Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A High-continuity and Annotated Tomato Reference Genome

**Xiao Su**
China Agricultural University
**Baoan Wang**
China Agricultural University
**Xiaolin Geng**
China Agricultural University
**Yuefan Du**
China Agricultural University
**Qinqin Yang**
China Agricultural University
**Bin Liang**
China Agricultural University
**Ge Meng**
China Agricultural University
**Qiang Gao**
Zhejiang University
**Wencai Yang**
China Agricultural University
**Yingfang Zhu**
Henan University
**Tao Lin** ( lintao35@cau.edu.cn )
China Agricultural University

**Research Article**

**Keywords:** de novo tomato genome, comparative genomics, high-density genetic map, QTL analysis

**Posted Date:** June 7th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-579393/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Version of Record:** A version of this preprint was published at BMC Genomics on December 1st, 2021. See the published version at https://doi.org/10.1186/s12864-021-08212-x.

# Abstract

**Background:** Genetic and functional genomics studies require a high-quality genome assembly. Tomato (*Solanum lycopersicum*), an important horticultural crop, is an ideal model species for the study of fruit development.

**Results:** Here, we assembled an updated reference genome of *S. lycopersicum* cv. Heinz 1706 that was 799.09 Mb in length, containing 34,384 predicted protein-coding genes and 65.66% repetitive sequences. By comparing the genomes of *S. lycopersicum* and *S. pimpinellifolium* LA2093, we found a large number of genomic fragments probably associated with human selection, which may have had crucial roles in the domestication of tomato. We also used a recombinant inbred line (RIL) population to generate a high-density genetic map with high resolution and accuracy. Using these resources, we identified a number of candidate genes that were likely to be related to important agronomic traits in tomato.

**Conclusion:**Our results offer opportunities for understanding the evolution of the tomato genome and will facilitate the study of genetic mechanisms in tomato biology.


# Background

Tomato (*Solanum lycopersicum*) is an important model plant for scientific researches on fruit development and quality[1]. The tomato cultivation area has increased by ~ 1 million hectares over the past decade, and the yield has increased from 155 million tons to 181 million tons (http://www.fao.org). As a nutritious vegetable that contributes to the human diet, tomato is reported to contain more health-promoting compounds such as lycopene than some other popular fruits. These compounds lower risk of cancer and maintain human health[2]. Tomato was originally found mainly in the Andean mountains of South America. Its fruit weight and quality differ markedly among different horticultural groups, and wild tomatoes have smaller seeds and lower yields than cultivars.

A draft genome of the tomato cultivar Heinz 1706 produced using shotgun sequencing technology was released in 2012[3] and widely used as a reference genome for scientific researches. However, the fragmented nature of this genome and the resulting incomplete gene models could hindered the discovery and functional analysis of important genes. The completeness, accuracy, and contiguity of genome assemblies depend mainly on sequencing technology and assembly strategy. In the current genomic era, single-molecule real-time (SMRT) sequencing technology and new assembly pipelines have remarkably improved the quality of genome assemblies such as those of rice[4], cucumber[5], and tomato[6]. Although these genome assemblies have accelerated some scientific researches, such as QTL mapping and transcriptome analysis, higher continuous and complete genome sequences are required for identification of large structural variations and gene mining.

Across over the past ~ 10,000 years, humans have selected consciously and unconsciously for beneficial traits that have made wild plants more suitable for human use. The impacts of human selection on crops have been recorded in their genomes and have a central role in crop improvement. The resulting

“domestication syndrome” comprises a common suite of traits useful for human needs and plant survival, such as larger fruits or grains, better taste, bigger seeds, and more robust plants overall[7, 8]. In crop species, bigger seeds can accumulate adequate nutrition for germination and produce more vigorous seedlings[9]. Seed size, as a domestication trait, has been elucidated in several crops, and genes such as *GS3*[10, 11], *DEP1*[12], *GW2*[13], *GW8*[14], and *qSW5*[15] in rice; *HvYrg* in barley[16], and *TaGW2* in wheat[17] have been characterized. In tomato, seeds are smaller in wild tomatoes than in cultivars. *Sw4.1* is one of the major QTLs which encode an ABC transporter responsible for phenotypic variation in tomato seed size[18, 19].

In this study, we generated a highly continuous and complete genome sequence of Heinz 1706 (version SLT1.0) that contains many fewer gaps and unplaced contigs and demonstrates better assembly of repetitive regions. By comparing the genomes of *S. lycopersicum* and *S. pimpinellifolium* LA2093, we found a large number of genomic fragments that appear to be likely involved in domestication. We also used a RIL population, which was derived from a cross between the cultivated line OH88119 and the wild line PI128216, to identify candidate genes and loci that control several important agronomic traits. Our work offers new opportunities for understanding the evolutionary history of the tomato genome and the genetic mechanisms that underlie complex traits in tomato breeding.

## Results

# High-quality genome assembly

We assembled a highly continuous and complete genome sequence of Heinz 1706 using an integrated genome sequencing approach that combined 131.78 Gb (168.52×) of SMRT data, 226.97 Gb (290.24×) of BioNano data, 140.52 Gb (179.70×) of Hi-C data, and 50.93 Gb (61.53×) of Illumina short-read data (Supplementary Table S1). The PacBio long reads with an N50 read length of 32.82 kb were assembled with CANU software[20], generating a 875.21-Mb genome with a contig N50 of 17.83 Mb (Table 1). To reduce fragmentation and fill in gaps, BioNano data and Hi-C data were used to assist with scaffold construction using Aigner and Assembler[21], HERA[22], and Juicer[23] software. A Hi-C-based physical heatmap comprising 12 groups was generated (Supplementary Fig. S1) and used to create 12 pseudo-chromosomes that anchor ~790.59 Mb of the genome and harbor 97.61% (33,562) of the predicted protein-coding genes. The genome assembly was polished with Illumina short reads for error homozygous SNPs or indels using Pilon software[24]. As a result, we generated a 799.09-Mb genome assembly, SLT1.0 (Fig. 1, Table 1).

Table 1

Genome assembly and annotation of SLT1.0

| | SLT1.0 | SL4.0 | SL3.0 |
|---|---|---|---|
| Genome assembly (Mb) | 799.09 | 782.52 | 828.08 |
| Non-N bases | 797,955,212 | 782,475,302 | 746,357,470 |
| Number of gaps | 210 | 286 | 22,700 |
| Number of total contigs | 1,615 | 504 | - |
| Longest contig length (Mb) | 47.16 | 26.29 | - |
| N50 of contigs (Mb) | 17.83 | 6.01 | - |
| Number of unplaced contigs | 112 | 176 | 4,374 |
| Unplaced contigs sequence length (Mb) | 8.50 | 9.64 | 20.85 |
| Number of genes | 34,384 | 34,075 | 35,768 |
| Percentage of gene length in genome (%) | 16.21 | 15.56 | 17.33 |
| Mean gene length (bp) | 3,766.53 | 3,572.44 | 4,011.09 |
| Gene density (per Mb) | 43.03 | 43.55 | 43.19 |
| Mean coding sequence length (bp) | 223.02 | 228.01 | 219.97 |
| Mean exon length (bp) | 310.11 | 275.03 | 308.36 |
| Mean intron length (bp) | 270.41 | 606.69 | 632.38 |
| Masked repeat sequence length (Mb) | 558.49 | 546.95 | 507.14 |
| Repeats percentage of genome size (%) | 69.89 | 69.90 | 61.24 |

The conserved genes from the Benchmarking Universal Single-Copy Orthologs (BUSCO) gene set[25] were used to gauge the accuracy and completeness of the SLT1.0 assembly. The results showed that the SLT1.0 assembly contained 97.70% complete genes and 0.30% fragmented genes. The value of the LTR Assembly Index (LAI) was 12.41, which was consistent with that of the previously released SL4.0 tomato reference genome (LAI 12.54). More than 99.88% of the genome assembly had greater than one-fold coverage with Illumina short reads. All these evidences demonstrated the high continuity and completeness of the SLT1.0 genome assembly.

# High-quality genome annotation

Except for *ab initio* prediction and protein-homology-based prediction, we also used transcriptome data, including the bulked RNA-seq data with a mapping rate of 99.73%, and previously-released RNA-seq data from various tissues[3] with a mapping rate of 97.97%, to facilitate gene annotation of the assembled genome. In total, we predicted 34,384 protein-coding genes with an average length of 3,766.53 bp and 6.55 exons per gene in the SLT1.0 genome (Table 1 and Supplementary Table S2). Gene completeness was estimated to be 98.20% based on the BUSCO gene set[25], and the protein-coding genes were unevenly distributed along the chromosomes (Fig. 1). Comparative analysis showed that 234 genes in the SLT1.0 genome corresponded to 488 genes in the SL4.0 genome (Supplementary Table S3). Gene collinearity analysis identified 33 collinear gene blocks between the SLT1.0 and SL4.0 genomes, harboring 28,892 (84.03%) and 28,389 (83.30%) homologous genes, respectively (Fig. 2A). Some unplaced contigs in the SL4.0 genome were successfully assigned to chromosomes in the SLT1.0 genome. These results highlight the high accuracy and completeness of the SLT1.0 genome assembly and gene models.

A comprehensive analysis of the genome sequences identified 965 collinear chromosomal blocks between the SLT1.0 and SL4.0 genomes. These blocks contained 32,922 and 32,554 genes, accounting for 95.75% and 95.54% of the SLT1.0 and SL4.0 genomes, respectively. However, we detected a 2.76 Mb inversion from 39.17 to 41.93 Mb on chromosome 2 of the SLT1.0 genome (Fig. 2B). The continuous interaction signals on the Hi-C heatmap, as well as PCR and Sanger sequencing, showed that this region was not misassembled (Fig. 2B-C, Supplementary Table S4). This result indicated that heterozygous variation may exist in the previously reported Heinz 1706 accession.

# Transposable element analysis

A total of 524.84 Mb of repetitive sequences were identified, accounting for 65.66% of the SLT1.0 genome assembly, which was similar to that reported in the SL4.0 genome (508.89 Mb, 65.03%) (Supplementary Table S5). Among these repetitive sequences, long terminal repeats (LTRs) were the predominant TE family, covering 50.25% (401.60 Mb) of the genome. *Gypsy*-type LTRs (344.52 Mb) were the most common subfamily and six times more abundant than *Copia*-type LTRs (50.09 Mb). We used a combination of methods, including LTR-FINDER[26], LTR-Harvest[27], and LTR-Retriever[28], to identify intact LTRs. A total of 3,220 LTRs were detected in the SLT1.0 genome assembly, including 1,553 *Gypsy*-type LTRs and 1,346 *Copia*-type LTRs. The estimated insertion time of the LTR retrotransposons showed that *Gypsy* and *Copia*-type LTRs had a recent and similar burst 0.60-1.00 million years ago (Mya) (Fig. 3A), and were enriched far from coding genes (Fig. 3B). These results indicated that the burst of *Gypsy*-type LTRs may be the major driving force for the expansion of the tomato genome.

To identify the centromere regions, we detected the top 12 TE subfamilies, including 11 *Gypsy* and one unknown-type subfamilies, which together comprised over 15.47% of the genome (Fig. 3C). The density of these TE subfamilies along all the chromosomes showed that only the *Unknown*-type rnd-1_family-4 subfamily (1.65% of the genome) was enriched near centromeres but absent from the rest of the genome

(Fig. 3D, Supplementary Fig. S2). In addition, we found that 65.21% of the unanchored Contig/Scaffold sequence length comprised highly repetitive regions. Overall, we predicted 12 potential centromeric regions ranging from 1.90 to 6.90 Mb on the 12 chromosomes.

# Comparison of the SLT1.0 and *S. pimpinellifolium* LA2093 genomes

Structural variations (SVs) between wild and cultivated species can cause many phenotypic differences in domestication traits such as fruit weight and quality[29]. Based on protein homologies between the SLT1.0 and LA2093 genomes, we found that 23,544 genes (68.47%) in the SLT1.0 genome had one-to-one collinear relationships with 23,474 genes (65.64%) in the LA2093 genome (Fig. 4A). In addition, genome collinearity analysis showed that syntenic genomic blocks occupied 95.63% of the SLT1.0 genome and 96.67% of the LA2093 genome, respectively. We also identified 6,647 SVs (more than 1 kb in length) between the SLT1.0 and LA2093 genomes, including 3,054 (45.95%) SVs in 2,862 genes (Fig. 4B). GO analysis showed that these genes were significantly enriched in the function of oxidation-reduction process, photosynthetic electron transport chain and proton-transporting ATP synthase complex (Supplementary Fig. S3). We also identified 4,493,889 SNPs and 2,459,597 indels between the two genomes (Fig. 4B), including 418,844 SNPs and 245,310 indels located in 29,862 genes. We noted that 45,229 nonsynonymous SNPs resided in 18,178 genes and 9,148 frameshift indels in 1,559 genes, including 7,788 located in domestication regions[30]. They were significantly enriched in macromolecular complex, pigment metabolic process, nutrient reservoir activity, and intracellular organelle parts (Fig. 4C), suggesting these genes may have contributed to disease resistance and fruit traits during tomato domestication.

# High-density genetic map construction

A high-quality tomato reference genome can provide new insights into the genetic basis of important agronomic traits. Here, we constructed a RIL7 population of 247 progenies derived from a cross between the cultivated line OH88119 (64.27×) and the wild line PI128216 (42.35×). Resequencing of these progenies generated 1.7 Tb of data with an average depth of 6.91× and 97.98% coverage of the SLT1.0 genome. After aligning the reads to the genome, we identified 4,739,716 SNPs between the parental lines, 2,818,901 (59.35%) of which were genotyped in the RIL7 population. To construct a high-density tomato genetic map, we employed a Hidden Markov Model (HMM) to infer all recombination events in the RIL population. A total of 17,726 recombination events were detected across the whole genome, with one recombination event at an average interval of 3.78 kb (Fig. 5A). We found that 1,477 crossovers per chromosome ranged from 1.01 kb to 7.92 Mb in size and that the recombination rate varied across the genome with an average of 11.22 cM/Mb. Furthermore, recombination rate increased with distance from the centromeric regions on chromosomes 5 and 9, but this trend was not evident on the other chromosomes (Fig. 1).

We defined 35,836 raw bins (an average physical length of 22.06 kb) across the entire RIL population and constructed a high-quality set of 17,741 bin markers with a physical length of greater than 1 kb and a minor allele frequency of greater than 0.05. Based on recombination rate, we created a genetic linkage map of 13,973 unique bin markers anchoring twelve linkage groups that corresponded to the twelve chromosomes (Fig. 5B). This genetic map spanned 786.81 Mb, representing more than 98.46% of the assembled SLT1.0 reference genome. Taken together, these analyses produced a high-quality genetic map for genetic research on important agronomic traits in tomato.

# QTL mapping of important agronomic traits in tomato

The RIL population exhibited diversity for several important agronomic traits. The cultivated line OH88119 has a larger fruit and seed size, both of which are key traits associated with human selection in tomato. To gauge the accuracy of the genetic bin map, we performed an association study on three agronomic traits (seed size, leaf architecture, and trichomes) and identified five significantly associated QTLs in the SLT1.0 genome (Fig. 5C, Supplementary Fig. S4). Plant glandular trichomes can produce secondary metabolites that defend against herbivores and pathogens[31, 32]. Previous studies have shown that a 618-bp *H* gene encoding a C2H2 zinc finger protein regulates the formation of multicellular trichomes in tomato leaves[33]. Here, the leaf trichome locus was localized to a smaller interval of 2.82 kb that included two bins on chromosome 10 (Supplementary Fig. S4), suggesting the high quality of the genetic bin map.

The seed is an important plant reproductive organ, but larger seeds affect fruit taste, reducing the economic value of the berry crop. We used the genetic bin map to identify two seed size QTLs above the LOD threshold associated with seed length (Fig. 5C). These major-effect QTLs were resolved into a 1.32-Mb region with 40 bins ranging from 5.75 to 7.07 Mb on chromosome 4 (*sl4.1*) and a 0.74-Mb region with 36 bins ranging from 4.08 to 4.82 on chromosome 11 (*sl11.1*). Intriguingly, we found that *sl11.1* on chromosome 11 was located in domestication sweep DS169[30], indicating that it may have played an essential role during human selection. We identified 11,709 parental SNPs in this region, including 303 nonsynonymous variations in 87 genes. Among these SNPs, 132 nonsynonymous variations in 65 genes were also present in the RIL population. After exploring their functions, we identified six candidate genes for seed length, including a TIFY domain protein, a lipid droplet-associated protein (LDAP)-interacting protein (*LDIP*), and three small auxin-up RNA (SAUR) proteins (Fig. 5D). In particular, one gene encoding an *LDIP*, *SlT11G005670*, appeared to be a strong candidate for seed length. Based on phylogenetic analysis of homologous genes, we found that the *LDIP* gene influences lipid droplet size and neutral lipid homeostasis in *Arabidopsis* seeds[34] (Fig. 5E).

# Discussion

A highly contiguous and complete genome assembly is a powerful tool for molecular genetic studies of agronomic traits in tomato. In this study, we combined PacBio, BioNano, and Hi-C data to produce the

high-quality SLT1.0 tomato genome. The 799.09-Mb assembly had an N50 of 17.83 Mb, and more than 98.94% of its sequences were anchored to 12 chromosomes. The SLT1.0 genome had more repeats were sorted and anchored to chromosomes than the previously released SL4.0 genome. Analysis of repeat subfamilies showed that a specific subfamily, rnd-1_family-4, was found in centromeric regions of the SLT1.0 genome. We could not find a similar reliable repeat family in the SL4.0 genome. Comparative genome analysis revealed that a 2.76-Mb inversion was present on chromosome 2 in SLT1.0 relative to SL4.0 (Fig. 2). The inversion was validated by Sanger sequencing and contained no functional genes in adjacent breakpoints, suggesting it is a continuous fragment that has no effect on the SLT1.0 genome. However, we must be cautious and further verify these different fragments between the SLT1.0 and SL4.0 genomes.

Because of the small size and poor taste of wild species, fruit mass and quality are important domestication traits in tomato. Human domestication and improvement have increased the size of modern tomato fruits about 100 times relative to their ancestors[30]. At the same time, the flavor and resistance of tomatoes have been greatly reduced[35]. Comparative genomic analysis of the SLT1.0 and *S. pimpinellifolium* LA2093 genomes identified structural variations in some important genes that affect tomato fruit weight, such as *SlT11G021690* (oxidoreductase activity). This result suggests that structural variations in breeds were more likely to have been fixed in cultivars during the process of human selection. Furthermore, these structural variations may be used as potential targets for future breeding programs to improve fruit mass and quality.

In the process of crop domestication, human beings have paid particular attention to yield, shelf-life, and resistance to biotic stresses[7, 36]. It seems that people are inclined to select larger seeds to improve the emergence rate, yield, quality, and other important traits, especially in edible seed crops such as rice, wheat, and corn. However, size and number of seed affect the taste and reduce the quality of edible fruits such as tomato and cucumber. Here, QTL analysis of 247 tomato RIL populations indicated that the *SlT11G005670* gene may have influenced tomato seed length during domestication. By extracting the SNP sites of the parents, we found that the gene contained a nonsynonymous mutation in the cultivated tomato line. We also found that *LDIP* genes from other species have high homology with *SlT11G005670*. However, more evidence are required to understand the potential molecular mechanisms by which this gene controls seed length.

# Conclusion

Overall, we produced a high-quality tomato genome that will facilitate the molecular dissection of important agronomic traits in tomato. We generated a high-density genetic map and detected five QTLs related to seed and leaf traits. We also identified six candidate genes in two genomic regions that appear to control differences in seed length. This high-quality genome and high-density genetic map will be powerful tools for tomato breeding and can deepen our understanding of tomato biology.

# Methods

# Plant materials and sequencing

Plants were grown in the greenhouse in China Agricultural University in Beijing, with a 16 h light/ 8 h dark cycle. The cultivated line OH88119 was crossed with the wild line PI128216 to create the $F_1$ progeny, and RIL population was developed through single-seed-descent. Fresh leaf tissue were collected from each line and resequenced on the Illumina NovaSeq 6000 platform. A PacBio SMRT library was constructed and sequenced on the PacBio Sequel  platform. A Hi-C library was prepared following the Proximo Hi-C plant protocol with Hind as the restriction enzyme for chromatin digestion. The Hi-C libraries were sequenced on the Illumina NovaSeq platform with a read length of 150 bp. For optical mapping, high-molecular-weight DNA was isolated and labeled using a Bionano Saphyr System.

# De novo genome assembly

The raw SLT1.0 SMRT reads were corrected and assembled into sequence contigs using CANU with default parameters. The contigs were used for HERA assembly with the corrected SMRT reads. To identify sequence overlaps, all contigs and corrected reads were aligned all-against-all using Minimap2[37] and BWA[38] with default parameters. The HERA-assembled super-contigs were combined with BioNano genome maps to generate hybrid maps using IrysView software (BioNano Genomics) with a minimum length of 150 kb. The resulting contigs were further clustered basing on the Hi-C data using 3D-DNA software[39]. Pilon[24] was used for further error correction.

# Repeat analysis and gene annotation

The integrity of the final genome assembly was assessed in conjunction with BUSCO (v4.1.4)[25] using Benchmarking Universal Single-Copy Orthologs. A combination of *de novo* and homology-based methods was used to identify interspersed transposable elements (TEs). A *de novo* repeat library was built using RepeatModeler (v2.0.1)[40] and LTR_retriever (v2.9.0)[28]. Both the *de novo* library and RepBaseRepeatMaskerEdition-20181026, which is the most commonly used repetitive DNA element database, were used to identify TEs with RepeatMasker (v4.1.0)[41].

The RNA-Seq reads from this study were used to predict protein-coding genes in the repeat-masked SLT1.0 genome[3]. The cleaned high-quality RNA-Seq reads were aligned to the assembled genome using HISAT2[42] with default parameters, and the read alignments were assembled into transcripts using StringTie[43]. The complete coding sequences (CDS) were predicted from the assembled transcripts by the PASA pipeline[44]. The BRAKER[45], GeneMark-ET[46], and SNAP[47] softwares were performed on *ab initio* gene predictions. Finally, high-confidence gene models were predicted by integrating *ab initio* predictions, transcript mapping, and protein homology evidence with the MAKER pipeline[48].

# Genome comparisons and SV identification

Genome comparisons between SLT1.0 and SL4.0 and between SLT1.0 and LA2093 were performed via whole-genome alignment using the MUMmer package (v3.23)[49]. The one-to-one alignment blocks were identified using delta-filter program. Then the show-snp tools were used to identify SNPs and indels using uniquely aligned fragments, and the show-diff tool statistics were used to screen for structural variations over 1 kb in length. The SnpEff[50] software was used to analyze the various SNPs and indel types on the chromosomes.

# Genetic map construction and QTL analysis

We identified SNPs across the 247 $F_7$ RILs and the two parents using BWA (v0.7.10)[38] and samtools (v0.1.19)[51] softwares. The high-quality SNPs were called by bcftools (v1.10.2)[52], and SNPs were further filtered to retain only those with different homozygous genotypes in both parents, of which the quality $\geq 30$, MQ $\geq 30$, $2 \leq$ AF1 $\leq 200$. We generated a genotype matrix from the 247 RILs , and genetic distances were calculated using MSTMap[53]. The resulting bin marker data were imported into MG2C (v2.0) (http://mg2c.iask.in/mg2c_v2.0) to construct the genetic map.

QTL analysis was performed using trichomes, leaf type, and seed size phenotype data from 247 RIL population samples. The maximum likelihood estimation method was used to calculate the recombination rate and LOD values between bin markers. The bin markers with LOD value greater than 3.0 were selected as QTLs. The candidate genes were identified based on non-synonymous SNP mutations in both parents and their homologs in other species. We extracted the homologs from *Arabidopsis*, *Persea americana* and *Capsicum annuum* from NCBI (https://ncbi.org) and constructed an evolutionary tree using MEGAX[54].

## Declarations

# Ethics approval and consent to participate

The authors declare that this study complies with current laws of China.

# Consent for publication

Not applicable.

# Availability of data and materials

All the raw sequencing data for genome assembly and annotation have been deposited into the Genome Sequence Archive (GSA) database (http://bigd.big. ac.cn/gsa) in BIG Data Center under Accession Number PRJCA004585. Information for the assembled genome SLT1.0 was deposited both into the

Genome Warehouse (GWH) database (https://bigd.big.ac.cn/gwh/) in the BIG Data Center under Accession Number GWHBAUD00000000.

# Competing interests

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

# Funding

This work was supported by the 111 Project (B17043), the Beijing Municipal Education Commission Construction of Beijing Science and Technology Innovation and Service Capacity in Top Subjects (grant CEFF-PXM2019_014207_000032), the National Key Research and Development Program of China (2019YFD1000300) and the National Natural Science Foundation of China (32072571).

# Authors' Contributions

T. L., Y. Z. and W. Y. conceived and designed the project. G. M. collected the plant materials. X. S., B. W., Y. D. and Q. Y. performed all the data analysis under the supervision of T. L.. X. G and B. L. designed and conducted the experiments. S. H. supervised the project and Q. G. modified the article. All authors contributed and approved the final paper.

# Author information

[1]State Key Laborary of Agrobiotechnology, Beijing Key Laboratory of Growth and Developmental Regulation for Protected Vegetable Crops, College of Horticulture, China Agricultural University, Beijing, China. [2]Genomics and Genetic Engineering Laboratory of Ornamental Plants, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China. [3]State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, Kaifeng, China.

# References

1. Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A, Elkind Y, Levy A: **A new model system for tomato genetics**. *Plant J* 1997, **12**(6):1465-1472.
2. Giovannucci E: **Tomatoes, tomato-based products, lycopene, and cancer: Review of the epidemiologic literature**. *JNCI-J Natl Cancer Inst* 1999, **91**(4):317-331.
3. The Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**(7400):635-641.

4. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, Ma B, Qi M, Li Y, Zhao X *et al*: **Sequencing and de novo assembly of a near complete indica rice genome**. *Nat Commun* 2017, **8**.
5. Li Q, Li H, Huang W, Xu Y, Zhou Q, Wang S, Ruan J, Huang S, Zhang Z: **A chromosome-scale genome assembly of cucumber (Cucumis sativus L.)**. *GigaScience* 2019, **8**(6).
6. Hosmani PS, Flores Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez Perez G, Peters S *et al*: **An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps**. *bioRxiv* 2019:767764.
7. Doebley JF, Gaut BS, Smith BD: **The molecular genetics of crop domestication**. *Cell* 2006, **127**(7):1309-1321.
8. Chen Y, Song W, Xie X, Wang Z, Guan P, Peng H, Jiao Y, Ni Z, Sun Q, Guo W: **A Collinearity-Incorporating Homology Inference Strategy for Connecting Emerging Assemblies in the Triticeae Tribe as a Pilot Practice in the Plant Pangenomic Era**. *Molecular Plant* 2020, **13**(12):1694-1708.
9. Galindez G, Ortega Baes P, Seal CE, Daws MI, Scopel AL, Pritchard HW: **Physical seed dormancy in Collaea argentina (Fabaceae) and Abutilon pauciflorum (Malvaceae) after 4 years storage**. *Seed Science and Technology* 2010, **38**(3):777-782.
10. Fan CH, Xing YZ, Mao HL, Lu TT, Han B, Xu CG, Li XH, Zhang QF: **GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein**. *Theor Appl Genet* 2006, **112**(6):1164-1171.
11. Mao H, Sun S, Yao J, Wang C, Yu S, Xu C, Li X, Zhang Q: **Linking differential domain functions of the GS3 protein to natural variation of grain size in rice**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(45):19579-19584.
12. Huang X, Qian Q, Liu Z, Sun H, He S, Luo D, Xia G, Chu C, Li J, Fu X: **Natural variation at the DEP1 locus enhances grain yield in rice**. *Nat Genet* 2009, **41**(4):494-497.
13. Song XJ, Huang W, Shi M, Zhu MZ, Lin HX: **A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase**. *Nat Genet* 2007, **39**(5):623-630.
14. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, Zeng R, Zhu H, Dong G, Qian Q *et al*: **Control of grain size, shape and quality by OsSPL16 in rice**. *Nat Genet* 2012, **44**(8):950–954.
15. Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M: **Deletion in a gene associated with grain size increased yields during rice domestication**. *Nat Genet* 2008, **40**(8):1023-1028.
16. Zombori Z, Nagy B, Mihaly R, Pauk J, Cseri A, Sass L, Horvath V G, Dudits D: **RING-Type E3 Ubiqitin ligase barley genes (HvYrg1-2) control characteristics of both vegetative organs and deeds as yield components**. *Plants-Basel* 2020, **9**(12):1693.
17. Su Z, Hao C, Wang L, Dong Y, Zhang X: **Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (Triticum aestivum L.)**. *Theoretical and Applied Genetics* 2011, **122**(1):211-223.
18. Doganlar S, Frary A, Tanksley SD: **The genetic basis of seed-weight variation: tomato as a model system**. *Theor Appl Genet* 2000, **100**(8):1267-1273.

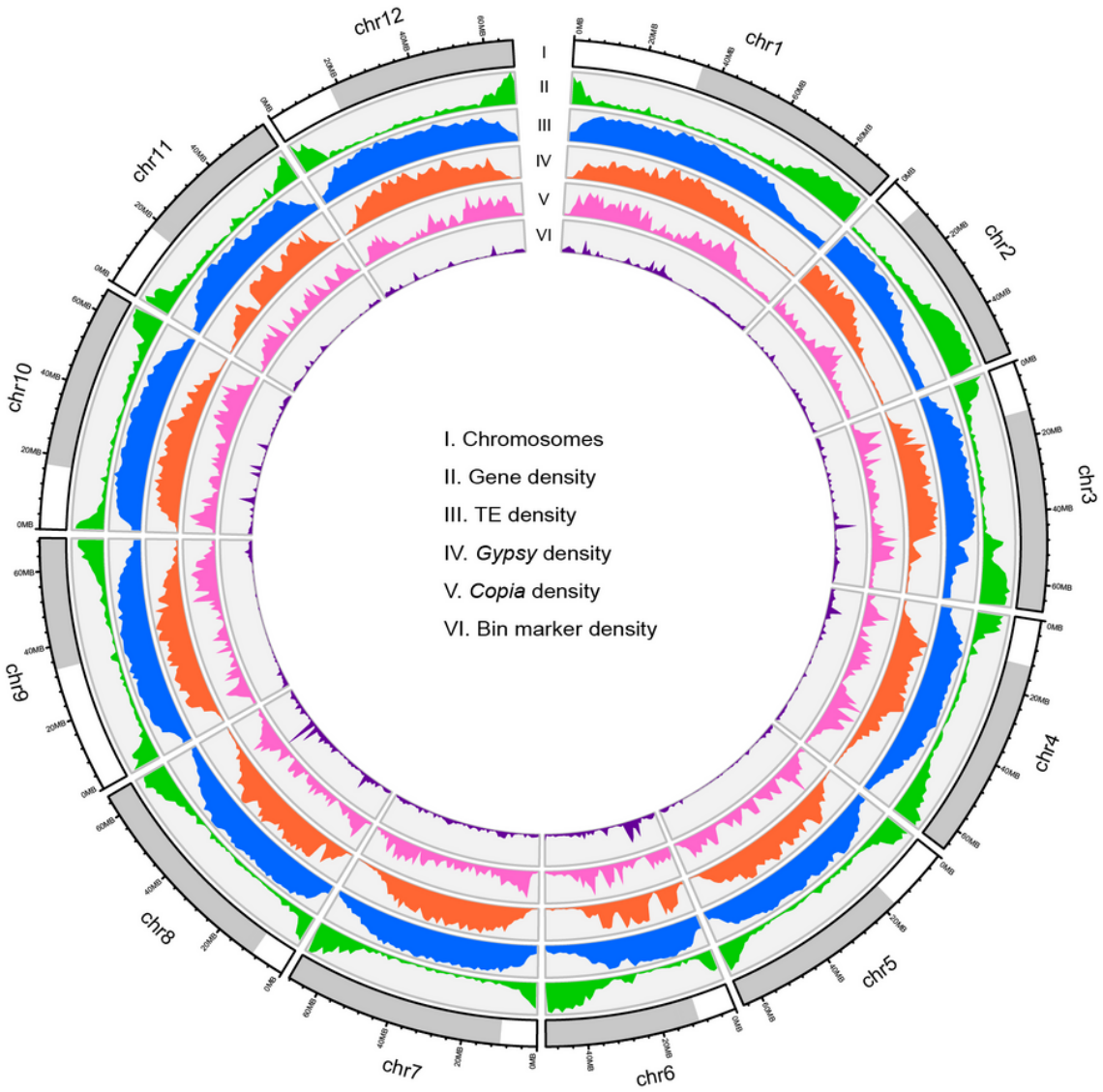
19. Orsi CH, Tanksley SD: **Natural variation in an ABC transporter gene associated with seed size evolution in tomato species**. *PLos Genet* 2009, **5**(1).
20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. *Genome Res* 2017, **27**(5):722-736.
21. Shelton JM, Coleman MC, Hemdon N, Lu N, Lam ET, Anantharaman T, Sheth P, Brown SJ: **Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool**. *BMC Genomics* 2015, **16**:734.
22. Du H, Liang C: **Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads**. *Nat Commun* 2019, **10**(1):5360.
23. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL: **Juicer orovides a one-click system for analyzing loop-resolution Hi-C experiments**. *Cell Systems* 2016, **3**(1):95-98.
24. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**(11).
25. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
26. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. *Nucleic Acids Res* 2007, **35**:265-268.
27. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons**. *BMC Bioinformatics* 2008, **9**(1):18.
28. Ou S, Jiang N: **LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons**. *Plant Physiology* 2018, **176**(2):1410-1422.
29. Jin L, Zhao L, Wang Y, Zhou R, Song L, Xu L, Cui X, Li R, Yu W, Zhao T: **Genetic diversity of 324 cultivated tomato germplasm resources using agronomic traits and InDel markers**. *Euphytica* 2019, **215**(4):69.
30. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X *et al*: **Genomic analyses provide insights into the history of tomato breeding**. *Nat Genet* 2014, **46**(11):1220-1226.
31. Aharoni A, Jongsma MA, Kim TY, Ri MB, Giri AP, Verstappen FWA, Schwab W, Bouwmeester HJ: **Metabolic engineering of terpenoid biosynthesis in plants**. *Phytochem Rev* 2006, **5**(1):49-58.
32. Kang JH, McRoberts J, Shi F, Moreno JE, Jones AD, Howe GA: **The flavonoid biosynthetic enzyme chalcone isomerase modulates terpenoid production in glandular trichomes of tomato**. *Plant Physiology* 2014, **164**(3):1161-1174.
33. Chang J, Yu T, Yang Q, Li C, Xiong C, Gao S, Xie Q, Zheng F, Li H, Tian Z *et al*: **Hair, encoding a single C2H2 zinc-finger protein, regulates multicellular trichome formation in tomato**. *Plant J* 2018, **96**(1):90-102.

34. Pyc M, Cai Y, Gidda SK, Yurchenko O, Park S, Kretschmar FK, Ischebeck T, Valerius O, Braus GH, Chapman KD: **Arabidopsis LDAP-interacting protein (LDIP) influences lipid droplet size and neutral lipid homeostasis in both leaves and seeds**. *Plant J* 2017, **92**(6):1182-1201.
35. Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, Zhang J, Mainiero S, Strickler SR, Catala C *et al*: **Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding**. *Nat Commun* 2020, **11**(1):5817.
36. Chen Q, Li W, Tan L, Tian F: **Harnessing knowledge from maize and rice domestication for new crop breeding**. *Molecular Plant* 2021, **14**(1):9-26.
37. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018(18):18.
38. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
39. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP *et al*: **De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds**. *Science* 2017, **356**(6333):92-95.
40. Bao, Z.: **Automated de novo identification of repeat sequence families in sequenced genomes**. *Genome Res* 2002, **12**(8):1269-1276.
41. Graovac MT, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences**. *Current Protocols in Bioinformatics* 2009, **25**(1).
42. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL: **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype**. *Nat Biotechnol* 2019, **37**(8):907–915
43. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nat Biotechnol* 2015, **33**(3):290–295.
44. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD *et al*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Res* 2003, **31**(19):5654-5666.
45. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M: **Whole-Genome Annotation with BRAKER**. In: *Gene Prediction: Methods and Protocols.* Edited by Kollmar M, vol. 1962; 2019: 65-95.
46. Alexandre L, Burns PD, Mark B: **Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm**. *Nucleic Acids Res* 2014(15):119-119.
47. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:9.
48. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M: **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes**. *Genome Res* 2008, **18**(1):188-196.
49. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biol* 2004, **5**(2):R12.
50. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff**. *Fly* 2012, **6**(2):80-

92.

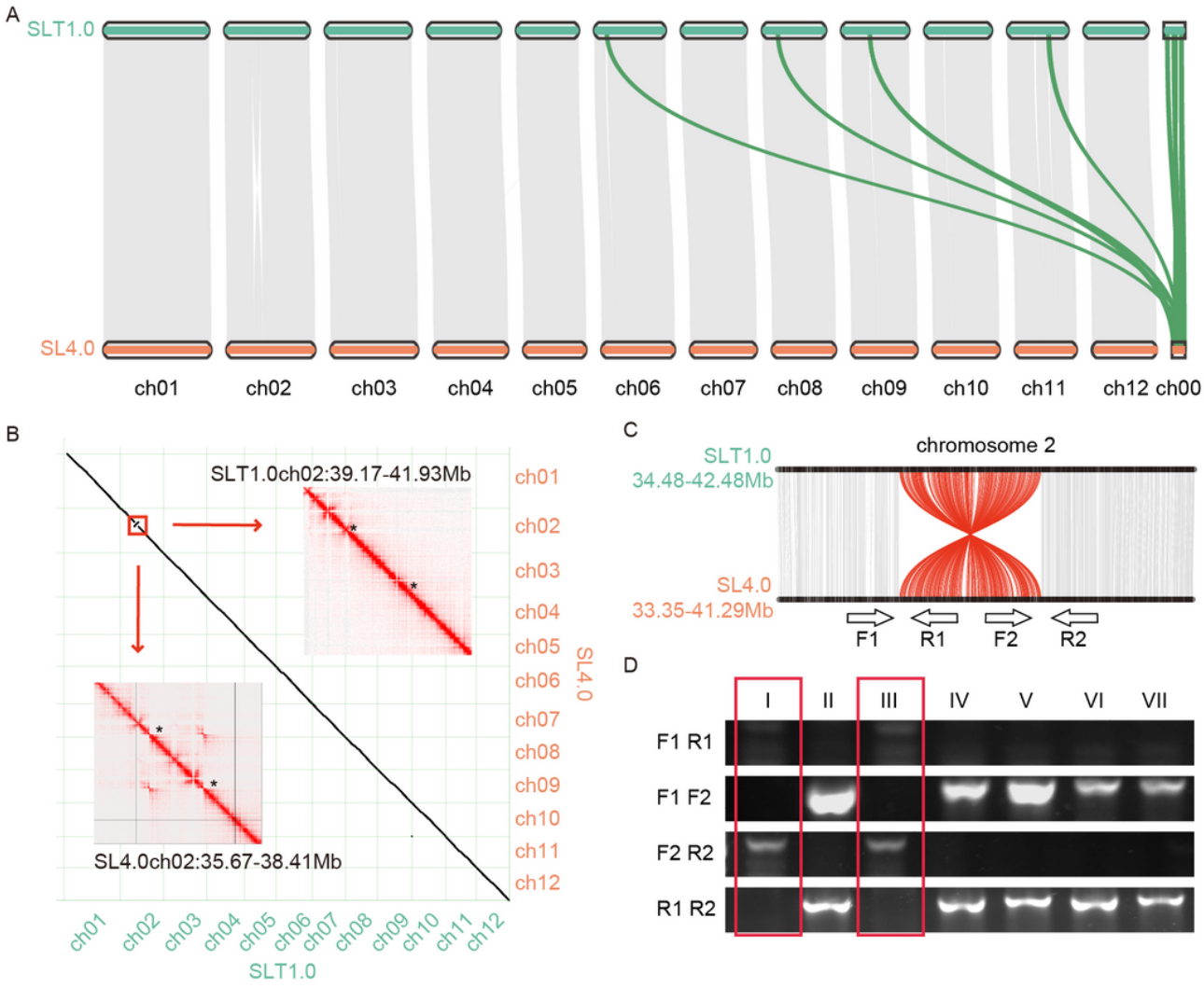
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
52. Vagheesh N, Petr D, Aylwyn S, Xue Y, Chris TS, Richard D: **BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data**. *Bioinformatics* 2016(11):1749-1751.
53. Wu Y, Bhat PR, Close TJ, Lonardi S, Kruglyak L: **Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph**. *PLos Genet* 2008, **4**(10):1000212.
54. Sudhir K, Glen S, Koichiro T: **MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets**. *Mol Biol Evol* 2016, **33**(7):1870-1874.

# Figures



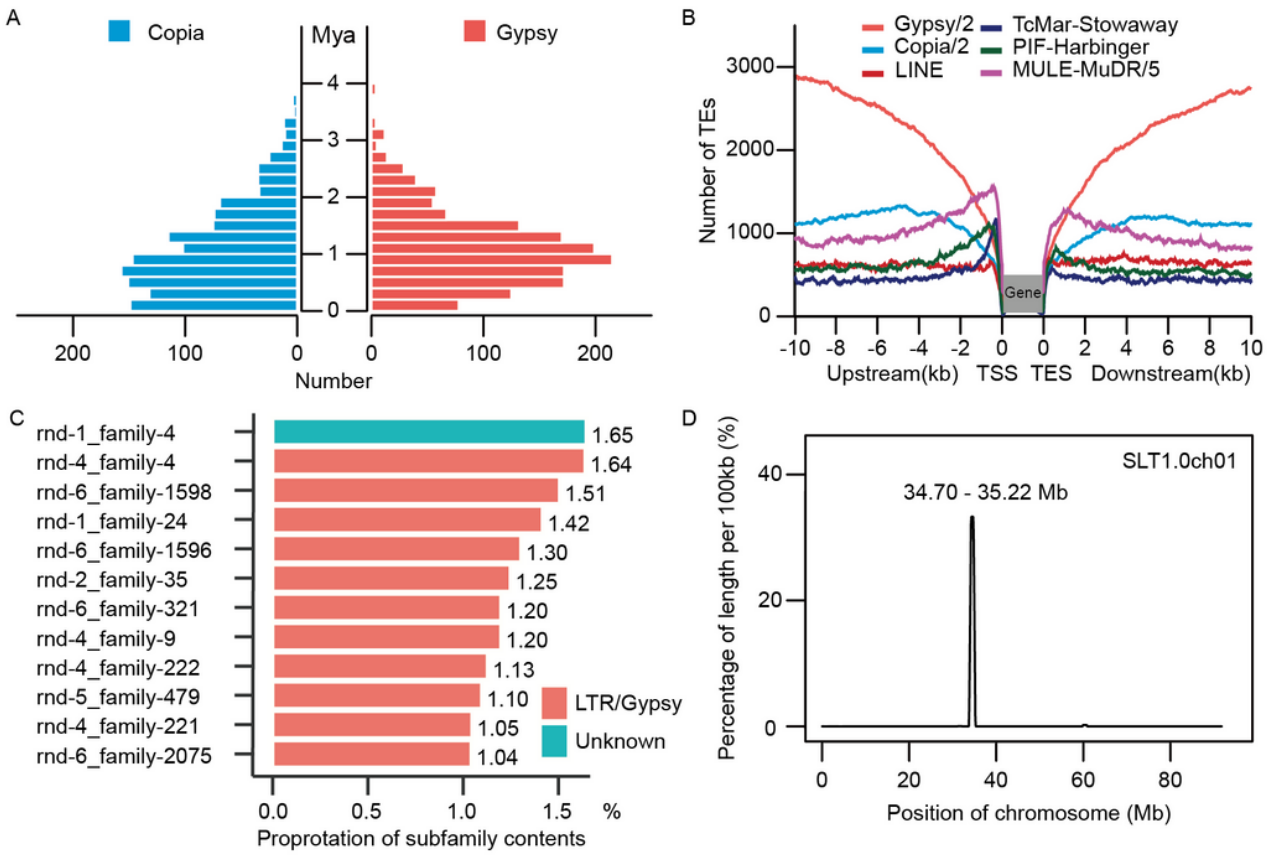

## Figure 1

Genomic landscape and structural variants of S. lycopersicum cv. Heinz 1706. (i) Ideogram of the 12 chromosomes with scale in Mb. (ii) Gene density (number of genes per Mb). (iii) Repeat content (% nucleotides per Mb). (iv) Gypsy content (% nucleotides per Mb). (v) Copia content (% nucleotides per Mb). (vi) Bin marker content (% nucleotides per Mb)



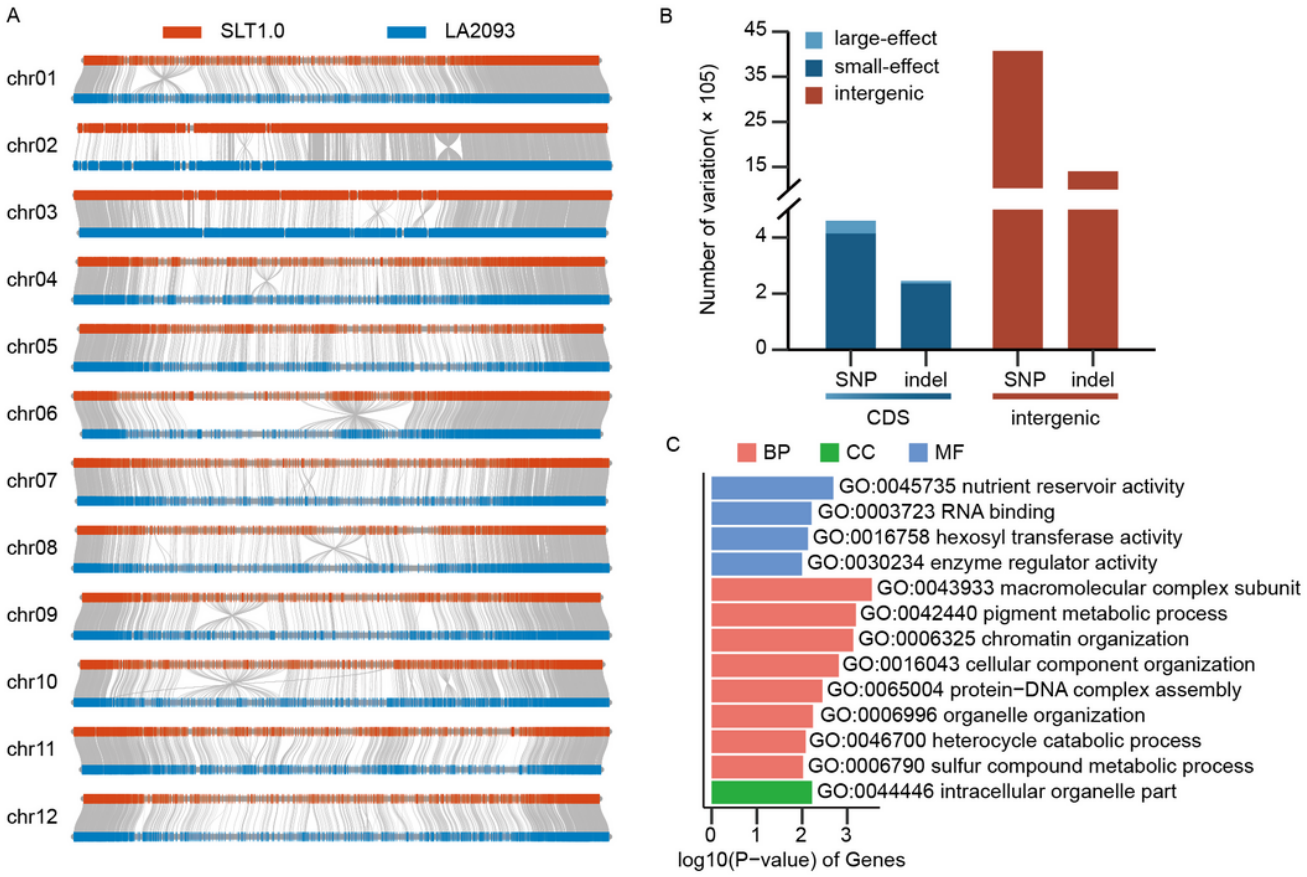

## Figure 2

Alignment between the Heinz 1706 SLT1.0 and SL4.0 genomes. A Genome collinearity analysis showed that four scaffolds from SL4.0 are placed on chromosomes of the SLT1.0 genome and that there is an inversion on chromosome 2. B The color intensity of the Hi-C heatmap represents the number of links between two 25-kb windows. The presence of an inversion is supported by high-density contacts indicated by two asterisks in the Hi-C heatmap generated from SL4.0 Hi-C reads (lower left), whereas no corresponding contact is found in the SLT1.0 Hi-C heatmap (upper right). C The inversion shown in red on chromosome 2. F1, R1, F2, and R2 are primers around the break points. D Seven Heinz 1706 individuals were identified, two of which (I, III) had inversions



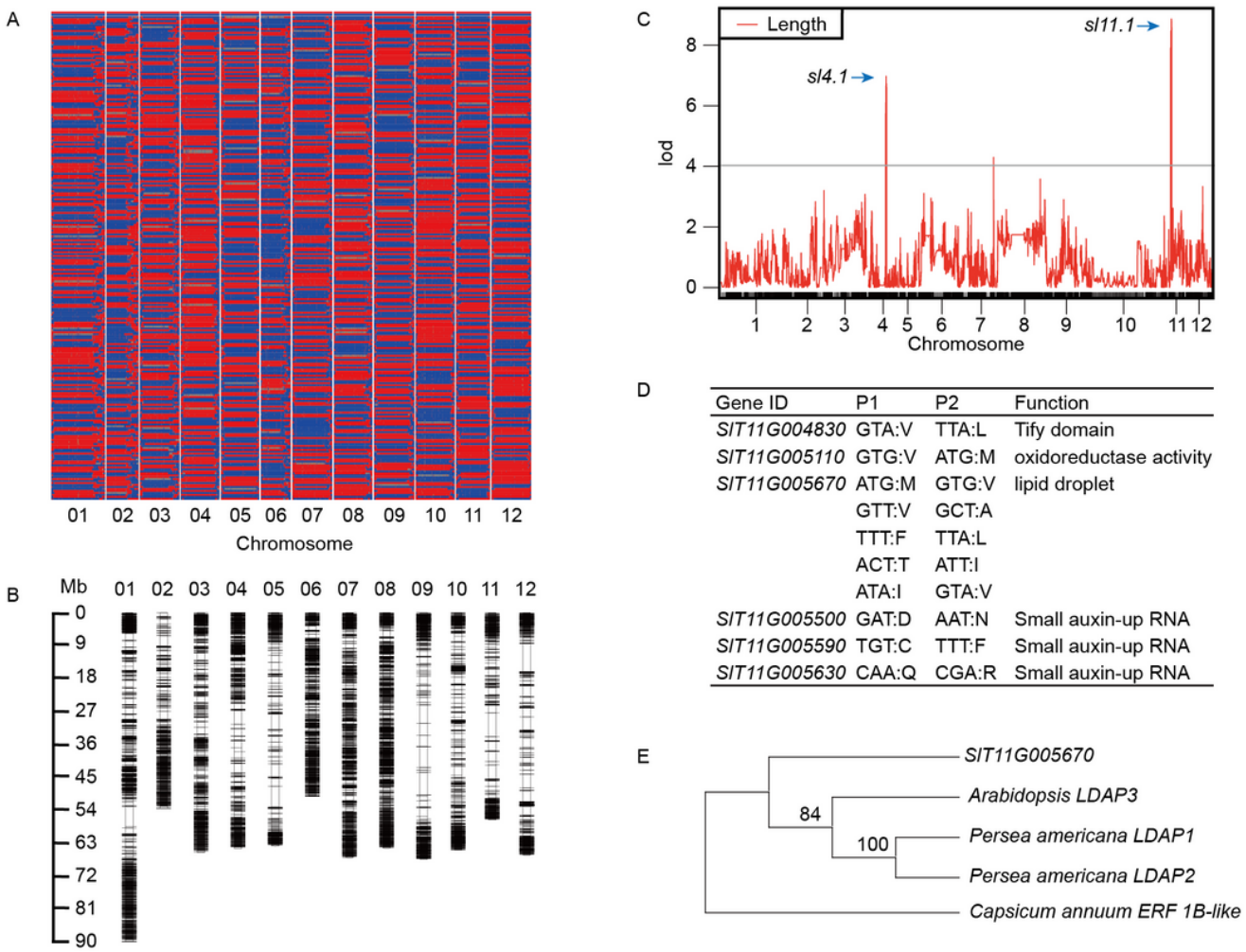

## Figure 3

Repetitive sequence analysis. A The estimated insertion time of LTR retrotransposons, showing Gypsy and Copia-type LTRs. B Frequencies of transposable elements (TE) in the vicinity of genes. C The top 12 TE subfamilies, including 11 Gypsy and one Unknown-type subfamily. D The Unknown-type rnd-1_family-4 subfamily was enriched towards the centromere of chromosome 1




## Figure 4

Alignment between the SLT1.0 and S. pimpinellifolium LA2093 genomes. A The red bar represents the SLT1.0 chromosome, and the blue bar represents the LA2093 chromosome. B Numbers of SNPs with nonsynonymous mutations (large-effect), SNPs with synonymous mutations (small-effect), and SNPs in intergenic regions, as well as the number of non-triple (large-effect) indels, triple (small-effect) indels, and indels in intergenic regions. C GO terms enriched in genes affected by SNPs and indels selected during domestication




| Gene ID | P1 | P2 | Function |
|---|---|---|---|
| *SlT11G004830* | GTA:V | TTA:L | Tify domain |
| *SlT11G005110* | GTG:V | ATG:M | oxidoreductase activity |
| *SlT11G005670* | ATG:M | GTG:V | lipid droplet |
| | GTT:V | GCT:A | |
| | TTT:F | TTA:L | |
| | ACT:T | ATT:I | |
| | ATA:I | GTA:V | |
| *SlT11G005500* | GAT:D | AAT:N | Small auxin-up RNA |
| *SlT11G005590* | TGT:C | TTT:F | Small auxin-up RNA |
| *SlT11G005630* | CAA:Q | CGA:R | Small auxin-up RNA |

**Figure 5**

Genetic map construction and QTL mapping of the RIL population. A Graphical representation of the re-sequencing-based mapping results of 251 RILs. B High-density genetic map. C QTLs for seed size (length). D Seed-size-related candidate genes. E Protein sequences of the SlT11G005670 gene and homologs from other species were used to construct an ML evolutionary tree

# Supplementary Files

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

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx

- TableS4.xlsx
- TableS5.xlsx
- FigureS1.png
- FigureS2.png
- FigureS3.png
- FigureS4.png