RESEARCH

Long-read sequencing of extrachromosomal circular DNA and genome assembly of a *Solanum lycopersicum* breeding line revealed active LTR retrotransposons originating from *S. Peruvianum* L. introgressions

Pavel Merkulov[1,2], Melania Serganova[1,2], Georgy Petrov[2], Vladislav Mityukov[3,4] and Ilya Kirov[1,2*]

Abstract

Transposable elements (TEs) are a major force in the evolution of plant genomes. Differences in the transposition activities and landscapes of TEs can vary substantially, even in closely related species. Interspecific hybridization, a widely employed technique in tomato breeding, results in the creation of novel combinations of TEs from distinct species. The implications of this process for TE transposition activity have not been studied in modern cultivars. In this study, we used nanopore sequencing of extrachromosomal circular DNA (eccDNA) and identified two highly active Ty1/*Copia* LTR retrotransposon families of tomato (*Solanum lycopersicum*), called *Salsa* and *Ketchup*. Elements of these families produce thousands of eccDNAs under controlled conditions and epigenetic stress. EccDNA sequence analysis revealed that the major parts of eccDNA produced by *Ketchup* and *Salsa* exhibited low similarity to the *S. lycopersicum* genomic sequence. To trace the origin of these TEs, whole-genome nanopore sequencing and de novo genome assembly were performed. We found that these TEs occurred in a tomato breeding line via interspecific introgression from *S. peruvianum*. Our findings collectively show that interspecific introgressions can contribute to both genetic and phenotypic diversity not only by introducing novel genetic variants, but also by importing active transposable elements from other species.

Keywords Extrachromosomal circular DNA, Long-read sequencing, Mobilome, Tomato, Retrotransposons, Insertions, Interspecific introgressions

*Correspondence:
Ilya Kirov
kirovez@gmail.com
[1]All-Russia Research Institute of Agricultural Biotechnology, 127550 Moscow, Russia
[2]Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia
[3]Skolkovo Institute of Science and Technology, 121205 Moscow, Russia
[4]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, 127051 Moscow, Russia

Introduction

Transposable elements are ubiquitous components of plant genomes. LTR retrotransposons (LTR-RTEs) are among the highest copy members of the plant mobilome [1, 2], accounting for 76% of the rye genome [3] and 50% of the tomato genome [4]. Uncontrolled TE reactivation can lead to genetic instability. Therefore, plants have a complex system of epigenetic control of TEs, including RNA-dependent DNA Methylation (RdDM) [5, 6] and *DDM1*-mediated DNA methylation [2]. TEs in plants carrying mutations in the genes involved in epigenetic regulation can be transcriptionally reactivated. For example, a threefold increase in TE transcription was observed in a triple mutant of *A. thaliana* with mutations in three key TE-controlling genes, *ddm1*, rdr6, and polV [7]. Similarly, transposition of some TE families has been observed in the DNA methylation-free *A. thaliana* mutant [8], which is likely due to the redistribution of histone modifications [9].

TE transposition can also occur in wild type plants under natural conditions. Such natural TE activity has contributed significantly to the evolution, adaptation, and domestication of plants [10]. Bursts in LTR retrotransposon activity have had a key impact on the size of the genomes of some plant species [11–13], whereas individual insertions have created functionally altered or new alleles of genes [14, 15]. More specific examples include the emergence of traits such as seedless apples [16], grape skin color [17], and red orange pulp [18]. In tomatoes, the insertion of elements of the *Rider* family resulted in an elongated fruit shape [19, 20], yellow flesh fruit [21], and lack of formation of a detachment zone in the peduncle [22]. A study on *A. thaliana* also provided an interesting example of the contribution of TE insertion to plant adaptation to a novel ecological environment. A rare variant of the *A. thaliana* FLC locus was found to contain an intronic insertion of a heat-inducible element in the *ONSEN* family, which may be an adaptation to flowering in the absence of vernalization [23]. Whole genome sequencing of *A. thaliana* ecotypes revealed that hundreds of TEs generate novel insertions [23]. Despite this, the ability to control the activation of specific TEs has only been possible for a small set of elements, such as the heat-induced *ONSEN* LTR retrotransposon of *A. thaliana* and the plant tissue culture-triggered Tos17 LTR retrotransposon of rice [24].

In addition to environmental factors, TE activity can be triggered by 'genomic shock' as proposed by McClintock [25]. Genomic shock can result from chromosomal rearrangements and interspecific hybridization. Transcriptional reactivation of LTR-RTE has been detected during long-distance hybridization in various plant species, including rice [26], wheat [27], *Arabidopsis* [28], and wild potato species [29]. In addition, although much less frequently, hybridization-induced LTR-RTE transposition in real time has been detected in rice [30], poplar [31], and potato [32]. The consequences of interspecific hybridization on TE composition in the genome have been well described for the *Solanum* genus. Interspecific hybridization is one of the main sources of genetic diversity in tomato (*Solanum lycopersicum* L.) breeding and domestication. A wide range of wild species has been implicated in this process, including *S. peruvianum* [33, 34], S. chilense [35], and *S. habrochaites* [36], *S. penellii* [37] and *S. pimpinellifolium* [38]. A recent study demonstrated that the TE composition of modern tomato cultivars is less divergent than that of wild species [39]. At the same time, the domesticated tomato *S. lycopersicum* var. cerasiforme shows minor losses in the number of mobile TE families [39], which could potentially be due to recurrent hybridization with its closest wild relative, *S. pimpinellifolium* [40]. Whether TEs located in interspecific introgressions maintain their activity during breeding and generate new insertions in the recipient genome has not been well studied.

In this study, we aimed to decipher inducible mobilome activity originating from TEs located at interspecific introgressions in the tomato genome. We performed a whole-genome analysis of TE activity using long-read nanopore sequencing of extrachromosomal circular DNAs from a tomato line. We found thousands of eccDNAs mapped on members of families that we called '*Ketchup*' and '*Salsa*'. The eccDNA sequence analysis revealed that the major parts of eccDNA produced by *Ketchup* and *Salsa* did not fully align with any tomato (SL3.0) reference TEs, but were similar to the TEs from the *S. peruvianum* genome. We performed whole-genome nanopore sequencing and assembly for our tomato line and revealed large interspecific introgressions carrying members of the *Ketchup* and *Salsa* TE families. Our results suggested that active TEs introduced by interspecific hybridization may serve as an additional source of genetic diversity during plant breeding.

Results

Mobilome-seq of a tomato breeding line

To unravel TEs capable of completing their life cycle, we performed nanopore (ONT) sequencing of extrachromosomal circular DNA (eccDNAs) of a tomato plant (Fig. 1). To collect more active TEs, we grew the plants in a special medium containing a mixture of zebularine and α-amanitin (A&Z). These chemicals lead to DNA methylation reduction and inhibition of Polymerase II, a major player in the PolII-RDR6 TE RdDM silencing pathway [41]. Plants grown in MS medium without A&Z were used as controls to statistically evaluate eccDNA peaks.

In total, we obtained 64,000 and 48,000 reads for the A&Z and Control plants, respectively. The eccDNA
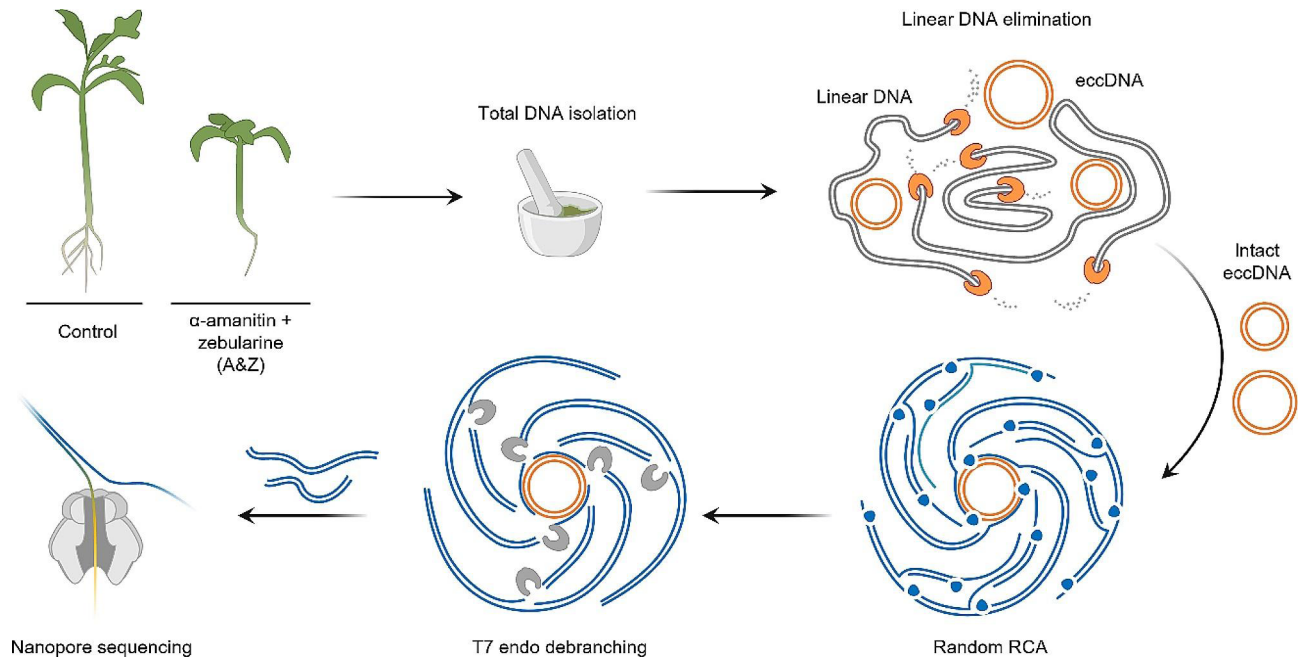
**Fig. 1** Overview of the eccDNA nanopore sequencing experiment

reads were mapped to the reference genome (SL3.0), followed by intersection with LTR-RTE coordinates (4206 annotated LTR-RTEs) and manual curation. We found 101 L-RTEs, for which > 10 ONT reads were accounted. Of these, 38 L-RTEs demonstrated significant overrepresentation of ONT reads from the A&Z sample compared with the control sample (Fisher's exact test with multiple correction p-value < 0.01) (Fig. 2A). Phylogenetic analysis of these LTR-RTEs based on their LTR sequences revealed that they belong to two families, that we named '*Salsa*' (a tomato-based sauce that dates back to the earliest tomato cultivators, the Aztecs and Mayans [42]) and '*Ketchup*' (Figure S1). According to the GyDB database classification, the *Ketchup* elements belong to the Tork clade and are almost identical to *CopiaSL_35* (average 98% identity with 99% coverage), whereas the *Salsa* family belongs to the Bianca clade and has some similarity to *CopiaSL_25* (average about 72% identity with 41% coverage). Among the elements of each family, we selected one RTE with the highest CPM value: RTE976 (Sly11:3,646,824.3,652,356) from *Salsa* and RTE511 (Sly01:68,869,427.68,874,507) from *Ketchup* family. We checked for the presence of open reading frames in the genomic sequences of selected RTEs in the SL3.0 genome assembly, and both elements were found to have one or two long ORFs (Figure S2A).

Thus, nanopore Mobilome-Seq revealed that under epigenetic stress conditions (A&Z treatment), tens of LTR-RTEs belonged to two distinct families of tomato lines producing eccDNAs.

RTEs producing eccDNAs originate from *S. peruvianum*

Surprisingly, however, a detailed analysis of LTR-RTE coverage by ONT eccDNA reads revealed numerous SNPs (> 50 for RTE511 and > 100 for RTE976), distinguishing the LTR-RTE sequences of our tomato line from the reference. In addition, LTRs of RTE976 were not covered by eccDNA reads, and LTRs of RTE511 possessed a > 20 bp deletion based on read mapping (Fig. 2B). These observations show that the eccDNA reads most probably originated from RTEs that were not present in the reference genome sequence (SL3.0). To verify this, we performed a BLAST search for the most similar LTR sequences in the genomic assemblies of wild relatives of *S. lycopersicum*, including *S. penellii*, *S. lycopersicum* var. cerasiforme, *S. pimpinellifolium*, *S. lycopersicoides*, *S. peruvianum*, *S. chilense*, *S. habrochaites*. We used consensus LTR sequences deduced from the eccDNA reads mapped to RTE976/RTE511 as queries for the BLAST search. This analysis revealed that the most similar sequences were found in the *S. peruvianum* (SP) genome with > 96% identity to the query LTR sequences (Fig. 3A).

We identified the full-length RTEs (5033 and 5500 bp) of SP genome with LTRs that have > 98% similarity to eccDNA deduced LTRs and named them *Ketchup-1* (SP05:7,068,171-7,073,205) and *Salsa-1* (SP01:93,430,320 − 93,435,819). Both RTEs had 99–100% LTR identity, suggesting their recent activity in the SP genome. Both RTEs possessed well-defined reading frames (1302 aa for *Ketchup-1*; 352 and 1173 aa for *Salsa-1*) encoding all the required domains, including GAG coat protein (GAG), aspartic proteinase (AP), integrase
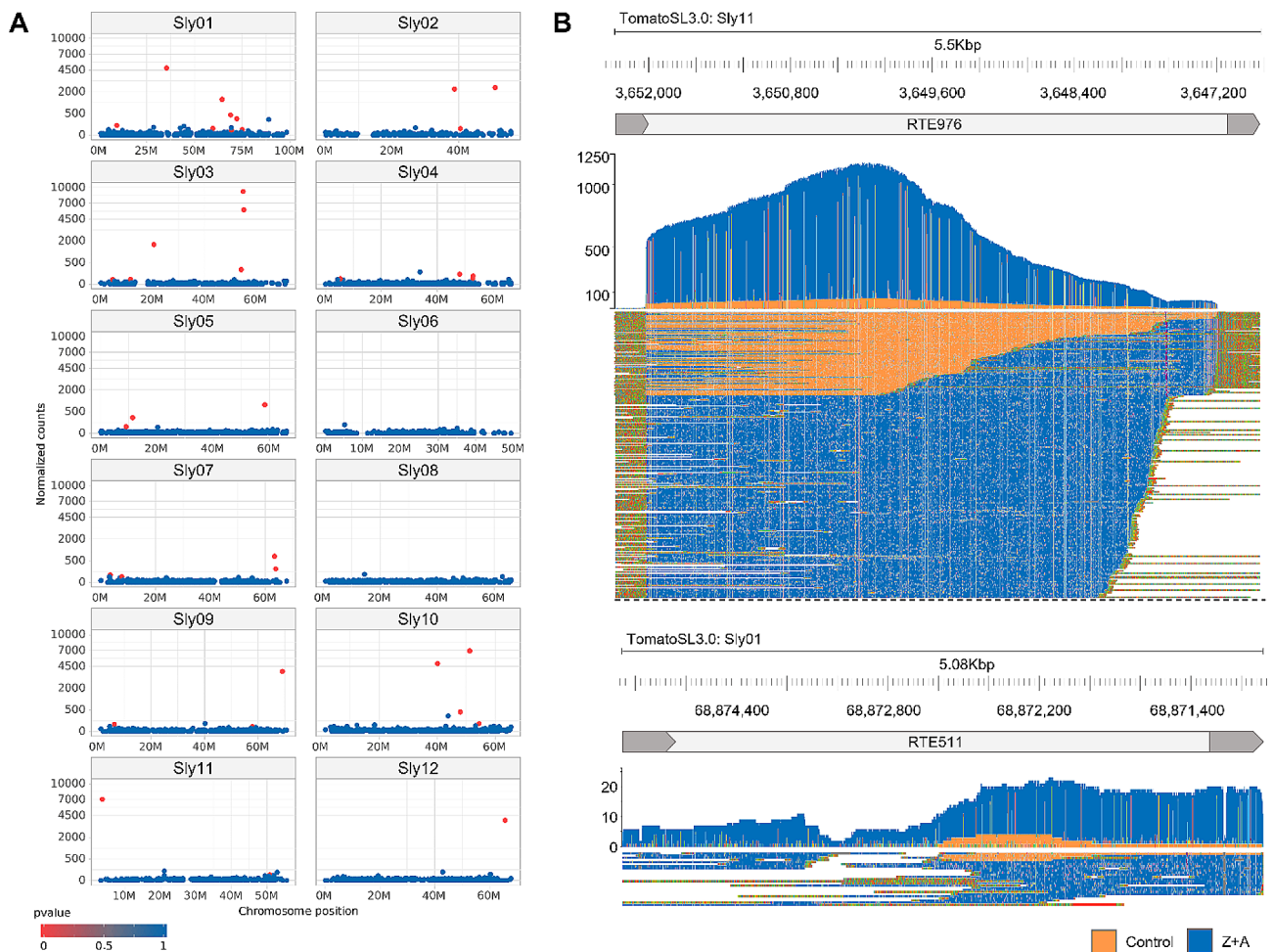
Fig. 2 Analysis of eccDNA production by LTR-RTEs assessed using ONT sequencing. (**A**) Normalized (reads per 100,000 ONT reads) count of eccDNA reads from Z&A samples for 4206 RTEs located on chromosomes of SL3.0 genome assembly; (**B**) Coverage of RTE976 (*Salsa* family) and RTE511 (*Ketchup* family) by eccDNA reads. Orange and blue colors indicate eccDNA reads from the Control and Z&A samples

(INT), reverse transcriptase (RT) and ribonuclease H (RNaseH) (Fig. 3B). We further checked the existence of IRES (an internal ribosome entry site) between two Salsa-1 ORFs. Using the IRESpy tool (https://irespy. shinyapps.io/IRESpy/) we detected an IRES (probability ∼ 0.597) in the region between the two ORFs of Salsa-1. These results suggest that Salsa-1 ORF2 may be translated by cap-independent mechanisms.

Ketchup-1 and *Salsa-1* produce full-length eccDNAs with one or two LTRs

Individual RTE eccDNAs may represent different structural variants covering only a small LTR part, as well as whole RTEs [43]. To assess the structure of eccDNAs produced by *Ketchup-1* and *Salsa-1*, we investigated the monomers of individual eccDNA reads possessing concatemers. We found that a significant portion of eccD-NAs of *Ketchup-1* contained only one LTR, with some eccDNAs also possessing small deletions (Fig. 4A). In turn, *Salsa-1* produced full-length eccDNA with one

or more LTR (Fig. 4B) and a small proportion of eccD-NAs containing truncated sequences (Figure S3). We then performed inverted PCR using specific primers for amplification of the LTR junction regions of eccDNA (Fig. 4C). For this experiment, genomic DNA from Control and A&Z samples before and after eccDNA enrichment and RCA were used. Weak and strong PCR products were obtained for solo-LTR eccDNAs produced by *Ketchup-1* and *Salsa-1* in control and Z&A samples, respectively (Fig. 4C). However, we were unable to detect extrachromosomal linear DNA (eclDNA) for *Salsa-1* and *Ketchup-1* in either the control or Z&A samples (Figure S4).

Altogether, the results demonstrate that *Ketchup-1* and *Salsa-1* RTEs produce eccDNAs with one or two LTRs under control and A&Z conditions.
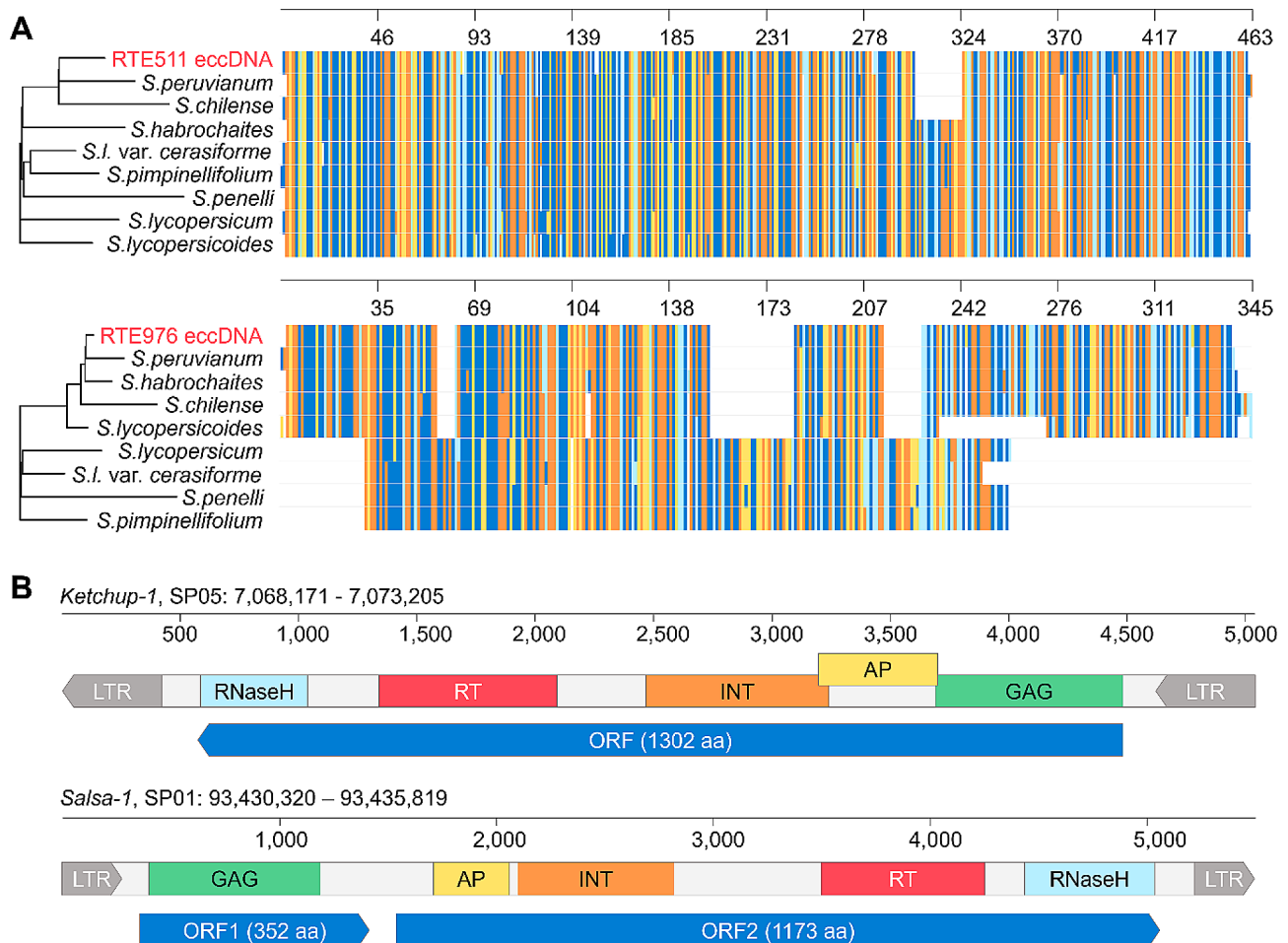
Fig. 3 Phylogenetic and structural analyses of *Ketchup* and *Salsa* RTEs. (**A**) Alignment of LTR sequences from eccDNA and the genome assemblies of diverse tomato species. (**B**) Structure and open reading frames of two RTEs (*Ketchup-1* and *Salsa-1*) in the *S. peruvianum* genome

Ketchup-1 and *Salsa-1* invaded *S. lycopersicum* genome via an interspecific introgression

To determine how SP RTEs occurred in the genome of our tomato line, we performed whole-genome nanopore sequencing. We obtained 4,417,781 reads with a total length of 57.7Gb corresponding to ~64x genome coverage of the tomato genome (1 C=900Mb [44]),. We performed SNP calling and found significant biases in SNP density along SL chromosomes. The results unambiguously demonstrated a significantly high density of SNPs along full-length chromosomes 6 (2 Mb–32 Mb) and 9 (2 Mb–64 Mb) (Fig. 5; Figure S5). These results indicated that the genome of the tomato line used in this study possessed large interspecific introgressions.

To gain further insight into the origin of *Salsa* and *Ketchup* in our tomato line, we performed whole-genome assembly using only WGS nanopore reads. The assembly was performed using NextDenovo [45]. The draft assembly resulted in an N50 around 16.7 Mb (16.721.217 bp) and total length of approximately 800 Mb (813,056,715 bp) that is ~100 Mb smaller than

the predicted genome size of cultivated tomato The latter suggested that some highly abundant genomic repeats (e.g. telomere and centromere satellite repeats) can be collapsed during the assembly procedure. The quality of the draft assembly was verified using the BUSCO software [46]. The percentage of the BUSCOs benchmark genes was high (>98%) (C:99.3%[S:98.8%,D:0.5%], F:0.5%,M:0.2%,n:425). A comparison of the assembled genome and reference SL3.0 revealed that SL chromosomes 2, 4, and 7 were almost completely covered by two assembled contigs, further suggesting the relatively high contiguity of the draft genome (Figure S5). In line with the SNP density distribution, the comparison also revealed a low alignment rate between our assembly and chromosomes SL6 and SL9, pointing to the genomic differences between SL3.0 and the genome of our breeding line.

After *de novo* prediction of LTR retrotransposons and eccDNA read mapping, we identified *Salsa-1* (Salsa-1-1) and *Ketchup-1* (*Ketchup*-1-1) original sequences on the contig ctg0009404853 (ctg000940:9,832,685.9,838,179)
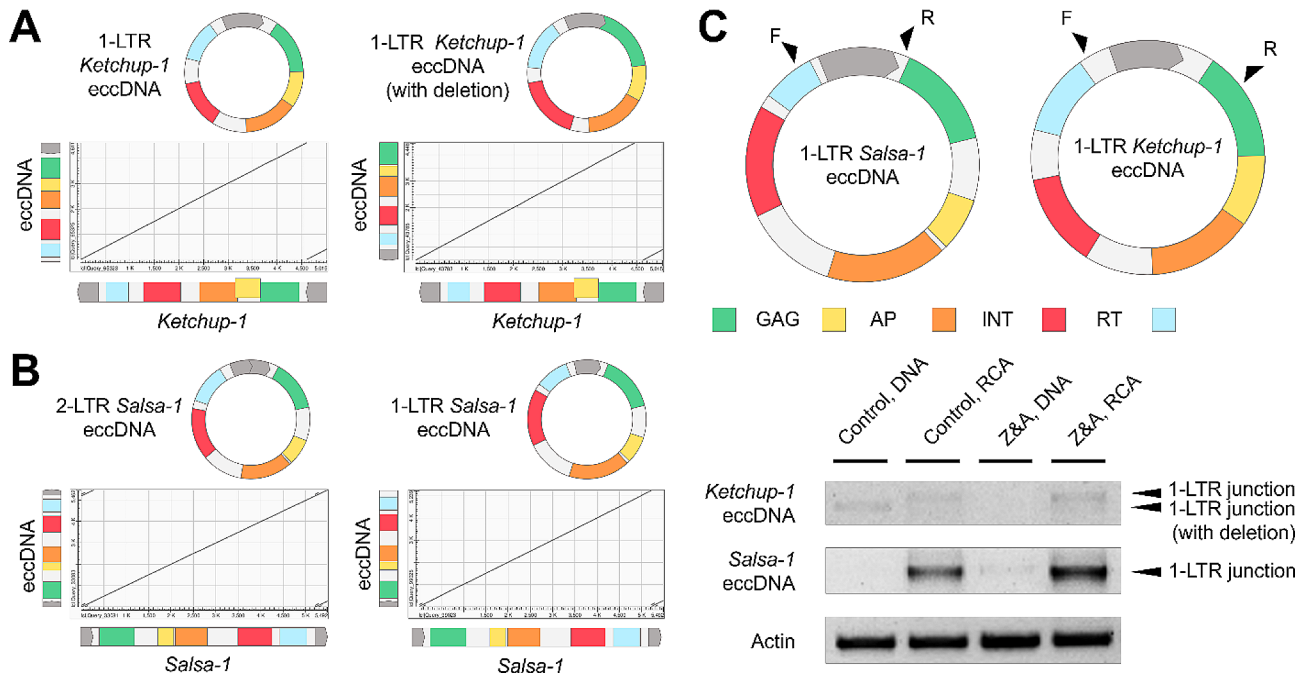
Fig. 4 Analysis of the structure of *Salsa-1* and *Ketchup-1* eccDNA. Dot plot from alignment of eccDNA deduced monomer sequences against full-length reference *Ketchup-1* (**A**) and *Salsa-1* (**B**) sequences. (**C**) Primer positions (top) and gel electrophoresis (bottom) for inverted PCR with genomic DNA from Control and A&Z samples before and after eccDNA enrichment and RCA

and ctg001010 (ctg001010:18,666,308.18,671,325) in our genome assembly, respectively. Additionally, we found one extra copy for *Salsa-1* (*Salsa-1-2*) on the contig ctg001270 (ctg001270: 8946770.8956279). Comparison of our draft genome assembly with SL3.0 identified three sites in the SL3.0 assembly corresponding to *Salsa-1* and *Ketchup-1* copies in our genome assembly: *Salsa-1-1* (Sly9:46,243,555) and *Salsa-1-2* (Sly09:42,506,399), as well as a single insertion of *Ketchup-1*, localized on chromosome 6 (Sly06:31,315,480).

Whole-genome analysis of TEIs using our ONT reads and SL3.0 references revealed four insertions in the genome of our breeding line generated by *Ketchup* (RTE5645) and *Salsa* (RTE5672) RTEs of *S.lycopersicum* origin (three TEIs for RTE5645 and one TEI for RTE5672) (Fig. 5): Sly02:51,094,794, Sly06: 2,425,515, Sly11: 9,441,876 and Sly02:46,224,366, respectively. All TEIs were validated using PCRs with primers targeting the flanking regions and TEs (Figures S6A and S6B). These results suggested that RTEs of the *Ketchup* and *Salsa* families from *S.lycopersicum* as well as from *S. peruvianum* genomes are transpositionally active. This is in good concordance with the high similarity of LTR sequences for these RTEs (99.3% for *Ketchup*-1-1, 99.64% for *Salsa*-1-1, 98.2% for *Salsa*-1-2, 97.3% for RTE5645 and 96% for RTE5672) and their insertions (94.75–99.3%). However, only RTEs of *S. peruvianum* origin (*Ketchup*-1-1, *Salsa*-1-1, and *Salsa*-1-2) produced eccDNA in our experimental conditions, suggesting that SL and SP members of these families acquired different strategies for transposition activation.

We next asked whether the introgressed SP RTEs in the genome of our tomato line are present in the same location in the SP genome. For this, we compared the RTEs with 2Kb flanking sequences between our assembly and SP, as well as SL (as a control) genome assemblies. This analysis revealed that *Ketchup*-1-1, *Salsa*-1-1, and *Salsa*-1-2 are not present in the same location in the SP and SL genomes (Figure S7). These results may suggest that the identified insertions in our genome assembly occurred during the plant breeding process. However, we cannot rule out the possibility that these RTEs were just not present in the sequenced SP plant. Thus, using WGS ONT data, we showed that *Salsa-1* and *Ketchup-1* occurred in the genome of our tomato line via interspecific hybridization and chromosomal introgression.

Discussion

Interspecific hybridization has been extensively used to introduce desirable genes from wild species into cultivated tomato [40]. It has been known for a long time that interspecific hybridization may trigger TE reactivation and transposition [25]. The TE composition of the tomato genome has been described previously [39]. Here, we explored whether interspecific introgression might bring novel active TEs from other species. Using nanopore sequencing of eccDNAs, we described real-time mobilome activity that occurred under epigenetic
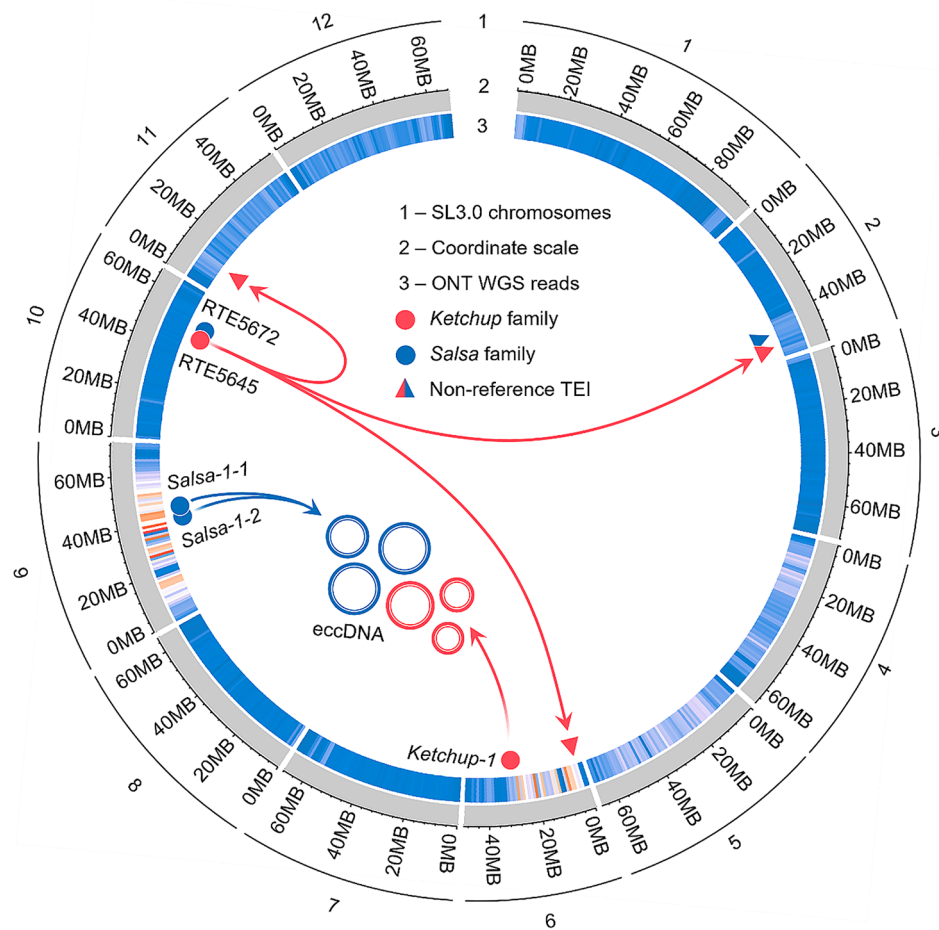
Fig. 5 Whole-genome nanopore sequencing of the analyzed tomato line. SNP density deduced from alignment of ONT WGS reads of the studied tomato line on SL3.0 genome assembly; circles and triangles indicate original TEs and their insertions; rings represent eccDNAs produced by *Ketchup* and *Salsa* of *S. peruvianum*

stress in a tomato breeding line. The sequences of individual eccDNA reads allowed us to accurately determine two families of active LTR retrotransposons: *Salsa* and *Ketchup*. Further elucidation of the newly obtained draft genome assembly for our breeding line revealed that the eccDNA-producing RTEs from these two families were introgressed from *S. peruvianum*.

Our results highlight how active transposons can be introduced into a new genome to maintain their activity for several generations. Indeed, hybridization-induced TE mutagenesis can be a major factor in the evolution of sexually reproducing organisms [47], and it has even been exploited for crop improvement [48]. Interestingly, the population of transpositionally active TEs in wild tomato species is significantly larger than that in cultivated *S. lycopersicum* [39]. Therefore, interspecific chromosomal introgressions in modern tomato varieties may carry active TEs.

Interestingly, we did not observe any novel insertions of introgressed RTEs in the genome of our breeding line.

This can be partially explained by the transposition of the original elements in the first stages of hybridization, followed by their subsequent elimination during backcrossing and selection. In contrast to the introgressed SP RTEs, we identified novel insertions for SL members of the *Salsa* and *Ketchup* families. It is interesting to speculate that the presence of active *Salsa* RTEs from *S. peruvianum* complemented SL RTEs to transpose, as has been shown for BARE-2 and Tos17 GAG-defective elements [49, 50].

Short read sequencing has been frequently used for eccDNA detection [51–53]. Utilization of long-read WGS and eccDNA sequencing allowed us to accurately determine the structure and full-length sequence of the eccDNAs. This allowed us to identify the positions of the active elements that are absent in the reference genome. Although the formation of eccDNA originating from RTEs has been considered a by-product of their activities [54], a recent study suggested that eccDNA is one of the key steps in the life cycle of RTEs [55]. The concatameric

structure of the eccDNA ONT reads allows distinguishing naturally occurring truncated sequences from DNA breaks that occur during the sample preparation procedure. This feature of eccDNA ONT data can help shed light on the composition and origin of eccDNA in cells [43, 56]. The authors showed that the *ONSEN* and *EVD* elements produced almost equal amounts of full-length (>5 Kb) and truncated (<1000 bp) eccDNAs in the *ddm1* background. Interestingly, growing *Arabidopsis* on A&Z medium resulted in a shift in eccDNA composition toward truncated eccDNAs [43]. These results are in contrast with our results for the *Salsa* and *Ketchup* elements. Here, we showed that *Salsa* and *Ketchup* TEs mainly produced full-length eccDNAs with one or two LTRs in tomato plants grown on A&Z media. These results suggest that eccDNA formation under similar growth conditions (for example, A&Z) may differ for different species and TEs.

In addition to the production of eccDNA under the relaxation of epigenetic control, *Salsa* and *Ketchup* also exhibited activity in the control sample, although to a lesser extent. EccDNA production poses a serious threat to genomic integrity and stability. The generation of eccDNA may result in genomic rearrangement via spontaneous reintegration into the genome, as has been shown for various types of eccDNA in eukaryotes [57]. In addition, it has been suggested that a high load of eccDNA may alter DNA repair pathways, leading to new genetic variations [56]. Additionally, eccDNAs may serve as a template for transcription of protein coding or non-coding RNAs further expanding the repertoire of possible consequences for the plant [58]. For inheritance to the next generation, eccDNA-mediated genetic changes need to be produced in the plant 'germ line' cells, such as meristematic cells of the shoot apical meristem (SAM), pollen, or egg cells. However, RTE transcription and transposition are limited in these cells through specific epigenetic mechanisms [59]. Thus, it remains an open question whether *Salsa* and *Ketchup* are capable of generating novel genetically inherited insertions, and whether their eccDNAs contribute to genome instability. This question could be answered by genomic analysis of M1 plants, which will be the subject of our future research.

Conclusion

Using nanopore whole-genome and eccDNA sequencing, we identified two novel families of tomato TEs, *Salsa* and *Ketchup*, that produce eccDNAs under both control and epigenetic stress conditions. We showed that these TEs occur in a tomato breeding line via interspecific introgression from *S. peruvianum*. Collectively, our results demonstrate that interspecific introgression may contribute to genetic and phenotypic diversity not only by

providing new genetic variants, but also by bringing new active TEs from other species.

Materials and methods

Plant material and in vitro growth conditions

Seeds of tomato line '812/18' used in this study were kindly provided by Tereshonkova Tatyana Arkadyevna (All-Russian Research Institute of Vegetable Production, Moscow, Russia). Tomato plants were grown on ½ MS medium supplemented with 4 mg/ml α-amanitin and 8 mg/ml zebularine for two weeks under a long-day photoperiod (16/8).

Total DNA isolation

Total DNA was isolated from two-week-old seedlings using the modified CTAB method described by Pucker (https://www.protocols.io/view/plant-dna-extraction-and-preparation-for-ont-seque-kxygxenmkv8j/v1).

eccDNA isolation and sequencing

For eccDNA isolation we used the techniques described by Lanciano et al. [51] and Wang et al. [60] with modifications. Briefly, to remove linear DNA, 1 μg of total DNA was treated with 1 μl (10 U/μl) of PlasmidSafe DNase supplemented with 2 μl of ATP (25 mM) and 5 μl of 10× PlasmidSafe buffer in a volume of 50 μl. The reaction was incubated for 72 h with additional reagents (0.1 μl enzyme, 0.2 μl ATP, 0.3 μl buffer) was added every 24 h, followed by incubation at 72 °C for 30 min. Precipitation of eccDNA was carried out by overnight incubation at -20 °C in the presence of 0.1 volume of 3 M sodium acetate (pH 5.2) and 2.5 × volume of absolute ethanol, followed by centrifugation at 12,000 × g for 30 min. The eccDNA pellet was washed with ice-cold 70% ethanol and dissolved in 10 μl of deionized nuclease-free water. For eccDNA amplification using random RCA, 2 μL phi29 polymerase (Thermo Scientific, EP0091), 2 μL 10× phi29 reaction buffer, 5 μL 10 mM dNTP, and 1 μL 500 μM exo-resistant oligo (NpNpNpNpNpSNpSN, where p is phosphodiester and pS is the phosphorothioate group) with the addition of nuclease-free water to a final volume of 20 μL. The reaction was preheated to 95 °C for 5 min, ramped to 30 °C at a 1% ramp rate on a thermocycler, and incubated for 36 h at 30 °C. The enzyme was inactivated by heating the mixture at 65*C for 10 min. For debranching, 500 ng of RCA amplicons were treated with T7 endonuclease 5 μL of 10× reaction buffer and 1 μL of T7 endonuclease I (New England Biolabs, M0302S) in a 50 μL reaction volume. After incubation at 37 °C for 15 min, the reaction was stopped immediately and purified by adding an equal volume of chloroform. The Debranched RCA product was precipitated by adding 1/10V 3 M sodium acetate (pH 5.2) and absolute ethanol (2.5 V),

Merkulov *et al. BMC Genomics*   (2024) 25:404

Page 9 of 11

followed by incubation at − 80 °C for 30 min and centrifugation at 12,000× g for 30 min. The pellet obtained was dissolved in nuclease-free water and used for nanopore sequencing.

Nanopore Library Preparation and sequencing
For eccDNA sequencing, library preparation was carried out with 500 ng of cDNA using Native Barcoding Expansion 1–12 (Oxford Nanopore Technologies (Oxford, UK), catalog no. EXP-NBD104), and the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies). Sequencing was performed using MinION equipped with an R9.4.1 flow cell.

For whole-genome sequencing, a fraction of short fragments was removed from 9 μg of total DNA using the Short-Read Eliminator Kit XL (PacBio, SKU 102-208-400), according to the manufacturer's recommendations. The library was prepared with 1 μg of long fragment-enriched DNA using the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies). Sequencing was carried out using PromethION P2 equipped with an R9.4.1 flow cell for 72 h. Basecalling was done using Guppy 6.4.6 (Oxford Nanopore Technologies, 2019). Adapters were trimmed from reads by Porechop 0.2.4 (Wick, n.d.) with default parameters.

Whole-genome sequencing and de novo assembly
The raw Nanopore long reads were assembled into sequence contigs using NextDenovo (version 2.5.2) [45] with the following parameters: 900 Mb of estimated genome size and for assembly minimap option, minimum overlap was set to 5000 bp, and other parameters were set by default. The program was run using a config file with the following parameters:

```
[General]
job_type=local.
job_prefix=nextDenovo_cherry.
task=all # 'all', 'correct', 'assemble'.
rewrite=yes # yes/no.
deltmp=yes.
rerun=3.
parallel_jobs=8.
input_type=raw.
read_type=ont.
input_fofn =./cherry.fofn.
workdir =./cherry_assembly.
[correct_option]
read_cutoff=1k.
genome_size=900 Mb.
pa_correction=2.
sort_options = -m 1 g -t 2.
minimap2_options_raw = -t 8.
correction_options = -p 10.
[assemble_option]
```

```
minimap2_options_cns = -t 8 --minlen 5000.
nextgraph_options = -a 1.
```

Draft assembly resulted in an N50 of approximately 16.7 Mb (16.721.217 bp) and a total length of approximately 800 Mb (813,056,715 bp), which was verified by BUSCO software (v5.5.0) [46] with both eukaryota and *viridiplatae* lineages, as well as both metaeuk and miniprot options, and by FastANI (version 1.33) [61] alignment on *Solanum lycopersicum* reference (genome assembly SL3.0). The percentage of complete BUSCOs ranged from 94.5% with miniprot option and eukaryota lineage; 97.8% with miniprot option and *viridiplantae* lineage; 98.0% with metaeuk option and eukaryota lineage to 99.3% with metaeuk option and *viridiplantae* lineage. SNPs were identified using Clair3 software [62] and their chromosome distribution was visualized by pycircos python package (https://github.com/ponnhide/pyCircos).

DNA amplification
Amplification was carried out using a Bio-Rad T100™ thermal cycler (Bio-Rad Laboratories, USA). A 25 μl reaction mixture contained: 1 μl DNA (25 ng), 2.5 μl 10× Taq Turbo buffer, 0.2 μl Hot Start Taq polymerase (5 units/μl), 1 μl 10 pmol of each primer, 0.5 μl dNTP (10 mM) and 18.8 μl nuclease-free water.

Validation of the insertions
For validation, 25 ng of total DNA/ecDNA was amplified using RTE-specific inverted PCR primers (Table S2).

eclDNA intermediates detection
For eclDNA amplification, the Sequence-Independent Retrotransposons Trapping (SIRT) method was used [63]. To form SIRT adaptors, equal volumes of 100 μM of SIRT_adaptor_1 (5′-GTAATACGACTCACTATAGGGCACGCGTCCACGACGGCCCGGGCTCCA-3′), and SIRT_adaptor_2 (5′-PO4-TGGAGCCC-3′) oligos were mixed and incubated at 95℃ for 10 min, followed by cooling to room temperature. The ligation mixture was prepared on ice with 300 ng of total DNA using 8 μl of adapters, 1.6 μl of 10× overnight buffer, and 1 μl of T4 ligase (100 U/μl), with nuclease-free water added to the final volume. 16 μl. Ligation was performed at 14℃ for 16 h, followed by enzyme inactivation at 65℃ for 10 min. The entire reaction volume was purified using 0.5 volumes of AMPure XP SPRI Reagent (Beckman Coulter, A63881) according to the manufacturer's instructions. DNA eluted in 30 μl was amplified using the adaptor-specific primer AP1 (5'-GTAATACGACTCACTATAGGGC-3') and the TE-specific primers listed in Table S1. The amplification program consisted of 95 °C for 3 min and 35 cycles of 95 °C for 10 s, 51 °C for 10 s, and 72 °C for

1 min. The resulting amplicons were separated on 1.5% agarose gel at 80 V for 60 min.

Bioinformatic analysis of eccDNA sequencing and data visualization

Raw eccDNA nanopore reads were mapped to the SL3.0 genome using the minimap2 software [64] with the following parameters: -ax map-ont -t 100. The obtained SAM files were converted to BAM format, sorted, and indexed using SAMtools [65]. To obtain the eccDNA peaks, the obtained sorted bam files were analyzed using the eccStructONT pipeline, as previously described [43].

For the evolutionary analysis, the genomes of *S. lycopersicum* var. *lycopersicum* (Heinz1706; ver. SL3.0) and *S. lycopersicum* var. *cerasiforme* (LA1673) were downloaded from https://solgenomics.net/. *S. lycopersicum* var. lycopersicum cv. M82, *S. lycopersicum* var. lycopersicum cv. ZY65, *S. penellii* (LA716), *S. pimpinellifolium* (LA1547), S. lycopersicoides (LA2951), *S. peruvianum* (LA0446), *S. corneliomulleri* (LA1331), *S. neorickii* (LA0247), *S. chmielewskii* (LA1028), S. chilense (LA1969), *S. habrochaites* (LA1777) and *S. galapagense* (LA0436) genome assemblies were downloaded from http://caastomato.biocloud.net/.

An alignment and tree visualisation were made using ggplot2 (version 3.4.4) [66], ggtree (version 3.8.2) [67] and ggmsa (version 1.6.0) [68].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10314-1.

Supplementary Material 1

Supplementary Material 2

Author contributions
Conceptualization, I.K. and P.M.; methodology, I.K. and P.M.; software, I.K.; validation, M.S. and G.P.; formal analysis, I.K. and P.M.; investigation, I.K., P.M., M.S., G.P. V.M.; resources, I.K.; data curation, P.M., V.M.; writing—original draft preparation, P.M, I.K.; writing—review and editing, I.K.; visualization, P.M., I.K.; supervision, I.K.; project administration, I.K.; funding acquisition, I.K. All authors have read and agreed to the published version of the manuscript.

Data availability
Sequence data that support the findings of this study have been deposited in NCBI under BioProject accession number: PRJNA1077878.

Declarations

Competing interests
The authors declare no competing interests.

Conflict of interest
The authors declare no conflict of interest.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

References
1. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3(5):329–41.
2. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, Harmer SL, Zilberman D. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell. 2013;153(1):193–205.
3. Li G, Wang L, Yang J, He H, Jin H, Li X, Ren T, Ren Z, Li F, Han X, et al. A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. Nat Genet. 2021;53(4):574–84.
4. Su X, Wang B, Geng X, Du Y, Yang Q, Liang B, Meng G, Gao Q, Yang W, Zhu Y, et al. A high-continuity and annotated tomato reference genome. BMC Genomics. 2021;22(1):898.
5. Cuerda-Gil D, Slotkin RK. Non-canonical RNA-directed DNA methylation. Nat Plants. 2016;2(11):16163.
6. Matzke MA, Kanno T, Matzke AJ. RNA-Directed DNA methylation: the evolution of a Complex Epigenetic Pathway in Flowering plants. Annu Rev Plant Biol. 2015;66:243–67.
7. Panda K, Slotkin RK. Long-read cDNA sequencing enables a Gene-Like transcript annotation of transposable elements. Plant Cell. 2020;32(9):2687–98.
8. He L, Huang H, Bradai M, Zhao C, You Y, Ma J, Zhao L, Lozano-Duran R, Zhu JK. DNA methylation-free Arabidopsis reveals crucial roles of DNA methylation in regulating gene expression and development. Nat Commun. 2022;13(1):1335.
9. Zhao L, Zhou Q, He L, Deng L, Lozano-Duran R, Li G, Zhu JK. DNA methylation underpins the epigenomic landscape regulating genome transcription in Arabidopsis. Genome Biol. 2022;23(1):197.
10. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49–61.
11. Kim S, Park J, Yeom SI, Kim YM, Seo E, Kim KT, Kim MS, Lee JM, Cheong K, Shin HS, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. 2017;18(1):210.
12. Penin AA, Kasianov AS, Klepikova AV, Kirov IV, Gerasimov ES, Fesenko AN, Logacheva MD. High-resolution transcriptome Atlas and Improved Genome Assembly of Common Buckwheat, Fagopyrum esculentum. Front Plant Sci. 2021;12:612382.
13. Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res. 2005;110(1–4):91–107.
14. Galindo-Gonzalez L, Mhiri C, Deyholos MK, Grandbastien MA. LTR-retrotransposons in plants: engines of evolution. Gene. 2017;626:14–25.
15. Cai X, Lin R, Liang J, King GJ, Wu J, Wang X. Transposable element insertion: a hidden major source of domesticated phenotypic variation in Brassica rapa. Plant Biotechnol J. 2022;20(7):1298–310.
16. Yao J, Dong Y, Morris BA. Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. Proc Natl Acad Sci U S A. 2001;98(3):1306–11.
17. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. Science. 2004;304(5673):982.
18. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell. 2012;24(3):1242–55.
19. Jiang N, Gao D, Xiao H, van der Knaap E. Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. Plant J. 2009;60(1):181–93.

20. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008;319(5869):1527–30.

21. Jiang NV, Wu S, Van Knaap S. Der: Rider Transposon insertion and phenotypic change in Tomato. Plant Transposable Elem Top Curr Genet 2012:297–312.

22. Roldan MVG, Perilleux C, Morin H, Huerga-Fernandez S, Latrasse D, Benhamed M, Bendahmane A. Natural and induced loss of function mutations in SIMBP21 MADS-box gene led to jointless-2 phenotype in tomato. Sci Rep. 2017;7(1):4402.

23. Quadrana L. The contribution of transposable elements to transcriptional novelty in plants: the FLC affair. Transcription. 2020;11(3–4):192–8.

24. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. Curr Opin Plant Biol. 2018;42:23–9.

25. McClintock B. The significance of responses of the genome to challenge. Science. 1984;226(4676):792–801.

26. Liu B, Wendel JF. Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome. 2000;43(5):874–80.

27. Kashkush K, Feldman M, Levy AA. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics. 2002;160(4):1651–9.

28. Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW, Martienssen R, Comai L. Genomic changes in synthetic Arabidopsis polyploids. Plant J. 2005;41(2):221–30.

29. Paz RC, Rendina Gonzalez AP, Ferrer MS, Masuelli RW. Short-term hybridisation activates Tnt1 and Tto1 Copia retrotransposons in wild tuber-bearing Solanum species. Plant Biol (Stuttg). 2015;17(4):860–9.

30. Wang HY, Tian Q, Ma YQ, Wu Y, Miao GJ, Ma Y, Cao DH, Wang XL, Lin C, Pang J, et al. Transpositional reactivation of two LTR retrotransposons in rice-Zizania recombinant inbred lines (RILs). Hereditas. 2010;147(6):264–77.

31. Usai G, Mascagni F, Vangelisti A, Giordani T, Ceccarelli M, Cavallini A, Natali L. Interspecific hybridisation and LTR-retrotransposon mobilisation-related structural variation in plants: a case study. Genomics. 2020;112(2):1611–21.

32. Gantuz M, Morales A, Bertoldi MV, Ibanez VN, Duarte PF, Marfil CF, Masuelli RW. Hybridization and polyploidization effects on LTR-retrotransposon activation in potato genome. J Plant Res. 2022;135(1):81–92.

33. Kaloshian I, Yaghoobi J, Liharska T, Hontelez J, Hanson D, Hogan P, Jesse T, Wijbrandi J, Simons G, Vos P, et al. Genetic and physical localization of the root-knot nematode resistance locus Mi in tomato. Mol Gen Genet. 1998;257(3):376–85.

34. Tanksley SD, Bernachi D, Beck-Bunn T, Emmatty D, Eshed Y, Inai S, Lopez J, Petiard V, Sayama H, Uhlig J, et al. Yield and quality evaluations on a pair of processing tomato lines nearly isogenic for the Tm2a gene for resistance to the tobacco mosaic virus. Euphytica. 1998;99(2):77–83.

35. Zamir D, Ekstein-Michelson I, Zakay Y, Navot N, Zeidan M, Sarfatti M, Eshed Y, Harel E, Pleban T, van-Oss H, et al. Mapping and introgression of a tomato yellow leaf curl virus tolerance gene, TY-1. Theor Appl Genet. 1994;88(2):141–6.

36. Yang X, Caro M, Hutton SF, Scott JW, Guo Y, Wang X, Rashid MH, Szinay D, de Jong H, Visser RG, et al. Fine mapping of the tomato yellow leaf curl virus resistance gene Ty-2 on chromosome 11 of tomato. Mol Breed. 2014;34(2):749–60.

37. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sorensen I, Lichtenstein G, et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet. 2014;46(9):1034–8.

38. Zhang C, Liu L, Wang X, Vossen J, Li G, Li T, Zheng Z, Gao J, Guo Y, Visser RG, et al. The Ph-3 gene from Solanum pimpinellifolium encodes CC-NBS-LRR protein conferring resistance to Phytophthora infestans. Theor Appl Genet. 2014;127(6):1353–64.

39. Dominguez M, Dugas E, Benchouaia M, Leduque B, Jimenez-Gomez JM, Colot V, Quadrana L. The impact of transposable elements on tomato diversity. Nat Commun. 2020;11(1):4058.

40. Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Diez MJ, Francis D, Causse M, van der Knaap E, Canizares J. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics. 2015;16(1):257.

41. Thieme M, Lanciano S, Balzergue S, Daccord N, Mirouze M, Bucher E. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biol. 2017;18(1):134.

42. Ibsen G, Nielsen J. The great Tomato Book. Ten Speed; 1999.

43. Merkulov P, Egorova E, Kirov I. Composition and structure of Arabidopsis thaliana Extrachromosomal circular DNAs revealed by Nanopore Sequencing. Plants (Basel) 2023, 12(11).

44. Tomato Genome C. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635–41.

45. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K et al. An efficient error correction and accurate assembly tool for noisy long reads. 2023:2023.2003.2009.531669.

46. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data Quality and Beyond. Curr Protoc. 2021;1(12):e323.

47. Fukai E, Yoshikawa M, Shah N, Sandal N, Miyao A, Ono S, Hirakawa H, Akyol TY, Umehara Y, Nonomura KI, et al. Widespread and transgenerational retrotransposon activation in inter- and intraspecies recombinant inbred populations of Lotus japonicus. Plant J. 2022;111(5):1397–410.

48. Paszkowski J. Controlled activation of retrotransposition for plant breeding. Curr Opin Biotechnol. 2015;32:200–6.

49. Tanskanen JA, Sabot F, Vicient C, Schulman AH. Life without GAG: the BARE-2 retrotransposon as a parasite's parasite. Gene. 2007;390(1–2):166–74.

50. Sabot F. Tos17 rice element: incomplete but effective. Mob DNA. 2014;5(1):10.

51. Lanciano S, Carpentier MC, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, Ghesquiere A, Panaud O, Mirouze M. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLoS Genet. 2017;13(2):e1006630.

52. Kwolek K, Kedzierska P, Hankiewicz M, Mirouze M, Panaud O, Grzebelus D, Macko-Podgorni A. Diverse and mobile: eccDNA-based identification of carrot low-copy-number LTR retrotransposons active in callus cultures. Plant J. 2022;110(6):1811–28.

53. Esposito S, Barteri F, Casacuberta J, Mirouze M, Carputo D, Aversano R. LTR-TEs abundance, timing and mobility in Solanum commersonii and S. tuberosum genomes following cold-stress conditions. Planta. 2019;250(5):1781–7.

54. Garfinkel DJ, Stefanisko KM, Nyswaner KM, Moore SP, Oh J, Hughes SH. Retrotransposon suicide: formation of Ty1 circles and autointegration via a central DNA flap. J Virol. 2006;80(24):11920–34.

55. Yang F, Su W, Chung OW, Tracy L, Wang L, Ramsden DA, Zhang ZZZ. Retrotransposons hijack alt-EJ for DNA replication and eccDNA biogenesis. Nature. 2023;620(7972):218–25.

56. Zhang P, Mbodj A, Soundiramourtty A, Llauro C, Ghesquiere A, Ingouff M, Keith Slotkin R, Pontvianne F, Catoni M, Mirouze M. Extrachromosomal circular DNA and structural variants highlight genome instability in Arabidopsis epigenetic mutants. Nat Commun. 2023;14(1):5236.

57. Arrey G, Keating ST, Regenberg B. A unifying model for extrachromosomal circular DNA load in eukaryotic cells. Semin Cell Dev Biol. 2022;128:40–50.

58. Peng H, Mirouze M, Bucher E. Extrachromosomal circular DNA: a neglected nucleic acid molecule in plants. Curr Opin Plant Biol. 2022;69:102263.

59. Nguyen V, Gutzat R. Epigenetic regulation in the shoot apical meristem. Curr Opin Plant Biol. 2022;69:102267.

60. Wang Y, Wang M, Zhang Y. Purification, full-length sequencing and genomic origin mapping of eccDNA. Nat Protoc. 2023;18(3):683–99.

61. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.

62. Luo J, Ding H, Shen Z, Zhai H, Wu Z, Yan C, Luo H. BreakNet: detecting deletions using long reads and a deep learning approach. BMC Bioinformatics. 2021;22(1):577.

63. Griffiths J, Catoni M, Iwasaki M, Paszkowski J. Sequence-independent identification of active LTR retrotransposons in Arabidopsis. Mol Plant. 2018;11(3):508–11.

64. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.

65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S: the sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

66. Wickham H. ggplot2: elegant graphics for data analysis. Springer International Publishing; 2016.

67. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. 2017, 8(1):28–36.

68. Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, Wu T, Huang H, Zhan L, Li L et al. Ggmsa: a visual exploration tool for multiple sequence alignment and associated data. Brief Bioinform 2022, 23(4).

Publisher's Note