Plant Communications
Correspondence

CellPress
Partner Journal

A telomere-to-telomere reference genome assembly of tomato cultivar 'Heinz 1706'

Dear Editor,

Tomato (*Solanum lycopersicum*) is one of the most important vegetable crops globally. As a model plant for genetic and genomic research, it has been extensively studied, particularly for insights into fruit development and the mechanisms behind biotic and abiotic resistance. While the first tomato genome assembly of the cultivar 'Heinz 1706' was released over a decade ago (Tomato Genome Consortium, 2012), the latest 'SL5.0' version, assembled using Pacific Biosciences HiFi sequencing data (Zhou et al., 2022), remains incomplete, with 31 gaps, 21 missing telomeres, and a lack of information about 45S ribosomal RNA genes (rDNAs). Recently, advances in ultra-long Oxford Nanopore Technologies (ONT) sequencing technology have enabled the assembly of many telomere-to-telomere (T2T) plant genomes. It is crucial to provide such a T2T tomato reference genome for genomic and evolutionary studies.

We obtain 95.7 Gb of ultra-long ONT reads (~98× coverage of high-quality reads with lengths > 50 kb and base accuracy > 99%, ~Q25 on average) and 11.8 Gb of Illumina short reads from 'Heinz 1706' (Supplemental Table 1; Supplemental Figure 1A). Using the high-quality ONT reads as input to Hifiasm (Cheng et al., 2024), we obtained an assembly that included 10 complete chromosomes, each with two telomeres, a nearly complete chromosome 3 with one telomere, a partial chromosome 2 with one telomere, and 4 additional contigs longer than 1 Mb that were mainly composed of 45S rDNAs (including a telomere in a single contig) (Supplemental Table 2). Using overlapping ONT reads, chromosome 3 was completed by extending the chromosome end with the missing telomere (Supplemental Figure 2A). Chromosome 2 was further built by scaffolding the 5 rDNA-containing contigs (Supplemental Table 2) with 4 gaps, each represented by 100 "N". After correction using the Illumina short reads for homozygous SNPs and small insertions or deletions (Supplemental Table 3), the final genome sequence (designated SL-T2T) had a total size of 831.45 Mb with a contig N50 size of 68.49 Mb. Composed of 12 chromosomes with all 24 telomeres, the genome was completely assembled except for the 45S rDNA region on chromosome 2 (Figure 1A; Supplemental Tables 4 and 5).

To examine the assembly quality, we first combined the previously released HiFi reads (~43× coverage) (Zhou et al., 2022) with the high-quality ONT reads to obtain an assembly consisting of 12 chromosome-level contigs using Verkko (Rautiainen et al., 2023). A comparison of SL-T2T, SL5.0, and the major contigs of the Verkko assembly showed only small discrepant regions (up to 46 kb) that were caused either by heterozygosity (probably due to the mixed seeds used in this study) or by sample differences between the ONT and HiFi sequencing (Supplemental Figures 2B and 2C; Supplemental Table 6), with the exception of the extra sequences present in SL-T2T. We next checked the mapping rate of ONT, HiFi, and Illumina reads on SL-T2T. These were 100%, 100%, and 98.15%, respectively, with uniform depth of coverage throughout the genome except in the 45S rDNA region (Supplemental Figure 3A). Visual inspection in the Integrative Genomics Viewer (https://igv.org/) of all satellite DNA regions confirmed the contiguous mapping of ONT reads. Finally, SL-T2T was assigned a Merqury quality value estimate of 53.39, a completeness score of 99.66% (k-mer length of 25 and k-mer multiplicity > 22) (Supplemental Figure 1B), and a Benchmarking Universal Single-Copy Orthologs score of 98.6%. Taken together, these results demonstrate the high base accuracy and completeness of SL-T2T.

Compared with SL5.0, SL-T2T contained a total of 29.67 Mb of additional sequences, the majority (84.51%) of which were 45S rDNAs and other tandem repeats. Among the 31 gaps in SL5.0 (Figure 1A), the largest was 362 kb (Supplemental Figure 4). All other gaps were shorter than 100 kb, including nine surrounded by overlapping repetitive sequences. Compared with SL-T2T, SL5.0 also contained two erroneously inverted (an example is shown in Supplemental Figure 4) and three erroneously translocated sequences with a total length of 371 kb and 23 kb, respectively. The gap-free assembly of all of these regions in SL-T2T, supported by contiguous ONT read mapping, highlights the superiority of the ultra-long, highly accurate ONT reads for assembling complex regions.

We annotated a total of 36,006 genes in SL-T2T, of which 22,784 were functionally annotated (including 254 genes transferred from SL5.0). Among the SL-T2T genes, 31,350 matched with SL5.0 annotation, and 23,294 matched with gene structures in SL5.0. We found that the SL-T2T genes that differed from SL5.0 usually had better transcriptome evidence (an example is shown in Supplemental Figure 3). Moreover, we identified 66.51% of the SL-T2T genome as repetitive sequences, including approximately 61.09% of transposons and 4.94% of tandem repeats, which included two satellite DNAs, SlSat181 and SlSat141, and two minisatellite DNAs, SlSat53 and SlSat35 (Supplemental Table 7; Supplemental Figures 1C and 5).

Using the previously reported tomato centromere-enriched repetitive sequence TGRIV (Chang et al., 2008), we identified 12 centromeres with a total length of 40.90 Mb in SL-T2T (Figures 1B and 1C). A total of 582 annotated genes (including 195 functionally annotated) were found in the centromeric regions, with a gene density far below the genome average (14.2 vs. 43.3 genes per Mb). SL-T2T contained only 1.24 Mb of tandem repeats in the centromeric regions, consistent with a previous report showing that the centromeric regions of Solanaceae species are primarily composed of long terminal repeat (LTR)

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

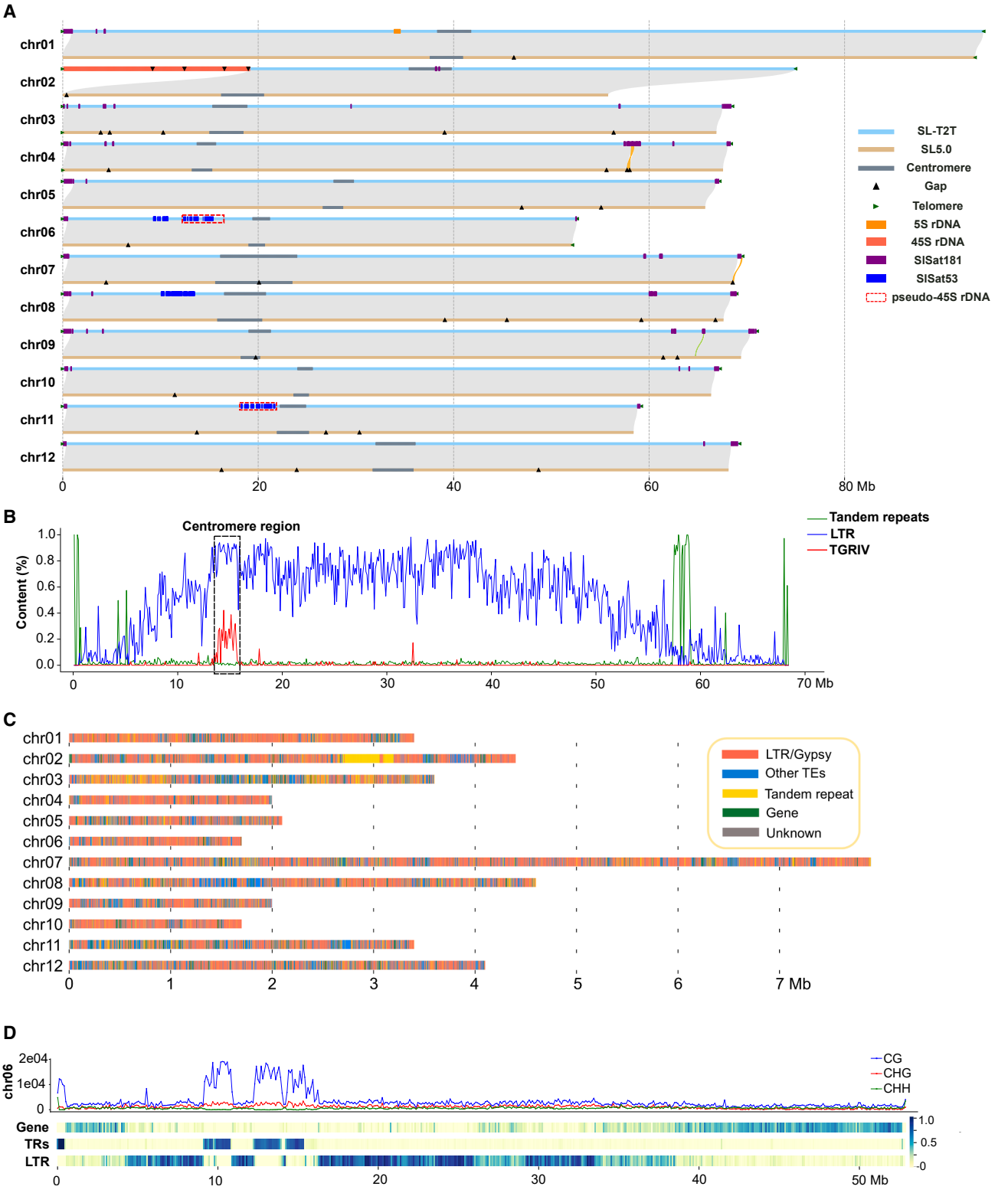

Figure 1. Summary of SL-T2T genome features.

(A) Comparison of the SL-T2T and SL5.0 genomes.

(B) Identification of the putative centromere on chromosome 4 using the centromere-enriched repetitive sequence TGRIV. The distribution of LTR retrotransposons and tandem repeats is also shown.

(C) Distribution of tandem repeats, transposons, and genes in the putative centromeres.

(D) Distribution of CG, CHG, and CHH methylation sites, as well as genes, tandem repeats (TRs), and LTR retrotransposons on chromosome 6.

retrotransposons (Figure 1B; Supplemental Table 8) (Chen et al., 2024).

The satellite sequence SlSat181, with a tandem repeat unit length of 181 bp, was identified in 21 of the subtelomeric regions with a tandem array size ranging from 100 kb to 1.1 Mb. SlSat181 was also found in the centromeres of chromosomes 2 and 5 with total lengths of 480 and 7 kb, respectively (Figure 1A). The 35-bp minisatellite DNA SlSat35 was widely distributed across the genome, including all 12 centromeric regions, repeated a total of 57,254 times (Supplemental Figure 1C). In addition, a single 5S rDNA array was found at 33.87–34.58 Mb on chromosome 1, containing 1,918 units and a single LTR retrotransposon.

It was previously known that the beginning of chromosome 2 contains a large amount of 45S rDNA, with an estimate of 2,300 units (Chang et al., 2008). Based on the total number of 45S rDNA units in the ONT data, we estimated that the 'Heinz 1706' genome contained approximately 3,400 complete 45S rDNA units. We found that the 45S rDNA array in SL-T2T, which adjoins the telomere, contained 2,105 complete units, with a total length of 19.11 Mb. All 45S rDNA units on chromosome 2 are possibly oriented in the same direction, as no adjacent units in the opposite direction were observed in the ONT reads. Outside of chromosome 2, a single unit of 45S rDNA was located at 24.38 Mb on chromosome 11. The incompleteness of the assembled 45S rDNAs in SL-T2T was evidently caused by the presence of large number of nearly identical 45S units, since the average sequence similarity was 99.92% between the assembled repeat units including the intergenic spacers (Supplemental Figure 6A), and the average sequence similarity of 98.24% between the repeat units in the ONT reads (Supplemental Figure 6A) was very close to the limit set by the ONT read accuracy of Q25 as well. To assess sequence variation among the 45S rDNA repeat units, we identified a total of 11 SNPs and 62 small insertions or deletions ($\leq$ 5 bp) at a minor allele frequency $\geq$ 0.05 (Supplemental Figures 6B–6E). These variations were consistent between the assembled sequence and the ONT reads but did not show characteristic patterns that could divide the repeat units into distinct subgroups. To better understand the evolutionary pattern of the 45S rDNA in tomato, we also computed the copy number (6–1,551) and average sequence similarity (89.45%–97.02%) of the 45S rDNA units in several recently published Solanaceae T2T genomes (Supplemental Table 9). The largest 45S rDNA size and unit similarity of SL-T2T among these species indicated a recent burst of expansion of the 45S rDNA in tomato after it diverged from potato.

In addition to the complete 45S rDNA units, we also identified two pseudo-45S rDNA regions containing only partial 18S or 28S rDNA sequences, located on chromosome 6 at 12.30–16.56 Mb (including six 18S units and 28 28S units) and on chromosome 11 at 18.06–21.90 Mb (including 44 18S units and 53 28S units) (Supplemental Figure 6F). These pseudo-45S regions on chromosomes 6 and 11 contained a high proportion of LTR retrotransposons (18.49% and 36.14% of the entire region, respectively), whereas the 45S rDNA region on chromosome 2 contained only miniature inverted-repeat transposable elements (12%). The 45S rDNA region also contained two types of tandem repeats in the intergenic spacers: SlSat141 with a unit length of 141 bp and SlSat53 with 53 bp (Supplemental Figure 5). The two pseudo-45S regions were composed mainly of the minisatellite SlSat53 (Supplemental Figure 6F), which had undergone extensive expansion compared to its presence in the 45S rDNA region. SlSat53 was also found on chromosome 8 at 10.06–13.49 Mb, athough no 45S rDNA sequences were identified in that region. These results suggest that the pseudo-45S rDNA regions evolved from the degradation of 45S rDNAs and the expansion of the minisatellite SlSat53.

DNA methylation at cytosine bases (5mC) in plant genomes can occur at CpG dinucleotide sites (designated mCG) or CHG (H = A, T, or C) and CHH trinucleotide sites (mCHG or mCHH). Using the ONT sequencing data, we identified a total of 48,498,625 methylation sites in SL-T2T (Supplemental Tables 10 and 11). The number and proportion of mCGs in the tandem repeats were markedly higher than the genome-wide averages (142.01 vs. 37.65 sites/kb and 98.23% vs. 92.09%, respectively) (Figure 1D; Supplemental Figure 7), consistent with methylation patterns observed in the human genome (Francastel and Magdinier, 2019) and suggesting a potential role for mCGs in maintaining genome integrity in tomato. The proportions of mCG and mCHG in LTR retrotransposons also increased substantially compared with genome-wide averages (98.65% vs. 92.09% and 59.98% vs. 43.74%, respectively), consistent with the established function of DNA methylation in transposon silencing to maintain genome stability (Zhang et al., 2018). Notably, the densities of mCG and mCHG in 45S rDNAs were the highest among all genomic features, possibly reflecting selective expression within the nucleolar organizing region (Woo and Richards, 2008). Similarly, mCHH density was elevated in gene upstream regions (within 500 bp) (8.71 vs. 3.86), consistent with a regulatory role for mCHH in gene expression (Henderson and Jacobsen, 2008). Together, these results highlight the distinct functions of the three types of methylation sites in the tomato genome, and the observed 5mC landscape aligns with their proposed biological functions.

In summary, we generated a highly accurate 831.45-Mb T2T genome assembly of the tomato cultivar 'Heinz 1706', providing a new reference genome to facilitate tomato genomics research and genetic improvement.

DATA AND CODE AVAILABILITY

All raw sequencing reads and SL-T2T genome sequences have been deposited in the Chinese National Genomics Data Center (https://ngdc.cncb.ac.cn/) under BioProject PRJCA034231 and Genome Warehouse accession GWHFILF00000000. HiFi sequencing data are available in GenBank (https://www.ncbi.nlm.nih.gov/) under BioProject PRJNA733299. Data related to methylation site identification can be accessed at https://tomato.mbkbase.org/slt2t. Genome sequences and gene annotations are also available at the tomato genome database (https://tomato.mbkbase.org) through JBrowse and/or BLAST. Scripts used for data analysis are available at https://github.com/chenys1998/SL-T2T-assembly.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant 31991181).

ACKNOWLEDGMENTS

We thank Prof. Sanwen Huang for support and helpful discussions. No conflict of interest is declared.

AUTHOR CONTRIBUTIONS

C.L. designed the research and supervised the analyses. Y.C., J.T., and Y.Z. performed the data analyses. J.Z. provided materials for sequencing and offered valuable suggestions for data analysis. C.L. and Y.C. wrote the manuscript. All authors read and approved the final manuscript.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: January 23, 2025
Revised: September 9, 2025
Accepted: November 17, 2025
Published: November 19, 2025

Yongshun Chen[1], Jiaoyang Tian[1], Yuhui Zhao[1], Jinzhe Zhang[2] and Chengzhi Liang[1,3,]*

[1]Key Laboratory of Seed Innovation, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[2]State Key Laboratory of Vegetable Biobreeding, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
***Correspondence: Chengzhi Liang (cliang@genetics.ac.cn)**
https://doi.org/10.1016/j.xplc.2025.101618

REFERENCES

Chang, S.B., Yang, T.J., Datema, E., van Vugt, J., Vosman, B., Kuipers, A., Meznikova, M., Szinay, D., Lankhorst, R.K., Jacobsen, E., and de Jong, H. (2008). FISH mapping and molecular organization of the major repetitive sequences of tomato. Chromosome Res. **16**:919–933.

Chen, W., Yan, M., Chen, S., Sun, J., Wang, J., Meng, D., Li, J., Zhang, L., and Guo, L. (2024). The complete genome assembly of Nicotiana benthamiana reveals the genetic and epigenetic landscape of centromeres. Nat. Plants **10**:1928–1943.

Cheng, H., Asri, M., Lucas, J., Koren, S., and Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat. Methods **21**:967–970.

Francastel, C., and Magdinier, F. (2019). DNA methylation in satellite repeats disorders. Essays Biochem. **63**:757–771.

Henderson, I.R., and Jacobsen, S.E. (2008). Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. Genes Dev. **22**:1597–1606.

Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M., and Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat. Biotechnol. **41**:1474–1482.

Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**:635–641.

Woo, H.R., and Richards, E.J. (2008). Natural variation in DNA methylation in ribosomal RNA genes of Arabidopsis thaliana. BMC Plant Biol. **8**:92.

Zhang, H., Lang, Z., and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. **19**:489–506.

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