frontiers | Frontiers in Plant Science

TYPE Data Report
PUBLISHED 02 November 2022
DOI 10.3389/fpls.2022.1052817

OPEN ACCESS

EDITED BY
Fei Shen,
Beijing Academy of Agricultural and Forestry Sciences, China

REVIEWED BY
Xuepeng Sun,
Zhejiang Agriculture and Forestry University, China
Zhenyu Huang,
Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences (CAAS), China

*CORRESPONDENCE
Lijin Lin
llj800924@sicau.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 24 September 2022
ACCEPTED 10 October 2022
PUBLISHED 02 November 2022

CITATION
Deng H, Zhang L, Liao M, Wang J, Liang D, Xia H, Lv X, Deng Q, Wang X, Tang Y and Lin L (2022) A PacBio single molecule real-time sequencing-based full-length transcriptome atlas of tree tomato (*Solanum betaceum* Cav.) and mining of simple sequence repeat markers.
Front. Plant Sci. 13:1052817.
doi: 10.3389/fpls.2022.1052817

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

Honghong Deng[1†], Lu Zhang[1†], Ming'an Liao[2], Jin Wang[1], Dong Liang[1], Hui Xia[1], Xiulan Lv[1], Qunxian Deng[1], Xun Wang[1], Yi Tang[1] and Lijin Lin[1*]

[1]Institute of Pomology and Olericulture, Sichuan Agricultural University, Chengdu, China, [2]College of Horticulture, Sichuan Agricultural University, Chengdu, China

KEYWORDS

tree tomato, PacBio single molecule real-time sequencing, full-length transcriptome, simple sequence repeat, molecular marker

Introduction

Tree tomato (*Solanum betaceum* Cav., syn. *Cyphomandra betacea* (Cav.) Sendtn.), or tamarillo, is a fast-growing fruit species of the family *Solanaceae* genus *Solanum* (Acosta-Quezada et al., 2015). Native to the Andean region of South America, tree tomato cultivation has spread to several countries of the tropics and subtropics (South America, New Zealand, Australia and India) (Ramírez and Kallarackal, 2019). Recently, tree tomato has received increasing attention, as a rich source of sugars, organic acids, minerals, vitamins (vitamin C and B_6), carotenoids, anthocyanins, and phenolics (Acosta-Quezada et al., 2015; Espin et al., 2016). Thus, it developed from a disregarded crop to a promising fruit crop (Pacheco et al., 2021).

Previous studies on tree tomato mainly focused on its biochemical properties (Acosta-Quezada et al., 2015; Espin et al., 2016), phenology (Acosta-Quezada et al., 2016), and reproductive biology, including flower and pollen morphology, physiology, fruit characteristics, intraspecific hybridization, and genetic diversity (Ramírez and Kallarackal, 2019). Despite the recent interest and research progress, reference genome and transcriptome of tree tomato are not available, which hampers in-depth functional genomics, molecular genetics and genetic-assisted breeding. Additionally, *de novo*

assembly of transcriptome sequence through old-fashioned second-generation short-read sequencing, has been challenging, without a well-annotated reference genome (Amarasinghe et al., 2020). The advent of PacBio long-read single-molecule real-time (SMRT) sequencing approach addressed these challenges and provided the opportunity to obtain reliable genome-wide full-length (FL) transcripts directly (Amarasinghe et al., 2020).

Transcriptome profiling has proved an effective approach for the genome-wide development of simple sequence repeat (SSR) markers in several non-model plants, at a large scale and low cost (Deng et al., 2018; Jia et al., 2020). SSRs, used for genetic mapping, serve as DNA fingerprinting markers to assess genetic diversity and population structure. Furthermore, SSRs can be useful to distinguish closely-related cultivars, due to their advantages of single locus, multiple allele variations, and abundant polymorphism (Liu et al., 2015). To date, only amplified fragment length polymorphism (AFLP) markers were used to evaluate the genetic diversity between different tree tomato varieties (Acosta-Quezada et al., 2012). Therefore, identification of SSR markers at genome-wide scale for tree tomato are highly desirable.

In this study, we constructed for the first time, to the best of our knowledge, an atlas of tree tomato's FL transcriptome and analysed the distribution of SSR motifs.

Value of the data

- Using PcaBio SMRT sequencing, we constructed for the first time, an atlas of the FL transcriptome of tree tomato. This will facilitate further study of genome annotation to this crop, opening an exciting avenue in transcriptome-based studies, such as posttranscriptional regulation events analyses.
- To the best of our knowledge, no SSR markers were available for tree tomato gene mapping, until now. The current study encompasses the first mining and development of SSR markers in tree tomato, which will be determinant for genetic studies and molecular marker-assisted breeding in this fruit crop.

Materials and methods

Plant materials

Five-year-old plants of tree tomato were grown at the experimental base of the College of Horticulture, Sichuan Agricultural University, Chengdu, China (30.71°N, 103.87°E). Seven tissues (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 (root tips, shoot tips, and leaves) of three seedlings were sampled and mixed. Seedlings were obtained by incubation of seeds at 22°C and 95% relative humidity.

Library preparation and PacBio sequencing

Total RNA was extracted using a PureLink RNA mini kit (Invitrogen, CA, USA), followed by DNase digestion and RNA purification using an on-column PureLink DNase kit (Invitrogen). RNA concentration and purity were determined using a NanoPhotometer spectrophotometer (Implen, CA, USA). RNA integrity was determined using an RNA Nano 6000 assay kit on a Bioanalyzer 2100 system (Agilent Technologies, CA, USA). RNA integrity number (RIN) > 7.0 and 2.0 < OD 260/280 < 2.2 were the RNA quality requirements for the RNA samples. Iso-Seq cDNA library was constructed and PacBio sequencing were performed at Novogene Co., Ltd. (Beijing, China). The mRNA was enriched using oligo-dT magnetic beads in 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 (Sage Science, MA, USA) protocol and sequenced on the PacBio sequel platform.

Reads processing and error collection

Raw data were processed using SMRTlink v5.0 software. Circular consensus sequencing (CCS) reads were yielded from subread Binary Alignment Map (BAM) files. 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). Short reads (shorter than 50 bp in length) were discarded. FLNC sequences were isoform-level clustered with iterative clustering and error correction (ICE) software, generating one consensus isoform (Gordon et al., 2015). The non-full-length CCSs were polished using the Quiver algorithm. High quality FL transcripts were defined with the criterial of//a minimum Quiver accuracy of 0.99.

Functional annotation and transcript analysis

Gene functional annotation was performed using the National Center for Biotechnology Information (NCBI) non-

redundant protein (Nr, E-value ≤1 e-5), NCBI non-redundant nucleotide (Nt, E-value ≤1 e-5), gene ontology (GO, E-value ≤1e-10), Kyoto encyclopedia of genes and genomes (KEGG, E-value ≤1e-3), eukaryotic orthologous groups (KOG, E-value ≤1e-3), Swissprot protein (E-value ≤1e-5^{-5}), and protein family (Pfam, E-value ≤0.01) databases.

Coding sequence (CDS) was predicted by ANGEL (Robust Open Reading Frame prediction) with default parameters (Shimizu et al., 2006). Transcription factors (TFs) were predicted using iTAK software ((http://itak.feilab.net/cgi-bin/itak/index.cgi) (Zheng et al., 2016). Long non-coding RNA (LncRNA) was firstly screened *via* coding-non-coding-index with default parameters (Kong et al., 2007) and Coding Potential Calculator with NCBI eukaryotes' protein database (E-value <1e-10) (Sun et al., 2013). Each transcript was then translated in three possible frames, and Pfam Scan with default parameters of -E 0.001 –domE 0.001 was employed to determine the existene of a known protein family domain. SSRs were identified by MISA program (https://pgrc.ipk-gatersleben.de/misa/).

Results

Full-length transcriptome of tree tomato

A total of 9.92G subreads base was obtained, comprising 9,877,631 subreads, with an average subreads length of 1,005 bp and an N50 length of 1,974 bp. Approximately 70.41% of the subreads fell within the size range of 200 to 1,000 bp. Of the 416,144 CCS isoforms, 308,699 were identified as consensus FLNC reads, with a mean length of 2,099 bp (Table 1).

A total of 140,327, 104,294, 135,138, 78,300, 53,520, 152,310 and 53,520 transcripts were functionally annotated by sequence similarity search against Nr, Swiss-Prot, KEGG, KOG, GO, Nt and Pfam databases, respectively (Figure 1A). Annotation of Nr homologous species distribution showed the best blast hit with tree tomato and *Solanum tuberosum* (52,712 isoforms), *Solanum pennellii* (21,171 isoforms), *Solanum lycopersicum* (16,666 isoforms), and *Capsicum annuum* (15,851 isoforms) (Figure 1B). Transcripts were classified into three GO categories, including biological process, cellular component, and molecular function, in which the most abundant subcategory was metabolic process (27,699 matched genes, 51.75%), cell (12,693 matched genes, 23.72%), and binding (30,712 matched genes, 57.38%), respectively (Figure 1C).

KOG analysis showed tree tomato transcripts were assignmented to a total of 26 categories. The largest group belonged to general function prediction only (15,323 matched genes, 19.57%), followed by posttranslational modification, protein turnover, chaperones (9,750, 2.45%) and signal transduction mechanism (8,614, 11.00%) (Figure 1D). KEGG functional classification showed a total of 5895 out of the 135,138 transcripts assigned to the signal transduction, thus making it the largest group (4.36%) amongst the major categories, followed by translation (5,233, 3.87%), folding, sorting and degradation (4,989, 3.69%), and carbohydrate metabolism (4,745, 3.51%) (Figure 1E).

Structure analysis and SSR identification

The frequencies for each length of CDS were evaluated with the most prevalent length of CDS ranged from 400 to 2,000 bp (Figure 2A). A total of 5,114 genes were predicted to be TFs belonging to different families, amongst which the most abundant 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%) (Figure 2B). A total of 43227, 42872, and 110333 noncoding RNAs candidates were predicted by CPC, CNCI, and Pfam databases, respectively. Among them, 29,453 transcripts were simultaneously identified by the three computational approaches (Figure 2C). A screen of the 79549 transcripts using MISA program yielded diverse SSR types, including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, hexanucleotide, and some complex nucleotides. Amongst these, mononucleotide repeats (63.97%) exhibited the highest frequency of occurrence, followed by dinucleotide (8.54%) and trinucleotide repeats (7.79%) (Figure 2D).

TABLE 1 Summary of circular consensus sequence of tree tomato (*Solanum betaceum* Cav.) 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

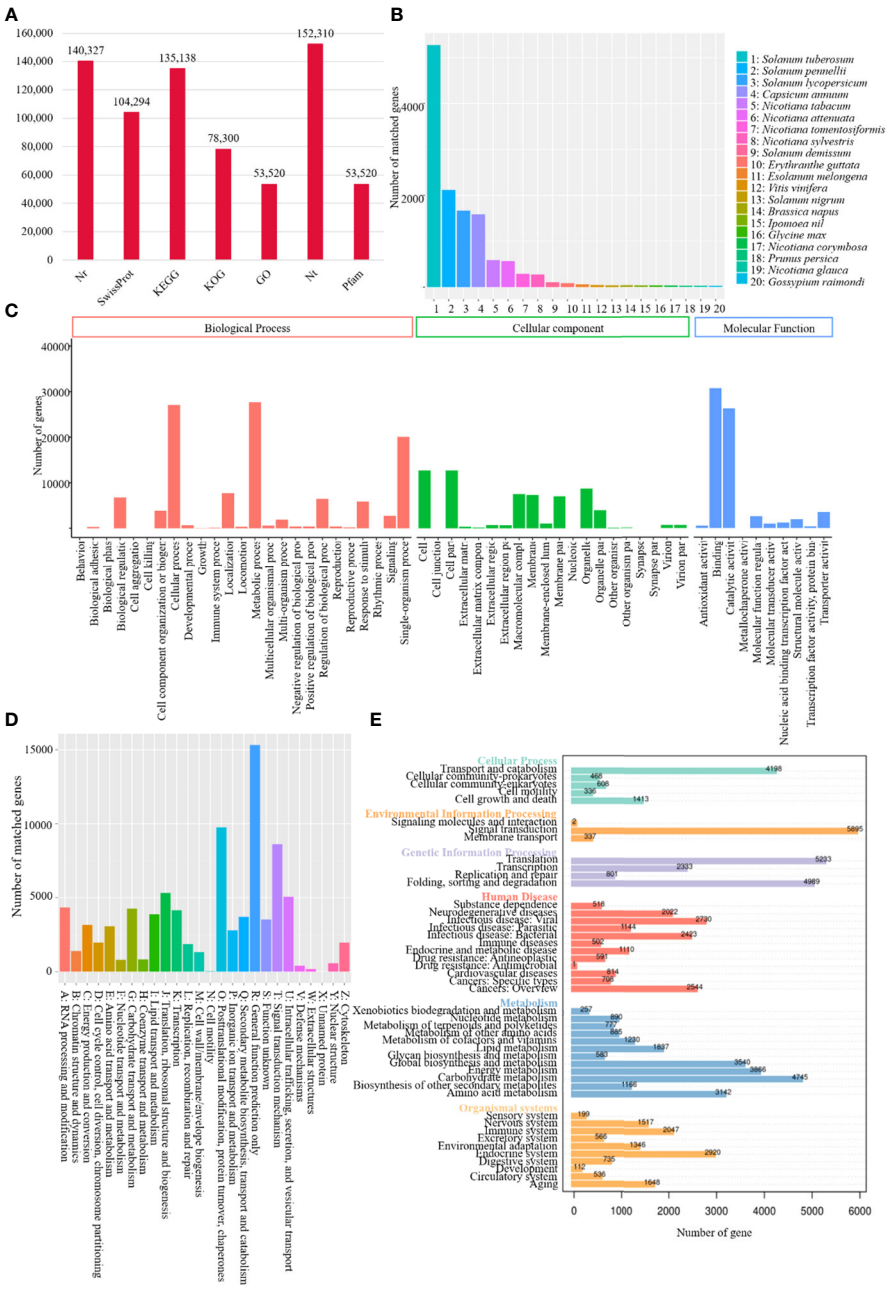

FIGURE 1
Functional annotation of tree tomato full-length transcriptome. **(A)** Bar chart demonstrating tree tomato annotated number of genes in different databases; **(B)** homologous species distribution of tree tomato transcripts annotated in Nr database; **(C)** distribution of GO terms for all tree tomato annotated transcripts; **(D)** KOG classification of tree tomato transcripts; **(E)** KEGG enrichment of tree tomato transcripts.

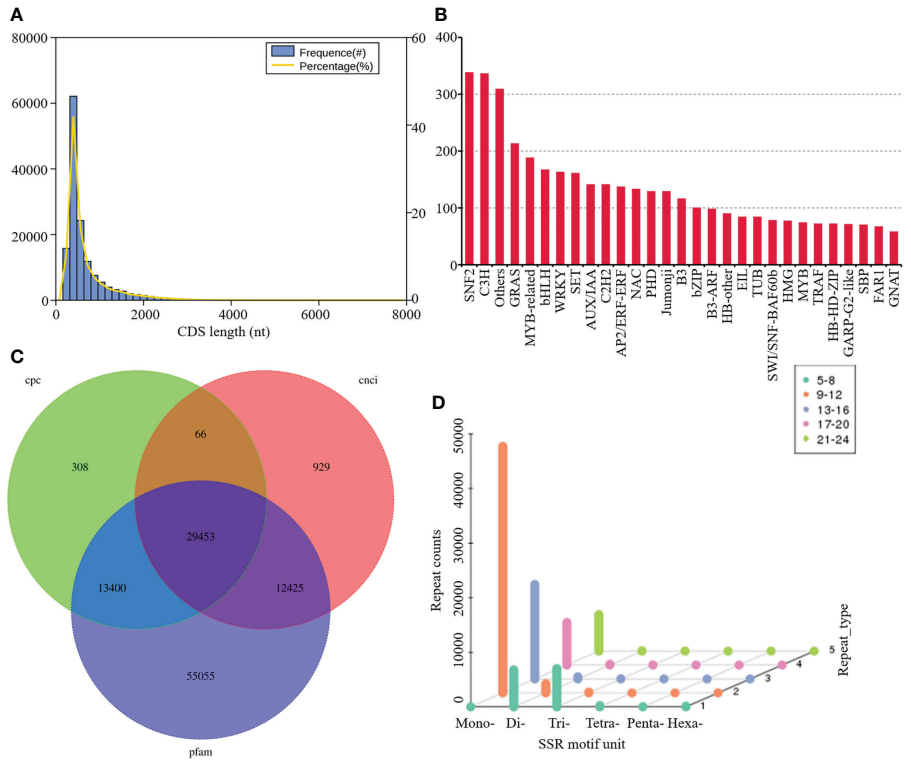

FIGURE 2
Structure Analysis and SSR Identification. **(A)** Frequency histogram depicting length distribution of CDS; **(B)** distribution of TFs; **(C)** Venn diagram summarizing lncRNAs number; **(D)** SSR motifs frequency.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA883812 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA883812). The function annotation, gene structure analysis and SSR identification have been deposited at the Figshare database with doi: 10.6084/m9.figshare.21200887.

Author contributions

LL conceived the idea and acquired funding; LZ, ML, JW, DL, and HX collected the samples and conducted the experiment; HD, XL, QD, XW, YT, and LL performed analysis on the data; HD wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the “Shuangzhi” program of Sichuan Agricultural University.

Conflict of interest

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.

Publisher’s note

References

Acosta-Quezada, P. G., Raigón, M. D., Riofrío-Cuenca, T., García-Martínez, M. D., Plazas, M., Burneo, J. I., et al. (2015). Diversity for chemical composition in a collection of different varietal types of tree tomato (*Solanum betaceum* cav.), an Andean exotic fruit. *Food Chem.* 169, 327–335. doi: 10.1016/j.foodchem.2014.07.152

Acosta-Quezada, P. G., Riofrío-Cuenca, T., Rojas, J., Vilanova, S., Plazas, M., and Prohens, J. (2016). Phenological growth stages of tree tomato (*Solanum betaceum* cav.), an emerging fruit crop, according to the basic and extended BBCH scales. *Sci. Hortic. (Amsterdam)* 199, 216–223. doi: 10.1016/j.scienta.2015.12.045

Acosta-Quezada, P. G., Vilanova, S., Martínez-Laborde, J. B., and Prohens, J. (2012). Genetic diversity and relationships in accessions from different cultivar groups and origins in the tree tomato (*Solanum betaceum* cav.). *Euphytica* 187, 87–97. doi: 10.1007/s10681-012-0736-7

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 1–16. doi: 10.1186/s13059-020-1935-5

Deng, K., Deng, R., Fan, J., and Chen, E. (2018). Transcriptome analysis and development of simple sequence repeat (SSR) markers in *Zingiber striolatum* diels. *Physiol. Mol. Biol. Plants* 24, 125–134. doi: 10.1007/s12298-017-0485-0

Espin, S., Gonzalez-Manzano, S., Taco, V., Poveda, C., Ayuda-Durán, B., Gonzalez-Paramas, A. M., et al. (2016). Phenolic composition and antioxidant capacity of yellow and purple-red Ecuadorian cultivars of tree tomato (*Solanum betaceum* cav.). *Food Chem.* 194, 1073–1080. doi: 10.1016/j.foodchem.2015.07.131

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, 1–15. doi: 10.1371/journal.pone.0132628

Jia, X., Tang, L., Mei, X., Liu, H., Luo, H., Deng, Y., et al. (2020). Single-molecule long-read sequencing of the full-length transcriptome of *Rhododendron lapponicum* l. *Sci. Rep.* 10, 1–11. doi: 10.1038/s41598-020-63814-x

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

Liu, Q., Song, Y., Liu, L., Zhang, M., Sun, J., Zhang, S., et al. (2015). Genetic diversity and population structure of pear (*Pyrus* spp.) collections revealed by a set of core genome-wide SSR markers. *Tree Genet. Genomes* 11, 1–22. doi: 10.1007/s11295-015-0953-z

Pacheco, J., Vilanova, S., Grillo-Risco, R., García-García, F., Prohens, J., and Gramazio, P. (2021). *De novo* transcriptome assembly and comprehensive annotation of two tree tomato cultivars (*Solanum betaceum* cav.) with different fruit color. *Horticulturae* 7, 431. doi: 10.3390/horticulturae7110431

Ramírez, F., and Kallarackal, J. (2019). Tree tomato (*Solanum betaceum* cav.) reproductive physiology: A review. *Sci. Hortic. (Amsterdam).* 248, 206–215. doi: 10.1016/j.scienta.2019.01.019

Shimizu, K., Adachi, J., and Muraoka, Y. (2006). Angle: A sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J. Bioinform. Comput. Biol.* 4, 649–664. doi: 10.1142/S0219720006002260

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41, e166. doi: 10.1093/nar/gkt646

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014