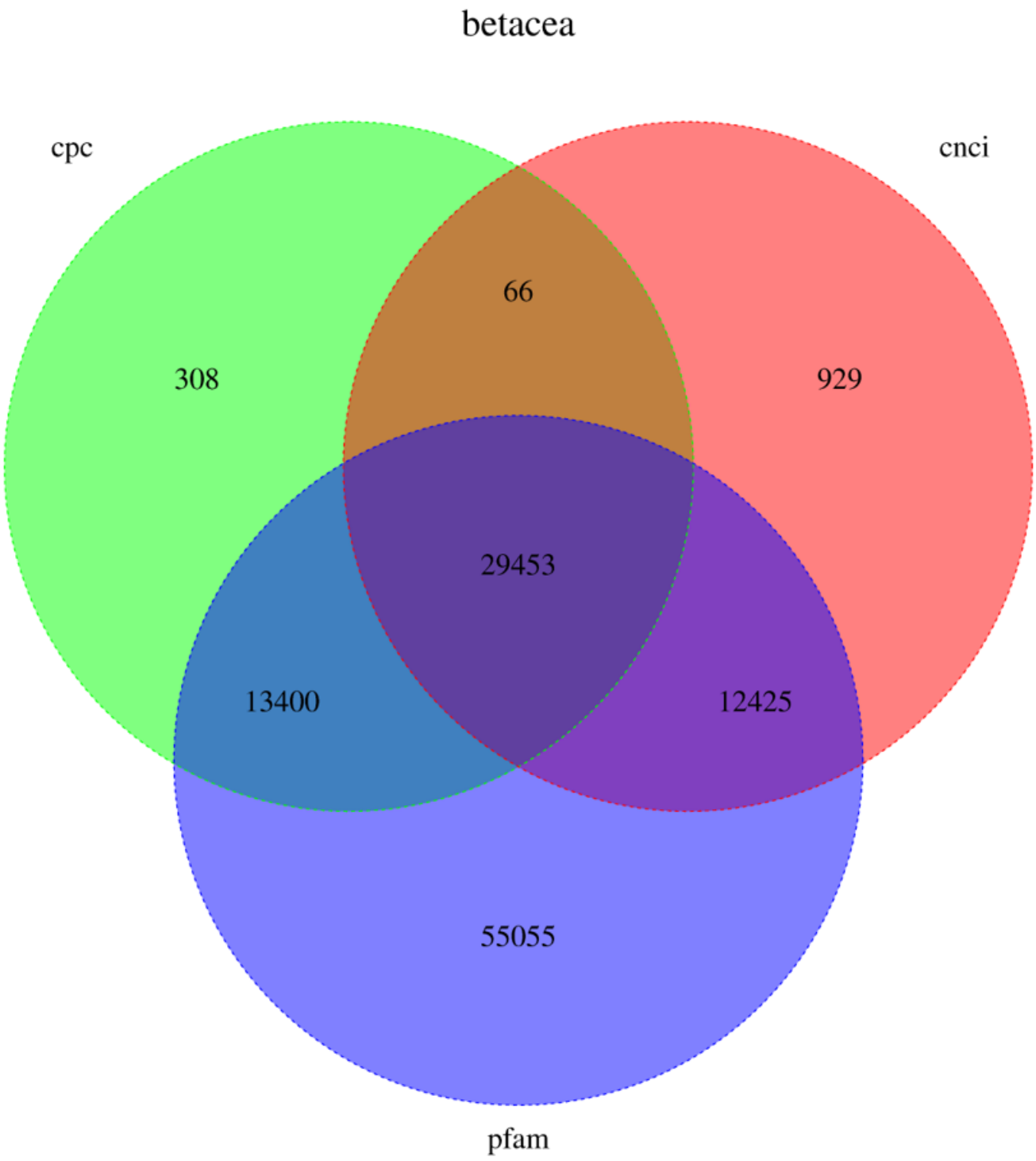
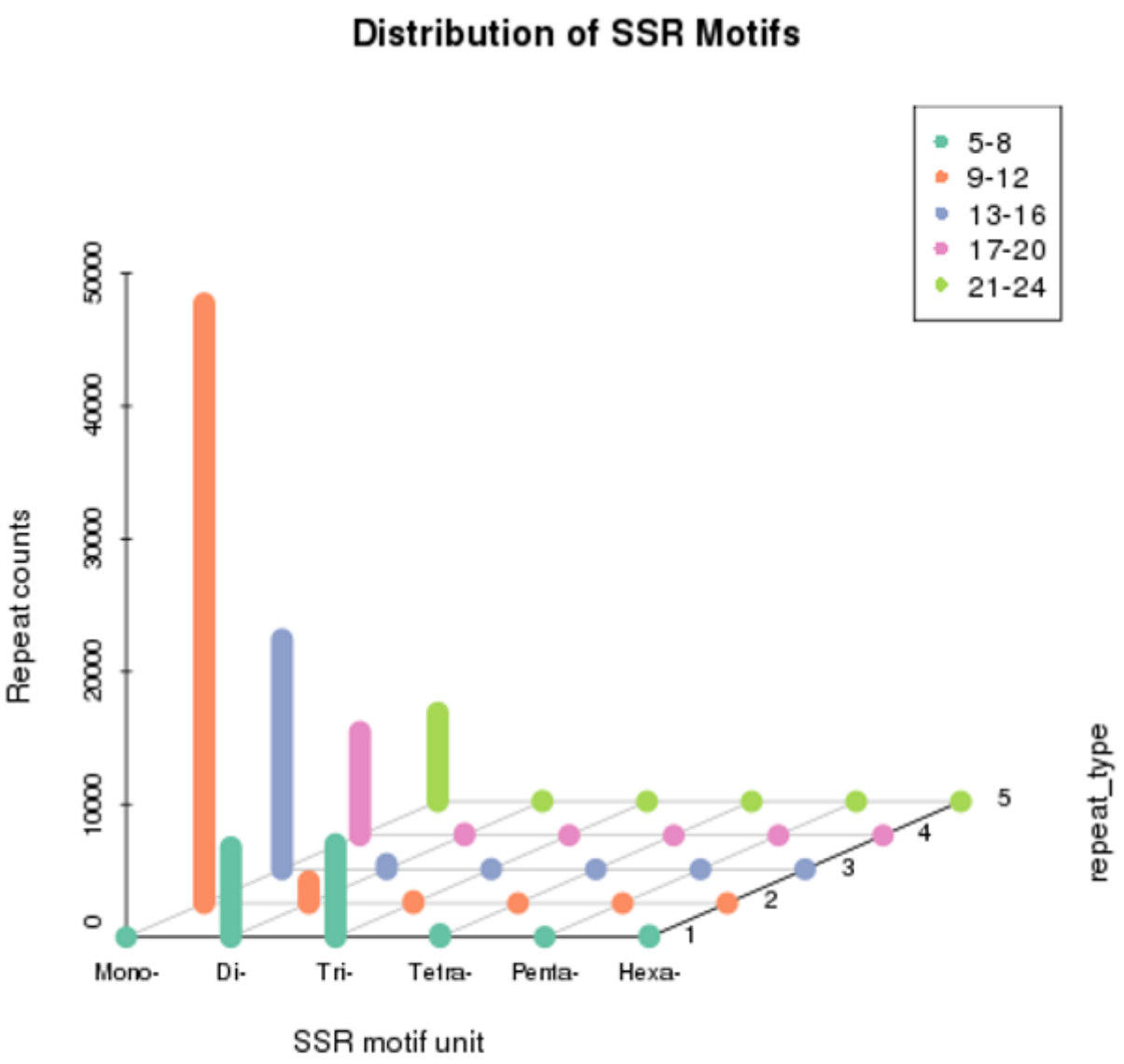



**Figure 9**

Venn diagram summarizing the lncRNA from full-length transcriptome of tree tomato (Cyphomandra betacea (Cav.) Sendtn.). Numbers within the Venn diagram indicate the number of sequences sharing among different databases.




## Figure 10

Frequency of simple sequence repeat (SSR) motifs in tree tomato (Cyphomandra betacea (Cav.) Sendtn.) full-length transcriptome.

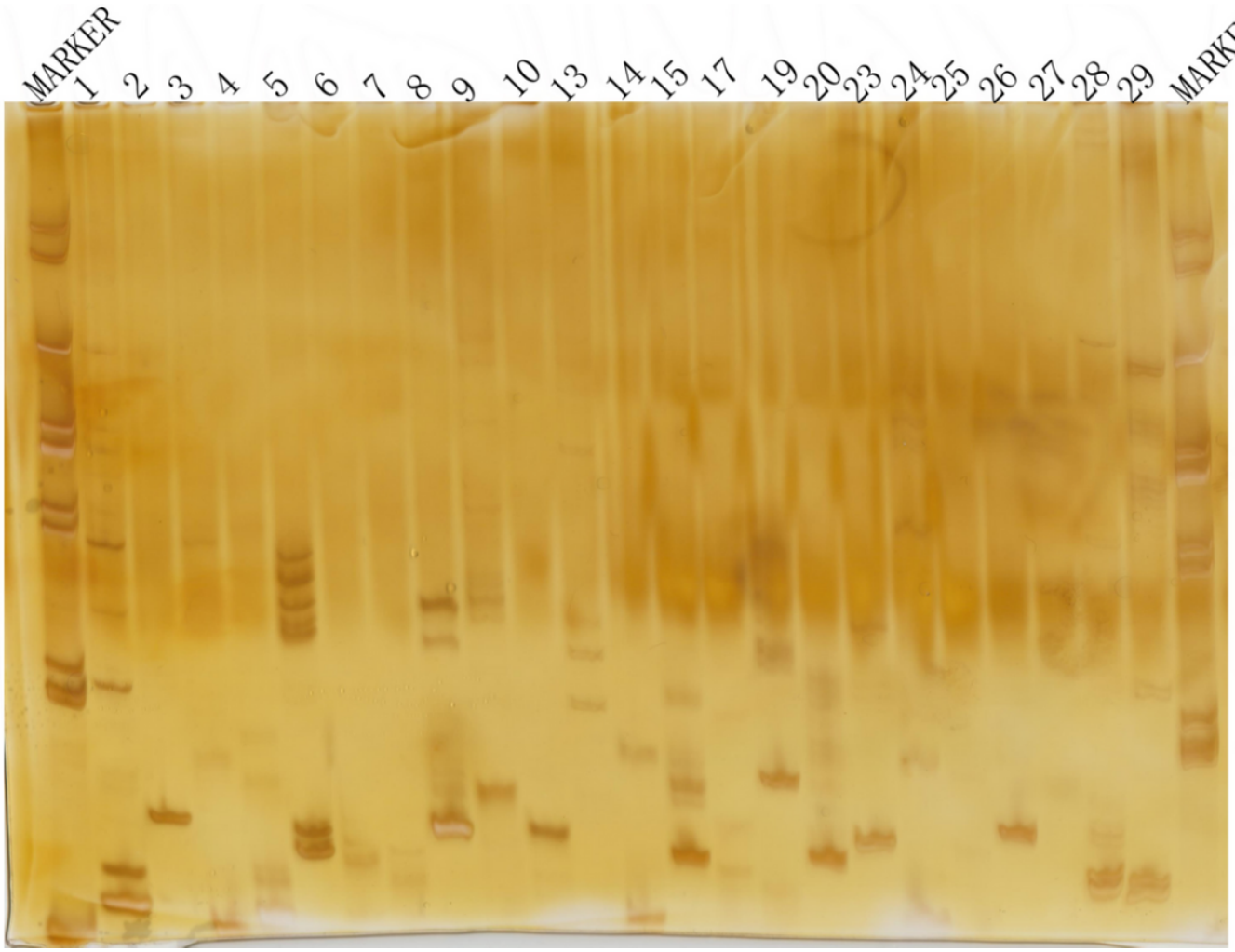



**Figure 11**

PCR amplification product of polymorphism of simple sequence repeat (SSR) primer pairs in in tree tomato (Cyphomandra betacea (Cav.) Sendtn.). The numbers of SSR primer pairs corresponded with the numbers in Additional file 12: Table S12.

## Supplementary Files

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

- Additionalfiles.xlsx