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.

Leveraging Structural Variants to Unlock Hidden Genetic Diversity for Tomato Crop Improvement

Reza Shekasteband

rshekas@ncsu.edu

North Carolina State University

Research Article

Keywords: Structural variants (SVs), Tomato genomics, Short-read sequencing, Marker-assisted breeding, Genetic diversity, Manta SV caller, Phylogenetic analysis

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9820447/v1

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

Additional Declarations: No competing interests reported.

Leveraging Structural Variants to Unlock Hidden Genetic Diversity for Tomato Crop Improvement

Reza Shekasteband[a, b, *]

[a] *Mountain Horticultural Crops Research and Extension Center, 455 Research Drive, Mills River, NC 28759, USA*

[b] *Department of Horticultural Science, North Carolina State University, Raleigh, NC 27695, USA*

[*]Correspondence: rshekas@ncsu.edu

ORCID iD: 0000-0001-8761-0650

Abstract

Structural variants (SVs)-large genomic alterations such as insertions, deletions, duplications, and translocations-are widespread in tomato genomes and play a critical role in phenotypic diversity. However, their detection has traditionally depended on expensive long-read sequencing technologies. Consistent with previous studies on structural variants, this study demonstrates a cost-effective approach for SV discovery using repurposed short-read sequencing data (150 bp), enabling integration of SVs into breeding workflows without additional sequencing investment. Using Illumina whole-genome data from 60 diverse tomato lines, including wild accessions, landraces, transgenic lines, and modern breeding lines, we identified over 71,000 high-confidence SVs, including a significant number of candidate private doubletons with the Manta SV caller, as well as 10.9 million short genetic variants. SVs were unevenly distributed across chromosomes, clustering in subtelomeric regions and co-localized with disease-resistance loci, with chromosomes 6, 7, and 9 showing the highest densities. Wild accessions harbored nearly twice as many SVs as cultivated lines, with deletions dominating wild genomes and insertions more prevalent in cultivated tomatoes than in the reference genome. Comparative phylogenetic analysis revealed strong concordance between SV-based and SNV/InDel-based trees (Baker's $\gamma = 0.95$), while SV data improved pedigree-consistent clustering and resolved ambiguous lineage relationships. These findings highlight SVs as hidden drivers of tomato diversity and valuable resources for marker-assisted selection, trait mapping, and genomic studies. Mining archived short-read datasets could offer breeding programs a scalable, low-cost strategy to unlock latent SV information, accelerate genetic improvement, and enhance genome-to-phenome insights.

Limitations include under-detection of complex rearrangements, warranting targeted validation for critical loci.

Keywords:

Structural variants (SVs), Tomato genomics, Short-read sequencing, Marker-assisted breeding, Genetic diversity, Manta SV caller, Phylogenetic analysis

1. Introduction

Tomato (*Solanum lycopersicum* L.), one of the most important vegetable crops in the world, has been scrutinized with next-generation sequencing (NGS) technologies and other genomic tools soon after the genome of the inbred processing tomato line, Heinz1706, was sequenced (Sato et al., 2012). In recent years, hundreds of wild, landrace, heirloom, and modern tomato genomes have been re-sequenced, and millions of short genetic variants (single-nucleotide variants, SNVs, and short insertions and deletions, InDels) were identified, underlining the vast genetic variation in the tomato clade within the genus Solanum (Gao et al., 2019). Ease of generating ample short genetic variation genotypic data from mapping populations and genome-wide association studies (GWAS) led to genetic mapping of numerous traits and disease resistance, QTL mapping, and consequently, functional gene discoveries in tomatoes (Pascual et al., 2016; Prasanna et al., 2023). Subsequently, breeding tomatoes for traits of interest and disease resistance has been accelerated significantly by exploiting DNA markers developed during the process (Prasanna et al., 2023).

Structural variants (SVs), which occur at a high number, are also an important part of the genome and have gained rapid research attention in the next-generation sequencing (NGS) era (Jobson & Roberts, 2022). Sizable variations in the genome's physical architecture, SVs, are common features that significantly impact an organism's phenotype. Any genomic variations larger than 50 base pairs fall into this category (Sudmant et al., 2015). Large deletions (DEL) and insertions (INS) are the predominant types of SVs, followed by duplication (DUP), breakends (BND), and other genome modifications (Alonge et al., 2020; Fuentes et al., 2019). Identification and precise physical mapping of SVs have been difficult due to their size and complex nature. Thus, modern genomic studies rely heavily on short genetic variants for population genetics, marker-trait association studies, GWAS, and QTL mapping. Recent advances in whole-genome resequencing technologies capable of capturing long reads, along with new computational tools, have enabled efficient SV discovery, opening a new frontier in genetic studies. For example,

Multiple studies show that SVs contribute significantly to human genomic variation (Huddleston et al., 2017; Pang et al., 2010). It has been shown that the impact of SVs on human population genetic variation is 3.4 times that of SNVs (Huddleston et al., 2017). The average genomic variation among individual human genomes published by 2010 due to SVs is estimated to be 15 times that due to SNVs, which are 1.5% and 0.1%, respectively (Pang et al., 2010). Significant structural variations within the plant genome frequently correlate with observable traits. For example, Fuentes et al. (2019) identified more than 63 million SVs in 3,000 rice genomes. Furthermore, this study found that long SVs are prevalent in promoter regions, whereas short SVs are prevalent in the 5′ UTR (Fuentes et al., 2019). Structural variants are present throughout the genome; however, studies show that they occur at high frequency in centromeric and subtelomeric regions, underscoring the need to map their physical locations (Turner et al., 2007). Thousands of SVs have been identified in tomatoes, emphasizing the importance of the considerable genomic variation in the genetic architecture of this vegetable crop (Alonge et al., 2020; N. Li et al., 2023; Liu et al., 2022). Additionally, multiple structural variants associated with plant growth habits, fruit size, flavor, disease resistance, and pathogen recognition have been identified in tomatoes (Gao et al., 2019; Jobson & Roberts, 2022; M. B. Lee et al., 2022; Wang et al., 2020). Present among modern tomato cultivars with low frequencies, SVs play a substantial role in the enormous genetic differences present between wild and modern tomato accessions (Alonge et al., 2020). Structural variants can extend our understanding of genome-to-phenome interactions, explaining the genetics behind phenotypes that are not possible to elucidate based on short genetic variants (Zhou et al., 2022). For example, in the human genome, structural variants affect 4,867 genes, of which around 24% would not be imputed by SNV-association (Pang et al., 2010).

It has been shown that using long reads, pangenomes, and super-pangenomes is the gold standard for SV calling and SV-based GWA studies (Alonge et al., 2020; Shi et al., 2026; Zhou et al., 2022). However, the substantial financial investment and specialized bioinformatic infrastructure required for these methods remain prohibitive for many research groups. Manta is widely regarded as one of the best-performing SV callers for short-read NGS data. It is known for its speed, low memory usage, and superior ability to detect insertions compared to other popular tools (Joe et al., 2024).To simplify the deployment of widespread SVs in tomato research, we examined the reliability and computational efficiency of the Manta SV caller for SV characterization and mapping, thereby bridging the gap between genomic discovery and practical breeding applications

at a small scale. This study used short genome reads (150 bp) generated by standard whole-genome sequencing to identify SVs with the Manta workflow, map them to the tomato genome, and assess their potential for marker-assisted tomato breeding and genomic studies. This study also demonstrates that structural variants contribute significantly to genetic diversity, making them a valuable resource for phylogenetic studies and highlighting the importance of SVs in the genetic architecture of modern tomatoes.

2. Materials and Methods:

2.1. Plant material

A core collection of 60 tomato lines, comprising wild accessions (3 lines), landraces (9 lines), specialty types (6 lines), transgenic lines (7 lines), and modern lines (35 lines), was used in this study. Compiled across multiple projects, the genome sequencing database is the result of 6 years of data collection, and the data were repurposed for this study. The wild tomato accessions and landrace tomatoes were obtained from the Tomato Genetics Resource Center (Davis, CA, USA) and the USDA's National Plant Germplasm System (NPGS). Modern and specialty-type tomato inbred lines have been developed and released by the tomato breeding program at North Carolina State University (NC State University, 2024). The tomato breeding program at NCSU is located at the Mountain Horticultural Crops Research and Extension Center (MHCREC) in Mills River, NC. In the last 5 decades, more than 70 elite tomato breeding lines and 35 F_1 fresh-market tomato cultivars have been developed, commercialized (e.g., Gardner, 1982; Gardner & Panthee, 2012), and transferred to public and commercial tomato research programs (NC State University, 2024). Two wild tomato accessions used in this study (LA1930 and LA1938) belong to *Solanum chilense* and have been used as sources of resistance to multiple viral diseases in cultivated tomatoes, including those in this study's core collection. The third wild tomato accession (LA2582) was selected for its rare ploidy level (autotetraploid) within the species *Solanum arcanum.* For the autotetraploid accession LA2582, variant calling was evaluated under a qualitative binary framework (presence/absence) rather than an allele-dosage model. On the other hand, the core cultivated tomato collection (57 lines) in this study encompasses a diverse range of horticultural characteristics, including fruit shape and size. The collection included 36 lines with large round fruits, 7 with small round fruits, 9 with plum (Roma-type) fruits, and 6 with grape-type fruits **(Supplementary Table 1)**.

2.2. Whole genome sequencing and short read alignment

Leaf samples were bulked from six 4-week-old seedlings for genomic DNA extraction. Genomic DNA was extracted with a DNeasy Plant Mini Kit (Qiagen, Germantown, MD). It was sent to a facility of Novogene Co., Inc. (Beijing, China) in California, USA, for sequencing library preparation and whole-genome sequencing. Standard Illumina DNA libraries were prepared with a 350 bp insert size and sequenced on an Illumina HiSeq instrument (Illumina, San Diego, CA), producing 150 bp paired-end reads and averaging 27 Gb of raw data per sample. The sequencing company trimmed raw sequencing reads to remove poor-quality bases and sequencing adapters, and to filter out reads that were too short, thereby improving the accuracy of downstream analyses such as alignment and variant detection. The alignment was run using a series of Bash commands to index the reference genome and map the paired-end reads. The published *Solanum lycopersicum* Heinz1706 reference genome [SL4.00, (Hosmani et al., 2019)]was indexed using Bowtie 2 [version 2.5.4 (Langmead & Salzberg, 2012)] under default parameters. Raw paired-end sequencing reads (FASTQ format) were subsequently aligned to the indexed reference genome. Alignment was performed using a local alignment strategy by invoking the `--local` parameter to accommodate soft clipping of non-matching read ends. The resulting sequence alignment stream was piped directly into SAMtools [version 1.9, (Li et al., 2009)], where reads were sorted by coordinate and written to a final binary alignment map (BAM) file. Lastly, coordinate-sorted BAM index files (`.bai`) were generated using SAMtools to facilitate fast random access for downstream variant analysis. The bioinformatic statistics for each sample's WGS data are presented in **Supplementary Table 1**.

2.3. Genetic variant calling

Short genetic variants (SNVs and short InDels) were identified individually in each sample using BCFtools [version 1.9, (Danecek et al., 2021)]. Putative variants were soft-filtered to remove sites with a quality score below 25. To ensure robust variant calling, individual SNV calls were restricted to sites with read depths between 10× and 100×, thereby removing low-coverage and hypervariable anomalies. To accommodate samples processed at different times, individual variant profiles were aggregated into a unified Variant Call Format (VCF) file using the `bcftools merge` utility. Downstream cohort curation, variant metric extraction, summary statistics

evaluation, and private doubleton classification were performed using the joint implementation of BCFtools and VCFtools [version 0.1.13, (Danecek et al., 2011)].

2.4. Structural variant calling

Structural variants (SVs), including deletions, duplications, insertions, and translocations, were identified using the Manta structural variant caller [version 1.6.0, (Chen et al., 2016)]. Manta was executed with default parameters to ensure a highly reproducible and standardized workflow, specifically configured to use BAM files generated by a local alignment strategy to achieve optimal breakpoint resolution. To balance sensitivity and specificity, downstream variant selection relied strictly on Manta's built-in quality control heuristics. Structural variants were retained only if they passed all internal quality filters, as indicated by a PASS status in the VCF filter column. This internal filtering guarantees high confidence calls by enforcing strict thresholds that can be found in Manta documentation. To preserve true positive variants, no additional downstream hard-filtering was applied to the PASS-validated SVs.

Manta can call SVs on individual or joint calls on a small set of samples. Joint calling is more desirable due to reduced workflow and downstream data merging; however, the number of samples that can be batched together for a single workflow is not well defined in Manta's documentation. The workflow quit running when more than 15 samples were batched in our experiment. Additionally, to examine the efficiency and sensitivity of SV calling, four independent rounds of the workflow were executed to call SVs from aligned short reads (150 bp) using the Manta structural variant caller with default settings. In the first round (R1), SV calling was performed on individual BAM files, generating a single VCF file per sample. Small sample sets were batched for the next three rounds (R2, R3, and R4) to run a single Manta workflow. Tomato lines with a similar pedigree, fruit shapes, and sizes were grouped to run a single round of SV calling (R2). During the third round (R3), lines were randomly assigned to a batch of 15 samples. The sample assignment for the fourth round (R4) was more complex. For this round, seven lines [NC 1 grape, NC 25P (plum line with early and late blight resistances), Mogeor (landrace inbred line with *Mi*, *Frl*, and *Tm-2a* genes and small round fruits), NC 123S (modern inbred line with *Sw-5b* and *I3* genes and large round fruits), NC 2GEM (modern inbred line with *Sw-5b* and *Tm-2a* genes and large round fruits), NC 233253 (modern inbred line with *Ty1* and extra-large round fruits), Kewalo (landrace inbred line with *Mi*, large round fruits)] were kept fixed in each batch. Then, eight

random samples were assigned to the batch to call and use SVs from. To call SVs from the three wild tomato accessions in this round, the seven fixed lines were batched only with the wild tomato accessions. Genomic variations greater than 50 base pairs were classified as SVs and categorized into four major categories. These included large deletions (DEL), insertions (INS), translocations (BND), and duplications (DUP). Chromosome translocations, BND-type SVs, were categorized into two subcategories based on the relocation of the segment within the same chromosome (intrachromosomal translocations, BNDa) or between nonhomologous chromosomes (interchromosomal translocations, BNDb). At the final step, the 'merge' option in BCFtools was used to merge individual VCF files, generating a single merged VCF file per round.

Line-specific structural variants (private doubletons) were identified using the `--singletons` module in VCFtools. Under this framework, private doubletons were defined as variants where the minor allele occurred exclusively in a single line in the homozygous alternative state (1/1, AC = 2), while all other lines were homozygous reference (0/0). To minimize short-read false positives in repetitive or low-mappability regions, candidate SVs were filtered to retain only high-confidence calls flagged as 'PASS' by Manta, with a minimum read-depth requirement (DP $\geq$ 5). Additionally, a representative subset of candidate line-specific SVs was visually verified using the JBrowse genome browser to confirm breakpoint alignment quality.

2.5. Genome Wide Variant Density and Statistical Correlation Analyses

To analyze the spatial distribution of genetic variation across the genome, each of the 12 assembled tomato chromosomes was binned into adjacent, non-overlapping 1-Mb genomic windows ($N = 778$ total windows across the genome). The density of structural variants (SVs) and short genetic variants (SNPs and InDels) within each 1-Mb window was quantified using BEDtools [version 2.31.1 9, (Quinlan & Hall, 2010)] and expressed as raw variant counts per 1-Mb window.

To evaluate spatial co-localization, non-parametric Spearman rank correlation analyses (ρ) were performed across all 1-Mb windows in R (v4.3.0). Correlations between SV density and SNP/InDel density were calculated separately for cultivated and wild tomato accessions. Additionally, to assess whether specific structural variant classes are concentrated in the same genomic regions, Spearman correlations were calculated between the density of each SV class (DEL, INS, BNDa, BNDb, and DUP) and the total SV density across all accessions. Two-tailed p-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg False

Discovery Rate (FDR) procedure ($\alpha = 0.05$). Additionally, the spatial distribution of all genetic variants was displayed in circular graphs using Circos software (Krzywinski et al., 2009).

2.6. Phylogenetic Inference and Hierarchical Clustering

Pairwise genetic distance matrices were calculated separately for short variants (SNVs and short InDels) and structural variants (SVs) across the 60 tomato accessions (57 cultivated *Solanum lycopersicum* lines and three wild *Solanum* spp. accessions) using the Identity-By-State (IBS) distance algorithm implemented in TASSEL v5.0 (Bradbury et al., 2007). Hierarchical clustering was performed directly on the resulting IBS distance matrices using the complete-linkage method to generate corresponding phylogenetic dendrograms for each variant class.

2.7. Cophylogenetic Tanglegram and Topological Congruence Analysis

To evaluate topological alignment between the short-variant and structural variant trees, a cophylogenetic tanglegram was constructed using the dendextend package (Galili, 2015) in R v4.3.1 (R Core Team, 2023). Major germplasm clades were annotated, and connector lines linking corresponding accessions were styled using distinct colors and line features to highlight topological similarities and localized branching discordances following the visual comparison framework described by De Vienne (2019). Topological congruence across the entire tree structure was quantitatively evaluated using the cophenetic correlation coefficient and Baker's gamma index (γ) (Baker, 1974; Galili, 2015). Additionally, the cophenetic correlation matrices of these dendrograms were calculated by the package "corrplot" [version 0.95, (Wei & Simko, 2024)]

2.8. Bootstrap Resampling and Permutation Testing

To evaluate the statistical confidence and stability of the tree-to-tree alignment without relying solely on single-node support metrics, a two-fold statistical framework was implemented in R ($N = 1{,}000$ iterations each):

2.8.1. Bootstrap Resampling for Tree Stability: Variant callsets were randomly resampled with replacement ($N = 1{,}000$ replicates) using dendextend (Galili, 2015) to generate an empirical bootstrap distribution of Baker's gamma index (γ). From this distribution, the 2.5^{th} and 97.5^{th} percentiles were extracted to define the 95% bootstrap confidence interval (95% CI).

2.8.2. Permutation Testing under Null Hypothesis (H_0): To test whether the observed topological agreement was significantly greater than expected by chance, a permutation test ($N =$

1,000 iterations) was conducted under H_0 (topological independence) by randomly shuffling accession labels across tree tips. A null distribution of correlation metrics was constructed, and the empirical p-value was calculated as the probability of obtaining a random correlation equal to or greater than the observed Baker's gamma index ($\gamma = 0.951$).

3. Results:

3.1. Short genetic Variants (SNVs and short InDels)

Over the course of 6 years, a whole-genome sequencing database of diverse tomato lines was established at NCSU and used in this study. Using Illumina next-generation sequencing (NGS) technology, 150-bp paired-end reads with an average of 26.81 Gb of raw data (35× sequencing depth) per sample were generated. With 97.96% of reads mapping primarily to the tomato reference genome, 179 million high-quality reads were generated per sample. The detailed statistics of the aligned reads for each sample are presented in the **Supplementary Table 1**.

In the cultivated tomato collection, short genetic variants were called for each sample individually and then merged into a single file, resulting in more than 10.9 million (10,953,613) variants. Of these, 10,886,815 genetic sites (10,044,343 SNVs and 842,472 short InDels) were polymorphic among the lines in the collection **(Supplementary Table 2a)**. In contrast to the reference genome, 66,798 monomorphic genetic variants (59,379 homozygous and 7,419 heterozygous) were present in all 57 cultivated tomato lines in the collection. Among the identified polymorphic genetic variants, 5,674,878 sites (5,200,573 SNVs and 474,305 short InDels) showed homozygosity across the lines. A high percentage of polymorphic genetic variants (47.9%) existed at the heterozygous stage in at least one and a maximum of 56 lines of the collection. In contrast, heterozygous monomorphic genetic variants accounted for a small percentage of the variation (0.07%), at only 7,419 genetic sites.

We also aligned sequencing reads from three wild tomato accessions to the reference genome (Heinz1706) to identify genetic variants. We identified 46,414,599 genetic variants (44,110,425 SNVs and 2,304,174 short InDels). Of those, 40,317,231 genetic sites showed polymorphism among the three wild tomato accessions, with 22,065,066 (20,477,542 SNVs and 1,587,524 short InDels) sites being homozygous and 18,252,165 (17,595,414 SNVs and 656,751 InDels) sites being heterozygous. Approximately 45.3% of the polymorphic genetic variants were heterozygous in at least one wild tomato accession. There were 6,097,368 monomorphic genetic variants among

the three accessions compared to the reference genome. Monomorphic variants at 5,876,446 sites (5,818,629 SNVs and 57,817 short InDels) showed homozygosity in all three wild tomato accessions. Meanwhile, the heterozygous monomorphic variants were observed at 220,922 genetic sites (218,840 SNVs and 2,082 short InDels) across the three accessions **(Supplementary Table 2b).**

3.2. Manta Workflow Impact on SV Profiling

During structural variant (SV) discovery using Manta, the initial alignment of raw sequencing reads by Bowtie2 required a critical parameter adjustment. By default, Bowtie2 performs end-to-end read alignment; however, BAM files generated using this default setting failed to yield any downstream SV calls. Realigning the raw reads across all samples using the local alignment parameter (--local) proved crucial for SV identification in Manta. The local alignment strategy enables Bowtie2 to map reads that flank and partially overlap SV breakpoints, largely due to altered minimum alignment score thresholds required to validate read mappings. Under default end-to-end settings, these hybrid breakpoint-spanning reads are classified as unmapped and lack genomic mapping coordinates, which are evident on BAM files when visualized on genome browsers. Supplementary Figures 1A-B). We identified five major types of structural variants: deletions (DEL), insertions (INS), tandem duplications (DUP), and two Inversions reported as break ends (BNDa and BNDb). Four approaches (R1, R2, R3, and R4) were implemented, resulting in increased SV calls from R1 to R4 in the wild accessions **(Figure 1A)** and most of the cultivated tomato lines **(Figure 1B)**. A total of 78,427 SVs were called when individual samples were subjected to independent variant calling by Manta (R1). Of those, 56,150 sites were flagged with 'PASS' in the filter column, indicating that the variants met Manta's thresholds for quality and reliability. There were 80,633, 87,328, and 89,459 SVs called in the next three rounds, R2, R3, and R4; of these, 62,779, 68,768, and 71,262 SVs were designated high quality with the 'PASS' flag in the filter column, respectively **(Figure 1C**). Further data analysis revealed that 35,985 SVs were common across all four workflows; however, each approach generated unique SVs that the others failed to identify. There were 10,941 SVs uniquely identified during the R4 workflow. The second highest number was achieved during the R3 workflow (8,676 SVs). In a similar manner, SV calling from individual samples (R1) or closely related lines (R2) resulted in 4,761 and 4,936 SVs, respectively. There were also SVs that were called by two or three

approaches concurrently (**Supplementary Table 3)**. The SVs from the fourth round (R4) with the 'PASS' flag (71,262 SVs) were used for downstream data analysis and visualization in this study.

Furthermore, data analysis revealed that 25,325 and 45,937 SVs were present in the genomes of the cultivated and three wild tomato accessions, respectively. Of those, 2,494 SVs (2088 homozygous and 406 heterozygous) were common across all lines in the collection (wild tomato accessions and cultivated tomatoes), potentially due to a lack of major disease resistance genes, absence of introgressions from wild accessions in the Heinz1706, or sequencing errors in the reference genome. Thus, there were 22,832 SVs present only in the cultivated tomatoes but not in any of the three wild tomato accessions in this study **(Supplementary Tables 4A and 4B)**. Among SVs identified in the cultivated tomato collection, 25,301 SVs (16,785 homozygous and 8516 heterozygous) were polymorphic across 57 lines. There were only 24 monomorphic SVs among the 57 cultivated tomato lines, setting them apart from Heinz1706, the old processing tomato type used as the reference genome **(Supplementary Table 4A)**. Among the three wild tomato accessions, 41,405 polymorphic SVs were identified, of which 31,372 and 10,033 SVs were homozygous and heterozygous, respectively **(Supplementary Table 4B)**. The investigation of allele frequencies revealed that a significant proportion of the SVs were present only at the heterozygous stage, accounting for 33% (8,522 SVs) in cultivated tomatoes, compared with 22% (10,247 SVs) in wild tomato accessions. The presence of heterozygous SV calls within highly inbred lines should be interpreted cautiously. These calls are hypothesized to stem primarily from repetitive DNA regions across the genome, where short-read alignment artifacts, reference bias, and mapping ambiguity in complex sequence contexts can falsely manifest as heterozygous structural variation.

3.3. Statistics of SV data

Across the cultivated tomato collection, a total of 25,325 high-confidence structural variants (SVs; 'PASS' filter) were identified in at least one accession. Of these, 3,018 sites (12%) were candidate private doubletons uniquely restricted to a single line. The remaining SVs (22,307 SVs) were present in two or more lines in the collection, with some SVs being present in all 57 lines **(Supplementary Table 5)**. The number of SVs identified in the grape tomato breeding lines developed at the NC tomato breeding program (NC 1-6 grape lines) was very high. Among them, 'NC 4grape' with 5,280 homozygous SVs harbors the highest number of SVs, and 'NC 5grape'

carries the lowest number of SVs in grape lines (4,462 SVs). Nevertheless, SVs in the 'NC 5grape' lines were still higher compared to the rest of the cultivated tomatoes in the collection **(see also Figure 1B)**. Old open-pollinated cultivars and landrace tomatoes, such as 'Thessaloniki' and 'Pearl Harbor', had the fewest SVs in their genomes, 522 and 326, respectively **(see also Figure 1B)**. The distribution of SVs on tomato chromosomes differed from a uniform distribution, with an average of 32.26 SVs per Mb of the reference genome (SL4.00, 782.5 Mb). The number of SVs on chromosome 9 (68.51 Mb) and chromosome 6 (47.26 Mb) was almost double the average, with 4,050 and 3,278 SVs, respectively. The fewest SVs, 794, were observed on chromosome 8 (64.00 Mb), with an average of 12.4 SVs per Mb of chromosome span **(Figure 2A)**. Additionally, each chromosome harbored significant numbers of line-specific private doubleton SVs, each present in a single line; the number per chromosome ranged from 119 to 624. Chromosome 6 contains the most private doubleton SVs (624), accounting for 19% of the SVs on this chromosome. Chromosomes 11 and 8, with 119 and 120 counts, showed the fewest private doubleton appearances among all 12 chromosomes **(Figure 2A)**. Investigation of SV classes revealed that 12,647 sites harbored insertion-type SVs in cultivated tomato genomes, accounting for 50% of all SVs when aligned to the Heinz1706 genome; deletions accounted for 41% of SVs (DEL = 10,321 SVs). We also identified 2,004 and 353 genetic sites with BND (8%) and DUP (1%) type SVs in the collection, respectively **(Figure 2B and 2C).** In cultivated tomatoes, 682 BNDa were identified; the remaining BNDs (BNDb = 1,322 SVs) were relocated onto nonhomologous chromosomes **(see also Supplementary Table 4A)**.

In the three wild tomato accessions, 45,937 SVs were called with the 'PASS' flag using a single Manta workflow from the short read (150 bp) aligned to the tomato reference genome, of which 23,242 SVs (51%) were private doubletons, present only in one wild tomato accession. The remaining SVs (22,695) were common across two or all three wild accessions in the study **(Supplementary Table 6)**. The distribution of total SVs across the chromosomes of the three wild tomato accessions was high and more uniform (58.7 SVs per Mb of the reference genome), with values ranging from 50.8 to 77.7 SVs per Mb. The number of SVs on chromosome 6, with 3,671 SVs and 77.7 SVs per Mb, was 32.4% higher than the expected average. Similarly, chromosomes 2 and 3 had 3,761 and 4,614 SVs, respectively, and 70.3 SVs on average, 19.8% higher than the genome average. In contrast, the distributions of 3,390 and 3,170 SVs on chromosomes 12 and 5 were lower than the genome average, with 50.8 and 50.9 SVs per Mb, respectively. The number

of SVs on the remaining six chromosomes was close to the average number, ranging from 54 to 64.2 SVs per Mb **(Figure 2D)**. Each chromosome in the three wild tomato accessions exhibited a high density of private doubleton structural variants, ranging from 1,253 on chromosome 5 to pronounced peaks on chromosomes 3 and 4, with 2,375 and 2,342 doubleton SVs, respectively **(Figure 2D)**. The deletion-type SVs were dominant in the genomes of the three wild tomato accessions, accounting for 25,626 sites and 56% of the genome architecture. Insertion-type SVs were present at 16,766 genetic sites and were the second dominant type (37%) of SV in the genomes of wild tomato accessions. The other two SV classes, BND and DUP, were observed at 2,951 and 594 genetic sites, accounting for 6% and 1% of the total SVs, respectively **(Figure 2E and 2F)**. Further classification of BND-type SVs revealed 1,104 BNDa and 1,847 BNDb-type SVs across the three wild tomato accessions **(see also Supplementary Table 4B)**.

On a per-line basis, the highest counts of line-specific private doubleton SVs occurred in the three wild tomato accessions. Moreover, these counts varied significantly both among the wild accessions themselves and relative to cultivated tomato lines **(Figure 3A and 3B)**. The highest number of candidate private doubleton SVs, 16,386, was observed in LA2582 (4n), while LA1938 (2n) had the lowest among the three lines (2,944). Similarly, the LA1930 (2n) wild tomato accession shows 3,912 candidate private doubleton SVs when aligned to the reference genome **(Figure 3A)**. A modern, large-fruited NC breeding line, NC 233253, possessed 317 candidate private doubleton SVs, followed by four lines with more than 200 candidate private doubletons in their genome (NC 6 Grape, Mogeor, NC 3 Grape, and LA4442). About half of the lines contained four or fewer private doubleton SVs, with six lines exhibiting none **(Figure 3B).**

3.4 Genome-Wide Distribution and Correlation of Structural and Short Variants

The physical locations of the identified short genetic variants, total SVs, and SV classes were mapped across the tomato genome, revealing distinct patterns of distribution and density across chromosomes. On average, there were 37 SVs per 1 Mbp span of the cultivated tomato genome; however, the distribution of SVs across chromosomes is highly uneven and inconsistent, with multiple regions exhibiting significantly higher numbers of SVs **(Figure 4A)**. In general, high densities of SVs were observed in the distal and subtelomeric regions of the chromosomes. Nevertheless, low-density SV coverage was present on the proximal and centromeric regions of all the chromosomes. On chromosomes 4, 9, 11, and 12, SV density increases significantly when

moving towards both subtelomeric regions; however, the rest of the chromosomes show a different distribution pattern with SVs clustered towards one of the distal regions. The distal part of the long arms of chromosomes 1, 2, 3, 5, and 7 contains many SVs, and, in contrast, the distal part of the short arms of Chromosomes 6, 8, and 10 shows highly dense SV presence **(Figure 4A)**.

Three regions, two on chromosome 6 and one on chromosome 7, exhibited the highest density of SVs in the genome of cultivated tomatoes. Nevertheless, there is a high density of SVs on the entire span of chromosome 6 (3278 SVs total), with 1128 SVs packed together in the very first 7 Mbp of the chromosome and 839 SVs clustered 20 Mbp away on another 7 Mbp span located at 20-27Mbp of the physical map. Chromosome 7, in contrast, harbors very low SV density compared to other chromosomes, except for 1,045 SVs clustered at the end of the chromosome at 57-64 Mbp (7 MbP). Other than the high peaks evident on the SV density histogram of the chromosomes, the density of the SV coverage on the entire span of chromosomes 6 and 9 was significantly higher than that of any other chromosome on the cultivated tomato genome. In contrast, chromosomes 8 and 10 exhibit low SV distribution across their entire length **(Figure 4A).**

Except for chromosome 2, the distribution of SVs called from the alignment of sequencing reads from 3 wild tomato accessions onto the cultivated tomato genome showed uniform patterns across all 12 chromosomes. A high distribution of SVs across the entire chromosome span was flanked by high-density peaks at the two ends of the chromosomes, creating a pronounced U-shape on 11 chromosomes. Chromosome 2 shows a different distribution pattern, with the major SV distribution confined to the end of the long arm **(Figure 4B).**

To test whether structural variation accumulates non-randomly across the genome, variant densities were calculated across 778 adjacent, non-overlapping 1-Mb genomic windows. Structural variant (SV) density exhibited a strong, positive spatial correlation with short genetic variant (SNV/InDel) density across the genome in cultivated tomato accessions ($\rho = 0.758$, $p < 0.001$, FDR-adjusted; Figure 5). In the three wild tomato accessions, however, short genetic variants were uniformly distributed across all 12 chromosomes, resulting in no spatial correlation with the U-shaped chromosomal distribution of SVs ($\rho = 0.036$, $p < 0.001$, FDR-adjusted; Figure 5). Through both cultivated and wild tomato accessions, spatial density distributions for all four individual SV classes were positively correlated with total SV density across 1-Mb windows. Insertions (INSs) and deletions (DELs) exhibited the strongest co-localization with total SV

density in cultivated and wild accessions, respectively: INSs ($\rho = 0.95$ and 0.94, $p < 0.001$), DELs ($\rho = 0.94$ and 0.97, $p < 0.001$), DUPs ($\rho = 0.43$ and 0.30, $p < 0.001$), and BNDs ($\rho = 0.60$ and 0.40, $p < 0.001$) (Figure 5). These results confirm that while individual SV classes consistently concentrate within shared hypervariable genomic hotspots across both germplasm groups, short genetic variants co-localize with SV hotspots exclusively in cultivated tomatoes.

3.5. High Topological Congruence Between Phylogeny Trees

To evaluate whether structural variants (SVs) reflect the same evolutionary histories as conventional short genetic variants (SNVs and InDels), we constructed parallel identity-by-state (IBS) complete-linkage dendrograms across all 60 tomato accessions (57 cultivated *Solanum lycopersicum* lines and three wild *Solanum* spp. accessions) (**Figure 6A**). Cophylogenetic tanglegram analysis revealed a striking degree of global tree congruence between the two call sets, yielding an overall cophenetic correlation coefficient of 0.747 and a Baker's gamma index (γ) of 0.951 (**Figure 6A**). Major germplasm groupings showed high lineage stability, clustering into eight distinct SV clades (SV-C1 through SV-C8) and seven corresponding short-variant clades (SNV-C1 through SNV-C7). Connecting lines between paired accessions demonstrated broad topology preservation across primary breeding and wild groups, with localized line crossings highlighting minor branching discordances attributable to fine-scale introgression or SV-specific evolutionary dynamics.

To confirm that the observed tree agreement was statistically robust and not an artifact of marker selection, we performed empirical bootstrap resampling ($N = 1{,}000$ replicates) (**Figure 6B**). The resulting bootstrap distribution of Baker's gamma index produced a narrow 95% confidence interval ($95\%\ \mathrm{CI} = [0.774, 1.000]$), demonstrating that the high topological alignment ($\gamma = 0.951$) remains stable across marker sub-samplings. Furthermore, permutation testing ($N = 1{,}000$ iterations) against the null hypothesis (H_0) of topological independence demonstrated that the observed correlation metrics fell far outside the randomized null distribution centered at zero ($p < 0.001$) (**Figure 6C**). Taken together, these results demonstrate that structural variant profiles strongly reflect the overall genome-wide evolutionary history established by short genetic variants, while providing diagnostic resolution for fine-scale sub-population discrimination.

The tomato lines in this study clustered into seven distinct clades in both phylogenetic trees; however, the clustering pattern on the SV-based phylogenetic tree was more consistent with the

known pedigree information of the tomato lines. A single line, 'NC 233253', formed a separate, eighth cluster (SV-C8) in the SV-based phylogenetic tree. One of the wild tomato accessions, LA2582, grouped with NC grape tomatoes in the short genetic variant phylogenetic tree, but clustered accurately with two other wild tomato accessions in the SV-based tree (SNV-C6 and SV-C7). While overall alignment efficiency was comparable across wild accessions, binary SV profiling captured a distinct set of structural variants that clearly resolved LA2582. The tomato lines within the remaining clades show minimal shuffling in two comparative dendrograms **(Figure 6A)**.

4. Discussion

4.1. A scalable path to SV-detection in tomato

Consistent with previous studies on structural variants in the genomes of tomatoes and other crops, this study demonstrates that the Manta SV caller can detect and map a substantial number of SVs using short-read sequencing (150 bp), with sufficient fidelity for downstream breeding decisions, significantly lowering barriers relative to long-read sequencing approaches in cost, turnaround time, and analytical complexity (Joe et al., 2024). It has been shown that by extracting SV information from existing whole-genome datasets, programs can convert sunk sequencing costs into new genetic insights that immediately inform trait-mapping, selection, and product-development pipelines (Zhou et al., 2022). The result is a practical, budget-conscious approach to integrating SVs alongside SNVs/InDels into routine marker-assisted workflows without new sequencing campaigns.

In this study, we successfully called SVs in the genomes of a collection of tomato lines using previously generated data from short-read sequencing. We also investigated their prevalence by SV type in different tomato lines, their distribution across chromosomes, and their importance in the tomato genome. The identification and study of short genetic variants have been the central focus of whole-genome sequencing with short-read data. The investigation of these short genetic variants was not the focus of this study; however, we explored the correlation between short genetic variants and SVs to estimate and validate the contribution of SVs to the genetic architecture and, consequently, to the genetic diversity of the tomato lines in this study. This finding will also encourage researchers to revisit and delve deeper into previously generated whole-genome sequencing data, potentially identifying the leading genetic cause of the phenotypes of interest in

gene discovery studies and improving the understanding of functional genome-phenome interactions, especially in plant breeding and gene discovery.

4.2. Workflow effect on SV calling

Beyond the biological implications, our study underscores a critical, often underestimated technical bottleneck in short-read SV discovery pipelines: the choice of read alignment strategy. Our analysis highlights a critical alignment adjustment required when coupling Bowtie2 with the Manta SV caller. The default end-to-end alignment strategy in Bowtie2 generated BAM files that yielded zero SV calls, as reads flanking or spanning SV breakpoints were penalized and marked as unmapped without coordinates. Transitioning to a local alignment strategy (`--local`) was essential, as it adjusts the minimum score thresholds used to determine valid read alignments. This allows Bowtie2 to retain hybrid reads that share partial sequence overlap with SV breakpoints, supplying Manta with the split-read mapping coordinates necessary to detect structural variants. According to the user manual, Manta can be used for SV calling of individual samples or for joint calling of a small set of samples; however, the number of samples that can be batched in a single workflow is not well documented. Joint calling was preferred due to increased workflow efficiency and subsequent data handling and aggregation. In this study, 15 samples were processed in a single Manta workflow without error, but processing more than 15 samples caused Manta to stop running without output, preventing us from calling SV from all 60 samples in a single workflow. Interestingly, joint calling led to a substantial increase in SV calls. Thus, to investigate the sensitivity of single versus joint SV calling with Manta, four different approaches (R1, R2, R3, and R4) were implemented, yielding more yet complementary SV discovery for most samples across R1-R4. In general, significantly more SVs were called when multiple samples were batched and subjected to a single Manta workflow. It is unclear to us why Manta calls more SVs when multiple samples are included in a single workflow; however, it appears that the number of SVs increases with greater sample diversity, as was evident in this experiment. Random numbers were generated and assigned to each of the 60 samples during the R3 workflow; meanwhile, an intentionally diverse set of seven samples was kept constant across workflows during the fourth round (R4), resulting in a large increase in general and unique SV calling. The outcomes from R4 were used for downstream analysis in this study because the Manta workflow conditions were similar across all samples; however, we believe that any of the approaches would have generated sufficient SVs for genotyping purposes. The outcomes from multiple runs can also be concatenated

and used after removing redundant SVs across multiple workflows if a larger number of SVs is needed. Additionally, multiple SVs called can be diploid to increase the discovery efficiency because it has also been shown previously that an ensemble approach to calling SVs, using multiple SV callers, can also improve the chance of SV identification and precision without substantially decreasing sensitivity (Zarate et al., 2021)

4.3. Prevalence of SV in the genome of tomato

A significant and satisfactory number of SVs were detected in the genomes of 60 tomato lines in the collection using short-read sequencing data. Compared with cultivated tomatoes, significantly more SVs were identified in wild tomato accessions; furthermore, LA2582, an autotetraploid wild tomato accession, has significantly more SVs than the other two wild accessions. The distinct variant landscape observed in LA2582 highlights both the substantial genetic diversity of Solanum arcanum and the practical value of SV profiling in wild germplasm. Because LA2582 is an artificially induced autotetraploid derived from a single diploid ancestor, its genomic loci remain fundamentally biallelic. Adopting a qualitative, binary framework enabled robust SV detection using standard diploid calling algorithms without introducing systemic caller artifacts, though this approach inherently treats varying allele dosage states (e.g., simplex vs. duplex) under a unified heterozygous classification. Furthermore, wild accessions like LA2582 harbor high levels of genetic diversity owing to their distinct evolutionary trajectories and historical underutilization in cultivated tomato breeding programs. While LA2582 maintains alignment performance comparable to other wild lines, its unique genetic background yields a distinct spectrum of structural variants. Our findings demonstrate that qualitative SV profiling is a robust approach for differentiating and characterizing distinct wild germplasm without being confounded by ploidy differences. The total number of SVs identified in this study was consistent with those reported in other studies (Alonge et al., 2020; Li et al., 2023; Liu et al., 2022). Alonge et al. (2020) identified 238,490 structural variants across 100 tomato genomes; however, they used long reads and included 100 tomato lines, with a greater proportion of wild tomato accessions. Deletion and insertion SV classes were slightly higher in wild tomato accessions and cultivated tomatoes, respectively, compared to the reference genome. This is evidence, but not proof, that the genome size of the fresh-market tomatoes in the collection is larger than that of the reference genome. Meanwhile, the reference genome is slightly larger than those of the wild tomato accessions used in this study.

Given that most of the cultivated breeding lines in this study are advanced modern inbred tomato lines, the presence of a significant number of SVs unique to each line and homozygous in that line (private doubletons) is noteworthy and warrants further exploration in integrated genome-wide studies for gene discovery and quantitative genetics. It was noteworthy that the presence of large numbers of private doubletons appears to be independent of fruit size and shape: NC 233253 and NC 9BS produce the largest fruit among the lines in the collection, yet NC 233253 has the highest number of doubletons, but no private doubletons were discovered in NC 9BS. The resistant allele of the Ty-1 gene with a large introgression linked to it on chromosome 6 is the only known and unique genetic characteristic of NC 233253. NC 9BS also carries two late blight resistance genes (Ph-2 and Ph-3 genes on chromosomes 10 and 9, respectively); however, multiple other lines in the collection carry one or both of the same genes, thus SVs co-localized with those two resistance genes were not categorized as private doubletons in NC 9BS or any other late blight resistance lines (NC 1CELBR, NC 161, NC 25P, NC 1GEM, NC 3GEM, NC7GEM, NC 9GEM, and NC 10GEM). Additionally, grape tomato lines and lines with nugget-sized fruits (Mogeor and LA4442) contain almost the same number of doubletons as large-fruited lines. While short-read SV callers are subject to false positives in structurally complex or repetitive regions, our filtering pipeline and JBrowes genome browser spot-checks indicate that candidate line-specific SVs (private doubletons) provide a useful proxy for capturing unique genetic diversity.

Different distribution patterns emerged when mapping the physical locations of SVs across tomato chromosomes. In the cultivated tomato collection, high-density SV regions co-localize with short genetic variant distributions, which in turn physically overlap with known disease-resistance loci and wild tomato introgressions. For instance, the concentration of SVs on chromosome 9 aligns closely with multiple disease-resistance genes, including *Frl* (Fusarium crown and root rot), *Tm-2a* (tomato and tobacco mosaic virus), *Ph-3* (late blight caused by *Phytophthora infestans*), and *Sw-5b* (tomato spotted wilt virus), with the *Tm-2a* locus spanning the largest introgression in lines carrying the resistant allele (Mogeor and NC 2GEM). Meanwhile, an extensive introgression overlapping the *Mi* and *Ty1* genes on chromosome 6 coincides with a high concentration of SVs on its short arm (Du et al., 2025; Foolad, 2007; Simko et al., 2021). Similarly, the pronounced SV peak on the distal portion of chromosome 7 coincides with the *I3* gene (Fusarium wilt, race 3) and is present exclusively in lines harboring the resistant allele (Chitwood-Brown et al., 2021). Beyond variations in fruit shape, size, and other horticultural traits, the cultivated tomatoes examined in

this study harbor the disease-resistance loci referenced above. DNA markers targeting SVs located within, within 1 kb of, or flanking these resistance genes have been designed and routinely used in HRM-PCR assays for marker-assisted selection in our tomato breeding program (unpublished data). Low SV density regions on chromosomes 8 and 10, as well as large segments on other chromosomes, might indicate genomic regions controlling fruit and horticultural characteristics of cultivated tomatoes, though these spatial patterns warrant further investigation. Structural variants scattered throughout the tomato genome have previously been linked to variation in plant growth habits, fruit size, flavor, disease resistance, and pathogen recognition (Gao et al., 2019; Jobson & Roberts, 2022; M. B. Lee et al., 2022; Wang et al., 2020). Overall, introgressions of disease resistance from wild relatives spatially correspond with the major high-density SV clusters in the tomato genome.

4.4. Cophylogenetic relationship

A cophylogenetic study using tanglegrams provided a visual and statistical comparison of rooted phylogenetic trees for two main genotyping approaches based on genetic variants. It has been shown that tanglegrams are a useful tool for comparing rooted phylogenetic trees (Scornavacca et al., 2011).However, it has been suggested that different colors and line styles should be used to highlight the differences and similarities between the trees compared (De Vienne, 2019). Additionally, Baker correlation coefficients of 0.95 indicate statistically similar clustering between phylogenetic trees generated from short genetic variants and from major SVs. While cophenetic correlation and Baker's gamma index demonstrate high topological similarity between short-variant and SV-based trees, these metrics evaluate congruence rather than absolute phylogenetic accuracy. Nevertheless, evaluated against known pedigree backgrounds (e.g., LA2582 and NC233252), SV-derived relationships effectively recapitulate expected lineage dynamics. Overall, these findings indicate that SV profiling provides valuable, complementary phylogenetic information that enriches standard short-variant-based trees.

Integrating the SV-derived distance matrix with short genetic variant genotypes can potentially tighten phylogenetic clustering and improve pedigree-consistent grouping, increasing confidence in selection and crossing decisions. The tanglegram analysis, supported by high Baker's gamma and strong cophenetic correlation, demonstrates that SV data enhances discriminatory power, uncovering genomic distinctions often missed by SNVs, such as correctly grouping a wild

accession with other wild types and isolating a breeding line outside the program pedigree. For commercial breeding teams, this improved resolution enables more accurate segment tracking, reduces misclassification, and ensures greater predictability in advanced lines.

Evidently, the frequency and physical map of SVs in tomatoes are co-localized with and apparently clustered around the distribution of short genetic variants in the genome and can be used in combination with short variants to improve the mapping resolution of genomic studies. Additionally, structural variants are excellent targets for developing DNA markers using high-resolution melting (HRM) curve-based methods. Genotyping by HRM curve provides the most affordable genotyping platform for small breeding programs; however, DNA marker development based on SNVs is challenging because most cases show insufficient shift in melting temperature to discriminate heterozygosity (Lee et al., 2018).

4.5. Limitations

While short-read SV calling (Manta) performed well across workflows, it may under-detect complex rearrangements or transgene insertion architectures, as evidenced by missed events in transformed lines with complicated breakpoints. While our optimized local-alignment pipeline achieved high calling precision, with manual inspection in JBrowse confirming no observable false-positive calls across audited loci, certain limitations remain inherent to short-read structural variant discovery. Notably, visual inspection of alignment tracks revealed that a subset of clearly discernible SVs remained uncalled by automated workflows, even when integrating union call sets across all pipeline iterations (R1–R4). This indicates that while our filtering protocol prioritizes high specificity, short-read algorithms still suffer from a sensitivity ceiling, particularly around complex or repetitive genomic regions. Resolving these remaining, algorithmically elusive variants will likely require future long-read sequencing (e.g., PacBio or Oxford Nanopore) and pangenome approaches to achieve complete structural resolution across the tomato genome. In four transgenic tomato lines with a “Moneymaker” background and two related backcrossed lines, we confirmed nine unique SVs (private doubletons) that Manta flagged as potential transgene insertion sites. Yet, Manta missed additional insertion sites, likely due to the complexity of these regions in the host genome. We employed an alternative method to identify six additional insertion sites (unpublished data), illustrating both the strengths and limitations of the Manta-based approach. Additionally, batching improves discovery but does not guarantee capture of all SV

classes; breakpoint-only detection can leave multi-break rearrangements partially unresolved. To mitigate risk in product development, tiered validation (visual inspection, orthogonal assays, and targeted long-read confirmation for critical loci) should be embedded prior to large-scale deployment. It has been pointed out that Manta and other SV callers only detect breakpoints, whereas multiple breakpoints are characteristic of complex genomic rearrangements in an organism (Shale et al., 2020). This study leveraged short-read sequencing to identify widespread structural variants (SVs) and map their positions on the tomato genome, highlighting the need for deeper investigations into complex genomic rearrangements and the genetic architecture of cultivated and wild tomatoes in genome–phenome interaction research.

Conclusion

In conclusion, this study demonstrates that short-read structural variant (SV) discovery can be reliably unlocked by optimizing alignment protocols, specifically adopting local alignment strategies to preserve breakpoint-spanning reads. Our findings reveal that while SV phylogenies broadly mirror traditional short-variant evolutionary trajectories across tomato accessions, SVs capture distinct structural variation that SNV panels overlook. Integrating this scalable, cost-effective SV detection framework into tomato pangenomics and breeding pipelines will facilitate the discovery of hidden genetic diversity, refine phylogenetic resolution, and enhance the power of marker-trait association studies for crop improvement.

Acknowledgements

We thank Dr. Dorith Rotenberg and Anna Whitefield (Professors at the Entomology and Plant Pathology Department, NCSU) for supplying the transgenic lines and Dr. Randy Gardner (NC State University, Emeritus Professor) for providing seeds from NC GEM tomato breeding lines.

Funding

This research was supported by multiple sources. An internal startup grant from the Department of Horticultural Science at North Carolina State University provided initial funding for project development. Additional support for next-generation sequencing (NGS) data generation was provided through federal and state programs, including:

- USDA National Institute of Food and Agriculture (USDA-NIFA), Grant No. 2022-67013-37076
- US Department of Agriculture Agricultural Marketing Service (USDA/AMS) and North Carolina Department of Agriculture and Consumer Services (NCDACS), Grant Nos. 23SCBPNC1202-00 and AM200100XXXXG061
- California Tomato Research Institute (CTRI), Grant No. 2021 - 294

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

All supporting processed data required to evaluate and reproduce the findings of this study, including filtered SV variant calls, 1-Mb window variant density tables, and filtering criteria, are provided within the article and its Supplementary Information files. Custom R scripts used for window calculations and correlation analyses are available from the corresponding author upon reasonable request. Raw whole-genome sequencing (WGS) datasets remain restricted due to ongoing research and proprietary breeding constraints.

References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., … Lippman, Z. B. (2020). Major Impacts of Widespread Structural Variation on Gene

Expression and Crop Improvement in Tomato. *Cell*, *182*(1). https://doi.org/10.1016/j.cell.2020.05.021

Baker, F. B. (1974). Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors. In *Source: Journal of the American Statistical Association* (Vol. 69, Number 346).

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chitwood-Brown, J., Vallad, G. E., Lee, T. G., & Hutton, S. F. (2021). Characterization and elimination of linkage-drag associated with Fusarium wilt race 3 resistance genes. *Theoretical and Applied Genetics*, *134*(7), 2129–2140. https://doi.org/10.1007/s00122-021-03810-5

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

De Vienne, D. M. (2019). Tanglegrams Are Misleading for Visual Evaluation of Tree Congruence. *Molecular Biology and Evolution*, *36*(1), 174–176. https://doi.org/10.1093/molbev/msy196

Du, M., Sun, C., Deng, L., Zhou, M., Li, J., Du, Y., Ye, Z., Huang, S., Li, T., Yu, J., Li, C. B., & Li, C. (2025). Molecular breeding of tomato: Advances and challenges. In *Journal*

of Integrative Plant Biology (Vol. 67, Number 3, pp. 669–721). John Wiley and Sons Inc. https://doi.org/10.1111/jipb.13879

Foolad, M. R. (2007). Genome mapping and molecular breeding of tomato. *International Journal of Plant Genomics*, *2007*. https://doi.org/10.1155/2007/64358

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., Hoz, J. F. D. La, Mohiyuddin, M., Wing, R. A., Mcnally, K. L., Tatarinova, T., Grigoriev, A., Mauleon, R., & Alexandrov, N. (2019). Structural variants in 3000 rice genomes. In *Genome Research* (Vol. 29, Number 5, p. 870). https://doi.org/10.1101/gr.241240.118

Galili, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., Knaap, E. Van Der, Huang, S., Klee, H. J., … Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. In *Nature Genetics* (Vol. 51, Number 6, p. 1044). https://doi.org/10.1038/s41588-019-0410-2

Gardner, R. G. (1982). *NC50-7 Breeding Line, "Cherokee", and "Mountain Pride" Tomato*.

Gardner, R. G., & Panthee, D. R. (2012). Tomato Spotted Wilt Virus-resistant Fresh-market Tomato Breeding Lines: NC 58S, NC 123S, NC 127S, and NC 132S. In *HORTSCIENCE* (Vol. 47, Number 4). http://www.ces.ncsu.edu/

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S., Fei, Z., Giovannoni, J. J., Mueller, L. A., & Saha, S. (2019). *An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps*. https://doi.org/10.1101/767764

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2017). Discovery

and genotyping of structural variation from long-read haploid genome sequence data. In *Genome Research* (Vol. 27, Number 5). https://doi.org/10.1101/gr.214007.116

Jobson, E., & Roberts, R. (2022). Genomic structural variation in tomato and its role in plant immunity. In *Molecular Horticulture* (Vol. 2, Number 1). https://doi.org/10.1186/s43897-022-00029-w

Joe, S., Park, J. L., Kim, J., Kim, S., Park, J. H., Yeo, M. K., Lee, D., Yang, J. O., & Kim, S. Y. (2024). Comparison of structural variant callers for massive whole-genome sequence data. *BMC Genomics*, *25*(1). https://doi.org/10.1186/s12864-024-10239-9

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lee, M. B., Shekasteband, R., Hutton, S. F., & Lee, T. G. (2022). A mutant allele of the flowering promoting factor 1 gene at the tomato BRACHYTIC locus reduces plant height with high quality fruit. *Plant Direct*, *6*(8). https://doi.org/10.1002/pld3.422

Lee, T. G., Shekasteband, R., Menda, N., Mueller, L. A., & Hutton, S. F. (2018). Molecular markers to select for the j-2–mediated jointless pedicel in tomato. *HortScience*, *53*(2). https://doi.org/10.21273/HORTSCI12628-17

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., Yang, T., Tang, Y., Yang, S., Aisimutuola, P., Xu, R., Hu, J., Jia, C., Ma, K., Li, Z., Jiang, F., Gao, J., Lan, H., Zhou, Y., … Yu, Q. (2023). Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nature Genetics*, *55*(5). https://doi.org/10.1038/s41588-023-01340-y

Liu, L., Zhang, K., Bai, J., Lu, J., Lu, X., Hu, J., Pan, C., He, S., Yuan, J., Zhang, Y., Zhang, M., Guo, Y., Wang, X., Huang, Z., Du, Y., Cheng, F., & Li, J. (2022). All-flesh fruit in

tomato is controlled by reduced expression dosage of AFF through a structural variant mutation in the promoter. *Journal of Experimental Botany*, *73*(1). https://doi.org/10.1093/jxb/erab401

NC State University. (2024, April). *https://mountainhort.ces.ncsu.edu/fresh-market-tomato-breeding/*.

Pang, A. W., Macdonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., & Scherer, S. W. (2010). *Towards a comprehensive structural variation map of an individual human genome*.

Pascual, L., Albert, E., Sauvage, C., Duangjit, J., Bouchet, J. P., Bitton, F., Desplat, N., Brunel, D., Le Paslier, M. C., Ranc, N., Bruguier, L., Chauchard, B., Verschave, P., & Causse, M. (2016). Dissecting quantitative trait variation in the resequencing era: Complementarity of bi-parental, multi-parental and association panels. *Plant Science*, *242*, 120–130. https://doi.org/10.1016/j.plantsci.2015.06.017

Prasanna, H., Rai, N., Hussain, Z., Yerasu, S. R., & Tiwari, J. K. (2023). Tomato: Breeding and Genomics. *Vegetable Science*, *50*(Special). https://doi.org/10.61180/vegsci.2023.v50.spl.02

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C., Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., … Gianese, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

Scornavacca, C., Zickmann, F., & Huson, D. H. (2011). Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*, *27*(13). https://doi.org/10.1093/bioinformatics/btr210

Shale, C., Baber, J., Cameron, D. L., Wong, M., Cowley, M. J., Papenfuss, A. T., Cuppen, E., & Priestley, P. (2020). *Unscrambling cancer genomes via integrated analysis of structural variation and copy number*. https://doi.org/10.1101/2020.12.03.410860

Shi, C., Chen, S., Wang, J., Chen, W., Sun, C., Guo, Q., Liao, S., Wang, H., Mu, Y., Shu, X., Meng, D., Zhao, J., Dong, L., Zhao, L., Hou, S., Guo, L., & Yang, C. (2026). A tomato telomere-to-telomere super-pangenome empowers stress resilience breeding. *Nature Genetics*. https://doi.org/10.1038/s41588-026-02508-y

Simko, I., Jia, M., Venkatesh, J., Kang, B. C., Weng, Y., Barcaccia, G., Lanteri, S., Bhattarai, G., & Foolad, M. R. (2021). Genomics and Marker-Assisted Improvement of Vegetable Crops. *Critical Reviews in Plant Sciences*, *40*(4), 303–365. https://doi.org/10.1080/07352689.2021.1941605

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H. Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S., & Hurles, M. E. (2007). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. In *Nature Genetics* (Vol. 40, Number 1, p. 90). https://doi.org/10.1038/ng.2007.40

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., Zhang, J., Mainiero, S., Strickler, S. R., Catala, C., Martin, G. B., Mueller, L. A., Vrebalov, J., Giovannoni, J. J., Wu, S., & Fei, Z. (2020). Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. In *Nature Communications* (Vol. 11, Number 1). https://doi.org/10.1038/s41467-020-19682-0

Wei, T., & Simko, V. (2024). *R package "corrplot": Visualization of a Correlation Matrix (Version 0.95). Available from https://github.com/taiyun/corrplot*.

Zarate, S., Carroll, A., Mahmoud, M., Krasheninina, O., Jun, G., Salerno, W. J., Schatz, M. C., Boerwinkle, E., Gibbs, R. A., & Sedlazeck, F. J. (2021). Parliament2: Accurate

structural variant calling at scale. *GigaScience*, *9*(12). https://doi.org/10.1093/gigascience/giaa145

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., … Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914). https://doi.org/10.1038/s41586-022-04808-9

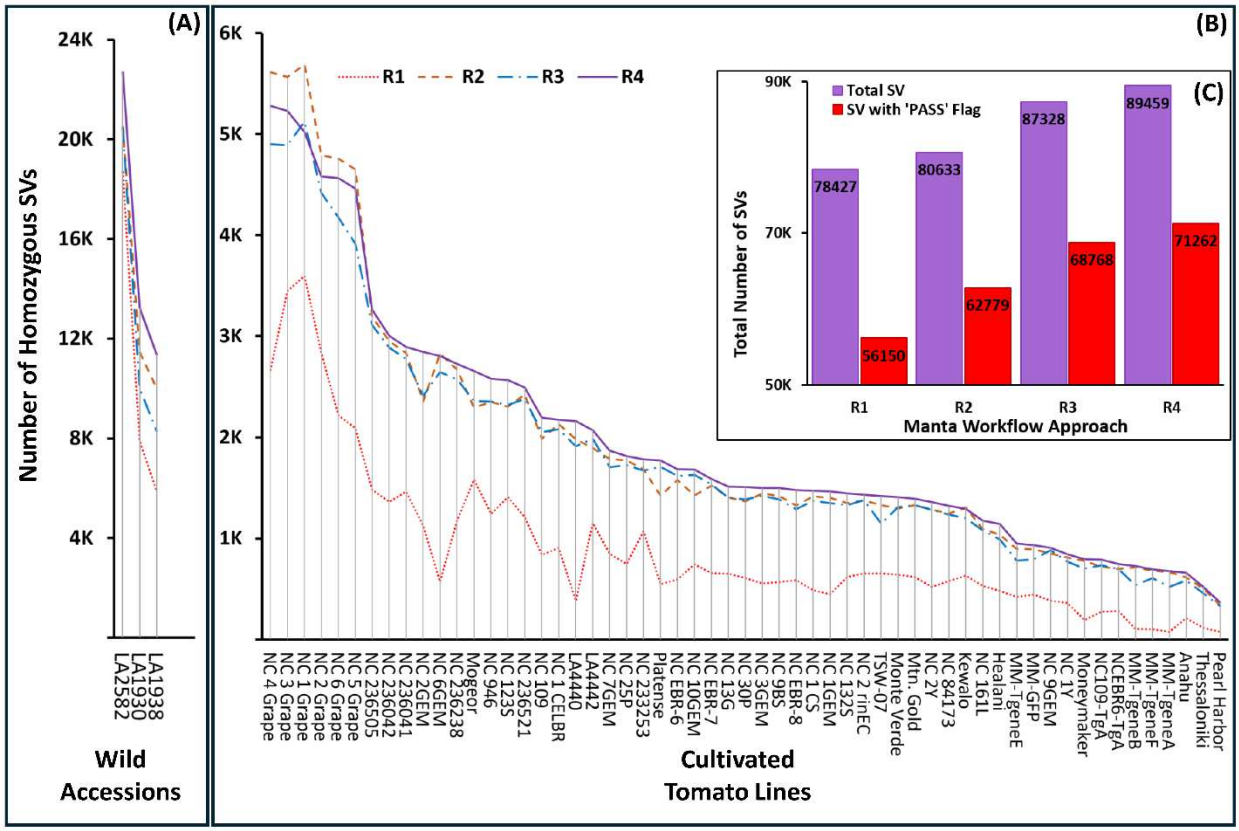


Figure 1. Evaluation of structural variant (SV) discovery yields across four Manta workflow strategies in 60 tomato accessions.

(A–B) Distribution of high-confidence ('PASS') homozygous SVs detected across (A) three wild tomato accessions and (B) 57 cultivated tomato lines using four workflow strategies: single-sample calling (R1) and 15-sample joint calling binned by pedigree (R2), randomized (R3), or genomic diversity (R4). (C) Comparison of total raw candidate SVs (purple) versus post-filtering high-confidence SVs (red) recovered across each workflow.

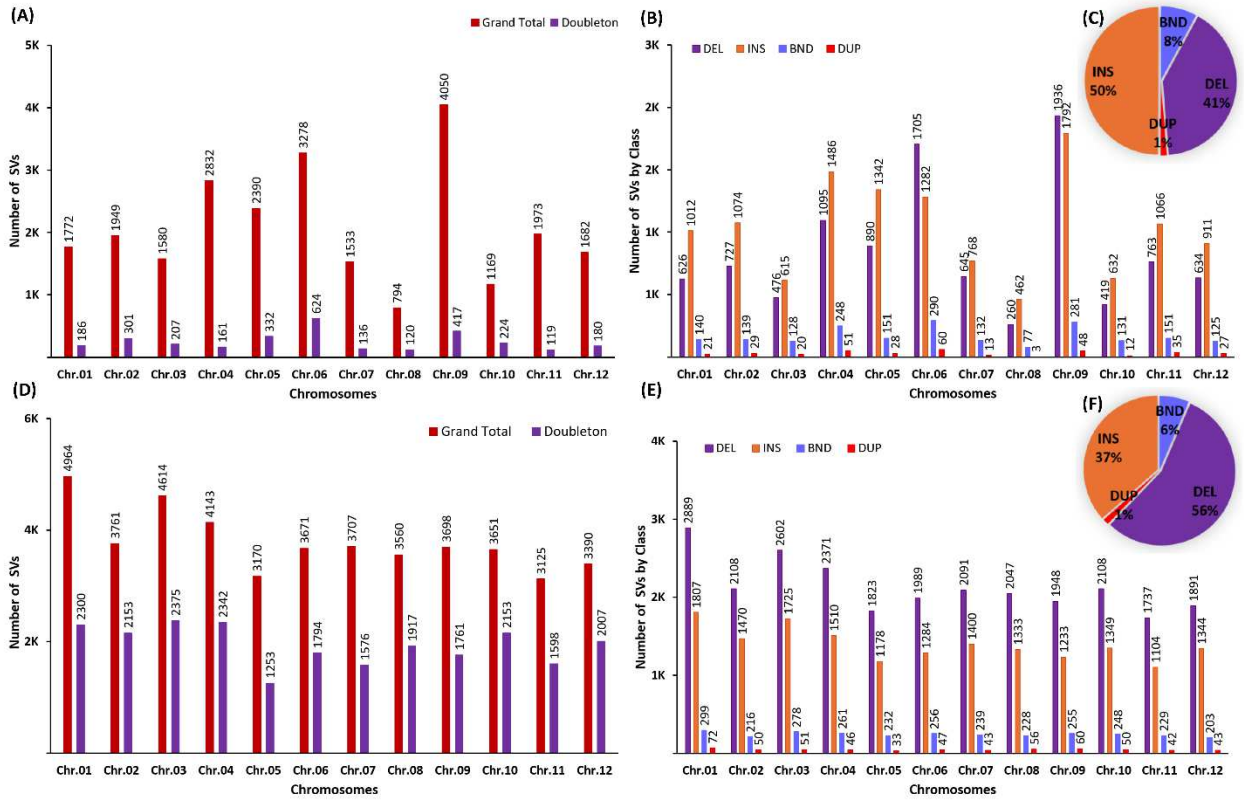


Figure 2. Chromosomal landscape and structural variant (SV) classification across cultivated and wild tomato accessions.

(A) Genome-wide distribution and density of total SVs (red) and line-specific private doubletons (blue) mapped across all 12 chromosomes in 57 cultivated tomato lines. (B–C) Absolute counts (B) and relative percentage proportions (C) of major SV classes, deletions (purple), insertions (orange), duplications (red), and breakends (blue), identified in the cultivated tomato panel. (D–F) Corresponding chromosomal distribution (D), total counts (E), and percentage breakdown (F) of SV classes across the three wild tomato accessions.

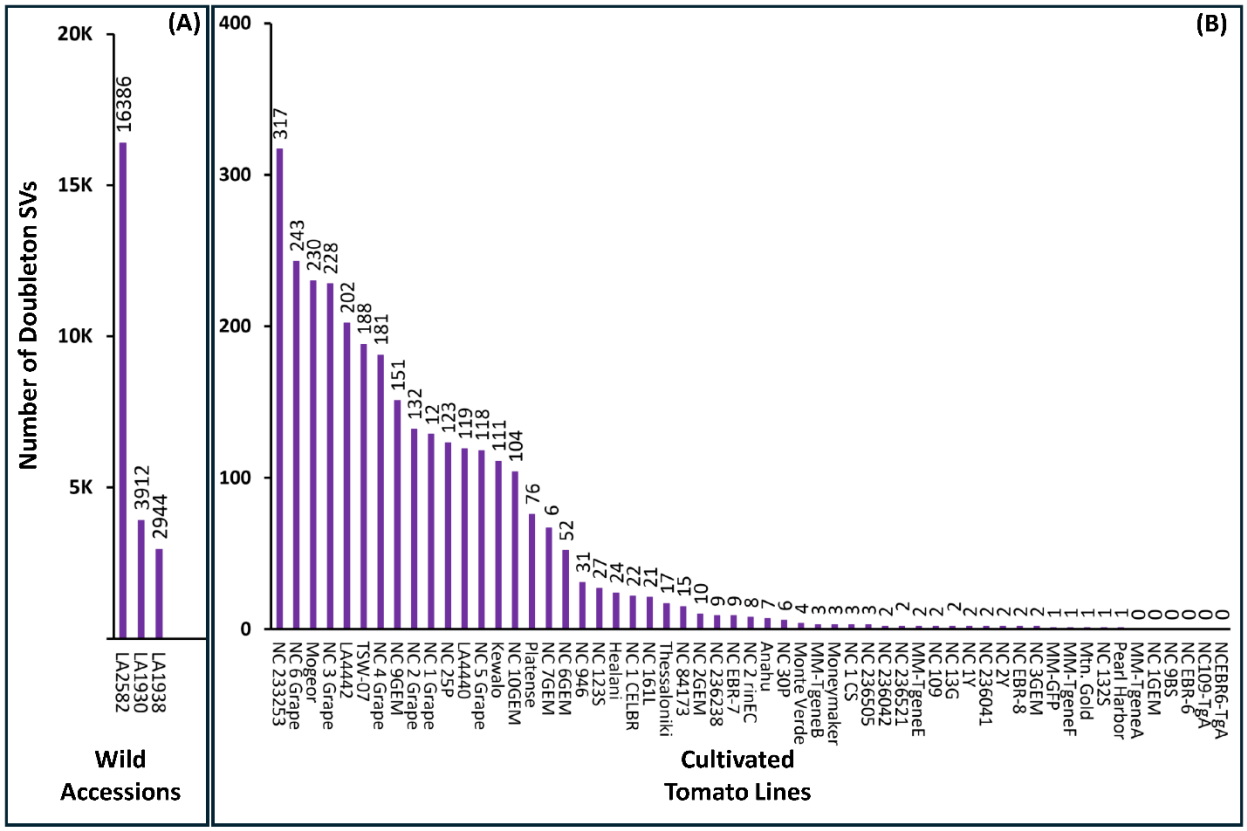
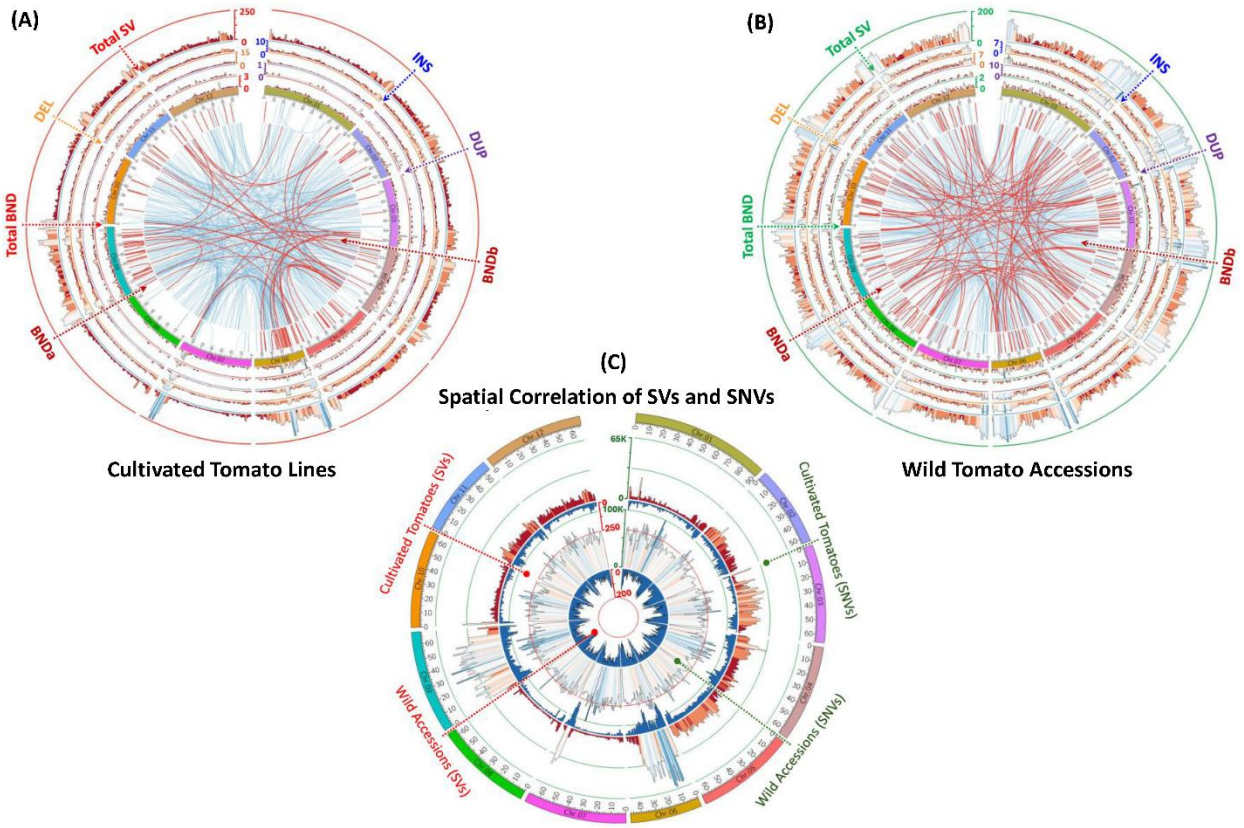


Figure 3. Comparative analysis of line-specific private doubleton structural variants (SVs) in wild and cultivated tomato accessions.

(A, B) Abundance and line-by-line distribution of quality-filtered (PASS) homozygous private doubleton SVs identified across (A) wild tomato species ($N = 3$) and (B) cultivated tomato accessions ($N = 57$). Private doubletons denote unique structural variants restricted to a single accession in the panel (homozygous allele count = 2 in one line), serving as diagnostic genomic fingerprints for variety identification and germplasm discrimination.



881

882 **Figure 4. Genome-wide distribution and spatial density profiles of structural variants and**
883 **short genetic variants.**

884 (A–B) Circular density plots displaying structural variant (SV) and short genetic variant (SNV and
885 InDel) distributions across the 12 tomato chromosomes for (A) 57 cultivated tomato lines and (B)
886 three wild accessions. Variant densities are calculated as raw variant counts per adjacent, non-
887 overlapping 1-Mb genomic window ($N = 778$ total windows across the SL4.00 reference
888 genome). Concentric tracks display overall SV density, specific structural variant classes
889 (deletions [DEL], insertions [INS], duplications [DUP], total translocation breakends [BND]), and
890 breakend types (BNDa, BNDb; red = homozygous, blue = heterozygous). (C) Spatial correlation
891 analysis comparing structural variant density relative to short genetic variant density across
892 cultivated and wild tomato genomes.

893

894

895

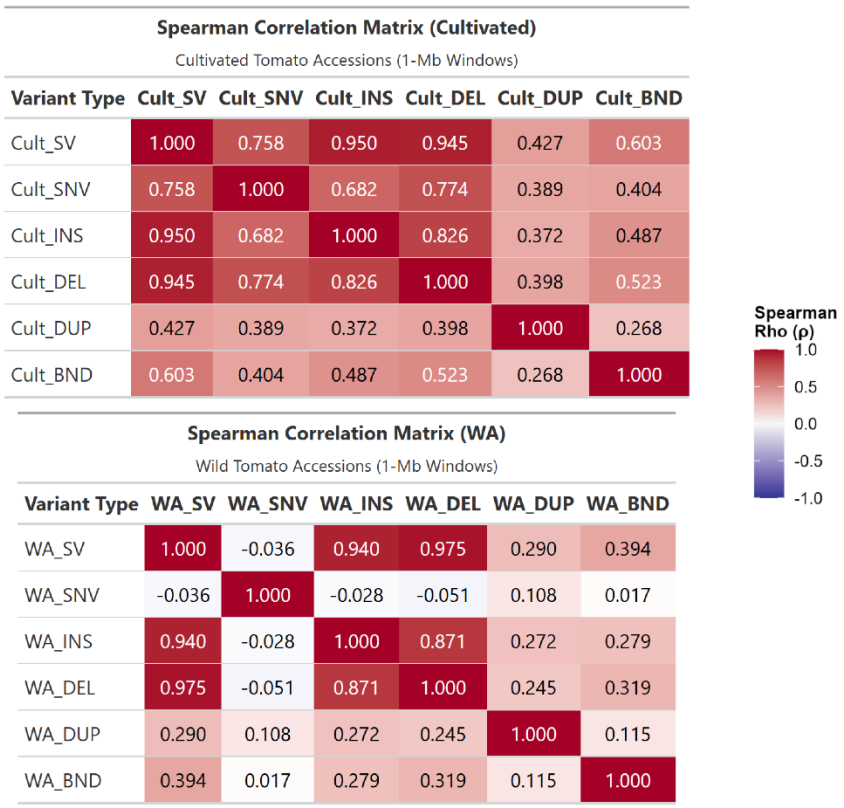


Figure 5. Non-parametric Spearman rank correlation matrix assessing spatial co-localization of structural variants and short genetic variants.

Heatmap matrix of non-parametric Spearman rank correlation coefficients (ρ) evaluating spatial co-localization across $N = 778$ non-overlapping 1-Mb genomic windows. Total structural variant (SV) density strongly correlates with short genetic variant density (SNVs and InDels) in cultivated tomato accessions ($N = 57; \rho = 0.758, p < 0.001$, FDR-adjusted), whereas wild accessions display localized divergence ($N = 3; \rho = 0.036, p < 0.001$). Strong spatial co-localization is observed between total SV density and specific SV classes (insertions and deletions; $\rho = 0.94 - 0.97, p < 0.001$), with moderate correlations observed for duplications and translocation breakends ($\rho = 0.30 - 0.60, p < 0.001$). Color intensity and internal cell values reflect pairwise Spearman ρ values. All two-tailed p-values were adjusted for genome-wide multiple testing using the Benjamini–Hochberg False Discovery Rate (FDR < 0.05).

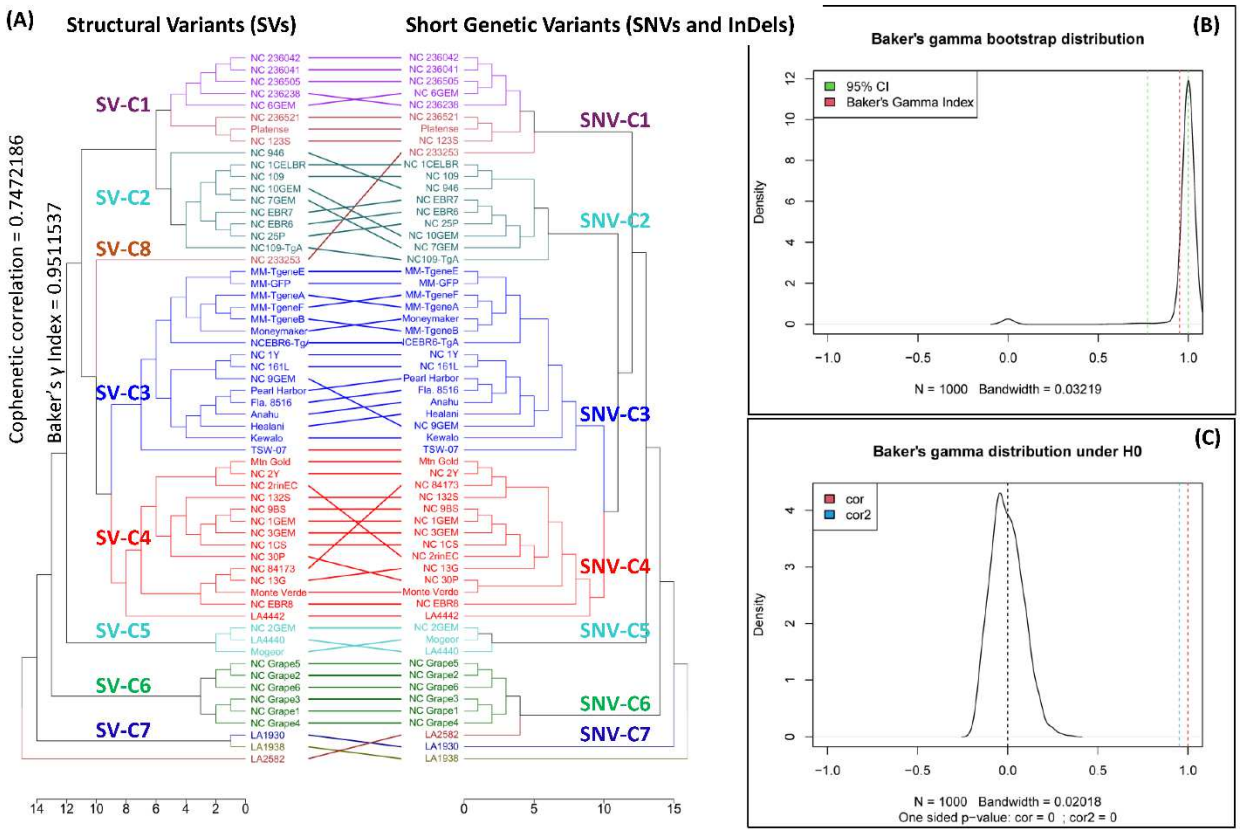
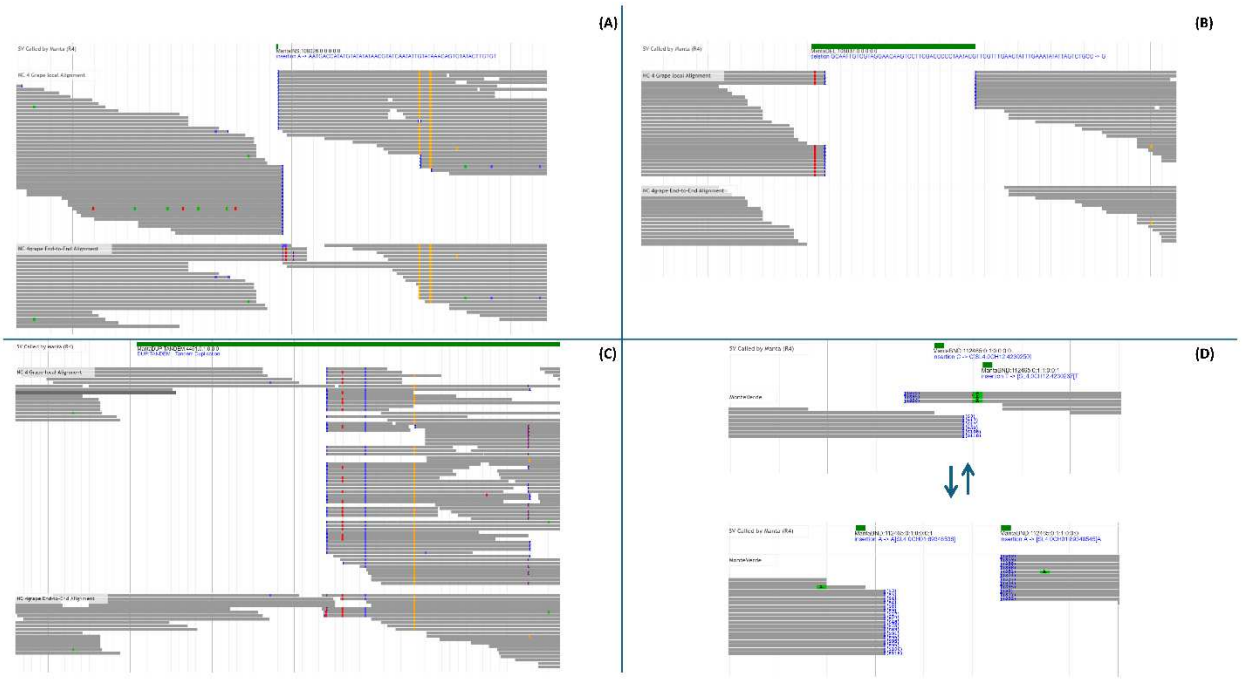


Figure 6. Cophylogenetic tanglegram and statistical testing of topological congruence between structural variant and short genetic variant phylogenies. (A) Cophylogenetic tanglegram comparing dendrograms constructed from structural variants (SVs; left) and short genetic variants (SNVs and InDels; right) across 60 tomato accessions. Dendrograms were generated from Identity-By-State (IBS) genetic distance matrices using complete-linkage hierarchical clustering. Branch colors and clade labels (SV-C1 - SV-C8; SNV-C1 - SNV-C7) delineate major germplasm groupings, while connector lines highlight lineage mapping and topological discordance across call sets. Overall tree congruence yielded a cophenetic correlation coefficient of 0.747 and a Baker's gamma index (γ) of 0.951. (B) Empirical Baker's gamma bootstrap distribution (N = 1,000 replicates) assessing topological stability. The red dashed line denotes the observed Baker's gamma index (γ = 0.951), and green dashed lines indicate the 95% bootstrap confidence interval (95% CI = [0.774, 1.000]). (C) Permutation-based null distribution (N = 1,000) testing the null hypothesis (H_0) of independence between tree topologies. Vertical dashed lines represent observed correlation metrics, which fall far outside the null expectation (empirical $p < 0.001$), confirming statistically significant cophylogenetic alignment.



Supplementary Figure 1. JBrowse 1 visualization of four structural variant (SV) classes comparing local and end-to-end alignment strategies.

Representative genomic alignment tracks in JBrowse 1 genome browser illustrating four major structural variant classes: (A) Insertion (INS), (B) Deletion (DEL), (C) Duplication (DUP), and (D) Breakend/Translocation (BND), depicting an inter-chromosomal translocation of a DNA fragment between chromosomes 1 and 12. Panels A–C feature representative alignments from tomato accession NC 4grape, while panel D features accession MonteVerde. Each vertical gray grid line represents 5 nucleotide bases (5 bp). For each panel of A-C, the top track displays raw reads aligned using the Bowtie2 local alignment strategy (`--local`), while the bottom track displays reads aligned using the default end-to-end strategy. Local alignment successfully retains soft-clipped split reads across SV breakpoints (top tracks), providing the required mapping coordinates for Manta SV identification. Conversely, end-to-end alignment fails to map reads spanning breakpoint junctions, resulting in unmapped reads, coverage gaps, and a complete loss of downstream SV calls (bottom tracks).

Supplementary Files

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

- SupplementrayFigure1.tif
- SupplementaryTables.xlsx