Data in Brief 55 (2024) 110567

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Whole genome sequence data of 516 F_2 plants of tomato (*Solanum lycopersicum*)

Tong Geon Lee

Horticultural Sciences Department, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Article history:
Received 18 April 2024
Revised 7 May 2024
Accepted 24 May 2024
Available online 29 May 2024

Dataset link: Genome sequence data of F2 populations of tomato (Original data)
Dataset link: F2 sequence note (Original data)

Keywords:
Genetic heterogeneity
Genetic recombination
Segregating population
Trait
Fresh-market tomato

ABSTRACT

The large-fruited fresh-market tomato cultivated in the U.S. represents a unique fruit market class of contemporary (modern) tomatoes for direct consumption. The genomes of F_2 plants from crosses between inbred contemporary U.S. large-fruited fresh-market tomatoes were sequenced. 516 F_2 individual plants randomly selected from five different bi-parental segregating populations were used for DNA extraction. The polymerase chain reaction (PCR)-free, paired-end (2×150 bp) sequencing libraries (350 bp DNA fragment length) were prepared, and sequenced on average 5 Gb for each plant using the Illumina next-generation sequencing technologies [1,2]. Raw Illumina reads with adapter contamination and/or uncertain nucleotides constitute (Ns, >10 % of either read; Q-score 5 or lower, >50 % of either read) were removed. This data article will contribute to improving our knowledge of the genetic recombination and variation in tomato.

E-mail addresses: tonggeonlee@ufl.edu, tonggeonlee@gmail.com
Social media: @realtonggeonlee

https://doi.org/10.1016/j.dib.2024.110567

Specifications Table

Subject	Genomics
Specific subject area	Applied Genetics, Bioinformatics, Breeding, Genetics, Horticulture
Type of data	Table, Raw, Analyzed, Filtered
Data collection	Illumina next-generation sequencing technologies (NovaSeq 6000, HiSeq)
Data source location	University of Florida, Gainesville, FL, USA
Data accessibility	Repository name: NCBI, figshare Data identification number: PRJNA1100794; 10.6084/m9.figshare.25556895 Direct URL to data: https://identifiers.org/ncbi/bioproject:PRJNA1100794; https://doi.org/10.6084/m9.figshare.25556895 Please see Table 1 and references [1–3] for the names and contents of FASTQ files.
Related research article	P. Bhandari, J.H. Kim, T.G. Lee, Genetic architecture of fresh-market tomato yield, BMC Plant Biol. 23 (2023) 18. https://doi.org/10.1186/s12870-022-04018-5. [2]

1. Value of the Data

- We report whole genome sequencing (WGS)-driven DNA sequence data of F_2 plants. The data can be an important resource to reveal diverse sequence variants including single nucleotide polymorphisms (SNPs) and structural variants (for example, [1-3]). Advances in bioinformatics technologies may uncover previously unidentified sequence variants and correct variants in error.
- Especially given the facts that segregating F_2 plants derived from crosses between inbred parental tomatoes were sequenced and these parental tomatoes were used to create released and commercialized F_1 hybrids [4,5], our data can be a rich DNA sequence resource for the discovery of genetic recombination/variations in the contemporary (modern) tomato (for example, intracultivar genetic heterogeneity found in soybean [6,7]).
- The data can be also used to sequence-resolve genomic sequence regions without molecular markers for both tomato genetics and breeding approaches because the genetic variation in this fresh-market tomato class has not been well-captured by relatively inexpensive genotyping platforms with the fixed number of polymorphic sites such as the tomato Illumina Infinium array [8] and its derivatives [e.g., Commercial Tomato PlexSeq™ SNP Panel (AgriPlex Genomics), Tomato Genotyping Library (LGC Biosearch Technologies)] (discussion in Bhandari and Lee [1] and Bhandari et al. [3]).

2. Background

The large-fruited fresh-market tomato (*Solanum lycopersicum*) (also called beefsteak tomato and round tomato in the U.S.) is a unique type of fresh-market tomato class for direct consumption [1] and it is selected for a high yield of large fruits (e.g., extra-large-sized fruit) to meet market demands [2]. WGS-driven DNA sequence data of biparental F_2 segregating populations had previously been used to construct the genetic map and linkage panel [1,3] and to study the genetic architecture of yields [2] for this economically important crop, but sequence datasets (i.e., FASTQ files created using the Illumina platform) of previous studies were not published. The objectives of this data article were to report the FASTQ files and a detailed description of data generation.

Table 1
Overview of data file/data set.

Name of data file/data set	File type (file extension)	Data content	Data repository
F_2 sequence note	PDF file (.pdf)	Name of FASTQ file to the corresponding population	figshare
Genome sequence data of F_2 populations of tomato	compressed FASTQ file (.fq.gz)	Quality controlled Illumina reads	NCBI

3. Data Description

FASTQ files were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (Table 1; Specifications table). The names of FASTQ files can be accessed on a figshare Dataset "F_2 sequence note" (Table 1; Specifications table). Populations A through D were used in Bhandari et al. [2], whereas population P was used in Bhandari and Lee [1] and Bhandari et al. [3].

4. Experimental Design, Materials and Methods

516 individual plants randomly selected from five different biparental F_2 segregating populations (progenies) (i.e., five crosses were made using 10 different inbred parental tomatoes; populations A, B, C, D, and P) were used for DNA extraction. For populations A through D, the leaf tissue was collected using the BioArk^{TM} Leaf collection kit (Product Code KBS-9370–001-L; LGC Biosearch Technologies) according to the manufacturer's instructions, and then the DNA extraction (GEN-9300–512) and quantification (GEN-9300–913) were performed. For population P, DNA was extracted using a DNeasy Plant Mini Kit (Qiagen). The PCR-free, paired-end (2 $\times$ 150 bp) sequencing libraries (350 bp DNA fragment length) were sequenced using the Illumina next-generation sequencing technologies (populations A through D and population P using the NovaSeq 6000 and the HiSeq, respectively). All raw Illumina reads were filtered using the quality control as follows: adapter contamination and/or uncertain nucleotides constitute (Ns, $>$10 % of either read; Q-score 5 or lower, $>$50 % of either read) were removed. Using the quality-controlled reads, we estimated genome coverage to be sequence-depth-of-coverage minimum 5 $\times$ on the SL4.0 reference genome assembly [9] for each plant. FASTQ files were examined on the University of Florida/Research Computing Linux server, HiPerGator 3.0 (https://www.rc.ufl.edu/about/hipergator) before data submission, and can be freely and openly accessed on NCBI BioProject PRJNA1100794.

Limitations

Although the current genome sequence data was generated using the PCR-free sequencing library, some sequence data might have gaps in coverage.

Ethics Statement

The author has read and follows the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

CRediT Author Statement

Tong Geon Lee: Writing.

Data Availability

Genome sequence data of F2 populations of tomato (Original data) (NCBI SRA)
F2 sequence note (Original data) (figshare).

Acknowledgements

The original research projects that created the data set were partially supported by the Florida Tomato Committee and the USDA National Institute of Food and Agriculture Hatch project [FLA-GCC-005550].

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Bhandari, T.G. Lee, A genetic map and linkage panel for the large-fruited fresh-market tomato, J. Amer. Soc. Hort. Sci. 146 (2021) 125–131, doi:10.21273/JASHS04999-20.
[2] P. Bhandari, J.H. Kim, T.G. Lee, Genetic architecture of fresh-market tomato yield, BMC Plant Biol. 23 (2023) 18, doi:10.1186/s12870-022-04018-5.
[3] P. Bhandari, R. Shekasteband, T.G. Lee, A consensus genetic map and linkage panel for fresh-market tomato, J. Amer. Soc. Hort. Sci. 147 (2022) 53–61, doi:10.21273/JASHS05110-21.
[4] J.W. Scott, E. Baldwin, H.J. Klee, J.K. Brecht, S.M. Olson, J.A. Bartz, C.A. Sims, Fla. 8153 hybrid tomato; Fla. 8059 and Fla. 7907 breeding lines, HortSci. 43 (2008) 2228–2230, doi:10.21273/HORTSCI.43.7.2228.
[5] J.W. Scott, S.M. Olson, H.H. Bryan, J.A. Bartz, D.N. Maynard, P.J. Stoffella, Solar Fire' hybrid tomato: Fla. 7776 tomato breeding line, HortSci 41 (2006) 1504–1505, doi:10.21273/HORTSCI.41.6.1504.
[6] W.J. Haun, D.L. Hyten, W.W. Xu, D.J. Gerhardt, T.J. Albert, T. Richmond, J.A. Jeddeloh, G. Jia, N.M. Springer, C.P. Vance, R.M. Stupar, The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82, Plant Physiol. 155 (2011) 645–655, doi:10.1104/pp.110.166736.
[7] T.G. Lee, B.W. Diers, M.E. Hudson, An efficient method for measuring copy number variation applied to improvement of nematode resistance in soybean, Plant J. 88 (2016) 143–153, doi:10.1111/tpj.13240.
[8] S.C. Sim, G. Durstewitz, J. Plieske, R. Wieseke, M.W. Ganal, A. Van Deynze, J.P. Hamilton, C.R. Buell, M. Causse, S. Wijeratne, D.M. Francis, Development of a large SNP genotyping array and generation of high-density genetic maps in tomato, PLoS ONE 7 (2012) e40563, doi:10.1371/journal.pone.0040563.
[9] N. Fernandez-Pozo, N. Menda, J.D. Edwards, S. Saha, I.Y. Tecle, S.R. Strickler, A. Bombarely, T. Fisher-York, A. Pujar, H. Foerster, A. Yan, L.A. Mueller, The sol genomics network (SGN)–from genotype to phenotype to breeding, Nucleic Acids Res. 43 (2015) D1036–D1041, doi:10.1093/nar/gku1195.