*https://doi.org/10.1093/nar/gkac794*

# Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line

**Shunhua Han [1,†], Guilherme B. Dias [1,2,†], Preston J. Basting [1], Raghuvir Viswanatha [3], Norbert Perrimon [3,4] and Casey M. Bergman [1,2,*]**

[1]Institute of Bioinformatics, University of Georgia, 120 E. Green St., Athens, GA, USA, [2]Department of Genetics, University of Georgia, 120 E. Green St., Athens, GA, USA, [3]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA, USA and [4]Howard Hughes Medical Institute, Boston, MA, USA

Received February 15, 2022; Revised July 21, 2022; Editorial Decision August 27, 2022; Accepted September 16, 2022



## ABSTRACT

**Animal cell lines often undergo extreme genome restructuring events, including polyploidy and segmental aneuploidy that can impede *de novo* whole-genome assembly (WGA). In some species like *Drosophila*, cell lines also exhibit massive proliferation of transposable elements (TEs). To better understand the role of transposition during animal cell culture, we sequenced the genome of the tetraploid *Drosophila* S2R+ cell line using long-read and linked-read technologies. WGAs for S2R+ were highly fragmented and generated variable estimates of TE content across sequencing and assembly technologies. We therefore developed a novel WGA-independent bioinformatics method called TELR that identifies, locally assembles, and estimates allele frequency of TEs from long-read sequence data (https://github.com/bergmanlab/telr). Application of TELR to a ∼130x PacBio dataset for S2R+ revealed many haplotype-specific TE insertions that arose by transposition after initial cell line establishment and subsequent tetraploidization. Local assemblies from TELR also allowed phylogenetic analysis of paralogous TEs, which revealed that proliferation of TE families *in vitro* can be driven by single or multiple source lineages. Our work provides a model for the analysis of TEs in complex heterozygous or polyploid genomes that are recalcitrant to WGA and yields new insights into the mechanisms of genome evolution in animal cell culture.**

## INTRODUCTION

Cell lines are commonly used in biological and biomedical research, however little is known about how cell line genomes evolve *in vitro*. For decades, it has been well-established that immortalized cell lines derived from plant or animal tissues often develop polyploidy or aneuploidy during routine cell culture (1–4). More recently, the use of DNA sequencing has further revealed that segmental aneuploidy and other types of submicroscopic structural variation are widespread in cell lines (5–14). Together, these observations indicate that cells in culture often evolve complex genome architectures that deviate substantially from their original source material. Resolving the evolutionary processes that govern the transition from wild-type to complex cell line genome architectures is important for understanding the stability of cell line genotypes and the reproducibility of cell-line-based research. However, the complexity of cell line genomes can impose limitations on efforts to perform *de novo* whole-genome assembly (WGA) (9,15,16) and thus limit the ability to study cell line genome structure and evolution using traditional WGA-based bioinformatics approaches.

Like many animal cell lines, Schneider-2 (S2) cells from the model insect *Drosophila* have undergone polyploidization (8,17), and display substantial small- and large-scale segmental aneuploidy (5,8,14). In addition, S2 and other *Drosophila* cell lines exhibit a higher abundance of transposable element (TE) sequences compared to whole flies (18–20), with TE families that are abundant in S2 cells differing from those amplified in other *Drosophila* cell lines (20–23). However, little is known about TE sequence variation in S2 cells or other *Drosophila* cell lines. For example, it is generally unknown whether the proliferation of particular TE families in *Drosophila* cell lines is caused by one or more source lineages (24). The lack of understanding about TE sequences in *Drosophila* cell lines is mainly due to previous studies using short-read sequencing data (14,20,22), which typically does not allow complete assembly of TE insertions or other structural variants (25–28).

Recent advances in long-read DNA sequencing technologies have substantially improved the quality of WGAs, including a better representation of repetitive sequences such as TEs (29). In *Drosophila*, long-read WGAs of homozy-

*To whom correspondence should be addressed. Tel: +1 706 542 1764; Fax: +1 706 542 3910; Email: cbergman@uga.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

gous diploid genomes such as those from inbred fly stocks can achieve high contiguity and permit detailed analysis of structural variation including TE insertions (29–36). However, successful WGA using long reads remains limited by complex genome features including polyploidy, heterozygosity, and high repeat content, all of which are present in cell lines such as *Drosophila* S2 cells (5,8,17–20,22). In fact, the state-of-the-art long-read assemblies of wild-type diploid genomes still suffer from the presence of repeats and heterozygosity, which may result in assembly gaps and haplotype duplication artifacts (37,38). Therefore, assembly of a complex *Drosophila* cell line genome is likely to result in substantially more fragmented WGAs than those generated from homozygous diploid fly stocks (Figure 1), and this degradation of assembly quality could impact the subsequent analysis of TE sequences.

To gain better insight into the role of transposition during genome evolution in animal cell culture, here we sequenced the genome of a commonly-used variant of S2 cells, the S2R+ cell line (39), using PacBio long-read and 10x Genomics linked-read technologies. As predicted, WGAs of S2R+ from long-read sequencing data were highly fragmented and yielded highly variable estimates of TE content using different assembly methods. To circumvent the limitations of WGA and characterize TE content in *Drosophila* cell lines, we developed a novel TE detection tool called TELR (Transposable Elements from Long Reads, pronounced 'Teller') that can predict non-reference TE insertions based on a long-read sequence dataset, reference genome, and TE library. Importantly, TELR can detect haplotype-specific TE insertions, reconstruct TE sequences, and estimate intra-sample TE allele frequencies (TAFs) from complex genomes that are not amenable to WGA. We applied TELR to our PacBio long-read dataset for S2R+ and similar datasets for a geographically-diverse panel of *D. melanogaster* inbred fly strains from the *Drosophila* Synthetic Population Resource (DSPR) (40). We discovered a large number of haplotype-specific TE insertions from a subset of LTR retrotransposon families in the tetraploid S2R+ cell line. We inferred that these haplotype-specific insertions came from transposition events that occurred *in vitro* after initial cell line establishment and subsequent tetraploidization (8,17). We also performed phylogenomic analysis on the full-length TE sequences that were assembled by TELR, which revealed that amplification of TE families in *Drosophila* cell lines can be caused by activity of one or multiple source lineages. Together, our work provides a novel computational framework to study polymorphic TEs in complex heterozygous or polyploid genomes and improves our understanding of the mechanisms of genome evolution during long-term animal cell culture.

## MATERIALS AND METHODS

### Cell culture

An initial sample of S2R+ cells, which we define as passage 0, was obtained from a routine freeze of cells made by the *Drosophila* RNAi Screening Center (DRSC). Cells from passage 0 were defrosted and recovered in Schneider's *Drosophila* medium (Thermo) containing 10% FBS (Thermo) and 1X Penicillin–Streptomycin (Thermo), then expanded continually for two additional passages in T75 flasks. Aliquots of cells from passage 3 flasks were frozen, and the remaining cells were expanded to 10 T75 flasks (passage 4A). Passage 4A cells were pooled and harvested to make DNA for PacBio libraries. A frozen stock was defrosted and expanded for two additional passages (passages 4B–5B). Passage 5B cells were harvested to make DNA for 10x Genomics libraries. The provenance of the cell line samples used in this study is depicted in Supplementary Figure S1.

### Fly stocks

A stock of *Drosophila melanogaster* strain A4 from the Drosophila Synthetic Population Resource (DSPR) (41) was obtained from Stuart Macdonald (University of Kansas) and reared on Instant Drosophila Medium (Carolina Biological, Cary, NC, USA) until used for DNA extraction.

### PacBio library preparation and sequencing

Cells from ten confluent T75 flasks from passage 4A were scraped into a 15 ml Falcon tube and centrifuged at 300 × g for 3 min. The pellet was washed in 10 ml of 1× PBS, then resuspended in 7 ml of 1× PBS containing 35 ul of 10 mg/ml RNAse A (Sigma). 200 ul of resuspended cells were aliquoted to 32 Eppendorf tubes containing 200 ul of buffer AL from the Qiagen Blood & Tissue kit, mixed gently by inversion, and incubated at 37°C for 30 min. 20 ul of Proteinase K solution from the Qiagen Blood & Tissue kit was then added to each tube and mixed gently by inversion. One volume of phenol:chloroform:isoamyl alcohol (24:24:1) was then added and inverted gently to mix for 1 min. Tubes were then spun for 5 min at 21 000 × g. 180 ul of the upper aqueous phase were then removed from each tube, and pairs of tubes were combined. 400 ul of chloroform was then added to each of the 16 tubes, shaken for 1 min to mix, and spun at max speed for 5 min. The top 300 ul was removed and pairs of tubes were combined. 600 uL of chloroform was added to each of the eight tubes, gently inverted 10 times to mix, and then spun at max speed for 5 min. 400 ul of the aqueous phase was removed and pairs of tubes were combined. 1/10 volume of 3M NaOAc was added to each of the four tubes, the remained of the tube was filled with absolute ethanol and then placed at –20°C overnight. Tubes were then spun 21 000 × g at 4°C for 15 min, and the supernatant was decanted over paper towels. 70% ethanol was then added to tubes, the pellet was gently resuspended with a P1000 tip, and then placed on ice for 10 min. Tubes were then spun 21 000 × g at 4°C for 15 min, and the supernatant was decanted over paper towels. The pellet was then resuspended in 50 ul of Buffer EB from the Qiagen Blood & Tissue kit, and gently pipetted with a P200 tip 5 times to resuspend. Purified S2R+ DNA was then used to generate PacBio SMRTbell libraries using the Procedure & Checklist 20 kb Template Preparation using BluePippin Size Selection protocol. The SMRTbell library was sequenced using 31 SMRT cells on a PacBio RS II instrument with a movie time of 240 minutes per SMRT cell, generating a total of 3,510,012 reads (~28.5 Gbp).

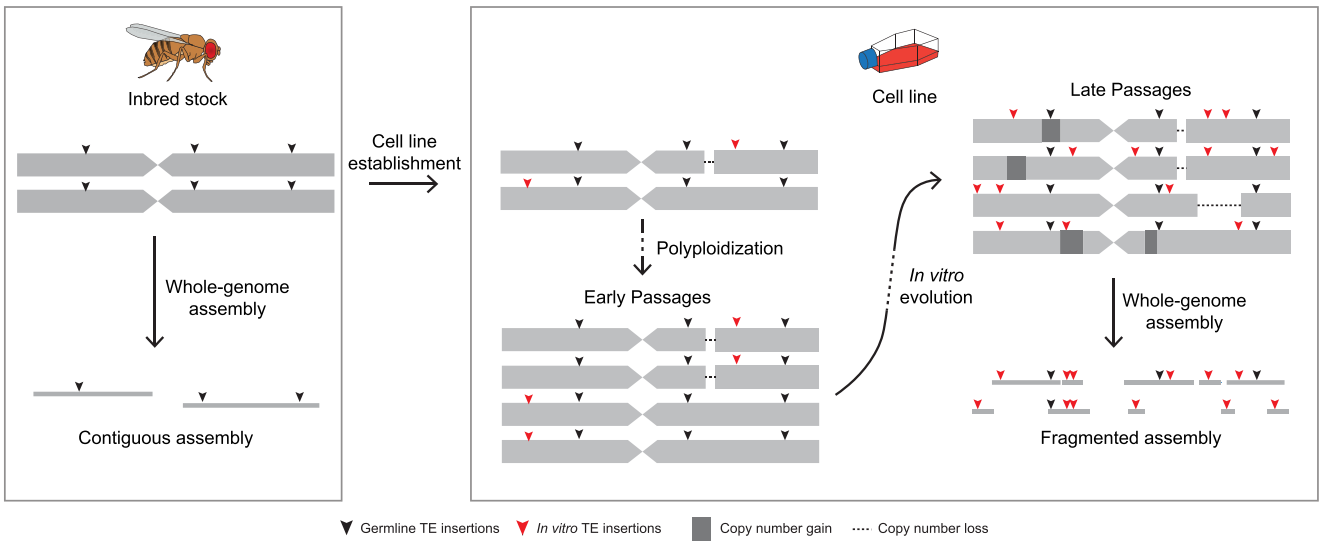


**Figure 1.** Complex genome architecture can hinder whole-genome assembly of long-term cultured cell lines. Inbred fly stocks have a highly homozygous diploid genome architecture that allows for contiguous whole-genome assembly (WGA). In contrast, cell lines established from such inbred fly stocks often undergo polyploidization and accumulate heterozygous variants including copy number alterations and haplotype-specific TE insertions during long-term culture. The complexity of cell line genome architecture is likely to lead to highly fragmented WGAs and, as a result, may limit the utility of using WGA-based approaches to study TE content and sequence evolution in animal cell lines.

### 10× Genomics library preparation and sequencing

Genomic DNA extraction of S2R+ cells followed the 10x Genomics 'Salting Out Method for DNA Extraction from Cells' protocol (https://support.10xgenomics.com/permalink/5H0Dz33gmQOea02iwQU0iK) adapted from (42). Genomic DNA for *D. melanogaster* strain A4 linked-read library was obtained from a single female fly following the 10x Genomics recommended protocol for DNA purification from single insects (https://support.10xgenomics.com/permalink/7HBJeZucc80CwkMAmA4oQ2). Purified DNA was precipitated by addition of 8 mL of ethanol and resuspended in TE buffer and size was analyzed by TapeStation (Agilent) prior to library preparation. Linked-read libraries were then prepared for both S2R+ and A4 after DNA size selection with BluePippin to remove fragments shorter than 15 kb. Libraries were prepared following the 10x Genomics Chromium Genome Reagent Kit Protocol v2 (RevB) using a total DNA input mass of 0.6 ng for each sample. The linked-read libraries were sequenced on an Illumina NextSeq 500 instrument mid-output flow cell with 150 bp paired-end layout, generating 95,280,430 reads for S2R+ (∼13.3 Gbp) and 127,009,398 reads for A4 (∼17.7 Gbp).

### Whole-genome assembly and QC

Raw PacBio reads from S2R+ (generated here; SRX7661404) and A4 from (30) (SRX4713156) were separately used as input for whole-genome assembly with Canu (v2.1.1; genomeSize=180m corOutCoverage=200 'batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50' -pacbio-raw), FALCON-Unzip (pb-falcon v0.2.6; seed coverage = 30, genome_size = 180000000), wtdbg2 v2.5 (-x rs -g 180m), and Flye (v2.8.2) (43–46). The reads were re-aligned to the resulting assemblies with pbmm2 (v1.3.0; –preset SUBREAD –sort) and the assemblies were polished with the Arrow algorithm from GenomicConsensus (v2.3.3) using default parameters. FALCON-Unzip performs read re-alignment and Arrow polishing automatically as part of its phasing pipeline.

10x Genomics linked-reads generated here were used as input for whole-genome assembly with Supernova (v2.1.1) for S2R+ (–maxreads=61508497) and A4 (–maxreads=77907944) (47). The optimal –maxreads parameter was calculated by Supernova in a previous run to avoid excessive coverage. Supernova assemblies were exported in pseudohap2 format and pseudo-haplotype1 was analyzed.

10x Genomics reads from S2R+ and A4 were also barcode-trimmed with LongRanger (v2.2.2; basic pipeline) (48) to create standard paired-end reads as input to SPAdes (v3.15.0) using default parameters (49).

All assemblies were filtered to remove redundancy using the sequniq program from GenomeTools (v1.6.1) (50). General assembly statistics were calculated with the stats.sh utility from BBMap (v38.83) (51). Assembly completeness was assessed with BUSCO (v4.0.6) (52,53) and the Diptera ortholog gene set from OrthoDB (v10) (54).

### Assessment of overall TE content

Transposable elements were annotated in all WGAs with RepeatMasker (v4.0.7; -s -no_is -nolow -x -e ncbi) (https://www.repeatmasker.org/RepeatMasker/) using v10.2 of the *D. melanogaster* canonical TE sequence library (https://github.com/bergmanlab/drosophila-transposons). TE abundance was calculated from RepeatMasker .out.gff files as the percentage of bases masked in each assembly. For this and subsequent analyses, we excluded the highly abundant and degenerate *INE-1* family since this family has been reported to be inactive in *Drosophila* for millions of years (55,56).

Barcode-trimmed 10× Genomics reads were also used as an assembly-free estimate of TE content in S2R+ and A4. Reads were filtered for adapters and low quality bases, and trimmed to 100 bp using fastp (v0.20.0; –max_len1 100 –max_len2 100 –length_required 100) (57). A random sample of 5 million read pairs (10 million reads) was extracted for each dataset using seqtk (v1.3; -s2) (https://github.com/lh3/seqtk) and masked using RepeatMasker (v4.0.7; -s -no_is -nolow -x -e ncbi) and v10.2 of the *D. melanogaster* canonical TE sequence library (https://github.com/bergmanlab/drosophila-transposons). Abundance for each TE family was calculated as the percentage of read bases that were RepeatMasked.

### Detection of non-reference TE insertions using long reads

Non-reference TEs were detected in PacBio long reads using a novel pipeline reported here called TELR (https://github.com/bergmanlab/telr). The TELR pipeline consists of four main stages: (i) general structural variant (SV) detection and filtering for TE insertion candidate, (ii) local reassembly and polishing of the TE insertion, (iii) identification of TE insertion coordinates and (iv) estimation of intra-sample TE insertion allele frequency.

In stage 1, PacBio or Oxford Nanopore long reads are aligned to the reference genome using NGMLR (v0.2.7) (58). The alignment output in BAM format is provided as input for Sniffles (v1.0.12) to detect structural variations (SVs) (58). TELR then filters for TE insertion candidates from SVs reported by Sniffles using the following criteria: (i) the type of SV is an insertion; (ii) the insertion sequence is available and (iii) the insertion sequences include hits to a user-provided TE library identified using RepeatMasker (v4.0.7; http://www.repeatmasker.org/).

In stage 2, all reads that support the TE insertion candidate locus based on Sniffles output are used as input for wtdbg2 (v2.5) (46) or flye (v2.8.3) (45) to assemble a local contig that covers the TE insertion for each candidate locus (46). Local assemblies are then polished using minimap2 (v2.20) (59) and wtdbg2 (v2.5) (46) or flye (v2.8.3) (45).

In stage 3, the TE library is aligned to the assembled TE insertion contigs using minimap2 and used to define TE-flank boundaries. TE family information in the TE region of each contig is annotated using RepeatMasker (v4.0.7). Sequences flanking the TE insertion are then re-aligned to the reference genome using minimap2 to determine the precise TE insertion coordinates and, if detected, the target site duplication (TSD) caused by the insertion on reference genome coordinates. If the locations of flanking sequences overlap on reference genome coordinates (up to a user-defined overlap threshold; default 20 bp), then the region of overlap defines the TSD and the insertion coordinates. If the locations of flanking sequences do not overlap on genome coordinates (up to a user-defined gap threshold; default 20 bp), then no TSD is reported and the gap between flanking sequences defines the insertion coordinates.

In stage 4, raw reads aligned to the reference genome are extracted within a 1 kb interval on either side of the insertion breakpoints initially defined by Sniffles. Extracted reads are then aligned to the assembled, polished contig to identify those that support the non-reference TE insertion and reference alleles, respectively, in following steps. (i) Reads are aligned to the forward strand of the contig, then the 5′ flanking sequence depth of coverage (5p_flank_cov) and 5′ TE depth of coverage (5p_te_cov) are calculated. (ii) Reads are then aligned to the reverse complement of the contig, and 5′ flanking sequence depth and 5′ TE depth in the reverse complement orientation are used to calculate corresponding values on the 3′ end of the insertion (3p_flank_cov and 3p_te_cov, respectively). Alignment to the reverse complement of the contig was performed to generate more accurate estimates of 3p_flank_cov and 3p_te_cov, since we found in simulated data that alignments of clipped and spanning reads beyond the 3′ junction of the TE insertion were under-reported by NGMLR in the forward orientation. (iii) The TE allele frequency (TAF) is estimated as (5p_te_cov/5p_flank_cov + 3p_te_cov/3p_flank_cov)/2. In all steps, reads that spanned both breakpoints of the TE insertion were counted towards coverage estimates at both the 5′ and 3′ ends.

In the current study, TELR (revision 80481c6d81efae62c624faf112278c6fbfbcab13) was applied to the S2R+ PacBio dataset and to a panel of 13 *D. melanogaster* strains from the *Drosophila* Synthetic Population Resource (DSPR) (Bioproject ID PRJNA418342) (40). The major arms of Release 6 of the *D. melanogaster* reference genome (chr2L, chr2R, chr3L, chr3R, chr4, chrX, chrY, chrM) (60) and v10.2 of the *D. melanogaster* canonical TE sequence library (https://github.com/bergmanlab/drosophila-transposons) were used for all TELR analyses. Local assembly was performed using wtdbg2 (46) and polishing of local assembles was performed using flye (45).

### Cross-validation of TELR results using short-read methods

To cross-validate results obtained by TELR, we employed two short-read TE detection methods implemented in McClintock (v2.0; revision 93369eff1c192132d8b27830310d149e53a2b608) (61) that output TAF values: ngs_te_mapper2 (22) and TEMP (62). 10× Genomics data obtained for S2R+ and A4 was barcode-trimmed with LongRanger (v2.2.2; basic pipeline) (48), de-interleaved, and trimmed to 100bp using fastp (v0.20.0; –max_len1 100 –max_len2 100 –length_required 100) (57). This data was downsampled to ∼50× mean mapped read depth for S2R+ (74,648,362 reads) and A4 (76,045,544 reads) before being used as paired-end input in McClintock to generate non-redundant non-reference TE insertion predictions.

### Construction of phylogenetic trees using TE sequences from TELR

TE sequences predicted, assembled, and polished by TELR on S2R+ and DSPR dataset were filtered for high-quality full-length TE sequences using the following criteria. (i) Predictions from DSPR strain A2 were excluded due to potential inversion-induced gain of heterozygosity (see RESULTS for details). (ii) Predictions from DSPR strain A7 were excluded due to potential sample contamination (see RESULTS for details). (iii) Sequences from chromosome X were excluded due to lower coverage compared

to autosomes and loss of heterozygosity (LOH) events. (iv) Sequences from low recombination regions were excluded using boundaries defined in (63) lifted over to dm6 coordinates. Normal recombination regions included in our analyses were defined as chrX:405967–20928973, chr2L:200000–20100000, chr2R:6412495–25112477, chr3L:100000–21906900, chr3R:4774278–31974278. We restricted our analysis to normal recombination regions since low recombination regions have high reference TE content which reduces the ability to predict non-reference TE insertions (64,65). (v) Only full-length TEs based on canonical sequence lengths were included. To do this, we first calculated the ratio between each TELR sequence length and the corresponding canonical sequence length. Next, we filtered TELR sequences for full-length copies using a 0.75–1.05 ratio cutoff for the *297* TE family and 0.95–1.05 ratio cutoff for other TE families. (vi) Only sequences with both 5′ and 3′ flanks mapped to reference genome were included. (vii) Only sequences from TE insertions with TAF estimated by TELR were included.

TELR sequences from each family that met these criteria were aligned with MAFFT (v7.487) (66). The multiple sequence alignments (MSAs) were filtered by trimAI (v1.4.rev15; parameters: -resoverlap 0.75 -seqoverlap 80) (67) to remove spurious sequences. The filtered MSAs were used as input to IQ-TREE (v2.1.4-beta; parameters: -m GTR+G -B 1000) (68) to generate maximum likelihood trees.

## RESULTS

### Fragmented assemblies yield variable estimates of TE content in the S2R+ genome

To better understand the process of TE amplification in the S2R+ cell line genome, we initially sought to use a *de novo* assembly-based approach by generating PacBio long-read (132X average depth) and 10x Genomics linked-read (89X average depth) sequencing data and assembled these data using a variety of state-of-the-art WGA software (43–47,49). All S2R+ whole-genome assemblies (WGAs) using long reads (Canu, FALCON-Unzip, wtdbg2, and Flye) or linked reads (Supernova) had better contiguities compared to a SPAdes assembly of standard Illumina paired-end short read data (Figure 2 A; Supplementary Table S1). However, S2R+ WGAs from different sequencing technologies and assemblers varied substantially in their contiguities and levels of duplicated BUSCOs (Figure 2A, B; Supplementary Table S1). Canu assembly of the S2R+ PacBio data displayed the highest level of BUSCO duplication (Figure 2B) and the longest total assembly length (Supplementary Table S1). We speculated that the high degree of BUSCO duplication in the Canu S2R+ assembly could be caused by haplotype-induced duplication artifacts in a partially-phased assembly that contained contigs from multiple haplotypes of the same locus (69,70). To test this, we took advantage of the fact that FALCON-Unzip leverages structural variants to phase heterozygous regions into a primary assembly ('FALCON-Unzip_p') and alternative haplotigs (43). Combining the primary FALCON-Unzip assembly with alternative haplotigs ('FALCON-Unzip_ph') resulted in a higher level of BUSCO duplication approaching those observed in the Canu assembly (Figure 2B). This result suggested that many regions of the S2R+ genome contain haplotype-specific structural variants that can lead to secondary haplotigs in the Canu and Falcon-Unzip assemblies, which consequently cause artifactual BUSCO duplication.

N50s for all S2R+ WGAs were less than 1 Mbp, which is more than ten-fold smaller than the size of assembled chromosome arms in the *Drosophila* reference genome (60). In support of this finding, poor contiguity has recently been observed for *de novo* assemblies of the related *Drosophila* S2 cell line using nanopore long-read data (71). To assess how S2R+ cell line WGAs compared to those from whole flies of inbred stocks, we also generated WGAs for a highly inbred *D. melanogaster* strain called A4 using available PacBio long-read data (110x average depth) from (40) and a 10x Genomics linked-read dataset for A4 generated in this study (118X average depth) using identical assembly software and parameters as we did for S2R+. We found that WGAs for A4 have reference-grade contiguities and exhibit lower variation in levels of BUSCO duplication than WGAs for the S2R+ cell line (Figure 2D, E; Supplementary Table S2). Given that the A4 strain is diploid homozygous (40), these results suggest that the highly fragmented WGAs for S2R+ are likely caused by polyploidy, aneuploidy, or heterozygosity in the S2R+ cell line genome rather than limitations caused by long- or linked-read sequence lengths or current assembly methods.

In addition to assembly quality, estimates of TE content in WGAs varied substantially across sequencing and assembly technologies for both S2R+ and A4 (Figure 2C, F; Supplementary Tables S1 and S2). Compared to unbiased estimates of TE content based on RepeatMasker analysis of unassembled short reads (dotted lines in Figure 2C, F) (72), long-read WGAs for both the S2R+ and A4 genomes typically gave similar or higher estimates of TE content, while short read WGAs always gave lower estimates. In particular, the Canu and Falcon-Unzip assemblies that we infer include alternative haplotigs gave the highest estimates of TE content relative to unassembled short read data, suggesting the possibility of haplotype-specific TE insertions in these assemblies. In addition to differences in overall TE content, we observed higher variation in the abundance of different TE families across sequencing and assembly technologies in WGAs for S2R+ (Supplementary Figure S2A) compared to A4 (Supplementary Figure S2B). This result indicates that WGA-based inferences about TE family abundance in S2R+ are highly dependent on sequencing and assembly technology. Despite this variation, higher estimates of overall TE content were observed in S2R+ WGAs relative to A4 WGAs for all sequencing or assembly technologies used (Figure 2C, F; Supplementary Tables S1 and S2). However, because of the relatively poor quality and high variation in estimates of TE content among WGAs generated from S2R+ long-read and linked-read data, we concluded that an alternative WGA-independent approach that is better suited to the complexities of cell line genome architecture was necessary to reliably study TE content in S2R+ cells.

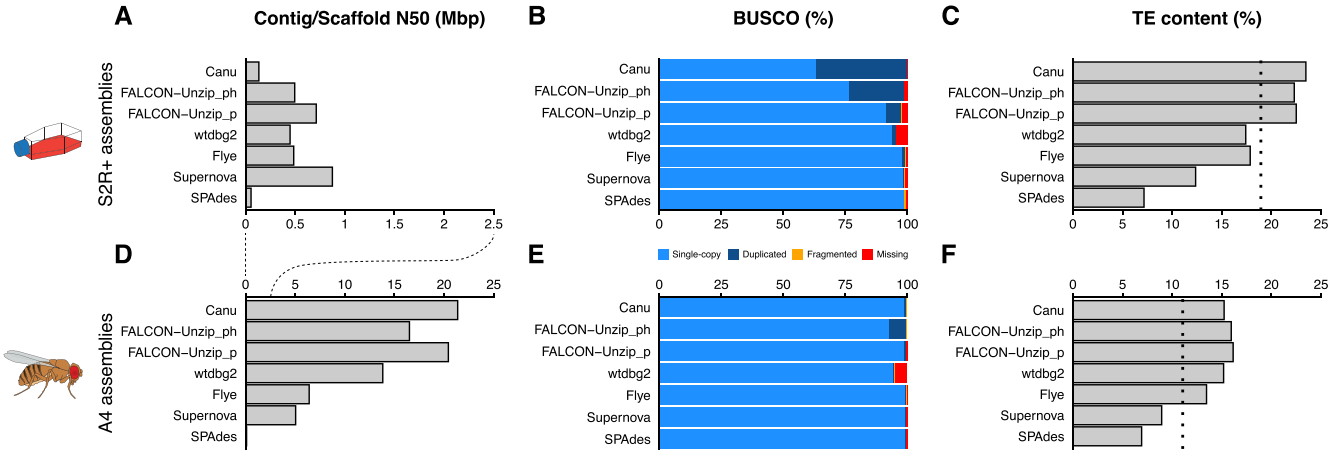


**Figure 2.** Contiguity, completeness, and TE content in whole-genome assemblies of S2R+ compared to those from an inbred fly strain. (**A** and **D**) Contig (Canu, FALCON-Unzip, and wtdbg2) and scaffold (Flye, Supernova, and SPAdes) N50 values for S2R+ and A4 whole-genome assemblies, respectively. (**B** and **E**) BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis with the Diptera gene set from OrthoDBv10 on S2R+ and A4 assemblies, respectively. (**C** and **F**) RepeatMasker estimates of TE content in WGAs of S2R+ and A4, respectively. Dotted lines in (**C**) and (**F**) represent RepeatMasker estimates of TE content from raw Illumina reads. 'FALCON-Unzip_p' represents primary contigs, 'FALCON-Unzip_ph' represents primary contigs + haplotigs. Note that the x-axis scale differs in (A) and (D).

### A novel long-read bioinformatics method reveals TE families enriched in S2R+ relative to wild-type *Drosophila* strains

To circumvent the impact of fragmented WGAs on the analysis of TE content in complex cell line genomes, we developed a new TE detection method called 'TELR' (Transposable Elements from Long Reads; https://github.com/bergmanlab/telr) that allows the identification, assembly, and allele frequency estimation of non-reference TE insertions using long-read data (Figure 3). Briefly, TELR first aligns long reads to a reference genome to identify insertion variants using Sniffles (58). The general pool of insertion variants identified by Sniffles is then filtered by aligning putative insertion sequences to a library of curated TE sequences to identify candidate TE insertion loci. For each candidate TE insertion locus, TELR then performs a local assembly using all reads that support the putative TE insertion event. Finally, TELR annotates TE sequence in each assembled contig, predicts the precise location of the TE insertion and (if detected) the TSD on reference genome coordinates, then remaps all reads in the vicinity of each insertion to the assembled TE contig to estimate TAF (see Materials and Methods for details). Evaluation on simulated *Drosophila* genomes demonstrated that TELR has high precision but variable recall to detect the location and family of non-reference TE insertions across different coverage, ploidies and zygosities (see Supplemental Text; Supplementary Table S3). For the ∼125X S2R+ dataset used here, TELR has >98% precision and >58% recall to detect non-reference TE insertions found in at least one haplotype of a tetraploid genome (Supplementary Table S3).

Applying TELR to regions of the *D. melanogaster* genome that are normally-recombining in whole flies, we identified 2408 non-reference TE insertions in S2R+, a ∼5-fold increase relative to the number identified in A4 ($n$ = 491; Figure 4A). These overall differences in non-reference TE abundance between S2R+ and A4 are unlikely to be caused by variation in coverage and read length between the S2R+ and A4 datasets, as shown by analysis of read length and coverage normalized datasets (Supplementary Figure S3). Despite a drop in the number of predictions in the normalized data relative to the full dataset, TELR still predicted substantially more TEs in S2R+ compared to A4 at all coverage levels (Supplementary Figure S3). This analysis also revealed that, unlike A4 which plateaued in the number of non-reference TE insertions at a normalized read depth of 50X, detection of non-reference TEs in S2R+ is likely not saturated even at 75×, consistent with the existence of TEs found at low allele frequency in the S2R+ sample. Therefore, in order to maximize TE prediction sensitivity, we used the complete non-normalized Pacbio data for S2R+ and all whole-fly strains in subsequent analyses.

Partitioning the number of non-reference TE insertions predicted by TELR in the complete S2R+ and A4 PacBio datasets by TE family revealed a subset of 14 TE families that are enriched in S2R+ relative to A4 (Figure 4B; Supplementary Figure S6). These S2R+ specific TE families consist mostly of long terminal repeat (LTR) retrotransposons from the *Gypsy*, *Pao* and *Copia* superfamilies, with the exceptions of *jockey* and *Juan* which are non-LTR retrotransposons in the *jockey* superfamily (Figure 4B; Supplementary Figure S6). The TE families revealed by TELR to be enriched in S2R+ relative to A4 were independently cross-validated using short-read sequences and two independent short-read TE detection methods (Supplementary Figure S4) (22,62).

We next used TELR to predict non-reference TEs in PacBio datasets for 13 geographically-diverse *D. melanogaster* inbred strains (including A4) from the DSPR project (40). This analysis revealed that S2R+ has more non-reference TE insertions than any of the DSPR strains surveyed (range: 445–660; Supplementary Figure S5). Partitioning TELR predictions by TE family reveals that only eight TE families account for ∼75% of non-reference insertions in S2R+, most of which are LTR retrotransposons

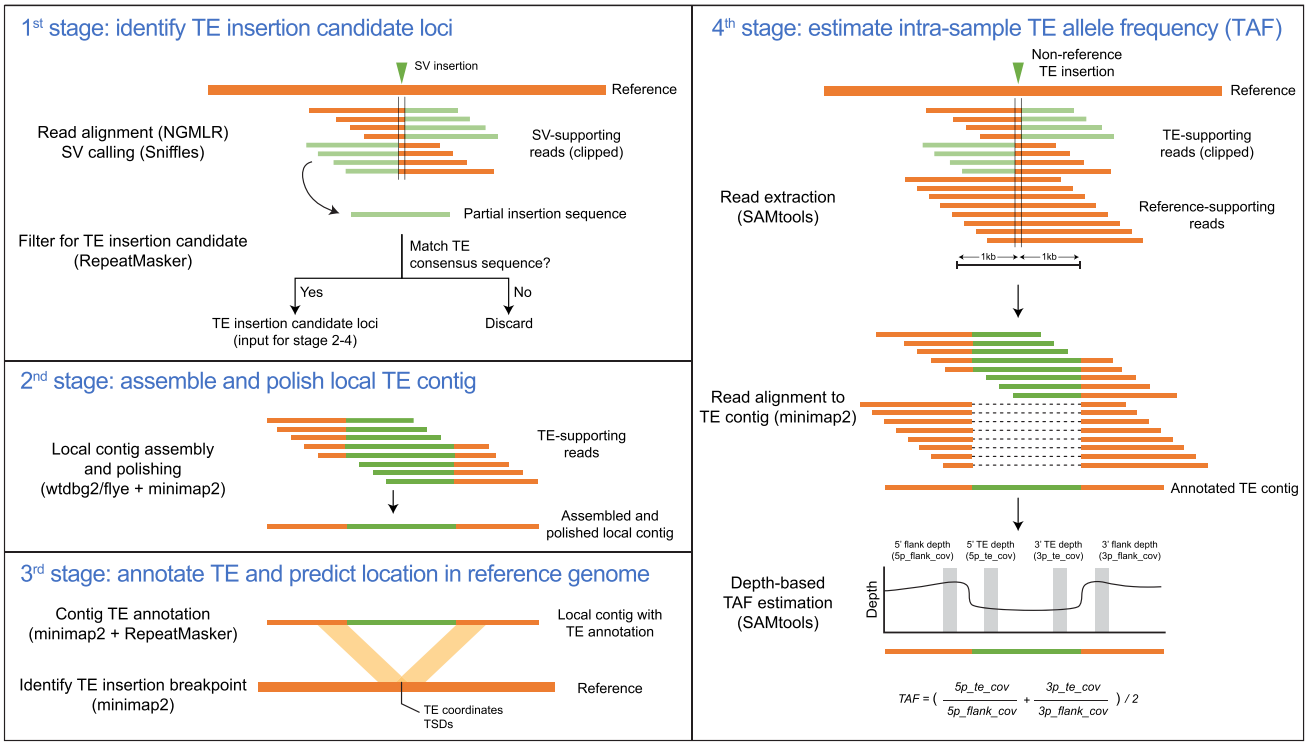


**Figure 3.** TELR workflow to predict non-reference TEs and estimate intra-sample TE allele frequency. TELR is a non-reference TE detection pipeline that uses long read sequencing data as input and consists of four main stages. In the first stage, TELR aligns long reads to a reference and identify insertion structural variants (SVs) using Sniffles (58). TELR then identifies candidate non-reference TE insertion loci by querying partial insertion sequences provided by Sniffles against a TE sequence library using RepeatMasker. In the second stage, TELR use all reads from Sniffles that support the insertion variant to assemble and polish local contigs using wtdbg2 (46) or flye (45). In the third stage, TE boundaries and family are annotated in the local contig using minimap2 (59) and RepeatMasker. Sequences flanking the TE in the local contig are then used to annotate coordinates and TSDs of the TE insertion on reference genome coordinates using minimap2. In the fourth stage, TELR determines the intra-sample allele frequency of each TE insertion by extracting all reads in a 2kb span around the insertion locus and aligning them to the TE contig. The mapped read depth over TE and flanking sequences are then used to calculate the intra-sample TE allele frequency (TAF).

(Supplementary Figure S6). In comparison, 10–16 TE families contribute ∼75% of all non-reference TE insertions in each of the DSPR strains, and they represent a more balanced distribution of LTR retrotransposons, non-LTR retrotransposons, and DNA transposons (Supplementary Figure S6). We also observed expansions of specific TE families in some DSPR strains, which we define as a greater than 3-fold increase in the number of non-reference TE insertions for a specific TE family relative to the mean values across all strains. For example, we see strain-specific expansions of *1360* ($n = 23$, mean = 7.2) in strain A2 (from Colombia), *hopper* ($n = 114$, mean = 18.5) in strain A6 (from USA), as well as *Doc* ($n = 112$, mean = 26.5) and *Quasimodo* ($n = 28$, mean = 7.1) in strain B2 (from South Africa) (Supplementary Figure S6).

Consistent with expected performance based on simulated genomes (Supplementary Table S3), between 46% and 55% of TELR predictions in S2R+ and DSPR samples were supported by TSD annotations (Supplementary Table S4). The distributions of TSD lengths for TELR predictions for the 20 TE families with greater than ten non-reference TEs in S2R+ were largely compatible with previous studies based on short read data (62,73,74). Specifically, TELR predictions for non-LTR retrotransposon families such as *Juan*, *jockey*, *F-element* and *Doc* exhibited variable TSD lengths generally in the 5-15 bp range, while LTR retrotransposon families typically exhibited tighter distributions with modal TSD lengths characteristic of their superfamily (*Gypsy*: 4 bp; *Pao*: 5 bp; *Copia*: 5 bp) (Supplementary Figure S7).

### Accurate estimation of intra-sample allele frequencies supports haplotype-specific TE insertion after tetraploidy in the S2R+ genome

An important feature of the TELR system is the ability to estimate the intra-sample allele frequency of non-reference TE insertions (Figure 3), which allowed us to observe drastic differences between S2R+ and A4 in genome-wide TAF patterns. TE insertions in S2R+ display a wide range of allele frequencies, with a striking difference in TAF distributions on the X chromosome relative to the autosomal arms (Figure 4C,D). In contrast, non-reference TEs in the highly-inbred strain A4 (41) are mostly enriched at TAF values ∼1 on all chromosome arms (Figure 4C,E). Broad-scale patterns of TAF distributions across the S2R+ and A4 genomes detected by TELR using long-read sequences were independently cross-validated using short-

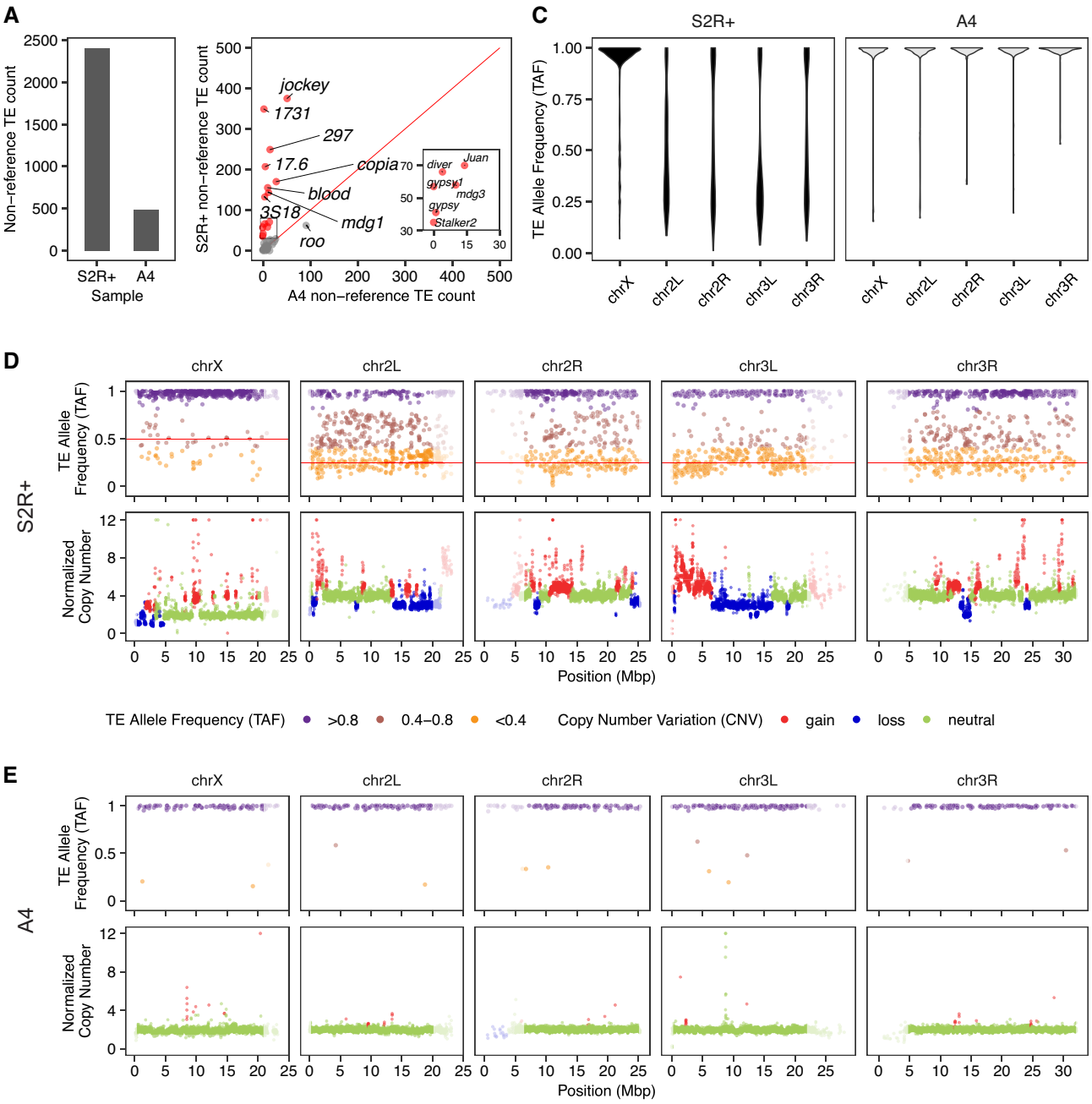


**Figure 4.** Abundance and allele frequency of non-reference TEs differs in the S2R+ cell line versus the A4 inbred fly stock. (**A**) Total number of non-reference TE predictions made by TELR for S2R+ and A4. (**B**) Number of non-reference TE predictions made by TELR for S2R+ and A4 partitioned by TE families. The 14 most enriched TE families in S2R+ relative to A4 highlighted in red. The insert zooms in on a set of six abundant TE families in S2R+ present in the black box in the main panel. (**C**) Genome-wide TE allele frequency (TAF) distribution for S2R+ and A4 partitioned by chromosome arm. (D, E) Genome-wide TAF and copy number profiles for S2R+ (**D**) and A4 (**E**). Low recombination regions in (D) and (E) are indicated by higher transparency.

read sequences and two independent short-read TE detection methods (Supplementary Figure S8) (22,62).

Like strain A4, non-reference TEs in other DSPR strains are mostly homozygous with TAF values enriched at the expected value of ∼1 for highly inbred diploid fly stocks (Supplementary Figure S9). However, our TELR analysis of DSPR datasets revealed two striking exceptions to this pattern. First, strain A2 displays mostly heterozygous TE insertions across chromosome arm 3R, which coincides with the presence of a known heterozygous chromosomal inversion in this strain (*In(3R)P*) that prevents full inbreeding (41). Second, TAF values in strain A7 are enriched at ∼0.25 and ∼0.75 across the whole genome (Supplementary Figure S9). This TAF pattern is unusual since A7 is thought to be fully inbred and devoid of large chromosomal inversions (41). We hypothesized that the bimodal TAF profile

in A7 could be indicative of contamination in the A7 data from a different fly strain in the DSPR project. Indeed, intersecting TELR predictions between A7 and other DSPR strains revealed an approximately 10-fold higher number of non-reference TE insertion overlaps between strains A7 and B3 relative to any other DSPR strain (Supplementary Table S5). Moreover, shared TE insertions between A7 and B3 have TAFs enriched at ∼0.25 in A7, which could be explained by ∼25% of the A7 dataset being contaminated by B3 sequences (Supplementary Figure S10). Our inference of contamination in the A7 dataset with reads from another DSPR strain can also explain the observations that A7 has the highest number of non-reference TEs in our TELR analysis (Supplementary Figure S5), and that the A7 WGA reported in (40) has the highest level of BUSCO duplication, longest assembly length, and most scaffolds of all DSPR strains in that study.

In S2R+, we observed a clear enrichment for TE insertions on the autosomes to have TAFs ∼0.25 (Figure 4C and D), which can be explained by haplotype-specific TE insertions that occurred after initial cell line establishment and subsequent tetraploidization (Figure 5A) (8,17). In contrast to the autosomes, TE insertions on the X chromosome in S2R+ are enriched at TAFs ∼1 (Figure 4C and D). The X chromosome in the tetraploid S2R+ genome has a baseline ploidy of two since the S2 lineage is thought to have been derived from a hemi-zygous male genotype (8). Thus, the enrichment of X-chromosome TE insertions with TAF ∼1 could be explained by a recent loss of heterozygosity (LOH) event in the X chromosome of S2R+ through mitotic recombination. This explanation is plausible since a recent study has shown that copy-neutral LOH events in cell culture can shape TAF profiles over large genomic regions in *Drosophila* cell lines (22).

Assuming uniform copy number throughout the genome, autosomal haplotype-specific TE insertions that occurred in S2R+ after tetraploidy are expected to have TAFs at ∼0.25. However, the extensive copy number variation observed in the S2R+ genome increases or decreases TAF estimates in affected segments relative to this expected value (Figure 4D). Additionally, we observed many TE insertions on the S2R+ autosomes that have intermediate TAFs between 0.25 and 1.0, suggesting the possibility of other mechanisms besides haplotype-specific post-tetraploid TE insertion to explain the observed TAF distribution. For example, ancestrally-heterozygous diploid TE insertions (either germline insertions in the Oregon-R lab strain that S2R+ was established from, or insertions during cell culture in the pre-tetraploid stage of the S2R+ lineage) could have undergone mitotic recombination events in the post-tetraploid state of the S2R+ lineage changing one haplotype from TE-present to TE-absent (22). Assuming that ancestral heterozygous diploid TE insertions would be randomly distributed on the two different haplotypes of the Oregon-R/pre-tetraploid state of S2R+, mitotic recombination in the post-tetraploid state would have the same probability of increasing or decreasing TAF.

To facilitate the interpretation of TAF values under varying copy number status and more rigorously test the 'haplotype-specific post-tetraploid TE insertion' (Figure 5A) versus 'ancestral TE insertion and post-tetraploid mitotic recombination' (Figure 5B) models, we developed a strategy to predict absolute TE allele copy number for non-reference TE on the autosomes. For each non-reference TE insertion, we multiplied TAF estimates generated by TELR by the local copy number estimated by Control-FREEC (75) in regions flanking the TE insertion, then rounded to the nearest integer value. This procedure generated accurate predictions of TE allele copy number on synthetic diploid and tetraploid genomes (see Supplemental Text; Supplementary Tables S6 and S7, Supplementary Figure S11). Our analysis revealed that a significant proportion of non-reference autosomal TE insertions from the 14 TE families that are amplified in S2R+ relative to A4 have a predicted TE allele copy number of one (Figure 5 C). A similar observation was recently reported in (71) using nanopore data in the related S2 cell line. Furthermore, we found that the number of TEs with predicted TE allele copy number of one is significantly higher than those with predicted TE allele copy number of three in autosomal regions of S2R+, in total (Figure 5 C; chi-squared = 391.47, df = 1, *P*-value < 2.2e–16) and for all but three S2R+ amplified TE families (*mdg3*, *Stalker2*, *17.6*). Thus, we conclude that the majority of insertions in TE families that are amplified in S2R+ are caused by haplotype-specific TE insertions that occurred after tetraploidization, rather than ancestral heterozygous insertions that were reduced in copy number after tetraploidization by mitotic recombination.

### TE expansions in *Drosophila* cell culture can be caused by one or more source lineage

While our results and prior work (18–20) clearly show that some TE families are amplified in *Drosophila* cell lines, it is generally unknown how many source copies or lineages contribute to proliferation of a TE family during cell culture. Using a PCR-based strategy, Maisonhaute *et al.* (24) previously concluded that all non-reference insertions for the *1731* family in the S2 cell line were derived from a single, strongly-activated source copy. However, only a single TE family was surveyed and the number of *1731* new insertions identified was likely underestimated due to the limitations of the PCR-based strategy used by Maisonhaute *et al.* (24). Moreover, it is difficult to conclude whether amplification is due to a single source copy or multiple closely-related copies from a single source lineage. As shown above, autosomal TE insertions with a copy number of one most likely occurred after tetraploidization during cell culture, and thus provide a rich set of TE sequences to study the general properties of TE expansion events during *in vitro* genome evolution.

To comprehensively test whether one or more source lineage is responsible for the amplification of all 14 TE families that expanded in S2R+ (Figure 4B), we took advantage of TELR's ability to assemble non-reference TE sequences and constructed phylogenies using data from S2R+ and 11 whole-fly strains from the DSPR panel (Figure 6; Supplementary Figure S12). Evaluation of TE sequences reconstructed by TELR using simulated datasets suggested that TELR produced high-quality local assemblies (see Supplemental Text; Supplementary Figures S13 and S14), and thus can be reliably used to infer the sequence evolution of TEs amplified in polyploid cell line genomes like S2R+.

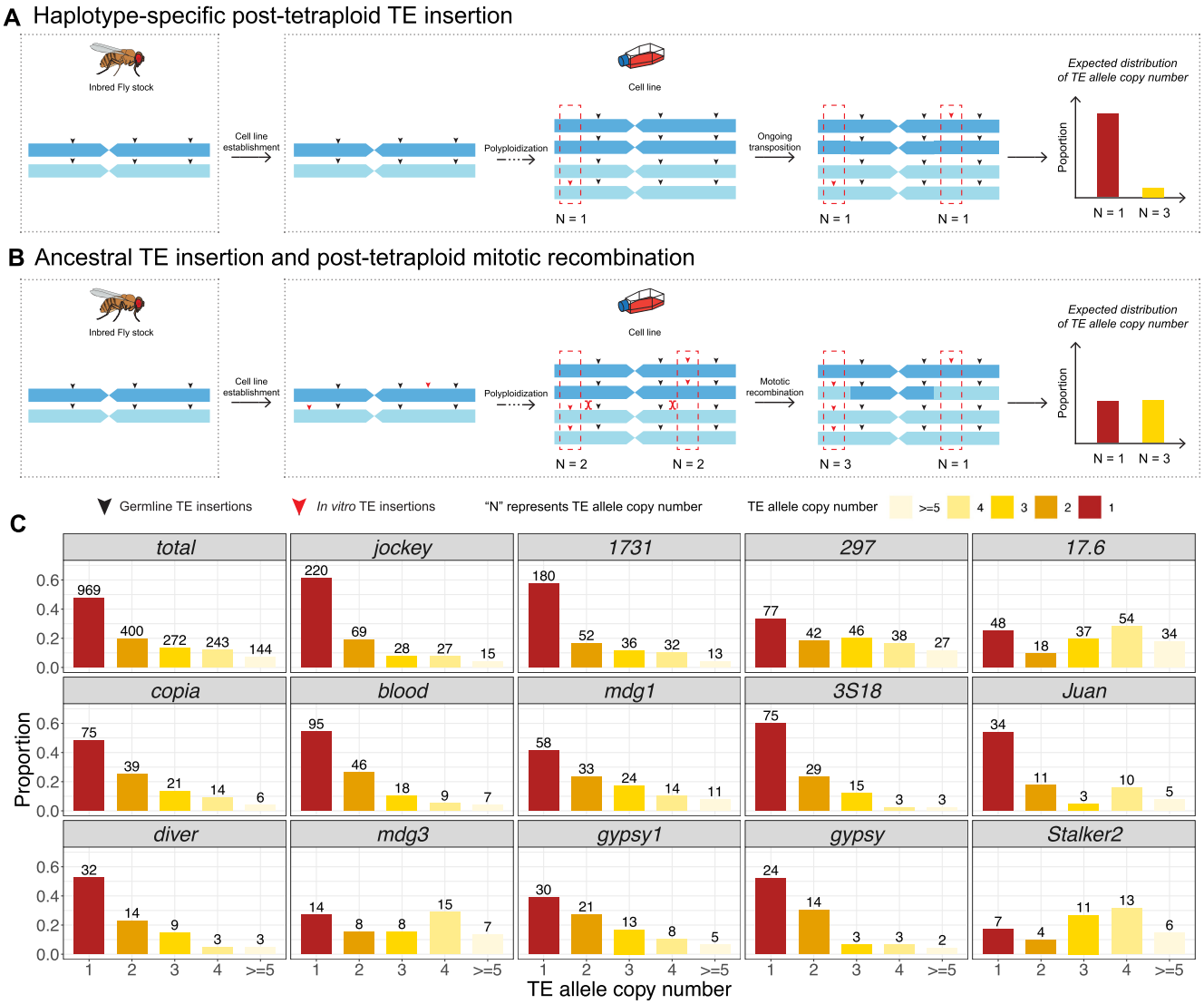


**Figure 5.** TE allele copy number distributions support haplotype-specific TE insertion after tetraploidy in the S2R+ genome. (**A, B**) Alternative hypotheses that could explain haplotype-specific TE insertions in the tetraploid S2R+ genome (see main text for details). (**C**) Distribution of TE allele copy numbers for all TEs combined and for the 14 TE families that are amplified in S2R+ relative to A4. The TE allele copy number is estimated based on TAF predicted by TELR and local copy number predicted by Control-FREEC (75). The histogram is colorized based on TE allele copy number. The number above each bar represents number of TEs under each TE allele copy number category.

Using the sequences of full-length TE insertions identified by TELR, we designed a set of criteria to identify TE expansion events in S2R+ that arise from a single source lineage. First, the TE expansion event should be marked by a monophyletic clade. Second, the monophyletic clade should include at least three post-tetraploid cell-line-specific TE insertions. Third, monophyletic clade should have at least 50% bootstrap support. Fourth, the proportion of post-tetraploid cell-line-specific TE insertions (i.e., TE allele copy number equal to one) within the clade should be equal to or higher than 20%. Finally, we only used TE sequences in autosomes for this analysis, given that TE allele copy number distribution in Chromosome X is different from the autosomes presumably due to a LOH event after tetraploidy (see above). Using these criteria, we annotated TE expansion events in the sequence phylogeny for each of the 14 TE families that are enriched in S2R+ relative to A4 (Figure 4B). We identified a single expansion clade for *1731*, *gypsy*, *gypsy1*, *mdg3* and *Stalker2* (Figure 6; Supplementary Figure S12), suggesting that proliferation of these TE families in the S2R+ cell line came from a single source lineage. We also identified multiple expansion clades for *jockey*, *Juan* and *3S18* (Figure 6; Supplementary Figure S12), suggesting multiple source lineages contribute to the amplification of these TE families in S2R+. Together, our results revealed that TE expansions in S2R+ can be caused by single or multiple source lineages, and that the pattern of source lineage activation in cell culture is TE family-dependent (Figure 6; Supplementary Figure S12).

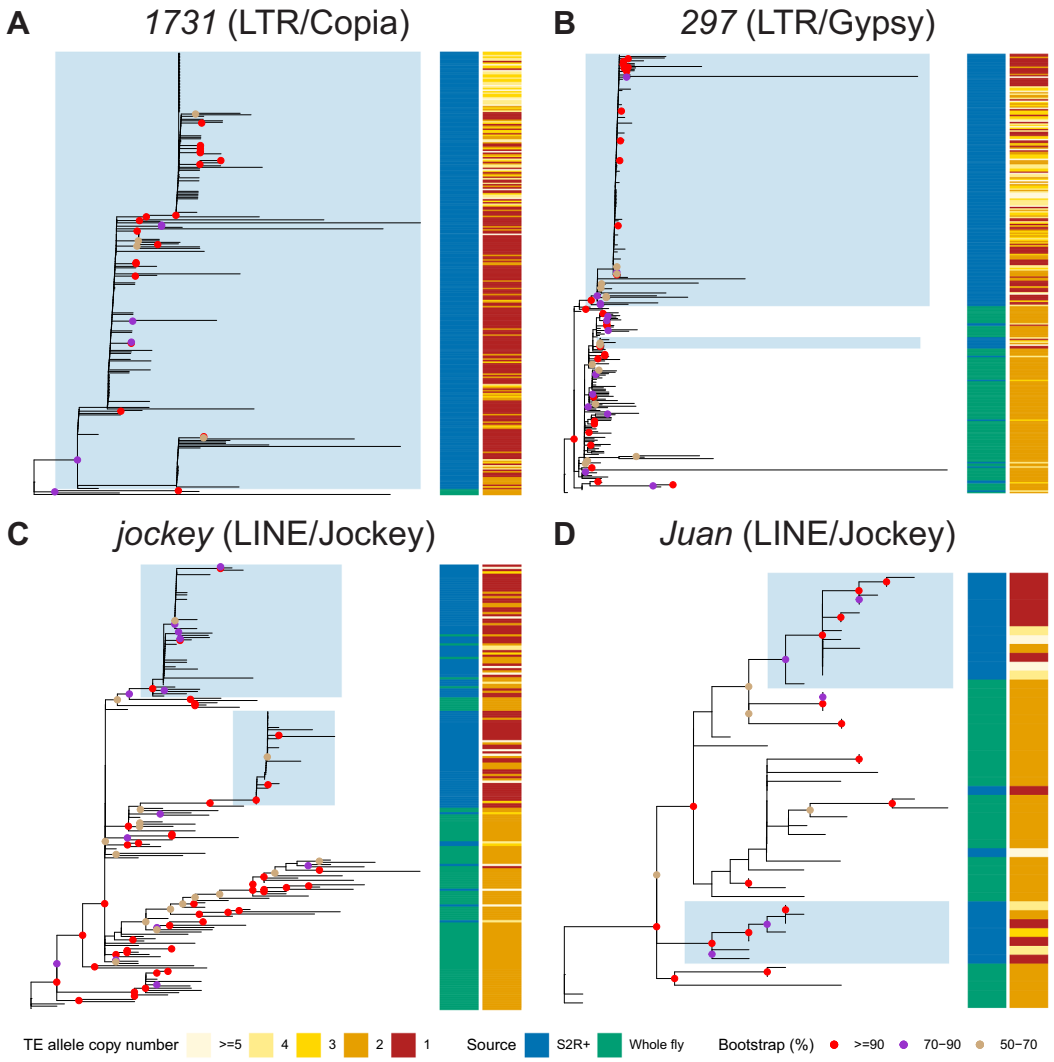


**Figure 6.** Amplification of TE families in the S2R+ genome can be driven by one or more source lineage. (**A–D**) Non-reference TE insertion sequences from S2R+ and 11 inbred *Drosophila* fly strains were predicted and assembled by TELR. Only high-quality full-length TE sequences in normal recombination autosomal regions were retained for this analysis (see Materials and Methods for details). TE sequences for each family were aligned using MAFFT (v7.487) (66). The multiple sequence alignments were used as input in IQ-TREE (v2.1.4-beta) (68) to build unrooted trees for *1731* (A), *297* (B), *jockey* (C) and *Juan* (D) elements using maximum likelihood approach. The sample source and TE allele copy number were annotated in the sidebars. Blue shading indicates a TE expansion event in S2R+ arising from a single source lineage based on the following criteria. (1) All sequences should form a monophyletic clade. (2) The monophyletic clade should include at least three post-tetraploid cell-line-specific TE insertions. (3) The bootstrap support for the clade should be equal to or higher than 50%. (4) The proportion of post-tetraploid cell-line-specific TE insertions (i.e. TE allele copy number equal to one) within the clade should be equal to or higher than 20%.

## DISCUSSION

Here, we report new long-read and linked-read sequence data and develop a novel bioinformatics tool to study the role of transposition during long-term *in vitro* evolution of an animal cell line. Our finding that the complexities of *Drosophila* S2R+ genome architecture preclude the ability to accurately study TE content using long-read or linked-read WGAs motivated the development of a novel WGA-independent TE detection system called TELR that can identify, locally assemble, and estimate allele frequency of TEs from long-read sequence data.

Using the TELR system, we found a significantly higher number of non-reference TEs in S2R+, a sub-line of the *Drosophila* S2 cell line (17,39) compared to whole flies of highly inbred strains from the DSPR project. Since TELR's false negative rates are higher in heterozygous samples with higher ploidies (like S2R+) relative to homozygous diploid samples (like those from the DSPR), the increased abundance of TEs observed in S2R+ cells relative to whole flies is unlikely to be caused by biases in TELR predictions. Moreover, our results using TELR predictions from PacBio sequences confirm related work in *Drosophila* cell lines based on classical molecular techniques and short-read genome sequences (18–20). The increased TE copy number we observe in S2R+ relative to wild type flies is contributed by a subset of mainly LTR and a few non-LTR retrotransposon families. Notably, TE families identified as enriched in

S2R+ by TELR using long-read sequences were also detected as having high activity at some point during the history of S2 cell line evolution in an independent analysis of short-read sequences for multiple sub-lines of S2 cells by Han *et al.* (14), providing cross-validation for both approaches. Future analysis of transcriptomic data could provide additional support for the activity of this subset of TE families in the S2R+ genome. In addition, TELR predicted that a significant proportion of the non-reference TE insertions identified in S2R+ have a TE allele copy number of one (see also (71)), which we interpreted as haplotype-specific insertions that occurred after initial cell line establishment and subsequent tetraploidization (17). This interpretation is consistent with the main conclusion from Han *et al.* (14) that TE amplification in *Drosophila* S2 cells is an ongoing, episodic process rather than being driven solely by an initial burst of transposition during cell line establishment.

Several WGA-independent bioinformatic methods in addition to TELR have recently been developed to detect non-reference TEs using long reads (76–81). These methods use a variety of strategies for TE detection and generate different information for predicted non-reference TEs (Supplementary Table S8). Importantly, none of these previously-reported methods for TE detection using long reads can estimate intra-sample TAF, a feature that we implemented in TELR specifically to identify haplotype-specific TE insertions and which enabled our analysis of post-tetraploid transposition in S2R+. Furthermore, TELR is the only WGA-independent long-read detection tool that outputs a polished assembly of the TE locus, providing a high-quality sequence of both the TE and its flanking regions. The polishing step in TELR is especially important to improve sequence quality when using long-read assemblers such as flye (45) or wtdbg2 (46) that do not error correct reads prior to the assembly step. High-quality sequences of predicted TE insertions generated by TELR allowed us to show that TE expansion in *Drosophila* cell culture could arise from a single or multiple source lineages, providing the first general insight into the sequence evolution of TE family expansions in an animal cell line. Further directions for improvement of the TELR system include investigation of the causes of its relatively low recall rate in low coverage or heterozygous samples, as well as implementation of a 'de novo' non-reference TE detection mode that eliminates the requirement for a user-supplied TE library. Future studies will also reveal if the TELR system can yield related insights into TE structure and evolution in complex heterozygous or polyploid genomes found in many other animal cell lines (8,9,82) or fungal and plant species (83,84), especially for multi-gigabase genomes with complex TE biology.

## DATA AVAILABILITY

PacBio and 10x Genomics whole genome sequences generated in this project are available in the NCBI SRA database under accession PRJNA604454. WGAs of long-read and linked-read sequence data for the S2R+ and A4 genomes are available in the EBI BioStudies database under accession S-BSST752. Datasets of TE insertions in the S2R+ and DSPR genomes predicted by TELR are available as Supplemental File 1. Datasets of TE insertions in the S2R+ and A4 genomes predicted by TEMP and ngs_te_mapper2 are available as Supplemental File 2. Multiple sequence alignments of TE insertion sequences identified by TELR in the S2R+ and DSPR genomes are available as Supplemental File 3. Tree files for phylogenies of TE insertion sequences identified by TELR in the S2R+ and DSPR genomes are available as Supplemental File 4.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Stuart Macdonald (University of Kansas) for providing fly stocks; Christina McHenry and Robert Lyons at the University of Michigan Biomedical Research Core Facilities for assistance with PacBio library preparation and sequencing; Noah Workman, Julia Portocarrero and Magdy Alabady at the University of Georgia Genomics and Bioinformatics Core for assistance with 10x Genomics library preparation and Illumina sequencing; the Georgia Advanced Computing Resource Center for computing time; and members of the Bergman Lab for helpful comments throughout the project.

## FUNDING

University of Georgia Research Education Award Traineeship (to P.J.B.); Howard Hughes Medical Institute (to N.P.); Human Frontiers of Science Program [RGY0093/2012 to C.M.B.]; Georgia Research Foundation (to C.M.B.). Funding for nopen access charge: University of Georgia Research Foundation.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Ford,D.K. and Yerganian,G. (1958) Observations on the chromosomes of Chinese hamster cells in tissue culture. *J. Natl. Cancer Inst.*, **21**, 393–425.
2. Hink,W. (1976) A Compilation of Invertebrate Cell Lines and Culture Media. In: Maramorosch,K. (ed). *Invertebrate Tissue Culture*. Academic Press, pp. 319–369.
3. Ogura,H. (1990) Chromosome variation in plant tissue culture. In: Bajaj,Y.P.S. (ed). *Somaclonal Variation in Crop Improvement I, Biotechnology in Agriculture and Forestry*. Springer, Berlin, Heidelberg, pp. 49–84.
4. Bairu,M.W., Aremu,A.O. and Van Staden,J. (2011) Somaclonal variation in plants: causes and detection methods. *Plant Growth Regul.*, **63**, 147–173.
5. Zhang,Y., Malone,J.H., Powell,S.K., Periwal,V., Spana,E., Macalpine,D.M. and Oliver,B. (2010) Expression in aneuploid Drosophila S2 cells. *PLoS Biol.*, **8**, e1000320.
6. Miyao,A., Nakagome,M., Ohnuma,T., Yamagata,H., Kanamori,H., Katayose,Y., Takahashi,A., Matsumoto,T. and Hirochika,H. (2012) Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing. *Plant Cell Physiol.*, **53**, 256–264.
7. Adey,A., Burton,J.N., Kitzman,J.O., Hiatt,J.B., Lewis,A.P., Martin,B.K., Qiu,R., Lee,C. and Shendure,J. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, **500**, 207–211.
8. Lee,H., McManus,C.J., Cho,D.-Y., Eaton,M., Renda,F., Somma,M.P., Cherbas,L., May,G., Powell,S., Zhang,D. *et al.* (2014)

DNA copy number evolution in Drosophila cell lines. *Genome Biol.*, **15**, R70.
9. Nattestad,M., Goodwin,S., Ng,K., Baslan,T., Sedlazeck,F.J., Rescheneder,P., Garvin,T., Fang,H., Gurtowski,J., Hutton,E. *et al.* (2018) Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.*, **28**, 1126–1135.
10. Ben-David,U., Siranosian,B., Ha,G., Tang,H., Oren,Y., Hinohara,K., Strathdee,C.A., Dempster,J., Lyons,N.J., Burns,R. *et al.* (2018) Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, **560**, 325–330.
11. Zhou,B., Ho,S.S., Greer,S.U., Zhu,X., Bell,J.M., Arthur,J.G., Spies,N., Zhang,X., Byeon,S., Pattni,R. *et al.* (2019) Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.*, **29**, 472–484.
12. Zhou,B., Ho,S.S., Greer,S.U., Spies,N., Bell,J.M., Zhang,X., Zhu,X., Arthur,J.G., Byeon,S., Pattni,R. *et al.* (2019) Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res.*, **47**, 3846–3861.
13. Liu,Y., Mi,Y., Mueller,T., Kreibich,S., Williams,E.G., Van Drogen,A., Borel,C., Frank,M., Germain,P.-L., Bludau,I. *et al.* (2019) Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.*, **37**, 314–322.
14. Han,S., Dias,G.B., Basting,P.J., Nelson,M.G., Patel,S., Marzo,M. and Bergman,C.M. (2022) Ongoing transposition in cell culture reveals the phylogeny of diverse Drosophila S2 sublines. *Genetics*, **221**, iyac077.
15. Miller,J.R., Koren,S., Dilley,K.A., Harkins,D.M., Stockwell,T.B., Shabman,R.S. and Sutton,G.G. (2018) A draft genome sequence for the Ixodes scapularis cell line, ISE6. *F1000Res*, **7**, 297.
16. Miller,J.R., Koren,S., Dilley,K.A., Puri,V., Brown,D.M., Harkins,D.M., Thibaud-Nissen,F., Rosen,B., Chen,X.-G., Tu,Z. *et al.* (2018) Analysis of the Aedes albopictus C6/36 genome provides insight into cell line utility for viral propagation. *Gigascience*, **7**, 1–13.
17. Schneider,I. (1972) Cell lines derived from late embryonic stages of Drosophila melanogaster. *J. Embryol. Exp. Morphol.*, **27**, 353–365.
18. Potter,S.S., Brorein,W.J., Dunsmuir,P. and Rubin,G.M. (1979) Transposition of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. *Cell*, **17**, 415–427.
19. Ilyin,Y.V., Chmeliauskaite,V.G., Ananiev,E.V. and Georgiev,G.P. (1980) Isolation and characterization of a new family of mobile dispersed genetic elements, mdg3, in Drosophila melanogaster. *Chromosoma*, **81**, 27–53.
20. Rahman,R., Chirn,G.-W., Kanodia,A., Sytnikova,Y.A., Brembs,B., Bergman,C.M. and Lau,N.C. (2015) Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Res.*, **43**, 10655–10672.
21. Echalier,G. (1997) In: *Drosophila Cells in Culture*. Academic Press, San Diego, Calif.
22. Han,S., Basting,P.J., Dias,G.B., Luhur,A., Zelhof,A.C. and Bergman,C.M. (2021) Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture. *Genetics*, **219**, iyab113.
23. Mariyappa,D., Rusch,D.B., Han,S., Luhur,A., Overton,D., Miller,D. F.B., Bergman,C.M. and Zelhof,A.C. (2022) A novel transposable element-based authentication protocol for Drosophila cell lines. *G3*, **12**, jkab403.
24. Maisonhaute,C., Ogereau,D., Hua-Van,A. and Capy,P. (2007) Amplification of the 1731 LTR retrotransposon in Drosophila melanogaster cultured cells: origin of neocopies and impact on the genome. *Gene*, **393**, 116–126.
25. Alkan,C., Sajjadian,S. and Eichler,E.E. (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods*, **8**, 61–65.
26. Tattini,L., D'Aurizio,R. and Magi,A. (2015) Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol.*, **3**, 92.
27. Kosugi,S., Momozawa,Y., Liu,X., Terao,C., Kubo,M. and Kamatani,Y. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, **20**, 117.
28. Zhao,X., Collins,R.L., Lee,W.-P., Weber,A.M., Jun,Y., Zhu,Q., Weisburd,B., Huang,Y., Audano,P.A., Wang,H. *et al.* (2021) Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Human Genet.*, **108**, 919–928.
29. Berlin,K., Koren,S., Chin,C.-S., Drake,J.P., Landolin,J.M. and Phillippy,A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotech.*, **33**, 623–630.
30. Chakraborty,M., VanKuren,N.W., Zhao,R., Zhang,X., Kalsow,S. and Emerson,J.J. (2018) Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nat. Genet.*, **50**, 20–25.
31. Bracewell,R., Chatla,K., Nalley,M.J. and Bachtrog,D. (2019) Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *eLife*, **8**, e49002.
32. Chang,C.-H., Chavan,A., Palladino,J., Wei,X., Martins,N.M.C., Santinello,B., Chen,C.-C., Erceg,J., Beliveau,B.J., Wu,C.-T. *et al.* (2019) Islands of retroelements are major components of Drosophila centromeres. *PLoS Biol.*, **17**, e3000241.
33. Mohamed,M., Dang,N.T.-M., Ogyama,Y., Burlet,N., Mugat,B., Boulesteix,M., Merel,V., Veber,P., Salces-Ortiz,J., Severac,D. *et al.* (2020) A transposon story: from TE content to TE dynamic invasion of Drosophila genomes using the single-molecule sequencing technology from Oxford Nanopore. *Cells*, **9**, 1776.
34. Ellison,C.E. and Cao,W. (2020) Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of Drosophila melanogaster. *Nucleic Acids Res.*, **48**, 290–303.
35. Hemmer,L.W., Dias,G.B., Smith,B., Van Vaerenberghe,K., Howard,A., Bergman,C.M. and Blumenstiel,J.P. (2020) Hybrid dysgenesis in Drosophila virilis results in clusters of mitotic recombination and loss-of-heterozygosity but leaves meiotic recombination unaltered. *Mob. DNA*, **11**, 10.
36. Wierzbicki,F., Schwarz,F., Cannalonga,O. and Kofler,R. (2022) Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. *Mol. Ecol. Resour.*, **22**, 102–121.
37. Rhie,A., McCarthy,S.A., Fedrigo,O., Damas,J., Formenti,G., Koren,S., Uliano-Silva,M., Chow,W., Fungtammasan,A., Kim,J. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.
38. Peona,V., Blom,M.P.K., Xu,L., Burri,R., Sullivan,S., Bunikis,I., Liachko,I., Haryoko,T., Jonsson,K.A., Zhou,Q. *et al.* (2021) Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol. Ecol. Resour.*, **21**, 263–286.
39. Yanagawa,S., Lee,J.S. and Ishimoto,A. (1998) Identification and characterization of a novel line of Drosophila Schneider S2 cells that respond to wingless signaling. *J. Biol. Chem.*, **273**, 32353–32359.
40. Chakraborty,M., Emerson,J.J., Macdonald,S.J. and Long,A.D. (2019) Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.*, **10**, 4872.
41. King,E.G., Merkes,C.M., McNeil,C.L., Hoofer,S.R., Sen,S., Broman,K.W., Long,A.D. and Macdonald,S.J. (2012) Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Res.*, **22**, 1558–1566.
42. Miller,S.A., Dykes,D.D. and Polesky,H.F. (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.*, **16**, 1215–1215.
43. Chin,C.-S., Peluso,P., Sedlazeck,F.J., Nattestad,M., Concepcion,G.T., Clum,A., Dunn,C., O'Malley,R., Figueroa-Balderas,R., Morales-Cruz,A. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, **13**, 1050–1054.
44. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
45. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.
46. Ruan,J. and Li,H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nat. Methods*, **17**, 155–158.
47. Weisenfeld,N.I., Kumar,V., Shah,P., Church,D.M. and Jaffe,D.B. (2017) Direct determination of diploid genome sequences. *Genome Res.*, **27**, 757–767.
48. Zheng,G.X.Y., Lau,B.T., Schnall-Levin,M., Jarosz,M., Bell,J.M., Hindson,C.M., Kyriazopoulou-Panagiotopoulou,S., Masquelier,D.A., Merrill,L., Terry,J.M. *et al.* (2016) Haplotyping

germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, **34**, 303–311.
49. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
50. Gremme,G., Steinbiss,S. and Kurtz,S. (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 645–656.
51. Bushnell,B., (2014) In: *BBMap: a fast, accurate, splice-aware aligner. Technical Report LBNL-7065E, Lawrence Berkeley National Lab. (LBNL)*. Berkeley, CA, United States.
52. Simao,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
53. Waterhouse,R.M., Seppey,M., Simao,F.A., Manni,M., Ioannidis,P., Klioutchnikov,G., Kriventseva,E.V. and Zdobnov,E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, **35**, 543–548.
54. Kriventseva,E.V., Kuznetsov,D., Tegenfeldt,F., Manni,M., Dias,R., Simao,F.A. and Zdobnov,E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.
55. Singh,N.D. and Petrov,D.A. (2004) Rapid sequence turnover at an intergenic locus in Drosophila. *Mol. Biol. Evol.*, **21**, 670–80.
56. Wang,J., Keightley,P.D. and Halligan,D.L. (2007) Effect of divergence time and recombination rate on molecular evolution of Drosophila INE-1 transposable elements and other candidates for neutrally evolving sites. *J. Mol. Evol.*, **65**, 627.
57. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
58. Sedlazeck,F.J., Rescheneder,P., Smolka,M., Fang,H., Nattestad,M., von Haeseler,A. and Schatz,M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**, 461–468.
59. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
60. Hoskins,R.A., Carlson,J.W., Wan,K.H., Park,S., Mendez,I., Galle,S.E., Booth,B.W., Pfeiffer,B.D., George,R.A., Svirskas,R. *et al.* (2015) The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res.*, **25**, 445–458.
61. Nelson,M.G., Linheiro,R.S. and Bergman,C.M. (2017) McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3*, **7**, 2749–2762.
62. Zhuang,J., Wang,J., Theurkauf,W. and Weng,Z. (2014) TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.*, **42**, 6826–6838.
63. Cridland,J.M., Macdonald,S.J., Long,A.D. and Thornton,K.R. (2013) Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. *Mol. Biol. Evol.*, **30**, 2311–2327.
64. Bergman,C.M., Quesneville,H., Anxolabehere,D. and Ashburner,M. (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol.*, **7**, R112.
65. Manee,M.M., Jackson,J. and Bergman,C.M. (2018) Conserved noncoding elements influence the transposable element landscape in Drosophila. *Genome Biol. Evol.*, **10**, 1533–1545.
66. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
67. Capella-Gutierrez,S., Silla-Martinez,J.M. and Gabaldon,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
68. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., von Haeseler,A. and Lanfear,R. (2020) IQ-TREE 2: New models and efficient methods for phylogenetic Inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.
69. Kelley,D.R. and Salzberg,S.L. (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biol.*, **11**, R28.
70. Dias,G.B., Altammami,M.A., El-Shafie,H. A.F., Alhoshani,F.M., Al-Fageeh,M.B., Bergman,C.M. and Manee,M.M. (2021) Haplotype-resolved genome assembly enables gene discovery in the red palm weevil Rhynchophorus ferrugineus. *Sci. Rep.*, **11**, 9987.
71. Lewerentz,J., Johansson,A.-M., Larsson,J. and Stenberg,P. (2022) Transposon activity, local duplications and propagation of structural variants across haplotypes drive the evolution of the Drosophila S2 cell line. *BMC Genom.*, **23**, 276.
72. Sackton,T.B., Kulathinal,R.J., Bergman,C.M., Quinlan,A.R., Dopman,E.B., Carneiro,M., Marth,G.T., Hartl,D.L. and Clark,A.G. (2009) Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biol. Evol.*, **1**, 449–65.
73. Linheiro,R.S. and Bergman,C.M. (2012) Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLOS One*, **7**, e30008.
74. Fiston-Lavier,A.-S., Barron,M.G., Petrov,D.A. and Gonzalez,J. (2015) T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.*, **43**, e22.
75. Boeva,V., Popova,T., Bleakley,K., Chiche,P., Cappo,J., Schleiermacher,G., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.
76. Disdero,E. and Filee,J. (2017) LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mob DNA*, **8**, 5.
77. Jiang,T., Liu,B., Li,J. and Wang,Y. (2019) rMETL: sensitive mobile element insertion detection with long read realignment. *Bioinformatics*, **35**, 3484–3486.
78. Zhou,W., Emery,S.B., Flasch,D.A., Wang,Y., Kwan,K.Y., Kidd,J.M., Moran,J.V. and Mills,R.E. (2020) Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.*, **48**, 1146–1163.
79. Ewing,A.D., Smits,N., Sanchez-Luque,F.J., Faivre,J., Brennan,P.M., Richardson,S.R., Cheetham,S.W. and Faulkner,G.J. (2020) Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell*, **80**, 915–928.
80. Chu,C., Borges-Monroy,R., Viswanadham,V.V., Lee,S., Li,H., Lee,E.A. and Park,P.J. (2021) Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.*, **12**, 3836.
81. Kirov,I., Merkulov,P., Dudnikov,M., Polkhovskaya,E., Komakhin,R.A., Konstantinov,Z., Gvaramiya,S., Ermolaev,A., Kudryavtseva,N., Gilyok,M. *et al.* (2021) Transposons hidden in Arabidopsis thaliana genome assembly gaps and mobilization of non-autonomous LTR retrotransposons unravelled by nanotei pipeline. *Plants*, **10**, 2681.
82. Talsania,K., Mehta,M., Raley,C., Kriga,Y., Gowda,S., Grose,C., Drew,M., Roberts,V., Cheng,K.T., Burkett,S. *et al.* (2019) Genome assembly and annotation of the Trichoplusia ni Tni-FNL insect cell line enabled by long-read technologies. *Genes*, **10**, E79.
83. Todd,R.T., Forche,A. and Selmecki,A. (2017) Ploidy variation in fungi: polyploidy, aneuploidy, and genome evolution. *Microbiol. Spect.*, **5**, https://doi.org/10.1128/microbiolspec.FUNK-0051-2016.
84. Meyers,L.A. and Levin,D.A. (2006) On the abundance of polyploids in flowering plants. *Evolution*, **60**, 1198–1206.