nature communications

Article

https://doi.org/10.1038/s41467-026-74784-5

Wild tomato genome assemblies reveal structural variants and repeat content act as recombination barriers

Received: 19 December 2025

Accepted: 16 June 2026

Published online: 28 June 2026

Check for updates

Willem M. J. van Rengs[1,8,11], Roven Rommel Fuentes[1,2,11], Zahra Zangishei[3,11], Elias Primetis[1,2,11], Yazhong Wang[1,9], Joiselle B. Fernandes[1], Tamara Susanto[1], Qichao Lian[1,10], Sieglinde Effgen[1], Bruno Huettel[4], Saleh Alseekh[5,6], Björn Usadel[3,7] ✉ & Charles J. Underwood[1,2] ✉

Crop wild relatives are used to improve cultivated plants and precise tracking of genetic introgression requires high-quality genome assemblies. Here we present de novo genome assemblies of two wild tomato species – the broadly stress-resistant *Solanum pennellii* (LA0716) and the salt-resistant *Solanum cheesmaniae* (LA1039). The improved *S. pennellii* genome adds 146 Mbp to the twelve chromosomes compared with the original reference. The alignment of the new assemblies with multiple gold-standard assemblies identified shared and species-specific structural variants. Analysis of repeat content demonstrates independent explosions of *Tekay* retrotransposons in *S. pennellii* and *S. peruvianum*. Genome sequencing of 709 recombinant plants derived from male and female backcrosses of three different hybrids reveals higher crossover rate in female meiosis. Conserved female-enhanced recombination regions were discovered and coldspots were attributed to megabase-scale inversions and insertion-deletion polymorphisms. Our *S. pennellii* and *S. cheesmaniae* genome assemblies reveal how repeat content diverged in nature and during breeding, and uncovers how both reproductive gender and structural variants dictate recombination landscapes in tomato hybrids.

Modern molecular plant breeding harnesses near-complete genome sequences to reveal gene markers that control traits of agronomic importance. The cultivated tomato is the world's number one vegetable crop and recent tomato improvement was catalyzed by the release of the first genome assembly of the *Solanum lycopersicum* variety Heinz1706 in 2012[1,2]. Furthermore, a combination of short-read sequencing, BAC sequencing and genetic map information resulted in a genome assembly of the stress-resistant distant tomato relative *Solanum pennellii* (accession LA0716)[3]. This assembly has been an important resource to elaborate on the genetic basis of important

[1]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, Cologne, Germany. [2]Department of Plant and Animal Biology, Radboud Institute for Biological and Environmental Sciences, Radboud University, Nijmegen, the Netherlands. [3]Faculty of Mathematics and Natural Sciences, Institute of Biological Data Science, CEPLAS, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [4]Max Planck Genome-center, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, Cologne, Germany. [5]Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg, Potsdam-Golm, Germany. [6]Center for Plant Systems Biology and Biotechnology, Plovdiv, Bulgaria. [7]Institute of Bio- and Geosciences IBG-4, Bioinformatics, BioSC, CEPLAS, Forschungszentrum Jülich, Jülich, Germany. [8]Present address: Keygene N.V., Wageningen, The Netherlands. [9]Present address: Laboratory of Advanced Breeding Technologies, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [10]Present address: Guangdong Basic Research Center of Excellence for Precise Breeding of Future Crops, Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Provincial Key Laboratory for the Development Biology and Environmental Adaptation of Agricultural Organisms, South China Institute for Soybean Innovation Research, College of Life Sciences, South China Agricultural University, Guangzhou, China. [11]These authors contributed equally: Willem M. J. van Rengs, Roven Rommel Fuentes, Zahra Zangishei, Elias Primetis. ✉e-mail: b.usadel@fz-juelich.de; cunderwood@mpipz.mpg.de

agronomic traits through the well-used *S. lycopersicum* cv. M82 x *S. pennellii* LA0716 recombinant inbred line population[4]. Although high-throughput short-read sequencing accelerated the discovery of natural genetic variants, it has also introduced an unavoidable bias: characterized variants are disproportionately skewed toward single-nucleotide polymorphisms (SNPs) and small indels[5,6]. Furthermore, short reads cannot accurately resolve complex genomic regions, repetitive sequences, and large structural variations, leading to fragmented genome assemblies that lack essential genomic context. Compared to Heinz1706 that has been re-assembled several times using newer technologies including long-read sequencing and chromatin conformation map[2,7–9], the *S. pennellii* LA0716 genome remained as it was, notwithstanding efforts to use long read sequencing for the *S. pennellii* accession LA5240 which is used for high resolution genetic mapping[10,11]. Despite the release of superpangenomes for the *Solanum* genus and the *Lycopersicon* clade[12,13] the absence of an improved *S. pennellii* LA0716 genome and an assembly of the stress-resistant *Solanum cheesmaniae* (accession LA1039)[14] hinders introgression efforts from these important sources of abiotic stress tolerance.

High-quality genomes enable the exploration of large and complex genetic variants, including structural variants (SVs) and transposable elements (TEs). Typically, SVs impose greater potential impact to gene structure, dosage and location than simple genetic variants, and SVs contribute to species evolution and diversity[12,13,15–20]. Recent studies on SVs have indicated they can be causal genetic variants for agronomically-relevant traits that are not accounted by simple variants like single nucleotide polymorphisms (SNPs) and small insertions or deletions[8,19,21]. For example, in the domesticated tomato an 83 kbp duplication at the *supressor of branching 3* locus was co-selected with a *jointless2* allele that improves fruit harvestibility[21]. Similar to SVs, TEs are among the major drivers of genome evolution and diversity in plants[22], often comprising the majority of genomic content in crop species such as maize and wheat[23,24]. Retrotransposons (class I TEs) are often activated under biotic and abiotic stress conditions, and developmental stages, contributing to genome plasticity and potentially facilitating plant adaptation and evolution[25–29]. In the *Solanaceae* family, the majority of Long Terminal Repeat (LTR) retrotransposons are from the *Ty3* (Gypsy) superfamily[12], and in the *Solanum* genus the *Tekay* (*Ty3*) and *Tork* (*Ty1*) lineages are most abundant[30,31]. The *Tekay* lineage is also a prevalently active LTR retrotransposon lineage in several other plant families[32–34]. It was previously estimated that repeats make up ~82% of the *S. pennellii* genome, 45% of which consist of retrotransposons[3].

After genetic linkage was demonstrated in *Drosophila melanogaster*[35], the cultivated tomato was one of the first plant species where genetic linkage was described[36]. Meiotic crossovers (COs) break genetic linkage and currently the interaction of genetic polymorphism with CO distribution is of interest to enable the precise introgression of alleles from wild relatives to domesticated tomato. COs are not distributed uniformly across plant genomes but occur in so-called hotspots often located in distal regions of the chromosomes[37–39]. This non-uniform crossover distribution limits the shuffling of alleles in regions of the genome that are devoid of COs, resulting in linkage between advantageous and negative traits (so-called linkage drag) and thereby inefficiency in breeding. CO hotspots are associated with gene-rich areas, whereas coldspots are observed in retrotransposon-rich, heterochromatic or inverted regions[40–43]. Studies in different species have linked recombination coldspots with structural variations[44–47], but it was only recently that the causal proof was provided by reverting an inversion through genome editing[48]. A relatively underexplored topic in meiotic CO control is the differential rate of meiotic recombination between male and female meiosis, known as heterochiasmy. Unlike in *Arabidopsis thaliana*, where meiotic COs are more frequent in male meiosis than female meiosis[49–52], there is evidence from genetic marker studies in a *S. lycopersicum* x *S. pennellii* hybrid that female recombination rate is higher than male[53]. Exploring tomato heterochiasmy in finer resolution and in multiple genetic backgrounds has yet to be carried out. Whole genome sequencing of bi-directional backcross populations could now enable not just the detection and comparison of crossovers in different hybrid crosses, but the identification of possible female-enriched crossover hotspots.

In this study, we generated a high-quality long-read based reference genome of *S. pennellii* LA0716 and the first assembly of *S. cheesmaniae* LA1039. We identified megabase-scale SVs and transposable elements variation between the new *S. pennellii* and *S. cheesmaniae* assemblies and high-quality genome assemblies of five cultivated tomato accessions (Heinz1706 SL5.0, Moneyberg-TMV, Micro-Tom, M82, Sweet100), *Solanum pimpinellifolium, Solanum galapagense*, *Solanum habrochaites*, and *Solanum tuberosum*[8,54–60] (Supplementary Table 1). Moreover, we explored the crossover distribution and heterochiasmy in tomato using re-sequencing data from six large backcross populations derived from three hybrids with variable divergence at the DNA sequence level. This new resource provides a compendium of variation between cultivated and wild tomatoes, and could enable the genome-guided introgression of wild traits to the cultivated tomato during the development of stress- and disease-resistant varieties.

Results

Chromosome-scale *Solanum pennellii* and *Solanum cheesmaniae* genome assemblies

We set out to assemble high-quality chromosome scale genome assemblies of *Solanum cheesmaniae* accession LA1039 and *Solanum pennellii* accession LA0716. To this end, we generated PacBio HiFi data using two different input sizes per genotype leading to datasets of 51.4 Gbp (*S. cheesmaniae* LA1039) and 49 Gbp (*S. pennellii* LA0716) (Supplementary Table 2). K-mer analysis of HiFi data revealed only a single clear peak for both accessions with residual heterozygosity estimated at ca. 0.2% indicating both accessions were highly homozygous (Supplementary Fig. 1). Due to the known higher repeat content of *S. pennellii*[3], we additionally generated 381.2 Gbp of ONT data (Supplementary Table 3) and filtered for high-quality reads that were longer than 90 kbp, resulting in 16.08 Gbp of high-quality long reads. We also generated chromosome confirmation capture data (Omni-C (Dovetail)) to provide datasets for reference-free scaffolding[61,62], including 41 Gbp for *S. cheesmaniae* LA1039 and 56 Gbp for *S. pennellii* LA0716 (Supplementary Table 4).

Due to the high homozygosity of both species, we assembled them as single genome haplotypes using Hifiasm[63]. This resulted in assemblies totaling 862 Mbp (*S. cheesmaniae*)–including the direct assembly of the complete *S. cheesmaniae* chromosome 6–and 1109 Mbp (*S. pennellii*) with respective N50 values 27.9 Mbp and 26.2 Mbp (Supplementary Table 5). Next, we used the Omni-C data for automated scaffolding of the raw assembled contigs. Contiguity of our de novo *S. cheesmaniae* and *S. pennellii* genome assemblies was further improved through multiple rounds of manual scaffolding and scaffolding correction, until chromosome-scale genomes were obtained (Fig. 1a, b and Table 1). The final assemblies were supported by chromatin conformation capture, as inter-chromosomal interaction was only observed between highly similar telomeric sequences, and none of the sequences were broken by re-running automated scaffolding (Fig. 1a, b). We observed highly co-linear alignments of the twelve assembled *S. cheesmaniae* chromosomes to *S. lycopersicum* cv. Moneyberg-TMV (hereafter MbTMV) chromosomes 1-12, whereas the chromosome set of the divergent relative *S. pennellii* showed more structural variations, as expected (Fig. 1c, d)[55]. Unplaced *S. cheesmaniae* and *S. pennellii* sequences did not show any further Hi-C interaction with any of the assembled chromosomes and were therefore not placed into the chromosomes (Fig. 1a, b).

https://doi.org/10.1038/s41467-026-74784-5

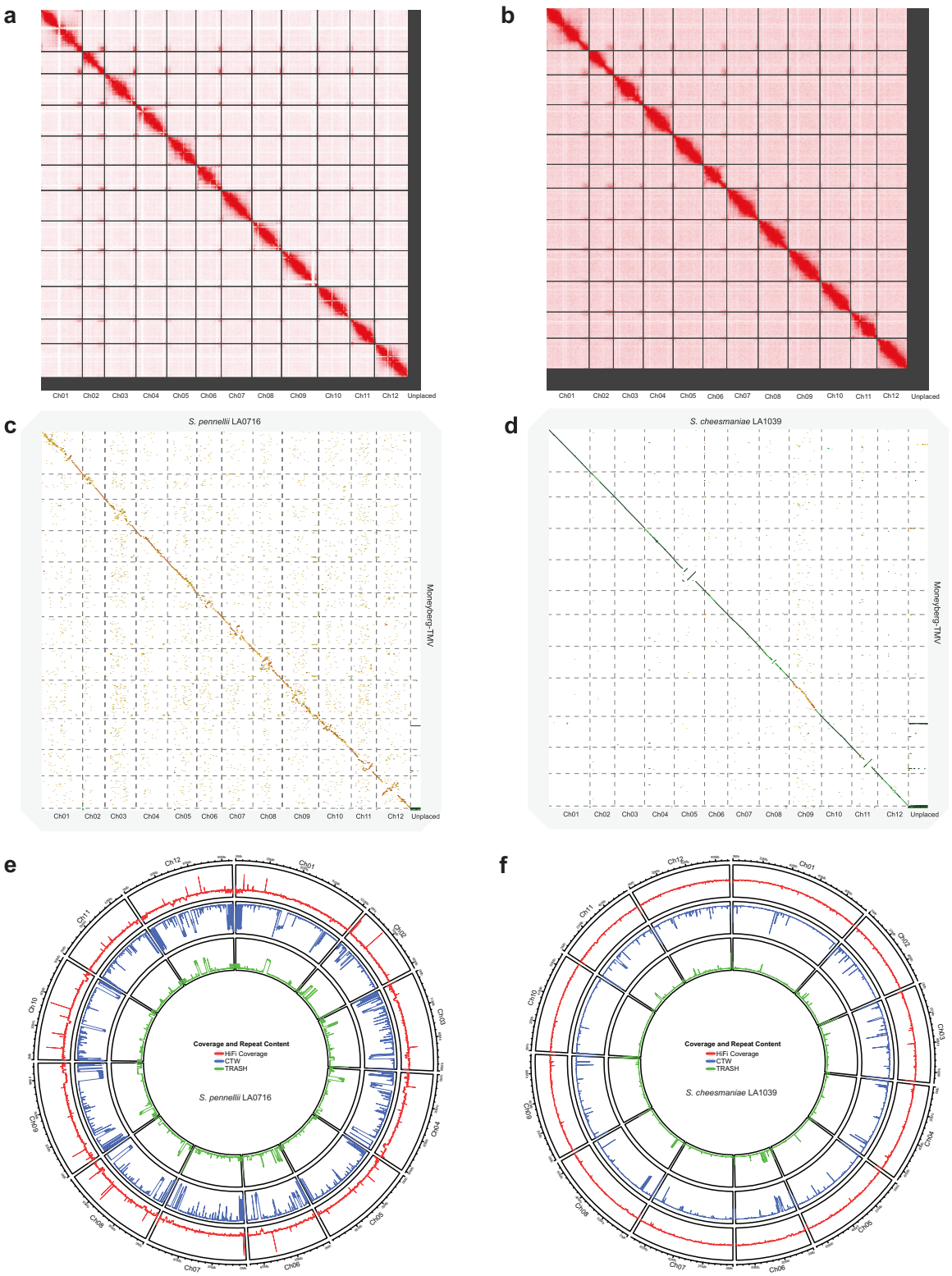


Fig. 1 | Chromosome-scale *Solanum pennellii* and *Solanum cheesmaniae* genome assemblies. a *S. pennellii* LA0716 Hi-C interaction plot. **b** *S. cheesmaniae* LA1039 Hi-C interaction plot. Hi-C interaction plotted by Juicebox, demonstrating strong interactions over the 12 assembled chromosomes of the *S. pennellii* (**a**) and *S. cheesmaniae* (**b**) genome assemblies. Gray lines depict sequence borders. **c** Whole genome comparison dotplot of *S. pennellii* (top axis) aligned to *S. lycopersicum* cv. Moneyberg-TMV (right axis). **d** Whole genome dotplot of *S. cheesmaniae* (top axis) aligned to *S. lycopersicum* cv. Moneyberg-TMV (right axis). Horizontal gray dashed lines represent Moneyberg-TMV chromosome borders. Vertical gray lines represent *S. pennellii* (**c**) and *S. cheesmaniae* (**d**) chromosome borders. Dotplots visualizations are created by D-genies (Cabanettes and Klop, 2018). Genome coverage and repeat content for *S. pennellii* (**e**) and *S. cheesmaniae* (**f**). From outer to inner layer (in 100 Kbp windows): logarithmically normalized and scaled HiFi genome coverage (red line), the log-probability of the CTW algorithm[67] (blue line) and the abundance of monomeric repeats using TRASH[68] (green line) for *S. pennellii* (**e**) and *S. cheesmaniae* (**f**).

https://doi.org/10.1038/s41467-026-74784-5

Table 1 | Genome assembly statistics of *S. pennellii* and *S. cheesmaniae* compared to published assemblies

Accession	Moneyberg-TMV	Micro-Tom	LA1039	LA0716	LA2093	LA0716
Species	*S. lycopersicum*	*S. lycopersicum*	*S. cheesmaniae*	*S. pennellii*	*S. pimpinellifolium*	*S. pennellii*
Reference	van Rengs et al., 2022	Wang et al., 2024	This study	This study	Wang et al., 2020	Bolger et al., 2014
Number of sequences	12	12	12	12	12	12
Number of sequences (>50 kb)	12	12	12	12	12	12
Cumulative size (Mbp)	824.45	812.44	814.47	1072.33	800.50	926.43
N50 (Mbp)	68.5	67.8	68.2	90.2	67.2	78
N90 (Mbp)	54.7	56.8	55.8	72.3	55.6	60.7
L50	6	6	6	6	6	6
L90	11	11	11	11	11	11
Longest sequence (Mbp)	96.5	96.3	96.7	123.5	95.5	109.3
Number of N's	0	4000	3300	4400	487215	71074709
Number of internal N-regions	0	40	33	44	373	41684
Raw LAI	9.67	8.42	8.64	12.43	8.49	3.24
LAI	15.3	14.05	14.27	18.06	14.12	8.87
Complete BUSCO (C)	5852 (98.35)	5851 (98.34)	5852 (98.35)	5857 (98.44)	5851 (98.34)	5827 (97.93)
Complete Single copy (S)	5747 (96.59)	5743 (96.52)	5741 (96.49)	5738 (96.44)	5744 (96.54)	5706 (95.9)
Complete Duplicated (D)	105 (1.76)	108 (1.82)	111 (1.87)	119 (2)	107 (1.8)	121 (2.03)
Fragmented (F)	12 (0.2)	12 (0.2)	12 (0.2)	14 (0.24)	12 (0.2)	15 (0.25)
Missing (M)	86 (1.45)	87 (1.46)	86 (1.45)	79 (1.33)	87 (1.46)	108 (1.82)
Total searched (Solanales)	5950	5950	5950	5950	5950	5950
QV	54.12	72.40	75.69	69.97	N.A.	N.A.
Completeness	99.10	99.22	99.17	99.17	N.A.	N.A.

Table 1 shows the summary statistics of the 12 chromosome scale sequences making up each genome assembly. BUSCO analysis of the completeness of gene content is using the Solanales odb10 benchmark set; % are shown in parentheses. QV and completeness are estimated by K-mer based analysis using HiFi data only.

To assess the quality of our chromosome-scale genome assemblies, we compared statistics of the assembled *S. cheesmaniae* and *S. pennellii* pseudochromosomes to the 12 pseudochromosomes of reference quality tomato assemblies including Moneyberg-TMV[55], the original *S. pennellii* assembly[3] and an *S. pimpinellifolium* assembly[59] (Table 1). We found our scaffolded assemblies performed favorably in terms of contiguity for *S. cheesmaniae* (N50 = 68.2 Mbp, N90 = 55.8 Mbp) and *S. pennellii* (N50 = 90.2 Mbp, N90 = 72.3 Mbp) and repeat content assembly (both with LAI above 14)[64], had few internal gap regions and performed similar in terms of gene completeness (BUSCO)[65] metrics (Table 1). K-mer based assembly evaluation using Merqury[66] showed completeness above 99% for the *S. cheesmaniae* and *S. pennellii* assemblies (QVs above 69) (Table 1), indicating the assemblies were of high-quality.

The genome assemblies were further validated by means of read coverage analysis and characterization of nucleotide composition and repeat elements, including Simple Sequence Repeats (SSRs) and LTRs (Fig. 1e, f, Supplementary Fig. 2 and 3). After remapping the long reads, we found mostly even coverage along the chromosomes with a few outlier regions. For both genomes, multiple windows with reduced coverage often overlaid drastic changes in nucleotide composition and reduced SSR and LTR densities (Fig. 1e, f, and Supplementary Fig. 2 and 3). We previously found such a behavior on the MbTMV chromosome 8, where these changes correspond to irregular genomic features including telomeres, 45S rDNA, and sub-telomeric Tomato Genome Repeat I (TGR1) sequences[55], suggesting a similar phenomenon may occur in the *S. cheesmaniae* and *S. pennellii* assemblies. Notably, these genomic features are extensively present in our *S. pennellii* assembly (Supplementary Fig. 3) and could be more completely assembled due to utilization of >90 kb ONT reads.

The short-read-based assembly of *S. pennellii* LA0716 has been an important platform for research into genetically encoded abiotic and biotic stress resistance in this wild relative of the cultivated tomato[3]. We performed an extensive comparison of the original and new *S. pennellii* genome assemblies and found a total increase in assembly size of 145.9 Mbp between the 12 assembled chromosomes (Supplementary Fig. 4 and Supplementary Table 6). The new *S. pennellii* assembly had higher gene completeness based on BUSCO analysis[65], the repeat content was more completely assembled based on LAI analysis[64] and more than 41,000 gaps were filled (Table 1). Whole genome alignment and pairwise comparison demonstrated highly syntenic regions especially in the distal ends of each chromosome. However, we also found several large differences including inversions (Chromosomes 1, 2, 3, 6, 7, 8) between the two assemblies (Supplementary Fig. 4) and expect that the absence of recombination could have hindered genetic map-based scaffolding in the original assembly. Re-mapping of raw sequencing reads to both assemblies indicated that the new assembly has a more even read coverage across the chromosomes compared to the short-read based assembly (Supplementary Fig. 4).

Next, we examined the distribution of repeats across the genomes of *S. cheesmaniae* and *S. pennellii* using complementary CTW (Context Tree Weighting)[67] and TRASH (Tandem Repeat Annotation and Structural Hierarchy)[68] analysis (Fig. 1e, f). In general, both tools identify more extensive repetitive regions in *S. pennelli* than *S. cheesmaniae* or MbTMV (Fig. 1e, f and Supplementary Fig. 5). We find that in *S. pennelli* repetitive regions are much more broadly distributed along the chromosomes than in *S. cheesmaniae* and MbTMV where punctate repetitive arrays are more common (Fig. 1e, f and Supplementary Fig. 5). The highly repetitive 5S rDNA (Chromosome 1) and tandem arrays of 45S rDNA derived satellite (Chromosomes 6, 8 and 11) were

https://doi.org/10.1038/s41467-026-74784-5

Table 2 | Gene models of *S. lycopersicum* cv. Moneyberg-TMV, *S. cheesmaniae* LA1039, and *S. pennellii* LA0716

Species	*S. lycopersicum*	*S. cheesmaniae*	*S. pennellii*
Accession	Moneyberg-TMV	LA1039	LA0716
Number of genes	34939	35519	31998
Complete BUSCO (C)	5787 (97.26%)	5839 (98.14%)	5854 (98.38%)
Complete Single copy (S)	5000 (84.03%)	5043 (84.76%)	5035 (84.62%)
Complete Duplicated (D)	787 (13.23%)	796 (13.38%)	819 (13.76%)
Fragmented (F)	19 (0.32%)	40 (0.67%)	20 (0.34%)
Missing (M)	144 (2.42%)	71 (1.19%)	76 (1.28%)
Total searched (Solanales)	5950	5950	5950
Complete HOGs	14784 (95.50%)	15033 (97.11%)	15087 (97.46%)
Single HOGs	13605 (87.88%)	13676 (88.34%)	13762 (88.90%)
Duplicated HOGs	1179 (7.62%)	1357 (8.77%)	1325 (8.56%)
Missing HOGs	697 (4.50%)	448 (2.89%)	394 (2.55%)
Total searched (Solanaceae)	15481	15481	15481
Consistent lineage placements	28162 (80.60%)	28924 (81.43%)	30806 (96.27%)
Inconsistent lineage placements	116 (0.33%)	155 (0.44%)	145 (0.45%)
Unknown	6661 (19.06%)	6440 (18.13%)	1047 (3.27%)
Contaminants	0 (0.00%)	0 (0.00%)	0 (0.00%)
PSAURON score	77.3	83.6	97.8

Table 2 shows the summary statistics of the representative gene models developed for the genome assemblies of Moneyberg-TMV (using Mikado), *S. cheesmaniae* (using Helixer) and *S. pennellii* (using Mikado and downstream fine-tuning). BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis of the completeness of proteome was carried out using Compleasm with the lineage dataset solanales_odb10. Measures of the proteome completeness with regard to the conserved homologs (HOGs) as well as proteome consistency with regard to known gene families of the ancestral lineage were carried out using OMArk. Gene model accuracy estimates were carried out using PSAURON.

identified in all three assemblies, while other repetitive regions of unknown origin were found by both tools on *S. pennellii* chromosomes 2, 3, 4, 9 and 12 (Fig. 1e, f and Supplementary Fig. 5). In most chromosomes the repetitive content of *S. cheesmaniae* and MbTMV was very similar, indicative of the closer evolutionary relationship of the two genomes compared to the more distant *S. pennelli* (Fig. 1e, f and Supplementary Fig. 5). Interestingly, some repetitive regions were not identified by both tools (Fig. 1e, f), reflecting the different methodological (mathematical and biological) approaches. In summary, our findings suggest that the *S. pennelli* genome has a more complex repetitive content than the genomes of *S. cheesmaniae* and the domesticated tomato.

Furthermore, structural annotation of the genome assemblies was performed using ab-initio predictions for *S. cheesmaniae,* whereas an integrated pipeline comprising both ab-initio and evidence-based approaches for MbTMV and for *S. pennellii* was used to control for repeat-driven mis-annotations. Annotation was followed by additional refinements (Supplementary Fig. 6), yielding consensus gene model sets for *S. cheesmaniae* (35,519), MbTMV (34,939), and *S. pennellii* (31,998) (Table 2). Benchmarking Universal Single-Copy Orthologs analysis indicated high completeness across all three annotated proteomes, at rates of 98.14% (*S. cheesmaniae*), 97.26% (MbTMV), and 98.38% (*S. pennellii*)[69]. Moreover, measures of the annotated proteome completeness and consistency regarding, respectively, the conserved homologs (HOGs) and known gene families of the ancestral lineage using OMArk confirmed the high quality of the predicted proteomes, identifying about 95.5–97.5% of expected HOGs and a large fraction of genes with consistent lineage placement (81.4% in *S. cheesmaniae*, 80.6% in MbTMV, and 96.3% in *S. pennellii*). Structural accuracy of the gene models, evaluated using PSAURON, further highlighted the robustness of the annotation (ranging from 77.3 in MbTMV to 97.8 in *S. pennellii*) and we identified less than 1% of transcript models were classified as potential TEs when looking for domains associated to TEs with GAQET[70,71]. Orthologous gene groups among the reference species *S. lycopersicum* (Heinz 1706 (ITAG4.0), *S. cheesmaniae* and *S. pennelli* were identified to extract orthologues (Supplementary Data 1 and 2). Finally, we performed a functional annotation and cross-species comparisons among MbTMV, *S. cheesmaniae* and *S. pennelli* compared to the reference *Solanum lycopersicum* ITAG4.0 gene set, providing a systematic assessment of the gene presence–absence variation and the copy number differences among the species (Supplementary Data 3).

Whole genome alignment of eleven high-quality *Lycopersicon* clade genomes reveals shared and species-specific structural variants

To analyze the structural variants of the new *S. cheesmaniae* and *S. pennellii* genome assemblies in the context of their closest extant relatives, we included ten additional high-quality genome assemblies that were constructed with at least PacBio long-read and Hi-C short-read data. These assemblies included five cultivated tomato accessions (Heinz1706 SL5.0, MbTMV, Micro-Tom, M82, Sweet-100), three additional wild tomato species (*S. pimpinellifolium* LA1589, *S. pimpinellifolium* LA2093, *S. galapagense* LA0317 and *S. habrochaites* LA0407), and *S. tuberosum* (DM8.1) as outgroup (Supplementary Table 1)[8,54–60]. We identified both small and large genomic variants, between our reference model accession *S. lycopersicum* cv. Moneyberg-TMV and the other genomes through whole genome pairwise comparison (Fig. 2 and Fig. 3, Supplementary Fig. 7 and 8).

The five *S. lycopersicum* genomes show a high degree of genome-wide synteny, making up more than 700 Mbp of their complete genome sizes (Figs. 2a, and 3b). Comparing *S. lycopersicum* MbTMV with the four other cultivated tomato accessions identified between 506,477-1,366,260 SNPs, and 74,706-179,406 SVs (Figs. 2a, and 3a, b). Notably, the distal, gene-rich, ends of all twelve chromosomes are fully syntenic in the five *S. lycopersicum* genomes apart from rare inversions on the short arms of chromosomes 9 (Micro-Tom, Heinz1706 and M82) and 12 (Micro-Tom and Sweet100) (Fig. 2a). A similar inversion on the short arm of chromosome 9 is also found in both *S. pimipinellifolium* accessions, *S. habrochaites* and *S. pennellii* (Fig. 2a). In all five cultivated tomatoes the complete chromosomes 6, 7 and 8, including the vast pericentromeric regions, are structurally almost identical (Fig. 2a). In the other nine chromosomes, different pericentromeric haplotypes were identified in specific accessions (Fig. 2a). For example, on chromosome 1 both Heinz1706 and M82 share a shorter pericentromeric haplotype compared to the other three *S. lycopersicum* accessions, while on chromosome 3 Micro-Tom contains two accession-specific large inversions in the pericentromeric region (Fig. 2a). Alternative pericentromeric haplotypes were also found on chromosomes 2 (Micro-Tom), 4 (M82 and Sweet100), 5 (Micro-Tom and M82), 9 (MbTMV and Sweet100), 10 (M82), 11 (MbTMV and Heinz1706) and 12 (Heinz1706 and Sweet100) compared to the other cultivated tomato species. Overall, the structural diversity in the cultivated tomato genomes is largely restricted to the pericentromeric heterochromatic regions.

The structural diversity between the five *S. lycopersicum* genomes can be considered modest in the context of the diversity of the five wild relatives (*S. pimpinellifolium*, *S. galapagense*, *S. cheesmaniae*, *S. habrochaites* and *S. pennellii*) and outgroup (*S. tuberosum*). Within the pericentromeric regions, complex structural variation was found

https://doi.org/10.1038/s41467-026-74784-5

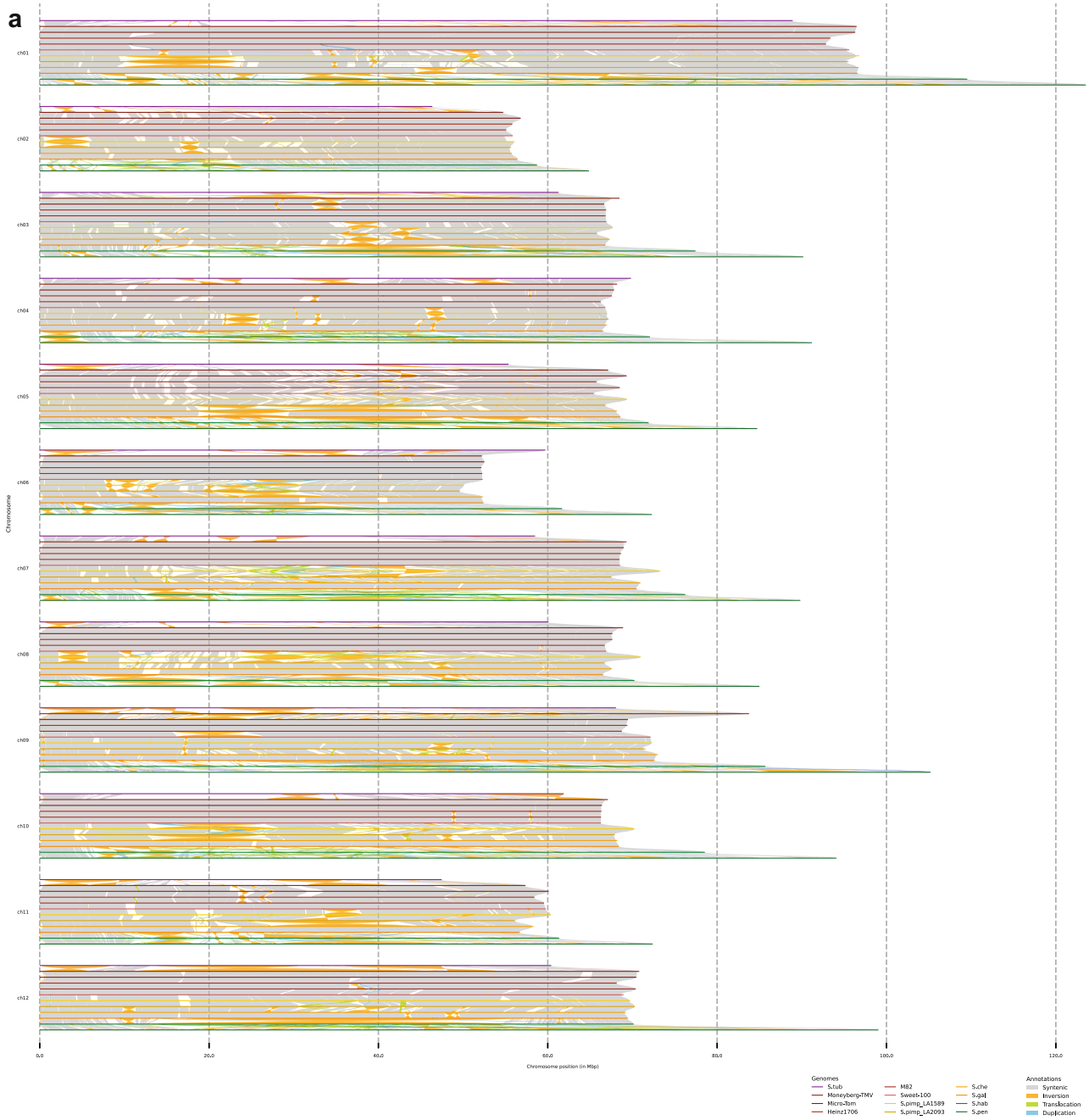


Fig. 2 | Whole genome alignment of eleven *Lycopersicon* clade genomes and a *S. tuberosum* genome reveals shared and species-specific structural variants. **a** Whole genome alignment visualization of eleven *Solanum* sect. *Lycopersicon* genomes including five *S. lycopersicum* varieties (Heinz1706, Moneyberg-TMV, Micro-Tom, M82, Sweet-100), two *S. pimpinellifolium* accessions (LA1589 and LA2093), the *S. cheesmaniae* accession LA1039 (S.che), with assemblies of *S. galapagense* (S.gal), *S. habrochaites* (S.hab), *S. pennellii* LA0716 (S.pen) and *S. tuberosum* (S.tub). All genomes were generated from at least *PacBio* (CCS) HiFi and Hi-C data, apart from the two 2 *S. pimpinellifolium* genomes which were generated with PacBio CLR data and Hi-C. The visualization was generated using SyRI and plotsr. Annotations between genomes include syntenic regions (gray), inversions (yellow), translocations (green) and duplications (blue). Not-aligned or highly diverse regions are shown in white.

between the *S. lycopersicum* genomes and *S. pimpinellifolium*, with substantial variation also present between the two *S. pimpinellifolium* accessions themselves (Fig. 2a). Outside of the pericentromeres, when compared to the five *S. lycopersicum* genomes private inversions were found in *S. pimpinellifolium* LA1589 at the start of chromosome 2, on the short arm of chromosome 8 and on the long arm of chromosome 11 (Fig. 2a). Due to the singleton nature of these SVs future observation in independent assemblies would add credibility to their validity. The independently assembled Galapagos tomato (*S. galapagense* and *S. cheesmaniae*) genomes show a high degree of synteny genome-wide sharing large inversions on chromosomes 5 (18-21 Mbp), 8 (7.4 Mbp) and 11 (16-17 Mbp) compared to the five *S. lycopersicum* accessions (Fig. 2a). Although over 600 Mbp of the Galapagos tomato genomes are syntenic to *S. lycopersicum*, a two-fold increase in the total number of variants was detected with respect to SNPs (3,473,617-3,510,695) and SVs (435,906-439,402) between *S. lycopersicum* and the two Galapagos tomato (Fig. 3a, b). Despite the similarities between the *S. cheesmaniae* and *S. galapagense* genomes, large inversions (greater

https://doi.org/10.1038/s41467-026-74784-5

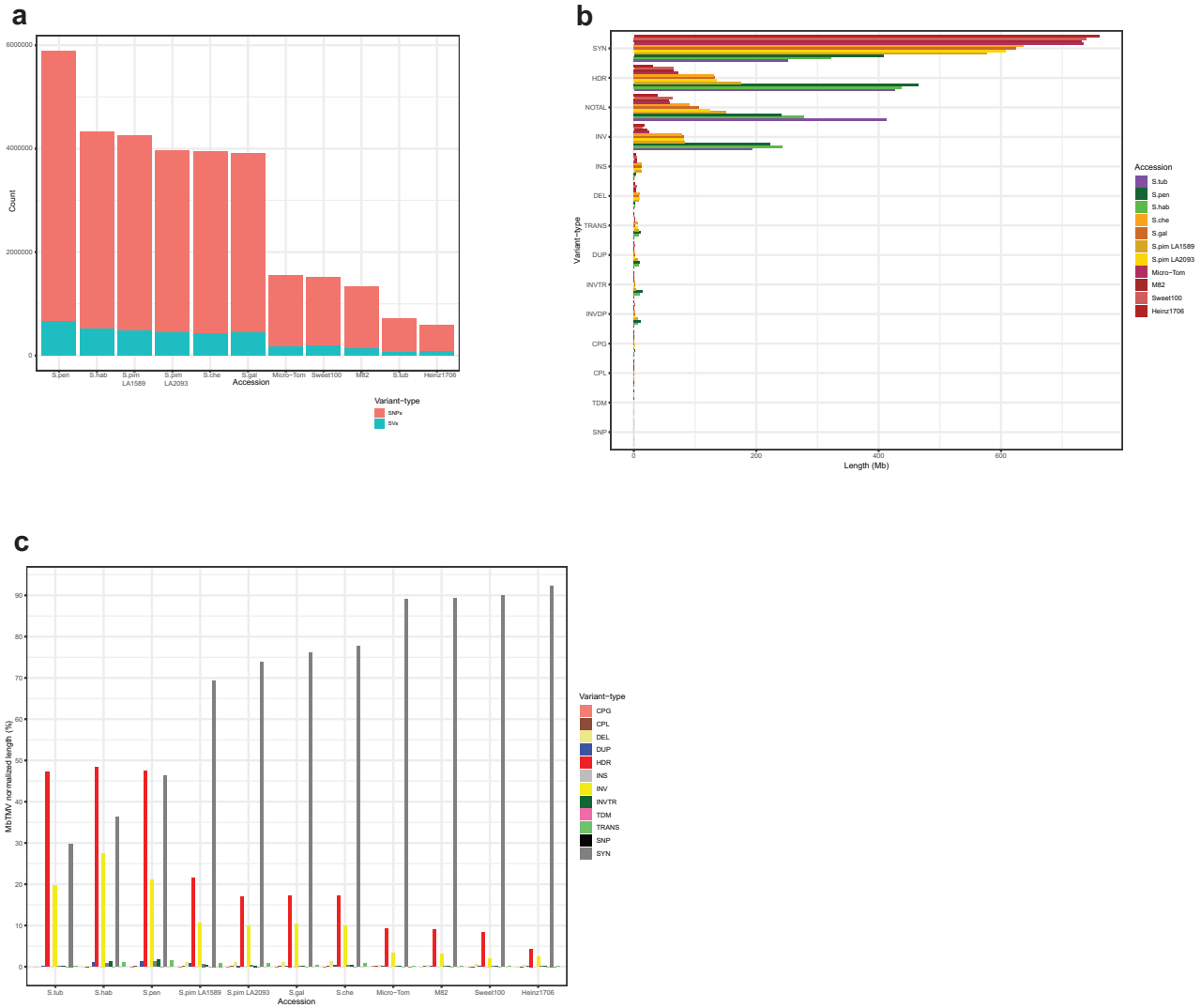


Fig. 3 | Quantification of genome-wide variation between eleven *Lycopersicon* clade genomes and a *S. tuberosum* genome. Here, the *S. lycopersicum* Moneyberg-TMV variety is used as the reference sequence for comparison with ten *Solanum* sect. *Lycopersicon* genomes including four *S. lycopersicum* varieties (Heinz1706, Micro-Tom, M82, Sweet-100), two *S. pimpinellifolium* accessions (LA1589 and LA2093) one *S. cheesmaniae* (LA1039; S.che), one *S. galapagense* (S.gal), one *S. habrochaites* (S.hab), one *S. pennellii* (LA0716; S.pen) and one *S. tuberosum* (S.tub) outgroup. **a** Total number of SNPs and Structural variants (SVs) identified from alignment blocks per line compared to the *Moneyberg*-TMV reference genome. SNP numbers exceed SV numbers in all lines. S.tub has low number of identified variants due to large non-aligned regions. **b** Summary of variant length per genome, based on the annotated alignment blocks compared to the Moneyberg-TMV reference genome. Besides syntenic (SYN) regions, Highly diverse regions (HDR) and inversions (INV) makeup the largest total length of SVs. A degree of overlap in variant length can be expected due to genome size differences and variant definitions, e.g., regions containing duplications (DUP) and HDRs may overlap. Not aligned regions (NOTAL) are inferred from regions not being part of any annotated alignment blocks and make up a large portion of each respective genome size. Inverted Duplicated region (INVDP), Insertion (INS), Deletion (DEL), Translocated region (TRANS), Inverted Translocated region (INVTR), Copy Gain region (CPG), Copy Loss region (CPL), Tandem Repeat region (TDM) and Single Nucleotide Polymorphism (SNP). **c** Length-wise genomic composition of each line compared to the Moneyberg-TMV reference genome. An increase of syntenic regions (SYN) is clearly visible in the *S. lycopersicum* varieties that are closely related to Moneyberg-TMV.

than 2 Mbp) on chromosomes 1, 3, 5 and 12 (Fig. 2a), and differences in the total variant lengths (Fig. 3b), demonstrate they are divergent from one another.

Next, we explored simple and structural divergence with the more divergent "Hirsutum" group species *S. habrochaites* and *S. pennellii*, and the outgroup *S. tuberosum*. Between *S. lycopersicum* and *S. habrochaites*/*S. pennellii*/*S. tuberosum* we identified SNPs (3,805,503/ 5,230,709/ 649,002) and SVs (536,721/ 682,778/ 74,706), with the lower number of variants with *S. tuberosum* being explained due to only 29.8% of the sequence being syntenic (Figs. 2a, and 3a, b). Consistent with previous reports we found complete inversions on the chromosome arms between *S. lycopersicum* and *S. pennellii* (short arms of chromosomes six and seven), and between *S. lycopersicum* and *S. tuberosum* (one arm inversion on chromosomes two, five, six, eight, nine, ten, eleven and twelve)[72–74]. Our analysis led to the discovery of two large unreported inversions on the chromosome arms. We found a 4.7 Mbp inversion between *S. lycopersicum* and *S. habrochaites* (on the short arm of chromosome 4) and a 2.2 Mbp inversion of the majority of short arm of chromosome 8 between *S. lycopersicum* and *S. pennellii* (Fig. 2a and Supplementary Fig. 8). The former inversion may be a technical artifact due to the presence of an assembly gap 11 kbp from the proximal inversion breakpoint, whereas the latter inversion between *S. lycopersicum* and *S. pennellii* could be validated by detailed Hi-C analysis (Supplementary Fig. 16). In total, we found 202 inversions between the *S. lycopersicum* and *S. pennellii* genomes making up over 200 Mbp in length (Fig. 3b). In addition, the new *S. pennellii* assembly

demonstrates the presence of a 7.1 Mbp inversion on the long arm of chromosome 3 that was previously reported between *S. lycopersicum* and many other wild relatives yet was thought to be absent in *S. pennellii*[3,13]((Fig. 2a and Supplementary Fig. 4 and 8). In general, the structure of the pericentromeric regions differs greatly between *S. habrochaites* and *S. pennellii* and yet a 300 kbp inversion common to the two "Hirsutum" group species was identified on the first arm of chromosome 9 (Fig. 2a) and is evidence of their close phylogenetic relationship. Finally, our qualitative assessment was validated by the normalization of total variant lengths compared to the MbTMV genome which demonstrated low synteny with *S. pennelllii* (46.3%) and *S. habrochaites* (36.2%) (Fig. 3c). In conclusion, high-quality genomes derived from long read sequencing and Hi-C data allowed us to uncover examples of genomic rearrangements in heterochromatic pericentromeric regions and euchromatic chromosome arms between related *Solanum* species.

Transposable element annotation of the *Lycopersicon* clade reveals independent explosions of *Tekay* retrotransposons

To elucidate the TE content of the *Lycopersicon* clade, we conducted a detailed annotation of retrotransposons (class I TEs) and DNA transposons (class II TEs) of the aforementioned tomato and wild relative genomes, and a single potato genome (Fig. 4, Supplementary Fig. 9, 10, Supplementary Table 1, and Table 3). Globally we found that the total TE content tends to be higher in wild tomato species compared to cultivated tomatoes, with nearly 65% of the *S. pennellii* genome composed of TEs, compared to just 54.74% of the *S. lycopersicum* MbTMV genome (Table 3). We noticed that the TE component is related to genome size, which is higher in non-domesticated accessions (Table 3). For example, the *S. pennellii* genome is 30.1% larger than the *S. lycopersicum* MbTMV genome (Table 3). We found that *Ty3* LTR retrotransposons are the most abundant TE superfamily across the ten species, representing 17.4% of the *S. galapagense* genome and up to 24.43% of the potato genome (Table 3). The genomic percentage of the *Ty1/Copia* Class I superfamily does not follow the same trend, with the highest percentage observed in *S. cheesmaniae* (6.26%) and the lowest in potato (5%) (Table 3). Other LTR retrotransposon classes account for 4.88-6.79% of the nine tomato genomes, whereas this grouping makes up 11.15% of the potato genome (Table 3). Class II TEs are less prominent in the studied genomes, where the *Mutator* family is most prevalent with its highest genomic percentage detected in *S. pennellii* (14.88%), while the lowest is found in potato (4.46%) (Table 3).

For most of the *Ty1/Copia* and *Ty3* lineages, the abundance of intact elements is stable across the domesticated tomato accessions, while it is higher in wild tomato species and highest in the most distant relative *S. pennellii*[3] (Fig. 4a). The *Ty1/Copia Tork* lineage is exemplary with 500 elements in MbTMV (domesticated tomatoes have between 480–505 *Tork* copies), while there are more than 1100 in *S. pennelli* (Fig. 4a). The abundant *Ty3 Tekay* lineage is represented by 1261–1380 elements in the four cultivated tomato accessions Micro-Tom, Heinz, M82 and Sweet100, while in *S. pennellii*, there are around 3200 *Tekay* elements, which makes it the most successful lineage in the ten species (Fig. 4a). We found that MbTMV has ~500 more *Tekay* elements compared to the other domesticated species (Fig. 4b). *Ty1/Copia* and *Ty3* elements are less abundant in potato than the tomato accessions, with the only exception being the prolific *Ale* lineage (Fig. 4a, b). Overall, the variation in specific *Ty1/Copia* and *Ty3* lineage abundances indicated that the aging of individual elements would allow us to understand how TE dynamics may have shaped the evolution of the *Lycopersicon* clade.

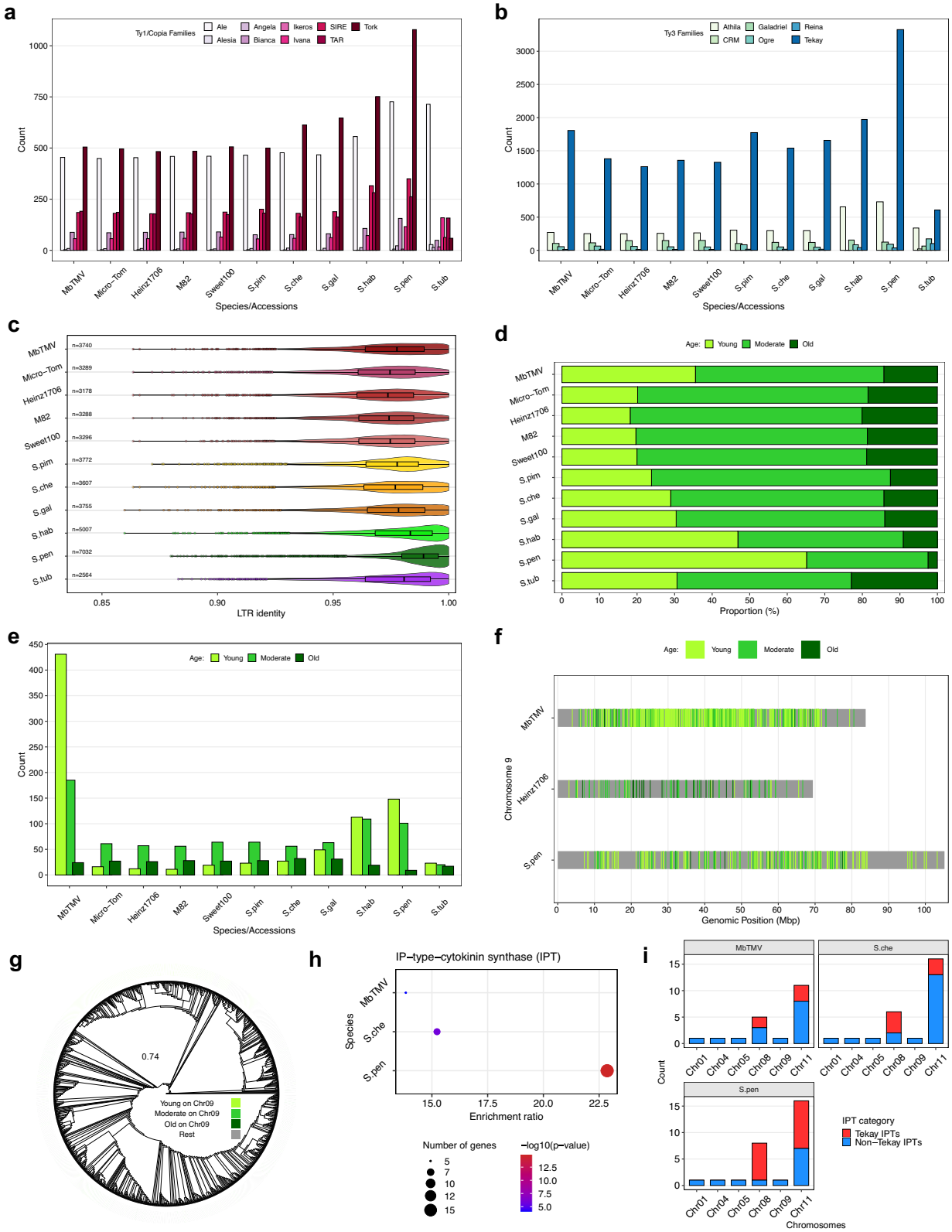
The age of LTR retrotransposons provides insights into transpositional activity in plant genomes[75]. Younger insertions suggest recent bursts of transposition whereas older elements are more degraded[76,77]. Estimating the timing of LTR insertions helps identify genome expansion and could indicate stress-induced mobilization over evolutionary periods[78,79]. We determined the age of the intact LTR retrotransposons by using the LTR identity metric, which shows how much the LTRs have altered since the insertion time. We found that LTR retrotransposons are younger and more recently integrated in wild tomato genomes compared to domesticated accessions (Fig. 4c). Moreover, MbTMV has more young elements compared to the other four domesticated accessions, while *S. pimpinellifolium*, *S.cheesmaniae* and *S.galapagense* have an intermediate age profile compared to *S. habrochaites* and *S. pennellii* that have the most recent integrations (Fig. 4c).

The high abundance of *Tekay* elements in MbTMV, together with the accumulation of many young LTR retrotransposons in this species, led us to investigate whether there is a connection between those characteristics. To this end, we examined the LTR identity of all the LTR retrotransposon lineages in MbTMV separately. We found the *Ogre* lineage is the youngest, while the *Tekay* lineage is the second youngest and contains the most elements (Fig. 4b and Supplementary Fig. 9b and 10a). Generally, *Tekay* is the most successful LTR retrotransposon lineage in the *Lycopersicon* clade[30,31]; hence, we intended to categorize *Tekay* elements based on their integration time into the genome (see Materials and Methods). Utilizing this age classification, we were able to quantify the proportion of different-aged elements in the ten species (Fig. 4d). Correspondingly, the majority of *Tekay* elements in *S. habrochaites* and *S. pennelli* are young, while for the rest of the species, they are of moderate age (Fig. 4d). Interestingly, the proportion of young elements in MbTMV is higher compared to the rest of the domesticated accessions (Fig. 4d). Finally, the fraction of old elements is always the lowest, illustrating the difficulty in identifying TEs after mutation accumulation.

Due to the presence of a 64.1 Mbp introgression from *S. peruvianum* on chromosome 9 of MbTMV that contains the *Tm2*^{2} TMV resistance gene[55], we suspected this could be the source of young *Tekay* elements within MbTMV. We located the young, moderate and old *Tekay* elements on chromosome 9 from the ten accessions, and we found a specific enrichment of more than two-thirds of young *Tekay* elements on MbTMV chromosome 9 that was not identified in the other accessions (Supplementary Fig. 10b) (Fig. 4e). We validated that the majority of the young *Tekay* elements in MbTMV are found in the pericentromeric region of chromosome 9 which is the location of the *S. peruvianum* introgression[55] (Fig. 4f). The abundance and location of young *Tekay* elements is different between the chromosome 9 copies of MbTMV, Heinz1706 and *S. pennellii* (Fig. 4e, f). Finally, a complete phylogenetic reconstruction of all *Tekay* elements in the MbTMV genome was carried out using the integrase gene of each element (Fig. 4g). We found that the majority of young elements on chromosome 9 are clustered together with a maximum likelihood value of 0.74, indicating those sequences were not randomly related (Fig. 4g). Summarizing, the introgression of the TMV resistance gene in MbTMV and all modern commercial tomato varieties has brought with it an abundance of *Tekay* TEs that likely exploded in the wild *S.peruvianum* genome in an independent event from *S. pennellii*.

Due to the independent explosions of *Tekay* LTR retrotransposons in tomato wild relatives we investigated if these events are co-located with specific genes in these species. We thus identified genes located within ±2 kb of, or intersecting with intact, solo LTR, or fragmented *Tekay* elements in MbTMV, *S. cheesmaniae* and *S. pennellii* and performed over-representation analysis of MapMan4 functional categories (Supplementary Table 7).This analysis revealed a pronounced enrichment of cytokinin biosynthesis *ISOPENTENYLTRANSFERASE* (*IPT*) genes[80] in *S. pennellii* (Fig. 4h). Similarly, in *S. cheesmaniae*, although with a lower magnitude compared to *S. pennellii*, *IPT* genes were enriched as well (Fig. 4h). In contrast, MbTMV only showed a non-significant tendency of enrichment of *IPT* genes (Fig. 4h). Further investigation revealed these *IPT* genes and *Tekay* elements are located in

https://doi.org/10.1038/s41467-026-74784-5



complex clusters on chromosomes 8 and 11 in the above species, and particularly the chromosome 11 cluster contributed to the proliferation of *IPT* genes co-located with *Tekay* elements in *S. pennellii* (Fig. 4i). Consistent with the differential expansion of *Tekay* retrotransposons across species—being highest in *S. pennellii*—*Tekay* bursts might have contributed to the genomic enrichment and distribution of *IPT* genes in tomato genomes.

Higher crossover rate in female meiosis underlies female-enhanced recombination regions in tomato intraspecific and interspecific hybrid backcross populations

To investigate the effect of sequence divergence and reproductive gender on crossover number and location, we pollinated *S. lycopersicum* cv. MbTMV with three other parents (*S. lycopersicum* cv. Micro-Tom, *S. cheesmaniae* LA1039 and *S. pennellii* LA0716), generating three

Fig. 4 | Transposable element annotation of the *Lycopersicon* clade reveals independent explosions of *Tekay* retrotransposons. a, **b** Quantification of the intact elements in the main *Ty1/Copia* and *Ty3* lineages. **c** Age of all the LTR retrotransposons (*Ty1/Copia* and *Ty3*) in all the species using the LTR identity *metric*. The boxes represent the interquartile range of the identity values and the thick black lines inside the boxes represent the median identity value. **d** The proportion of *Tekay* elements distinguished in the different age categories (young <0.5 mya, 0.5 > = moderate and old> 1.5 mya) in all the species. For panels **a**–**d** S.pim refers to the *S. pimpinellifolium* accession LA1589. **e** The abundance of the different age categories of *Tekay* elements on chromosome 9 of each species. **f** The distribution of the *Tekay* elements on chromosome 9 of *S. lycopersicum* cv. Moneyberg-TMV (MbTMV), *S. lycopersicum* cv. Heinz (Heinz1706) and *S. pennellii* LA0716 (S.pen). Different colors represent the age category of each element. **g** Phylogenetic tree of the high-quality integrases of the *Tekay* elements. The heatmap on the outer layer shows the age of the element. The tree is rooted using a *Ty1/Copia Tork* element from S.tub (purple block on the right of the heatmap). The bootstrap value for the main cluster of young elements is shown. **h** Over-representation of the IP-type-cytokinin synthase (*IPT*) genes surrounded (-/+ 2 kb) or intersected by intact, fragments or soloLTRs of *Tekay* elements on the same or the opposite strand for *S. lycopersicum* cv. Moneyberg-TMV (MbTMV), *S. cheesmaniae* LA1039 (S.che) and *S. pennellii* LA0716 (S.pen). Over-representation analysis was carried out using one-sided Fisher's exact test. The *p*-values are 7.09E-05 (MbTMV), 1.33E-06 (S.che) and 1.96E-15 (S.pen). **i** The abundance of previously mentioned *IPT* genes in different chromosomes of *S. lycopersicum* cv. Moneyberg-TMV (MbTMV), *S. cheesmaniae* LA1039 (S.che) and *S. pennellii* LA0716 (S.pen).

Table 3 | Genomic percentage of the TE superfamilies in ten *Lycopersicon* accessions and the potato reference genome DM8.1

Species	*Ty1/Copia*	*Ty3*	Unknown LTR	*CACTA*	*Mutator*	PIF/ Harbinger	*TC1/ Mariner*	*hAT*	*Helitron*	Total TE	Genome size scaffolded on chromosomes (Mbp)
***S.lyc*-MbTMV**	5.77	18.63	5.99	4.53	8.66	1.10	1.58	1.13	7.53	54.73	824.45
***S.lyc*-MicroTom**	5.88	18.51	5.15	5.01	9.86	1.48	1.19	0.94	8.50	56.53	812.44
***S.lyc*-Heinz1706**	6.16	18.98	4.88	4.30	8.85	1.74	1.48	1.01	8.19	55.59	800.12
***S.lyc*-M82**	5.87	18.89	5.26	5.25	9.47	1.56	1.18	0.85	8.42	56.74	801.96
***S.lyc*-Sweet100**	6.12	18.41	6.26	4.69	9.26	1.30	1.39	1.21	7.80	56.45	805.19
***S.pim*-LA1589**	5.70	20.40	7.13	5.10	7.24	1.08	0.83	0.96	8.29	56.73	823.74
S.che	6.26	17.78	6.35	6.14	7.83	1.76	1.22	0.92	9.06	57.32	814.47
S.gal	5.42	17.40	6.52	6.53	8.34	1.49	1.34	0.98	9.76	57.77	811.44
S.hab	6.10	19.86	5.69	6.93	9.74	2.15	1.12	1.17	6.90	59.64	893.57
S.pen	5.48	22.13	6.79	4.11	14.88	1.12	1.07	1.37	8.04	64.99	1072.33
S.tub	5.00	24.43	11.15	1.93	4.46	0.39	2.07	1.92	5.53	56.88	737.87

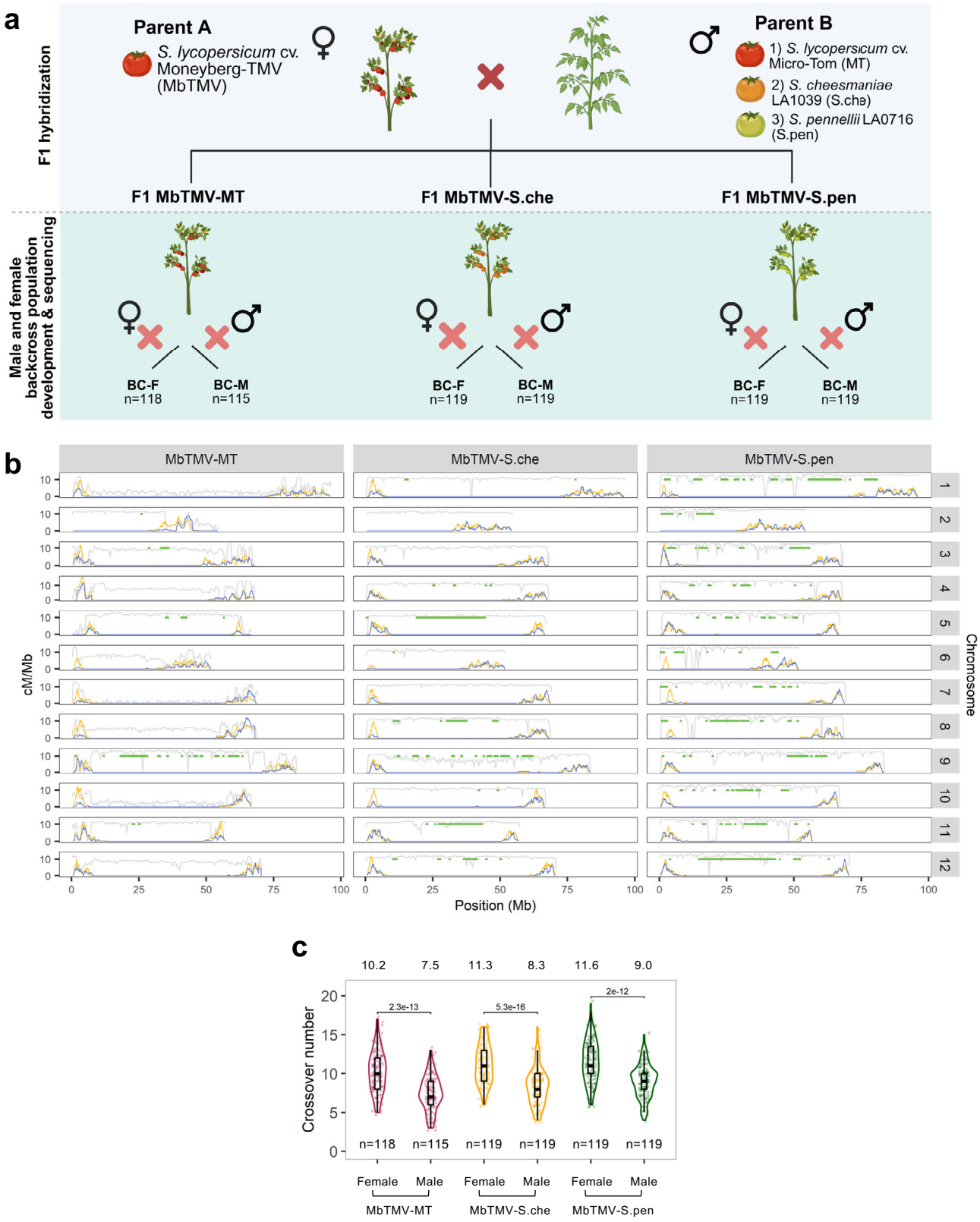
Table 3 shows the percentage of the scaffolded genome made up of the main TE superfamilies, including LTR retrotransposons and DNA transposons, in ten *Lycoperiscon* accessions and the potato reference genome DM8.1. In addition, the total TE percentage and scaffolded genome size in Mbp for each accession are presented. The TE annotation was performed using EDTA (v.2.0.1) (Ou et al.).

sequence-divergent F1 hybrids with increased amounts of genomic variation (Fig. 5a, Supplementary Table 8). Next, we generated three male and three female backcross populations derived from the intra- and interspecific F1 hybrids (See Materials and Methods). From the six backcross populations a total of 709 individuals were subjected to low-coverage sequencing resulting in 2.5-5 Gbp data per individual sample (Supplementary Fig. 11-13).

A total of 6853 crossovers were detected from all the samples, combining the female and male populations that represent recombination in female and male meiotic cells, respectively. We analyzed recombination landscapes for the six backcross populations (Fig. 5b) and found that recombination is skewed to the distal euchromatic arms in all cases, consistent with previous observations in tomato[81–83]. We defined recombination coldspots in each population (Supplementary Data 4) and found that certain functional gene classes were over-represented in recombination coldspots in some populations, including some RNA biosynthesis and cytokinin biosynthesis related terms (Supplementary Table 9). We found that all three female backcross populations have significantly more crossovers than their counterpart male backcross populations (Fig. 5c), which is consistent with previous findings on heterochiasmy in a *S. lycopersicum* x *S. pennellii* hybrid[53]. The respective female populations have 36% (MbTMV x Micro-Tom; hereafter MbTMV-MT), 35% (*S. lycopersicum* x *S. cheesmaniae*; hereafter MbTMV-S.che) and 28% (*S. lycopersicum* x *S. pennellii;* hereafter MbTMV-S.pen) higher CO number than the counterpart male populations (Fig. 5b). The consistently higher CO rate in the three female populations suggests that heterochiasmy is conserved in intraspecific hybrid tomato and minimally divergent interspecific hybrid tomato. Although we observed higher CO numbers in interspecific than intraspecific populations, some COs in MbTMV-MT backcross offspring may not be detected due to the absence of SNPs in some genomic regions that are identical by descent (Fig. 5a, b). Compared to the previous study[53], we now have high-resolution mapping of crossover sites that enable examination of recombination sites in finer detail, revealing female-enhanced recombination regions (FRRs) such as those in the short arms of chromosomes 1, 6, 7, 8 and 10 (Fig. 5b, Supplementary Fig. 14 and Supplementary Data 5). Comparing the overall recombination landscapes, heterochiasmy in tomato appears to be confined to specific genomic regions and is not a genome-wide effect. We selected the MbTMV-S.pen female population and examined the allele ratios in the offspring populations and found no evidence of segregation distortion across the whole genome. We found that certain functional gene classes were over-represented in FRRs (Supplementary Table 10). Overall, our analysis identified megabase-scale FRRs in three different tomato hybrids that appear to be over-represented on short chromosome arms.

Apart from the effect of reproductive gender on crossover number, we found that the genetic differences between the parental genomes also influence the distribution of COs along the chromosomes. Between 67% (MbTMV-MT) and 80% (MbTMV-S.pen) of the tomato genome is suppressed of recombination, limiting the genetic shuffling and resulting in linkage drag (Fig. 6a). Among these CO-suppressed regions, which we call coldspots, about 64% (529 Mb) are conserved across all three hybrid backgrounds. Conserved coldspots, spanning

https://doi.org/10.1038/s41467-026-74784-5



tens of megabases, are mostly in the pericentromeric regions while smaller cross-specific coldspots are located at the distal chromosomal ends (Fig. 6b). There is a significant depletion of COs in regions with structural variants (p value < 0.0001, Fisher's exact test), concordant with previous studies in *A. thaliana* and tomato[44,45,83]. The elevated proportion of coldspots in both MbTMV-S.pen backcross populations is partly due to the presence of large inversions between the *S. lycopersicum* and *S. pennellii* genomes (Supplementary Fig 15). The short arms of chromosome 6 and 7 contain inversions between the *S. lycopersicum* and *S. pennellii* genomes that suppressed recombination, consistent with previous observations[74,83]. As reported above, we discovered a 2.2 Mbp inversion in chromosome 8 between the eight *Lycopersicon* genomes and *S. pennellii* and found that in the MbTMV-S.pen population where this inversion is heterozygous no COs occur. In

Fig. 5 | Higher crossover rate in female meiosis underlies female-enhanced recombination regions in tomato intraspecific and interspecific hybrid backcross populations. a Generation of female and male backcross populations for genome sequencing. Created in BioRender. van Rengs, W. (2026) (https://BioRender.com/fz42vck). **b** Recombination landscape in female (yellow) and male (blue) backcross populations. Gray lines indicate the distribution of SNP (*log* scaled), while green *horizontal* lines show the locations of large inversions that are larger than 100 kbp. Some distal regions in MbTMV-MT such as chromosome 2 and 5 have low SNP density, limiting crossover detection. **c** Female backcross populations have significantly higher number of crossovers than male populations (two-sided Wilcoxon Rank Sum Test with Bonferroni correction). Box plots show the median (center line), the 25th and 75th percentiles (box limits), and whiskers extending to the most extreme observations within 1.5× the interquartile range. Individual samples are shown as points. The mean crossover number per population is specified on top of the graph and *p*-values are reported above the three horizontal bars.

contrast, the same region in the other two hybrids, where no inversion is present, meiotic COs can occur (Fig. 2a, Fig. 6c and Supplementary Fig. 16). This inversion overlaps 268 *S. lycopersicum* and 278 *S. pennellii* genes, implying that both agronomically desirable and detrimental alleles are inherited together. The assembly of *S. pennellii* genome not only enables the discovery of many other unknown SVs relative to *S. lycopersicum* but also reveals massive linkage drag that may result from introgressing *S. pennellii* segment into the domesticated tomato.

Retrotransposon expansions and genome inversions associated with recombination coldspots in a divergent tomato hybrid

Similar to SVs, retrotransposons such as *Gypsy* and *Copia* have been associated with CO suppression as well[84–89]. To further determine the influence of SVs and TEs in patterns of COs across all three populations, we identified the genomic positions of the coldspots relative to the assemblies of each parent. We found that some coldspots have different lengths based on one of the parental genomes, implying that the genome may contain deletions or insertions or may have undergone repeat expansion or contraction (Fig. 6d and Supplementary Fig. 17). MbTMV-MT coldspots are located in regions where MbTMV and MicroTom are syntenic but have high density of *Gypsy* and *Copia*. On the contrary, several MbTMV-S.che and MbTMV-S.pen coldspots overlap regions with chromatin expansions or inversions between the parents. Compared with Mb-MT and MbTMV-S.che, the retrotransposon-associated coldspots in MbTMV-S.pen contain significantly more retrotransposons (Supplementary Fig. 18), consistent with our results above about the expansion of *Tekay* elements in *S. pennellii* compared to other tomato and wild species. Apart from repeat elements, some coldspots contain insertions or translocations that may affect synapsis during pachetene, which in turn prohibits recombination just like in *A.thaliana*[44,45,90]. Moreover, some of these insertions overlapping Mb-Pen coldspots contain hemizygous genes (Fig. 6e), which is important for breeding because these genes may confer new functions but at the same time are tightly linked. Utilizing our high-quality genomes, we identified large structural genomic variants and differential TE composition between parental genomes that directly suppressed recombination. Our results add further resolution and insights to previous findings on heterochiasmy, tightly linked genes[53] and suppressed recombination between *S. lycopersicum* and other wild species[74,91].

Discussion

Here, we have assembled de novo chromosome scale genomes of *S. cheesmaniae* LA1039 and *S. pennellii* LA0716. We compared our two new assemblies with high-quality assemblies of ten related accessions and species, and extensively analysed the TE content of all genomes. Finally, three different hybrids (*S.pennellii*, *S. cheesemaniae* and *Solanum lycopersicum* cv. Micro-Tom crossed with *Solanum lycopersicum* cv. Moneyberg-TMV) were used to develop large recombinant offspring populations to reveal the genomic basis of heterochiasmy in hybrid tomato.

An improved *S. pennellii* LA0716 assembly and a *S. cheesmaniae* LA1039 assembly

Our improved assembly of *S. pennellii* LA0716 integrated an additional 145.9 Mbp to the 12 chromosomes including in regions that are rich in repetitive elements (Fig. 1). Sequencing data coverage analysis showed that the improved assembly had fewer deviating regions, unmapped and supplementary reads compared to the original assembly[3]. The improved *S. pennellii* assembly contains inversions at the start of chromosomes 6 and 7 that were not present in the original assembly but are confirmed by BAC-FISH results[74]. The improved assembly also includes a 7.1 Mbp inversion on chromosome 3 and a 2.2 Mbp inversion at the start of chromosome 8 that are validated by Hi-C and our recombination mapping results (Fig. 1a, Fig. 5b, Supp). The *S. cheesmaniae* LA1039 assembly gives new genomic insights into an understudied species that contains useful salinity stress tolerance that could be utilized for the improvement of the cultivated tomato[14]. As a relatively close relative of the cultivated tomato the *S. cheesmaniae* assembly shows that in the euchromatic arms are largely syntenic to *S. lycopersicum* but we did discover a 35.7 Mbp inversion in the pericentromeric region of chromosome 5 (Fig. 2a). Overall, the *S. pennellii* LA0716 and *S. cheesmaniae* LA1039 genomes improve structural information and gene identification in those accessions, and can facilitate breeding in interspecific crosses.

Long reads as a lens for the structural evaluation of the *Lycopersicon* clade

Over the last ten years, developments in long-read sequencing technologies have revolutionized the field of genomics[19,92–95]. PacBio HiFi produces highly accurate reads, limited in size (generally 20 kb), whereas ONT reads are somewhat less accurate but generally longer (up to 200 kb). Structural variants can be detected directly from long-read sequencing[96], however, genome assembly is needed to properly detect and characterize SVs in the megabase scale[19,55]. Recently, the construction of a tomato super-pan genome assembled from legacy PacBio data has led to insights in SVs within the whole tomato clade and discovery of a wild tomato gene with the potential to increase yields[13]. However, our results show PacBio HiFi and ONT ultra-long sequencing can lead to improved genome assemblies, and it is likely ONT ultra-long sequencing will prevail in challenging repetitive genomes in future. High quality de novo genomes have been assembled from HiFi[8,97] or ONT[10,19], however a combination of both approaches can overcome extended repetitive regions and generate de novo assemblies of the highest quality[94,95,98].

We have used a combination of long-read sequencing data with chromosome confirmation capture data yielding high-quality chromosome scale reference genomes which enabled the accurate detection of large structural variants spanning several megabases. Thus, we identified several megabase-scale inversions, including a roughly 20 Mbp inversion on chromosome 5 that is specific to the Galapagos tomato species *S. galapagense* and *S. cheesmaniae* (Fig. 2). Our finding that the complete chromosomes 6, 7 and 8, inclusive of the pericentromeric regions, are structurally almost identical in all five cultivated tomatoes is consistent with the thesis that these chromosomes contain a key set of agriculturally relevant genes. Short-read sequencing has shown chromosomes 6, 7 and 8 have low nucleotide diversity in fresh consumption and modern processing tomatoes[99], while low numbers of SVs were detected in cultivated tomatoes in the pericentromeric regions of the same chromosomes[19]. Notably chromosome 7 contains several sweep loci including *fw7.2*, while chromosome 6 additionally contains a fruit weight locus embedded within the pericentromere[100].

https://doi.org/10.1038/s41467-026-74784-5

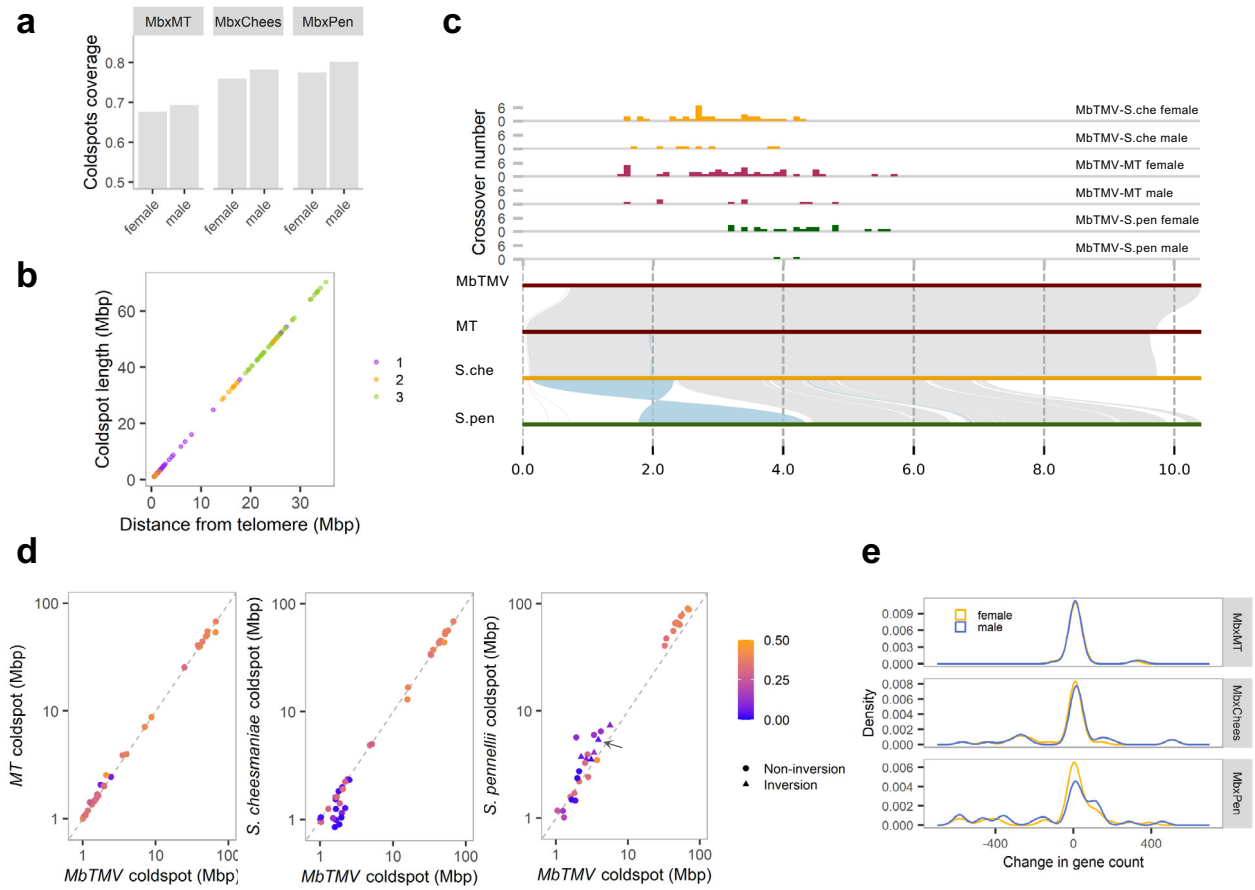


Fig. 6 | Retrotransposon expansions and genome inversions associated with recombination coldspots in a divergent tomato hybrid. a Proportion of coldspots in different hybrid crosses expressed as a percentage of the MbTMV genome. **b** Relative position of cross-specific and conserved coldspots across the three hybrid genotypes. The color indicates the number of hybrid genotypes where the coldspot is observed. **c** A 2.2 Mbp *inversion* suppresses recombination between MbTMV and *S. pennellii* chromosome 8 short arm, whereas other hybrid populations exhibit recombination in the same region. **d** Size of male coldspot regions relative to each parental genome. Deviation from the diagonal line indicates expansion or contraction of genomic segments, while the color indicates the total proportion of *Gyspy* and *Copia* elements relative to the MbTMV genome. We define inversion-associated coldspot as having more than 50% inversion coverage. Orange dots in the upper right of each plot represent large coldspots in the pericentromeres. The arrow points to the chromosome 8 coldspot illustrated in (**b**). **e** Change in coldspot gene counts between the parental genomes (Δ *gene count* = *MbTMV* − Parent 2). Cross-specific expansions and translocations in the chromosome arms contributed to the changes in the distributions in MbTMV-S.pen.

Our results have shown pairwise comparison of de novo reference quality chromosome scale genomes leads to full characterization of genomic variants between species along the chromosomes. Our high-quality genomes further the recently constructed tomato super-pan genome and facilitate the characterization of genomic variation between cultivated tomato and wild related species[12,13].

Diversity of TE and repeat profiles in the *Lycopersicon* clade

Transposons and repeats play a crucial role in plant evolution, diversity, and adaptation. We demonstrated that the repeat component is higher in wild relatives of the domesticated tomato (Fig. 1e, f and Supplementary Fig. 5), and this includes TE content regardless of class (Fig. 4a, b and Supplementary Fig. 9). Most of the LTR retrotransposons in these species have recently been integrated into the genome (Fig. 4c and Supplementary Fig. 10a), and the *Tekay* lineage is the most abundant and one of the youngest across the *Lycopersicon* clade (Supplementary Fig. 10a). The *S. lycopersicum* cv. MbTMV has more *Tekay* elements compared to the other domesticated accessions specifically on chromosome 9 (Fig. 4d and Supplementary Fig. 10b). MbTMV is known for resistance to tobacco mosaic virus (TMV), attributed to the *S. peruvianum* introgression on chromosome 9[55]. Consequently, along with the resistance trait, a significant number of *Tekay* elements were introduced (Fig. 4f), and phylogenetic analysis indicated that most of the young elements on chromosome 9 share a common origin (Fig. 4g). When analysing the co-location of *Tekay* elements with genes we found a strong over-representation of, and increase in copy number of, cytokinin biosynthesis *IPT* genes in the wild species, especially in *S. pennellii* (Fig. 4h, i). Young *Tekay* elements were identified close to *IPT* genes in *S. pennelli* but not in MbTMV and *Solanum cheesemaniae*. Cytokinin biosynthesis IPT proteins have been identified as rate-limiting factors in the production of cytokinin phytohormones and the modification of their expression can improve abiotic stress tolerance in diverse crop species, including cultivated tomato[80,101,102]. Given the variation in *IPT* gene copy number and TE presence/absence and insertion time in regulatory space, it will be interesting to explore the functional role of this amplified gene family in the abiotic stress tolerance of *S. pennellii* (accession LA0716)[3] and *Solanum cheesmaniae* (accession LA1039)[14].

Overall, we found that the total amount of TEs and repeat content is lower in cultivated tomato accessions compared to their wild relatives. This may be linked to the domestication of tomato and increased inbreeding[103]. Similarly, wild rice species exhibit extensive variation in TE content that correlates with genome size differences and LTR retrotransposon expansions, compared to the domesticated rice, which

has more constrained TE landscapes reflecting both TE amplification in wild lineages and TE reduction during the domestication and selection process[104]. Targeted comparisons between cultivated and wild rice suggest that artificial selection during domestication may have purged TE insertions that negatively influence gene function or expression stability[105]. In essence high mobilome activity may compromise the long-term viability of cultivated accessions. Conversely, introgression of genetic material from wild germplasm may serve to inadvertedly introduce young TEs into the breeding germplasm, as in the case of MbTMV, and this warrants further investigation.

Recombination analysis in sequence-divergent tomato hybrids

Understanding where meiotic crossovers (COs) occur is crucial for understanding the evolution of sexually reproducing species and for optimizing crop breeding. Hybridization of three sequentially divergent species (*S. lycopersicum* cv. Micro-Tom, *S. cheesmaniae* LA1039 and *S. pennellii* LA0716) to *S. lycopersicum* cv. Moneyberg-TMV and generation of six derived male and female backcrosses populations (Fig. 5a) allowed for investigation of genomic polymorphisms on recombination distribution and heterochiasmy. We have found that large recombination coldspots are mostly located in heterozygous inversions or retrotransposon-rich pericentromeric regions. Through offspring sequencing we can only detect meiotic recombination events that are inherited through viable gametes and it is plausible that crossovers within inversions could produce unviable gametes due to non-reciprocal exchanges. The divergent MbTMV-S.pen hybrid populations have less sites for recombination, consequently limiting the shuffling of alleles in the short arms of chromosome 6, 7 and 8 that are coldspots and contain agronomically relevant genes (Fig. 5b). Knowing the patterns of recombination in a specific hybrid cross enables careful inspection of potential regions to introgress from wild species into breeding lines, revealing possible consequences of tight linkage such as those due to large inversions.

The heterochiasmy observed in all three tomato F1 hybrids where meiotic CO rate is higher in female meiosis than male meiosis suggests it is likely universal throughout the tomato clade, and is consistent with a previous report in a *S. lycopersicum* x *S. pennellii* hybrid[53]. Opposite to tomato, *A. thaliana* has higher male recombination rate than female, in both inbreds and hybrids[49–52], whereas heterochiasmy is not observed in the maize B73 x Mo17 hybrid[106,107]. We noticed that female-enhanced recombination regions (FRRs) were more abundant on the physically shorter chromosome arms in the tomato genome (Fig. 5b). In male meiosis, the longer, more gene-rich, arms accumulate crossovers more easily, whereas in female meiosis, where total crossover number is higher, the shorter arms also begin to accumulate crossovers. This observation is potentially consistent with the recently proposed coarsening model for meiotic crossover interference where well-spaced crossover-promoting HEI10 foci determine the crossover landscape[108–110]. In future, the role of *HEI10* expression level, synaptonemal complex and chromatin structure could provide insight into the basis of heterochiasmy in tomato plants. Moreover, it is clear from our results that tomato introgression breeding should clearly focus on using hybrid (or partially hybrid) genotypes as mother, and the recurrent inbred parent as father, in order to maximize crossover number and exploit female-enhanced recombination regions to expedite breeding.

Methods

Plant materials and growth

Seeds of S. lycopersicum cv. Moneyberg-TMV, *S. lycopersicum* cv. Micro-Tom and *S. cheesmaniae* LA1039 were pre-soaked in water for 2 hours, followed by 15 min incubation in Na_3PO_4, 30 min incubation in 1:1 (volume:volume) with water-diluted household bleach (DanKlorix 2.8 g NaOCl per 100 g liquid) and at least 3 subsequent washes with sterilized water. Seeds of *S. pennellii* LA0716 were soaked in 700 ppm GA3 for 20 hours, followed by 30 min incubation in 1:1 (volume:volume) with water-diluted household bleach (DanKlorix 2.8 g NaOCl per 100 g liquid) and at least 3 subsequent washes with sterilized water. All seeds were sown on 0.8% agar and incubated in growth chambers with 16 h daylight, and 25 °C daytime and 21 °C nighttime temperatures. For seed origin information, please see the acknowledgements.

Seedlings of *S. lycopersicum* cv. Moneyberg-TMV, cv. Micro-Tom and *S. pennellii* LA0716 were transplanted to MPIPZ greenhouses and grown during late spring of 2020 under natural light supplemented with artificial light to ensure 16 h light per day. Seedlings of *S. cheesmaniae* LA1039 were transplanted to Bronson growth chambers and grown in short-day conditions (12 hours of daylight) with 28 °C daytime, 22 °C night-time and 60-80% relative humidity, to mimic native growth conditions.

Plant hybridization

Immature unopened green to yellowish Moneyberg-TMV flowers were carefully emasculated using forceps and pollinated with freshly extracted pollen of the other parent (e.g., Micro-Tom, *S. cheesmaniae* LA1039 and *S. pennellii* LA0716) 1 to 3 days after emasculation. Pollinated Moneyberg-TMV inflorescences were labeled with crossing dates and left to develop fruits. Mature fruits were collected, after which seeds were extracted and cleaned in a 1:1 (volume:volume) pulp to 2% HCl solution for 1 hour, followed by rinsing with excess water and drying at room temperature.

To overcome possible immature fruits, Moneyberg-TMV x *S. pennellii* LA0716 were collected 28-33 days after pollination, sterilized in 6% NaOCl for 10 minutes, followed by 3 consecutive washes in sterile water. Embryos were carefully removed from the endosperm using a binocular and transferred to 0.5 MS including vitamins (Sigma) agar plates containing 10 g sucrose/L, followed by incubation in growth chambers with 16 h daylight, and 25 °C daytime and 21 °C nighttime temperatures.

To generate backcross populations, the three F1 hybrid genotypes were cultivated in the greenhouse and used for reciprocal crosses with *S. lycopersicum* cv. Moneyberg-TMV, apart from the *S. lycopersicum* Moneyberg-TMV x *S. pennellii* LA0716 female population where *S. pennellii* LA0716 was used as pollen donor (*S. lycopersicum* Moneyberg-TMV x *S. pennellii* LA0716 does not accept pollen from *S. lycopersicum* Moneyberg-TMV due to unilateral interspecific incompatibility)[111].

High Molecular Weight (HMW) DNA extraction and sequencing

HMW DNA of *S. cheesmaniae* LA1039 and *S. pennellii* LA0716 was isolated from 1.5 g of young leaf material with a NucleoBond HMW DNA kit (Macherey Nagel, Düren, Germany). DNA quality was assessed with a FEMTOpulse device (Agilent, Santa Clara, CA, USA), and quantity was measured using a Quantus Fluorometer (Promega, Madison, WI, USA). HiFi libraries were prepared according to the manual "Procedure & Checklist – Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0" with an initial DNA fragmentation by Megaruptor 3 (Diagenode) and final library size binning into defined fractions by SageELF (Sage Science). Size distribution was again controlled by FEMTOpulse (Agilent). Size-selected libraries with expected insert sizes of 20 and 27 Kbp for LA1039 and 19 and 25 Kbp for LA0716 were independently sequenced on a single SMRT cell of a Pacific Biosciences Sequel II on Sequel device at the MPGC Cologne with Binding Kit 2.0 and Sequel II Sequencing Kit 2.0 for 30 h (Pacific Biosciences).

K-mer analysis of HiFi data was performed using the k-mer Analysis Toolkit (KAT) hist v2.4.1 with options -m 21[112] and plotted using R v.4.2.3 (https://cran.r-project.org/) and packages from the Tidyverse collection v2.0.0[113] and utilizing the maximum number of distinct k-mers within the homozygous peak to scale the *y*-axis. Additionally K-mer analysis of HiFi data was carried out using GenomeScope2[114] v2 with the parameters k-mer length 21, ploidy 2 and max k-mer coverage

−1. The average k-mer coverage for polyploid genome parameter was set to 29, 32 and 23 for *S. lycopersicum* cv. Micro-Tom, *S. cheesmaniae* LA1039 and *S. pennellii* LA0716, respectively, based on their sequencing coverage.

High-molecular weight DNA that was used for PacBio sequencing was size selected using the Circulomics Short-Read Eliminator XL Kit (Circulomics Cat# SKU SS-100-111-01). The size-selected DNA was used as an input for 10 library preparations with the Oxford Nanopore SQK_LSK112 ligation sequencing kit. Sequencing was performed on an Oxford Nanopore PromethION on 10 Flowcells FLO-PRO112 (R10.4) and sequenced for up to 72 h with reloading and washing where needed. ONT-basecalling was performed with "super accuracy" model (dna_r10.4_e8.1_sup) using guppy_6.0.1/2 or guppy_6.1.1 (https://nanoporetech.com/software/other/guppy/).

Chromatin conformation capture libraries were prepared using 0.5 grams of young leaf material as the input. All treatments were according to the recommendations of the kit vendor (Omni-C, Dovetail) for plants. As a final step, the Illumina-compatible libraries were prepared (Dovetail) and test sequenced (paired-end 2 ×150 bp) on an Illumina NextSeq 2000 device at the MPGC in Cologne, followed by sequencing on a NovaSeq 6000[115] at Novogene for increased coverage.

DNA extraction and sequencing of BC populations

Young true leaf material was collected and lyophilized overnight in an Alpha 1-4 freeze dryer (Martin Christ GmbH). DNA extraction was performed on the KingFisher robot (Thermofisher) using the NucleoMag Plant Kit (Macherey and Nagel, P/O 744400.4) and eluted in 200 µl buffer. Transposase-based libraries (Tn5, Nextera) were prepared according to[44] and test sequenced on an Illumina NextSeq 2000 in the MPGC Cologne. BC samples were outsourced to Novogene and sequenced on the NovaSeq 6000, aiming to generate ~3, 4 Gb of paired-end data per sample. Sequencing data were provided as deduplicated and adapter-removed reads in zipped FASTQ format.

Sequencing data statistics were obtained using seqkit stats v2.0.0[116] with options -a and FastQC with default settings (Andrews S. FastQC: a quality control tool for high-throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Genome assembly

Assembly: Hifiasm v0.16.1-r375[63] was used with option -l0 to assemble the HiFi reads of LA1039. LA0716 ONT reads were filtered for length (> 90 Kbp) and quality (q > 90) using Filtlong-0.2.1 (--min_length 90000 --min_mean_q 90). Hifiasm v0.17.6-r460[63] was used to assemble HiFi and filtered ONT reads with options -ul.

Scaffolding: The *S. cheesmaniae* LA1039 and *S. pennellii* LA0716 assemblies were scaffolded following the esrice/hic-pipeline (https://github.com/esrice/hic-pipeline). First, Omni-C (Dovetail) paired-end reads were mapped separately using Burrows-Wheeler Aligner v0.7[117] and the filter_chimeras.py script. Next, the mapped reads were combined using the combine_ends.py script, with three iteration steps and a minimum mapping quality value of 20 (default), followed by adding read mate scores, sorting and removing duplicate reads using SAMtools v1.9[118]. The resulting BAM file was converted to a BED and sorted (-k 4) using BEDtools v2.30[119]. Per assembly, we ran one single round of Salsa v2.2[120] with optional settings: -e DNASE -m yes -p yes. A modified version of the convert (convert.sh) script was used to convert the Salsa2 output to a Hi-C file, which was utilized within a local installation of Juicebox v.1.11.08 (https://github.com/aidenlab/Juicebox) to generate a Hi-C contact plot.

D-genies alignment to MbTMV: The *S. cheesmaniae* LA1039 and *S. pennellii* LA0716 genome assemblies and scaffolded genome assemblies were aligned to MbTMV[55] using Minimap2 2.24-r1122[121] with option -x asm5 to generate pairwise alignment format files. Used reference and query genomes were indexed using a Python-based custom D-genies index script, followed by generating dotplots using D-GENIES version 1.4.0[122].

Manual fine-tuning: Assemblies scaffolded with the automated Salsa2 pipeline were further fine-tuned, whereby smaller scaffolds were manually placed within larger scaffolds, based on Hi-C interaction and alignment to MbTMV. Fused chromosome-scale scaffolds, observed for LA0716 (chromosomes 1 and 2), likely a result of interaction between chromosome-wide repetitive elements (e.g., LTR-RTs) or specific repeat regions (e.g., TGR1) present in tomato[55,123], were manually broken at the ends of each chromosome based on Hi-C interaction and alignment to MbTMV. Multiple rounds of fine-tuning were performed, whereby changes in each round were kept as AGP files, followed by conversion into FASTA files using RagTag v2.1.0 agp2fasta[54] with gap sizes of 100 N's connecting scaffolds. After each round, the generated FASTA files were checked by alignment to MbTMV and Hi-C interaction as described previously. Broken contigs that were scaffolded after another were checked for continuity by alignment of HiFi reads using pbmm2 v1.7.0 (https://github.com/PacificBiosciences/pbmm2) and ONT reads in the case of LA0716 using Minimap2 version 2.24-r1122[121] as described below and repaired if supported by read continuity.

Chloroplast genome: The *S. lycopersicum* MbTMV, *S. cheesmaniae* LA1039 and *S. pennellii* LA0716 chloroplast genome assemblies were generated using the tool ChloroFORGE v1.0 (https://github.com/usadellab/ChloroFORGE).

Validation of genome assemblies

Statistics: For the validation of the 12 assembled chromosomes, we checked assembly statistics, BUSCO completeness, LTR assembly index, and both base quality and k-mer completeness. Quast v5.0.2[124] and GAAS v1.1.0 (https://github.com/NBISweden/GAAS) were used to calculate statistics on fasta files. BUSCO was calculated using BUSCO v5.2.1[125] depending on hmmsearch v3.1, and metaeuk v4.a0f584d was used with lineage datasets solanales (https://busco-data.ezlab.org/v5/data/lineages/solanales_odb10.2020-08-05.tar.gz) to obtain evolutionarily informed expectations of gene content. To assess the LAI of each genome assembly, LTR retriever v2.9.0[64] was run with recommended default settings invoking GenomeTools v1.6.2 (https://github.com/genometools/genometools) with options -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes and LTR_FINDER_parallel v1.2 (https://github.com/oushujun/LTR_FINDER_parallel) with options -threads 10 -harvest_out -size 1000000 -time 300. Base accuracy (QV) and k-mer completeness were assessed using Merqury v1.3[66] with default settings, using the read databases and genomes. The read database for the Merqury run was generated using the meryl algorithm v1.3, with k-mer size equal to 20 and default parameters, which is also part of the Merqury genome evaluation tool[66].

Read alignments & coverage analysis: PacBio HiFi data were mapped to each corresponding genome using pbmm2 (version 1.7.0) with the preset 'CCS' (https://github.com/PacificBiosciences/pbmm2). ONT data were mapped to each corresponding genome using Minimap2 version 2.24-r1122[121] with the preset 'map-ont' and 'eqx', followed by converting, sorting and indexing of the output file using Samtools v1.9[118]. Primary read coverage (default) was extracted in 100-kbp windows using Mosdepth v0.3.2[126] and plotted in the middle of each interval using R v.4.2.3 (https://cran.r-project.org/) and packages from the Tidyverse collection v2.0.0[113].

Re-mapping Omni-C: Following the esrice/hic-pipeline (https://github.com/esrice/hic-pipeline), we separately mapped Omni-C (Dovetail) paired-end reads to each genome using Burrows-Wheeler Aligner v0.7[117] and the filter_chimeras.py script. Mapped reads were combined using the combine_ends.py script, with three iteration steps and a minimum mapping quality value of 20 (default), followed by adding read mate scores, sorting and removing duplicate reads using

SAMtools v1.9[118]. The resulting BAM file was converted to a BED file and sorted (-k 4) using Bedtools v2.30[119]. We ran one single round of Salsa v2.2[120] with optional settings: -e DNASE -m yes -p yes. A modified version of the convert.sh script (https://github.com/marbl/SALSA) was used to convert the Salsa2 output to a Hi-C file, which was used within a local installation of Juicebox v.1.11.08 (https://github.com/aidenlab/Juicebox) to generate a Hi-C contact plot.

Alignment to MbTMV: The finalized LA1039 and LA0716 genomes were aligned to MbTMV[55] using Minimap2 2.24-r1122[121] with option -x asm5 to generate pairwise alignments. Used reference and query genomes were indexed using a Python-based custom D-genies index script, followed by generating dotplots using D-GENIES version 1.4.0[122].

Annotation of sequence features

Nucleotide composition: To obtain nucleotide composition in 100kbp windows of each genome, we first converted the FASTA index file to a genome.txt file by extracting only the first two columns, including the sequence identifier and length. Next, we used BEDtools makewindows command v.2.29.0[119] to generate 100kbp windows in BED format. Finally, we extracted the nucleotide composition using seqtk comp command v.1.3 (https://github.com/lh3/seqtk) with each respective FASTA file and 100kbp window BED file as inputs. Nucleotide composition was converted to a ratio relative to each other nucleotide and plotted on the midpoint of each interval using R v.4.0.3 (https://cran.r-project.org/) and packages from the Tidyverse collection v2.0.0[113].

Simple sequence repeats: Simple sequence repeats (SSR's) were identified using MISA (https://github.com/schellt/msat-loci-identification) running perl (v.5.24.1) invoking the following default settings: definition (unit_size, min_repeats): 1-10 2-6 3-5 4-5 5-5 6-5, interruptions (max_difference_between_2_SSRs): 100, GFF: true. Generated GFF files were modified using a series of Linux and R commands, where identified repeats were summarized in 100Kbp windows per class (e.g., monomer, dinucleotide repeat, trinucleotide repeat, etc.) and plotted on the midpoint of each interval using R v.4.0.3 (https://cran.r-project.org/) and packages from the Tidyverse collection v2.0.0[113].

Long terminal repeat retrotransposons (LTR-RT): The reported LTRs by the LTR-retriever and LAI pipeline "pass.list.gff3" were extracted using a custom pipeline invoking grep and awk commands. In short, we selected for reported LTRs on chromosomes 1 to 12 and extracted reported chromosome information, repeat region, starting position and end position which were used for length calculation before plotting the distribution over the genomic position using R v.4.0.3 (https://cran.r-project.org/) and packages from the Tidyverse collection v2.0.0[113].

Gene annotation: An integrated pipeline composing both ab-initio and evidence-based approaches was employed for MbTMV and LA0716 structural annotation.

The ab-initio predictions were performed for MbTMV, LA1039 and LA0716 genomes using Helixer v0.3.3 with the model land_plant_v0.3_a_0080.h5, integrating deep learning and Hidden Markov Models to enable accurate ab-initio gene annotations[127], with subsequence-length set to 504000. For the post-processing into primary gene models with HelixerPost, the parameters window_size 100, edge_thresh 0.1, peak_thresh 0.80, and min_coding_length 60 were applied. For LA0716, HelixerPost was utilized using peak_thresh 0.60.

For evidence-based gene predictions for MbTMV and LA0716 genomes, a multi-step approach which involved both short- and long-read datasets was employed. In brief, publicly available short-read RNA-seq data (Supplementary Table 11) were mapped to the genomes using Hisat2 v2.2.1[128] with the parameter –dta. PacBio Isoseq reads (Supplementary Table 12) were mapped to the respective genome using minimap2 v2.28[121] with parameters -ax splice:hq and -uf. Subsequently, all of the mapped reads were merged using SAMtools v1.20[118] and sorted with -m 5 G and 3 G for MbTMV and LA0716, respectively. Finally, transcript assemblies were generated using StringTie v2.2.3[129] with --mix -L -j 3 -f 0.1 as well as -c 10 and -c 50 for MbTMV and LA0716, respectively.

Subsequently, for the filtering of valid splice junctions from pre-aligned RNA-seq data for MbTMV and LA0716 genomes, Portcullis 1.2.4[130] was used.

Finally, evidence from both the ab-initio and homology-based gene models, together with validated splice junctions, was integrated using Mikado v2.3.3[131]. For MbTMV, weighting factors favoring ab-initio evidence (three for Helixer and one for StringTie) resulted in a reliable consensus gene model with high accuracy and completeness. For LA0716, weighting was balanced for ab-initio and homology-based evidence (one for Helixer and one for StringTie). Furthermore, the LA0716 gene model was subjected to the downstream fine-tuning steps through validation techniques to filter out potential false positives as follows.

A Pfam-based validation where the high-confidence sequences were identified by searching against the Pfam database using MMseqs2 v16-747c6[132]. This step yielded a "Pfam-verified" protein set comprising 10,154 entries with protein domains. Secondly, validation was conducted against a curated custom database built from UniProtKB entries taxonomically restricted to *Viridiplantae* (taxonomy ID 33090, reviewed as of February 18, 2025). This database was merged with proteins validated by PSAURON v1.0.0[70] from *Arabidopsis thaliana* (Araport11, https://www.arabidopsis.org/) and *Oryza sativa* (osa1r7, https://rice.uga.edu/). The resulting "CustomDB I-verified" set consisted of 33,023 high-confidence proteins. Thirdly, proteins were checked with PSAURON v1.0.0, generating a "PSAURON-verified" set containing 49,488 reliable proteins. Finally, all the verified proteins from the previous steps were re-evaluated against a second custom reference database. This "CustomDB II" included PSAURON v1.0.0-validated proteins from *Solanum lycopersicum* (ITAG4.0: https://solgenomics.net/) and *Solanum tuberosum* (PGSC6.1, https://solgenomics.net/), using MMseqs2 with a stringent coverage threshold -c 0.9. This final validation yielded a refined set of 37,099 verified proteins (comprising 31,998 representative sequences), culminating in a highly reliable consensus gene model characterized by both high accuracy and completeness.

The agat_sp_keep_longest_isoform.pl script from AGAT v1.4.0 (https://github.com/NBISweden/AGAT/releases/tag/v1.4.0) and Gffread v0.12.7[133] were used, respectively, to retain the longest isoform per gene locus and to extract the corresponding protein or coding sequences (CDS) from the gene models.

The quality of the gene repertoires annotated from genomic sequences was further validated based on the BUSCO completeness calculated using compleasm v0.2.6[69] with the lineage datasets solanales_odb10 and estimations for the proportion of accurate and erroneous gene models in the proteomes were performed using OMArk v0.3.0[134] with the OMA orthology database, in which exploits orthology relationships. The DETENGA database from GAQET pipeline was used to further evaluate the coding sequences[71].

Orthologous gene groups across the reference species *Solanum lycopersicum* Heinz 1706 (ITAG4.0) and the wild species, LA1039 and LA0716, were identified using OrthoFinder v3.1.3. To obtain a detailed description of the genes function, gene models were subjected to Mapman4 functional annotation using the online tool Mercator4 v8[135]. This enabled further comparison of Moneyberg -TMV, LA1039, LA0716 as well as *Solanum lycopersicum* ITAG4.0, representing the presence/absence and/or copy number variation of the genes across the species.

Characterization of genomic variation

To identify homozygous SNPs between Moneyberg-TMV and Micro-Tom, LA1039 and LA0716 respectively, we aligned HiFi reads of Micro-Tom, LA1039 and LA0716 to MbTMV using pbmm2 (version 1.7.0) with

the preset 'CCS' (https://github.com/PacificBiosciences/pbbioconda). Next, we used GATK HaplotypeCaller v4.2.5.0, HTSJDK v2.24.1 and Picard v2.25.4[136] with default options to identify variants. To remove possible variants due to mapping artifacts and obtain only homozygous SNPs, we filtered the identified variants for INFO/DP = > 5% percentile, INFO/DP < = 95% percentile, INFO/MQ = > 5% percentile, GT = "1/1" and TYPE = "snp" using BCFtools v1.15[118].

Synteny and Rearrangement Identifyer and Plotsr: Chromosome scale *S. cheesmaniae* LA1039, *S. pennellii* LA0716 genomes and publicly available genomes (Supplementary Table 1) were aligned to MbTMV[55] using Minimap2 2.24-r1122[121] with options -ax asm5 --eqx, followed by running SyRI v1.6[137] with default settings and the BAM file as input, to identify and call polymorphisms and structural variations and generate a VCF file per comparison. LA1039 and LA0716 synteny and rearrangement plots were created using plotsr v0.5.4[138] with default options. LA0716 genomes were compared following similar methods as described above by whole genome alignment and SyRI/plotsr.

For pairwise genome comparison, step-wise alignments were generated following the aforementioned method whereby each query genome was used as reference for the subsequent alignment, followed by generating synteny and rearrangement plots using plotsr v0.5.4[138] with options -H < 1-9> and -W < 1-9> and –genomes.txt.

Repeat identification using two distinct algorithms

The repetitive regions of *S.lycopersicum*-MbTMV, *S. cheesmaniae* and *S. pennellii* genomes were detected using two distinct methods. First, the mathematical Context Tree Weighting (CTW) algorithm[67] from the Bayesian Context Trees (BCT) R package v1.2[139] was used to estimate per-base sequence probabilities based on variable-length nucleotide contexts. Regions with low CTW probabilities indicate high sequence predictability, allowing the identification of biologically relevant repetitive or low-complexity regions without prior alignment or annotation. The second approach was to detect the repeat regions of those two genomes utilizing the Tandem Repeat Annotation and Structural Hierarchy (TRASH v2) pipeline[68]. TRASH also annotates tandem repeats within the genomic sequences and organizes them into a hierarchical structure, enabling the identification of complex or nested repeat architectures[68]. The CTW probabilities and the abundance of the detected monomeric repeats from TRASH were calculated in 100 Kbp windows and plotted using the R package circlize v.0.4.16[140].

Transposable element annotation and lineage classification

The transposable elements of all species included in this study (Supplementary Table 1) were identified using the Extensive De novo TE Annotator pipeline v.2.0.1[141] with the parameters –anno 1 –cds. For the five domesticated accessions of tomato, the *S.lyc*-Heinz1706 cds file[7] was used, to purge the gene coding sequences in the transposable element annotation, while for the *S. tuberosum*, the cds file was retrieved from the same database with the genome[57]. Additionally, for *S. galapagense* and *S. habrochaites* the cds files were obtained from a different publication from their genomes, which included the cds file for those two species[13]. For *S. cheesmaniae* the cds file was created using liftoff v1.6.1 to liftover ITAG4.0 gene annotation[7] and non-reference gene annotation of the *Lycopersicon* clade[6] to the *S. cheesmaniae* genome with default settings, while for *S. pennellii* the cds file from Bolger et al. was used[3]. The summary output of each species was used to create the genomic percentage table (Table 3). Intact and fragmented *Ty1/Copia* and *Ty3* LTR retrotransposons were further classified into lineages using TEsorter v.1.4.6[142] with parameters -db rexdb-plant. For the downstream analysis, all LTR retrotransposons classified into different superfamilies using Extensive De novo TE Annotator and TEsorter (for example, an element needed to be classified as *Ty1/Copia* by both pipelines) were excluded.

The ATHILAfinder v1.0 algorithm[143] was modified to detect more and higher-quality intact and soloLTR *Tekay* elements in all the genomes used in this study. The signature sequences and parameters of the ATHILAfinder were adjusted based on the characteristics of the *Tekay* elements that have been identified and classified by EDTA and TEsorter, respectively. These parameters can be found in Supplementary Table 13.

Age of LTR retrotransposons, phylogenetic analysis and gene proximity of *Tekay* elements

The age of the LTR retrotransposons was estimated based on the sequence divergence between the two LTRs of an intact element. This LTR identity metric is included in the EDTA outputs. By using the LTR identity, we were able to calculate the age of the LTR retrotransposons by utilizing the following formula:

$$Age = \frac{1 - LTR\,identity}{2\,(1.3\,x\,10^{-8})}$$

where 1.3×10^{-8} is the mutation rate in tomato and 2 is the number of different sequences in which the mutation rate is applied and in this case are the two LTRs of the LTR retrotransposon.

For the phylogenetic analysis, we retained *Tekay* elements that contained all six genes (gag, protease, reverse transcriptase, RNaseH, integrase, and chromodomain) in the correct order according to TEsorter. Those elements were further filtered based on statistics (coverage equal to or above 70% and e-value equal to or smaller than 1e-100) of their integrase domains. The amino acid sequences of integrase domains that passed the above filtering were aligned using MAFFT v.7.526[144] with parameters –-auto. We employed FastTree v.2.1.11[145] with parameters -wag and -gamma to generate maximum-likelihood trees.

All the *Tekay* elements detected with EDTA and ATHILAfinder in MbTMV, *S. cheesemaniae* and *S. pennellii* were examined to determine if they are found in proximity (+/- 2Kb) or within introns on the same or the opposite strand of the annotated genes in those species. This was possible using BEDTools v.2.31.1[119] window with parameters –a gene gff3 file –b TE annotation BED file –w 2000 and then performed over-representation analysis of MapMan4 functional categories of genes with one-sided Fisher's exact test using the online tool Mercator4 v8[135].

Downstream analysis was conducted and plotted using the R language and environment (R Core Team, 2025), utilizing packages included in the Tidyverse collection v.2.0.0[113], data.table v.1.16.4 (https://r-datatable.com/), ape v.5.8.1[146], ggtree v.3.10.1[147], circlize v.0.4.16[140], and ComplexHeatmap v.2.18.0[148].

Recombination analysis

We used the initial list of SNPs detected between parental genomes using HiFi reads. Subsequently, Illumina reads from a MbTMV x S. pennellii F1 plant were used to detect SNPs relative to the MbTMV reference. The Illumina reads of the F1 plant were trimmed using trimgalore v0.6.6 (https://github.com/FelixKrueger/TrimGalore) and aligned to Moneyberg-TMV reference genome using BWA-MEM v0.7.17[117]. SNPs were detected from the alignment output using GATK HaplotyperCaller v4.1.9.0[136] and were subsequently filtered by performing GATK hard-filtering and retaining only the heterozygous variants. Also, a SNP is excluded if it is found within a region with excessive SNP density (5-kb window; 95th percentile), within or near homopolymers (distance <5 bp), or having excessive read coverage (95th percentile). SNPs from the F1 plant with an unusual allele frequency ratio ($allele_1$ – $allele_2 > 0.8$) were used to further filter the *S.pennellii* SNPs, resulting in the final set of markers distinct between the two parental genomes. The same marker detection was performed on the other two hybrid crosses (MbTMV x Micro-Tom, and MbTMV x *S. cheesemaniae*).

BC Illumina data from all populations were aligned to MbTMV[55] using BWA-MEM[117] with default parameters followed by converting SAM to BAM and sorting of BAM files using SAMtools v1.9[118]. Duplicates

were removed using Sambamba markdup v0.8.1[149] with the option -r. For each BC sample, bcftools (v1.18) mpileup was used to detect SNPs. Afterwards, we ran the Meiotic CrossOver Finder (MeiCOFi v1.0)[150] to detect crossovers with window size = 500 kb, step size = 250 kb, window count = 6, minimum SNPs = 30.

Coldspot comparison

We first computed the number of crossovers using a sliding 50Kbp window with 10Kbp step size. All overlapping windows with zero crossovers were merged, and only those with at least 1-Mb total length are reported as coldspots. We excluded telomeric coldspots if the length is below 1.5 Mb or if they are from MbxMT male or female populations, where the density of SNPs is low, and detection of telomeric crossovers is inaccurate. The resulting coldspot locations were compared to the SV annotations to determine the putative crossover suppressor, such as large inversions.

We ran hometools mapbp v1.0 (https://github.com/mnshgl0110/hometools) on the previously generated Syri output to convert the coldspot coordinates from MbTMV to the other three parental genomes. If hometools cannot map the locations, we convert the location of the syntenic region closest to the coldspots boundaries. After generating the coldspot coordinate mapping, we compared the relative lengths of the coldspots and estimated the genomic expansion or contraction. We also used the TE and gene annotations of each parental genome to compute the difference in TE composition and gene count within the coldspot regions.

Over-representation analysis of genes in coldspots and FRRs

We extracted the female recombination coldspots and female-enhanced recombination regions in each hybrid and performed over-representation analysis of MapMan4 functional categories of genes using the online tool Mercator4 v8[135] with one-sided Fisher's exact test.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw HiFi, Omni-C and ONT sequencing data has been uploaded to the European Nucleotide Archive (ENA) under project codes: PRJEB62444 (*S. cheesmaniae* LA1039) and PRJEB62445 (*S. pennellii* LA0716). The genome assemblies and gene annotations can be found at the ENA under accessions CDWKDZ020000000 (*S. cheesmaniae* LA1039) and CDWKEA020000000 (*S. pennellii* LA0716). Illumina short-read sequencing data is available at the ENA for *S. lycopersicum* Moneyberg-TMV x *S. lycopersicum* Micro-Tom backcross populations PRJEB101732, *S. lycopersicum* Moneyberg-TMV x *S. cheesmaniae* backcross populations PRJEB101733 and *S. lycopersicum* Moneyberg-TMV x *S. pennellii* backcross populations PRJEB101734. The *S. lycopersicum* Moneyberg-TMV Isoseq and *S. pennellii* LA0716 Isoseq data has been uploaded to the ENA under project codes PRJEB104324. Genomic variants between *S. lycopersicum* cv Moneyberg-TMV and the eleven other genomes as identified by SyRI, along with EDTA-based TE annotations, are freely available at https://doi.org/10.60534/1txz5-xjw51 hosted by PLANTdataHUB[151]. Source data are provided with this paper.

References

1. Mueller, L. A. et al. The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comp. Funct. Genom.* **6**, 153–158 (2005).
2. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
3. Bolger, A. et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nat. Genet* **46**, 1034–1038 (2014).
4. Eshed, Y. & Zamir, D. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* **141**, 1147–1162 (1995).
5. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet* **17**, 333–351 (2016).
6. Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet* **51**, 1044–1051 (2019).
7. Hosmani, P. 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* https://doi.org/10.1101/767764 (2019).
8. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
9. Koren, S. et al. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res.* **34**, 1919–1930 (2024).
10. Schmidt, M. H.-W. et al. De Novo Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. *Plant Cell* **29**, 2336–2348 (2017).
11. Torgeman, S. et al. *Solanum pennellii* (LA5240) backcross inbred lines (BILS) for high resolution mapping in tomato. *Plant J.* **119**, 595–603 (2024).
12. Benoit, M. et al. Solanum pan-genetics reveals paralogues as contingencies in crop engineering. *Nature* **640**, 135–145 (2025).
13. Li, N. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet* **55**, 852–860 (2023).
14. Pailles, Y. et al. Diverse Traits Contribute to Salinity Tolerance of Wild Tomato Seedlings from the Galapagos Islands. *Plant Physiol.* **182**, 534–546 (2020).
15. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet* **12**, 363–376 (2011).
16. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
17. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
18. Fuentes, R. R. et al. Structural variants in 3000 rice genomes. *Genome Res.* **29**, 870–880 (2019).
19. Alonge, M. et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145–161.e23 (2020).
20. Lye, Z. N. & Purugganan, M. D. Copy Number Variation in Domestication. *Trends Plant Sci.* **24**, 352–365 (2019).
21. Soyk, S. et al. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. *Nat. Plants* **5**, 471–479 (2019).
22. Underwood, C. J., Henderson, I. R. & Martienssen, R. A. Genetic and epigenetic variation of transposable elements in Arabidopsis. *Curr. Opin. Plant Biol.* **36**, 135–141 (2017).
23. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet* **14**, 49–61 (2013).
24. Zhou, S.-S. et al. A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Sci Data* **8**, (2021).
25. Benoit, M. et al. Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLoS Genet* **15**, e1008370 (2019).
26. Cheng, X., Zhang, D., Cheng, Z., Keller, B. & Ling, H.-Q. A New Family of Ty1- *copia* -Like Retrotransposons Originated in the Tomato Genome by a Recent Horizontal Transfer Event. *Genetics* **181**, 1183–1193 (2009).

27. Deneweth, J., Van De Peer, Y. & Vermeirssen, V. Nearby transposable elements impact plant stress gene regulatory networks: a meta-analysis in A. thaliana and S. lycopersicum. *BMC Genomics* **23**, 18 (2022).
28. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet* **9**, 397–405 (2008).
29. Galindo-González, L., Mhiri, C., Deyholos, M. K. & Grandbastien, M.-A. LTR-retrotransposons in plants: Engines of evolution. *Gene* **626**, 14–25 (2017).
30. De Assis, R. et al. Genome relationships and LTR-retrotransposon diversity in three cultivated Capsicum L. (Solanaceae) species. *BMC Genomics* **21**, 237 (2020).
31. Yañez-Santos, A. M., Paz, R. C., Paz-Sepúlveda, P. B. & Urdampilleta, J. D. Full-length LTR retroelements in Capsicum annuum revealed a few species-specific family bursts with insertional preferences. *Chromosome Res.* **29**, 261–284 (2021).
32. Li, S.-F. et al. Impact of LTR-Retrotransposons on Genome Structure, Evolution, and Function in Curcurbitaceae Species. *IJMS* **23**, 10158 (2022).
33. Vangelisti, A. et al. LTR-retrotransposon dynamics in common fig (Ficus carica L.) genome. *BMC Plant Biol.* **21**, 221 (2021).
34. Zagorski, D. et al. Characterization and Dynamics of Repeatomes in Closely Related Species of Hieracium (Asteraceae) and Their Synthetic and Apomictic Hybrids. *Front. Plant Sci.* **11**, 591053 (2020).
35. Morgan, T. H. Sex Limited Inheritance in. *Drosoph. Sci.* **32**, 120–122 (1910).
36. Jones, D. F. Linkage in Lycopersicum. *Am. Naturalist* **51**, 608–621 (1917).
37. Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. & Grelon, M. The Molecular Biology of Meiosis in Plants. *Annu. Rev. Plant Biol.* **66**, 297–327 (2015).
38. Lambing, C., Franklin, F. C. H. & Wang, C.-J. R. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiol.* **173**, 1530–1542 (2017).
39. Wang, Y. & Copenhaver, G. P. Meiotic Recombination: Mixing It Up in Plants. *Annu. Rev. Plant Biol.* **69**, 577–609 (2018).
40. Choi, K. et al. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet* **45**, 1327–1336 (2013).
41. Wijnker, E. & Schnittger, A. Control of the meiotic cell division program in plants. *Plant Reprod.* **26**, 143–158 (2013).
42. Wang, Y., Van Rengs, W. M. J., Zaidan, M. W. A. M. & Underwood, C. J. Meiosis in crops: from genes to genomes. *J. Exp. Bot.* **72**, 6091–6109 (2021).
43. Peters, S. A. & Underwood, C. J. Technology-driven approaches for meiosis research in tomato and wild relatives. *Plant Reprod.* **36**, 97–106 (2023).
44. Rowan, B. A. et al. An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics* **213**, 771–787.
45. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
46. Fernandes, J. B. et al. Structural variation and DNA methylation shape the centromere-proximal meiotic crossover landscape in Arabidopsis. *Genome Biol.* **25**, 30 (2024).
47. Jing, J., Lian, Q., Durand, S. & Mercier, R. Maximizing meiotic crossover rates reveals the map of Crossover Potential. *Nat. Commun.* **16**, 5306 (2025).
48. Schmidt, C. et al. Changing local recombination patterns in Arabidopsis by CRISPR/Cas mediated chromosome engineering. *Nat. Commun.* **11**, 4418 (2020).
49. Drouaud, J. et al. Sex-specific crossover distributions and variations in interference level along Arabidopsis thaliana chromosome 4. *PLoS Genet* **3**, e106 (2007).
50. Capilla-Pérez, L. et al. The synaptonemal complex imposes crossover interference and heterochiasmy in*Arabidopsis*. *Proc. Natl. Acad. Sci. USA*. **118**, (2021).
51. Giraut, L. et al. Genome-Wide Crossover Distribution in Arabidopsis thaliana Meiosis Reveals Sex-Specific Patterns along Chromosomes. *PLoS Genet* **7**, e1002354 (2011).
52. Lian, Q. et al. The megabase-scale crossover landscape is largely independent of sequence divergence. *Nat Commun* **13**, (2022).
53. De Vicente, M. C. & Tanksley, S. D. Genome-wide reduction in recombination of backcross progeny derived from male versus female gametes in an interspecific cross of tomato. *Theor. Appl. Genet.* **83**, 173–178 (1991).
54. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23**, (2022).
55. Van Rengs, W. M. J. et al. A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. *Plant J.* **110**, 572–588 (2022).
56. Wang, Y. et al. Harnessing clonal gametes in hybrid crops to engineer polyploid genomes. *Nat. Genet* **56**, 1075–1079 (2024).
57. Yang, X. et al. The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. *Mol. Plant* **16**, 314–317 (2023).
58. Yu, X. et al. Chromosome-scale genome assemblies of wild tomato relatives *Solanum habrochaites* and *Solanum galapagense* reveal structural variants associated with stress tolerance and terpene biosynthesis. *Horticulture Res.* **9**, uhac139 (2022).
59. Wang, X. et al. Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nat Commun* **11**, (2020).
60. Han, H. et al. Chromosome-level genome assembly of Solanum pimpinellifolium. *Sci. Data* **11**, 577 (2024).
61. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, (2017).
62. Yamaguchi, K. et al. Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Mol. Ecol.* **30**, 5923–5934 (2021).
63. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
64. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
65. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
66. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, (2020).
67. Kontoyiannis, I., Mertzanis, L., Panotopoulou, A., Papageorgiou, I. & Skoularidou, M. Bayesian Context Trees: Modelling and exact inference for discrete time series. *J. R. Stat. Soc. Ser. B* **84**, 1287–1323 (2022).
68. Wlodzimierz, P., Hong, M. & Henderson, I. R. TRASH: Tandem Repeat Annotation and Structural Hierarchy. *Bioinformatics* **39**, (2023).
69. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).

70. Sommer, M. J., Zimin, A. V. & Salzberg, S. L. PSAURON: a tool for assessing protein annotation across a broad range of species. *NAR Genomics Bioinforma.* **7**, lqae189 (2025).
71. Garcia-Carpintero, V., De Martín, I., García-Juan, S. & Bombarely, A. Structural Genome Annotation QC with GAQET2 v2. Preprint at https://doi.org/10.17504/protocols.io.8epv5k5rnv1b/v1 (2025).
72. Gebhardt, C. A physical map of traits of agronomic importance based on potato and tomato genome sequences. *Front. Genet.* **14**, 1197206 (2023).
73. Peters, S. A. et al. Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *Plant J.* **71**, 602–614 (2012).
74. Szinay, D. et al. Chromosome evolution in *Solanum* traced by cross-species BAC-FISH. *N. Phytologist* **195**, 688–698 (2012).
75. Jedlicka, P., Lexa, M. & Kejnovsky, E. What Can Long Terminal Repeats Tell Us About the Age of LTR Retrotransposons, Gene Conversion and Ectopic Recombination? *Front. Plant Sci.* **11**, (2020).
76. Platt, R. N., Vandewege, M. W. & Ray, D. A. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res.* **26**, 25–43 (2018).
77. Quesneville, H. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. *Mobile DNA* **11**, (2020).
78. Hayward, A. & Gilbert, C. Transposable elements. *Curr. Biol.* **32**, R904–R909 (2022).
79. Kijima, T. E. & Innan, H. On the Estimation of the Insertion Time of LTR Retrotransposable Elements. *Mol. Biol. Evolution* **27**, 896–904 (2010).
80. Nguyen, H. N., Lai, N., Kisiala, A. B. & Emery, R. J. N. Isopentenyltransferases as master regulators of crop performance: their function, manipulation, and genetic potential for stress adaptation and yield improvement. *Plant Biotechnol. J.* **19**, 1297–1313 (2021).
81. Demirci, S. et al. Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between *Solanum lycopersicum* and *Solanum pimpinellifolium*. *Plant J.* **89**, 554–564 (2017).
82. Rommel Fuentes, R. et al. Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *Plant J.* **102**, 480–492 (2020).
83. Fuentes, R. R. et al. A catalogue of recombination coldspots in interspecific tomato hybrids. *PLoS Genet* **20**, e1011336 (2024).
84. Marand, A. P. et al. Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol.* **18**, 203 (2017).
85. Pan, Q., Deng, M., Yan, J. & Li, L. Complexity of genetic mechanisms conferring nonuniformity of recombination in maize. *Sci. Rep.* **7**, 1205 (2017).
86. Choi, K. et al. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Res.* **28**, 532–546 (2018).
87. Underwood, C. J. & Choi, K. Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. *Chromosoma* **128**, 279–296 (2019).
88. Marand, A. P. et al. Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *Plant Cell* **31**, 645–662 (2019).
89. Fuentes, R. R., De Ridder, D., Van Dijk, A. D. J. & Peters, S. A. Domestication Shapes Recombination Patterns in Tomato. *Mol. Biol. Evolution* **39**, msab287 (2022).
90. Zapata, L. et al. Chromosome-level assembly of*Arabidopsis thaliana*Lerreveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci. USA.* **113**, (2016).
91. Sim, S.-C. et al. Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato. *PLoS ONE* **7**, e40563 (2012).
92. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol* **20**, (2019).
93. Mahmoud, M. et al. Utility of long-read sequencing for All of Us. *Nat Commun* **15**,.
94. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
95. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet* **55**, 1221–1231 (2023).
96. Yuan, Y., Bayer, P. E., Batley, J. & Edwards, D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **19**, 2153–2163 (2021).
97. Mascher, M. et al. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
98. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, (2021).
99. Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet* **46**, 1220–1226 (2014).
100. Zhao, J., Sauvage, C., Bitton, F. & Causse, M. Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition. *Horticulture Res.* **9**, uhab009 (2022).
101. Žižková, E. et al. Tomato (Solanum lycopersicum L.) SlIPT3 and SlIPT4 isopentenyltransferases mediate salt stress response in tomato. *BMC Plant Biol.* **15**, 85 (2015).
102. Wang, X. et al. Evolution and roles of cytokinin genes in angiosperms 1: Do ancient IPTs play housekeeping while non-ancient IPTs play regulatory roles? *Hortic. Res.* **7**, 28 (2020).
103. Domínguez, M. et al. The impact of transposable elements on tomato diversity. *Nat. Commun.* **11**, 4058 (2020).
104. Long, W. et al. Genome evolution and diversity of wild and cultivated rice species. *Nat. Commun.* **15**, 9994 (2024).
105. Li, X. et al. Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. *BMC Genomics* **18**, 55 (2017).
106. Hasan, M. M., Yin, L., Wang, M., Pawlowski, W. P. & Zhao, M. mop1 affects maize recombination landscapes by modulating methylation of MITEs near genes in open chromatin. *Nat. Commun.* **16**, 10476 (2025).
107. Kianian, P. M. A. et al. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat Commun* **9**, (2018).
108. Durand, S. et al. Joint control of meiotic crossover patterning by the synaptonemal complex and HEI10 dosage. *Nat Commun* **13**, (2022).
109. Morgan, C. et al. Diffusion-mediated HEI10 coarsening can explain meiotic crossover positioning in Arabidopsis. *Nat Commun* **12**, (2021).
110. Ziolkowski, P. A. et al. Natural variation and dosage of the HEI10 meiotic E3 ligase control *Arabidopsis* crossover recombination. *Genes Dev.* **31**, 306–317 (2017).
111. Li, W. & Chetelat, R. T. Unilateral incompatibility gene *ui1.1* encodes an S-locus F-box protein expressed in pollen of *Solanum* species. *Proc. Natl. Acad. Sci. Usa.* **112**, 4417–4422 (2015).
112. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
113. Wickham, H. et al. Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).
114. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

115. Modi, A., Vai, S., Caramelli, D. & Lari, M. The Illumina Sequencing Protocol and the NovaSeq 6000 System. in *Bacterial Pangenomics* (eds Mengoni, A., Bacci, G. & Fondi, M.) vol. 2242 15–42 (Springer US, New York, NY, 2021).
116. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).
117. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
118. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
119. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
120. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol.* **15**, e1007273 (2019).
121. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
122. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
123. Jouffroy, O., Saha, S., Mueller, L., Quesneville, H. & Maumus, F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics* **17**, (2016).
124. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
125. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. in *Gene Prediction* (ed. Kollmar, M.) vol. 1962 227–245 (Springer New York, New York, NY, 2019).
126. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
127. Holst, F. et al. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat Methods* https://doi.org/10.1038/s41592-025-02939-1 (2025).
128. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
129. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
130. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience* **7**, giy131 (2018).
131. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, 1–15 (2018).
132. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
133. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res* **9**, 304 (2020).
134. Nevers, Y. et al. Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.* **43**, 124–133 (2025).
135. Schwacke, R. et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **12**, 879–892 (2019).
136. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* https://doi.org/10.1101/201178 (2017).
137. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, (2019).
138. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
139. Ioannis Papageorgiou, Valentinian Mihai Lungu, Ioannis Kontoyiannis. B. C. T.: Bayesian Context Trees for Discrete Time Series. 1.2 https://doi.org/10.32614/CRAN.package.BCT (2020).
140. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
141. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, (2019).
142. Zhang, R.-G. et al. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9**, (2022).
143. Primetis, E. & Bousios, A. ATHILAfinder: a tool to detect ATHILA LTR retrotransposons in plant genomes. *BioRxiv* https://doi.org/10.64898/2026.03.20.713144 (2026).
144. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evolution* **30**, 772–780 (2013).
145. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).
146. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
147. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
148. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
149. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
150. Fuentes, R. R., Fernandes, J. B., Susanto, T., Wang, Y. & Underwood, C. J. MeiCOFi: Meiotic CrossOver Finder in haploid, diploid, polyploid and hyper-recombinant genomes. *BioRxiv* https://doi.org/10.64898/2026.04.29.721680 (2026).
151. Weil, H. L. et al. PLANTDATAHUB: a collaborative platform for continuous FAIR data sharing in plant research. *Plant J.* **116**, 974–988 (2023).

Acknowledgements

S. cheesmaniae LA1039 and *S. pennellii* LA0716 seeds were obtained from TGRC (UC Davis, California, USA). Moneyberg-TMV seeds were kindly provided for research purposes by Cilia Lelivelt and Maarten Verlaan (Rijk Zwaan, Netherlands). Micro-Tom seeds were originally sourced from Tomato Growers Supply company (Ft. Myers, Florida, USA). We thank Marianne Harperscheidt and Christine Saenger for their help with maintenance of plant materials, Andre Marques for his help interpreting Hi-C contact plots, Manish Goel for his help with run-case scenarios in whole genome comparison, Rainer Schwacke for his help with the functional annotations and Marie Bolger for her help with our computational data maintenance.

Author contributions

C.J.U. and W.M.J.R. designed the study. S.E. maintained all plant materials with contributions of W.M.J.R. and sampled plant materials of backcross populations. S.E. performed hybridization with contributions

of W.M.J.R. B.H. performed library preparation and sequencing of backcross populations. B.H. performed high molecular weight DNA isolation and library preparation for Omni-C and HiFi sequencing. Z.Z. and B.U. performed ONT library preparation and ONT sequencing. W.M.J.R. generated the LA1039 genome assembly. Z.Z. generated the mixed ONT and HiFi LA0716 assembly. W.M.J.R. scaffolded the LA1039 and LA0716 genome assemblies and performed validation analyses with contributions of C.J.U., R.R.F., E.P., Z.Z. and B.U. W.M.J.R. performed pairwise comparison of genome assemblies and sequencing data with contributions of C.J.U., R.R.F. and E.P. Iso-seq data for Moneyberg-TMV was generated by YW, WMJR and BH. Iso-seq data for LA0716 was generated by SA and analysed by Z.Z. Gene annotation was performed by Z.Z. and B.U. W.M.J.R. characterized genomic variation with contributions of R.R.F., E.P. and C.J.U. E.P. performed the repeat content and transposable element analysis. RRF performed crossover analysis with contributions from W.M.J.R., J.F., T.S., C.J.U. and Q.L. W.M.J.R., R.R.F., E.P. and C.J.U. wrote the manuscript and generated the figures with contributions of all authors.

Funding

This study was funded by the Max Planck Society (Core funding to C.J.U.), the Radboud University executive board (Core funding to C.J.U.), the German Research Foundation (DFG. grant 465339501 to C.J.U. and 452682775, EXC 390686111 to B.U.) and the European Research Council (ERC starting grant 101076355, 'AsexualEmbryo', to C.J.U.). Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-74784-5.

Correspondence and requests for materials should be addressed to Björn Usadel or Charles J. Underwood.

Peer review information *Nature Communications* thanks Aureliano Bombarely, and Xianwen Meng for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026