OXFORD GENETICS

GENETICS, 2026, iyag101
https://doi.org/10.1093/genetics/iyag101
Advance Access Publication Date: 17 April 2026
Investigation

Leveraging long-read assemblies and machine learning to enhance short-read transposable element detection and genotyping

Austin Daigle [1,2,*] Logan S. Whitehouse,[1,2] Roy Zhao,[3] J.J. Emerson [3] Daniel R. Schrider [1,*]

[1]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599
[2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC 27599
[3]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697

*Corresponding authors: Austin Daigle, Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, Chapel Hill, NC 27599, USA. Email: adaigle@unc.edu; Daniel R. Schrider, Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, Chapel Hill, NC 27599, USA. Email: drs@unc.edu

Transposable elements (TEs) are parasitic genomic elements that are ubiquitous across the tree of life and play a crucial role in genome evolution. Advances in long-read sequencing have allowed highly accurate TE detection, though at a higher cost than short-read sequencing. Recent studies using long reads have shown that existing short-read TE detection methods perform inadequately when applied to real data. In this study, we use a machine learning approach (called TEforest) to discover and genotype TE insertions and deletions with short-read data by using TEs detected from *Drosophila melanogaster* long-read genome assemblies as training data. Our method first uses a highly sensitive algorithm to discover potential TE insertion or deletion sites in the genome, extracting relevant features from short-read alignments. To discriminate between true and false TE insertions, we train a gradient-boosted decision tree model with a labeled ground-truth dataset for which we have calculated the same set of short-read features. We conduct a comprehensive benchmark of TEforest and traditional TE detection methods using real data from *D. melanogaster* and humans, finding that TEforest identifies more true positives and fewer false positives across datasets with different read lengths and coverages, while also accurately inferring genotypes and the precise breakpoints of insertions. By learning short-read signatures of TEs previously only discoverable using long reads, our approach bridges the gap between large-scale population genetic studies and the accuracy of long-read assemblies. This work provides a user-friendly tool to study the prevalence and phenotypic effects of TE insertions across the genome.

Keywords: transposable elements; variant calling; short-read sequencing; *Drosophila melanogaster*; genomics

Introduction

Transposable elements (TEs) are genetic sequences capable of replicating themselves throughout the genome and that shape genome structure, gene expression, and evolutionary dynamics (Bourque et al. 2018; Drongitis et al. 2019; Makałowski et al. 2019; Wells and Feschotte 2020). TE insertions can impact phenotypes through the direct disruption of DNA sequences (Finnegan 1992), TE-induced chromosomal rearrangements (Montgomery et al. 1987, 1991) and changes in gene expression (Feschotte 2008; Lee 2015). Though most insertions leading to phenotypic changes are thought to be deleterious (Charlesworth and Langley 1989; Lynch 2007), numerous examples of adaptations driven by TE insertions have accumulated (reviewed in Casacuberta and González (2013) and Schrader and Schmitz (2019)).

To comprehensively quantify the impact of TEs on evolutionary processes, studies usually rely on tools that can detect and genotype TE insertions using genome re-sequencing data. Although recent breakthroughs in long-read sequencing have dramatically improved our ability to pinpoint and characterize TE insertions with unprecedented detail (Ewing et al. 2020; Zhou et al. 2020; Kirov et al. 2021; Han et al. 2022; Hoyt et al. 2022; Rech et al. 2022; Mohamed et al. 2023), these technologies remain more costly and less scalable than their short-read counterparts for the time being. Consequently, most large-scale population genetic studies rely heavily on short-read data, requiring robust computational methods for detecting and genotyping TEs from these more accessible, but inherently limited, resources.

Though numerous TE detection methods have been developed since the advent of whole-genome sequencing, most TE callers that use short-read sequence data (e.g. Illumina) follow a similar general strategy: identifying discordantly mapping read pairs, where one read maps to the reference genome and the other to a TE sequence, to identify locations of TE insertions (see Fig. 1a). Additionally, some methods also use split reads, where part of a single read maps to the reference genome and another portion of that read shares homology with a TE, to precisely identify TE insertion breakpoints. This process is followed by a series of filtering steps to limit the prediction of false positives (reviewed in Makałowski et al. (2019)). These approaches can be complicated by the difficulty of designing filters that apply to all read lengths and insert sizes, work for diverse types of TE sequences, and do not filter out TE insertions in repetitive regions or nearby other

Received on 12 January 2026; accepted on 09 April 2026

structural variations relative to the reference genome. While these approaches have proven useful, their accuracy often varies substantially according to the properties of the input data, including read length, sequencing coverage, or the degree of repetitiveness in the genome of the species being examined (Vendrell-Mir et al. 2019; Yu et al. 2021; Verneret et al. 2025). Furthermore, because most previous studies comparing the performance of TE detectors relied on benchmark sets of simulated TE insertions (Nelson et al. 2017; Chen et al. 2023; Verneret et al. 2025) or a small number of genomes assembled with long reads (Rishishwar et al. 2017; Vendrell-Mir et al. 2019; Yu et al. 2021) to perform these assessments, our understanding of the shortfalls and relative efficacy of TE detection methods is incomplete. Simulated data in particular may not adequately represent the complexity and messiness of short-read mapping patterns around true TE insertions. On the other hand, large high-quality datasets of known TE insertions, such as the long-read-based population-scale *Drosophila melanogaster* assemblies and TE annotations generated by Rech et al. (2022), open the possibility for more rigorous testing of TE detection algorithms. Crucially, when both long- and short-read data are available for the same genomes, short-read-based detection methods can also be benchmarked using a set of high-confidence TE insertions. Moreover, the availability of such data allows us to reframe short-read TE detection as a machine learning problem—rather than designing a method to detect what we think the patterns of short-read mapping around a TE insertion *ought* to look like, we can train a machine learning classifier to detect the actual signatures of known TE insertions in real empirical datasets.

Here, we present TEforest, a machine learning method that enhances short-read TE detection and genotyping by learning predictive features directly from high-confidence TE insertions previously detected from long-read assemblies. First, TEforest uses a sensitive initial scanning algorithm to identify a large set of potential TE insertions. TEforest then employs a LightGBM classifier to simultaneously discriminate between true and false TE candidates and genotype the insertions as heterozygous or homozygous. By training TEforest to examine a rich suite of features—drawn from multiple read-mapping signatures and tested across variable read lengths and coverages—we are able to not only achieve higher performance than existing tools but also provide more reliable genotype predictions, precise breakpoint predictions, and accurate allele frequency estimates. TEforest is freely available at https://github.com/SchriderLab/TEforest.git.

Methods

Algorithm overview

TEforest accepts as input (i) paired-end short-read fastq files, (ii) a reference genome in fasta format, (iii) a TE consensus library in fasta format, and (iv) a BED file detailing reference TE locations (Fig. 1b). The algorithm first identifies genomic regions that may contain nonreference TE insertions by finding read pairs that map to TE consensus sequences and TEs annotated in the reference genome. After a small number of filters are applied to the candidate insertion sites, a comprehensive set of features summarizing read alignments within each candidate region are computed and transformed into feature vectors. These vectors are then classified by a decision tree ensemble model as either a homozygous TE insertion, a heterozygous TE insertion, or no insertion. These feature vectors can also be used for training a model if the true genotypes are available. Finally, the algorithm attempts to pinpoint precise breakpoint locations using split-read evidence. For TEs annotated in the reference genome, we use an additional decision tree ensemble model trained to detect presence/absence using the same feature vectors of the nonreference model.

Algorithm to detect nonreference insertions

Discovery of regions with candidate TE insertions

To discover regions of the genome with potential TE insertions, the fastq reads are preprocessed with *fastp* v 0.24.0 to trim adaptors and low-quality sequences (Chen 2023) and then mapped to the TE consensus sequences as well as sequences annotated as TEs in the reference genome using BWA-MEM 2 v2.2.1, with default settings (Li 2013; Vasimuddin et al. 2019). Reads partially mapping or with a mate mapping to TE sequences are then mapped back to the reference genome using the same software/settings. Nested TE sequences, where an annotation of 1 TE overlaps with another TE, are omitted from the TE sequence database to prevent errors where sequences aligning to 1 TE are misattributed to another nested TE. For each TE family, the locations of read pairs where 1 of the 2 reads maps to the TE sequence (whether it be the canonical sequence or an annotated copy of that TE family found in the reference genome) are used to identify candidate regions for TE insertions; we refer to these pairs as "TE-mapping read pairs." Initially, the candidate region consists of any contiguous stretch of sites with coverage by reads in a TE-mapping read pair. To avoid confounding reads mapped to reference TEs with nonreference insertions, candidate regions overlapping with a reference TE of the same family as well as those contained entirely within a reference TE of a different family are filtered out.

Because we observed that many false-positive regions are very short in length, regions less than the read length of the dataset plus 5 base pairs are also filtered out to improve computational efficiency. This length is slightly larger than the largest read length used in our dataset and was chosen because many of these false-positive regions are the length of a single read. Shortening this filtering length did not improve performance for shorter read lengths (not shown). To ensure that our candidate region fully encompassed all TE-mapping read pairs around a putative TE insertion, all remaining regions are expanded a further 200 bp in either direction. Because many nonreference TE insertion sites contain gaps in coverage of TE-mapping read pairs (Fig. 1c), we next merge any overlapping candidate regions for the same TE family. The implementation of this algorithm benefitted from open-source bioinformatics packages for working with short-read alignments and genomic coordinate ranges, including SAMtools v1.6 (Danecek et al. 2021), SeqKit2 v2.7.0 (Shen et al. 2024), and GenomicRanges v1.60.0 (Lawrence et al. 2013).

Calculation of feature vectors

Here, we expand on concepts described by Hill and Unckless (2019), who proposed the use of a feature vector as input for machine learning models to detect structural variants. This feature vector representation summarizes many aspects of the alignment of short reads to the genome calculated on a per-base level (all features are defined in Supplementary Table 1). For each candidate region, we compute the same set of features from 2 alignment contexts: (i) the BAM containing genome-wide alignments of all reads to the reference and (ii) a TE family–specific BAM containing only those read pairs that were flagged as TE associated during candidate region detection (e.g. reads that partially map to a TE sequence and/or whose mate partially maps to a TE

sequence; these include both TE sequences in the reference and those in the consensus TE database). For each site in the candidate region, a per-base pair sum is calculated for all features listed in Supplementary Table 1 by extracting information from the CIGAR string, bitwise flag, coverage, and other alignment information reported by the BAM file containing the alignment of all reads to the reference genome. For each site, the value of each of these features is divided by the total number of reads partially mapping or with a mate mapping to that site. Next, for each feature, the normalized per-base signal across the candidate region is converted to an empirical cumulative distribution and then binned into 100 equally sized windows. This step standardizes regions of different physical lengths so that each region contributes the same number of values per feature before summary statistics are calculated. A concise vector summary of these features for a candidate region is created using the mean, standard deviation, median, and interquartile range (IQR) for each statistic across all sites in the candidate region, resulting in $21 \times 2 \times 4 = 168$ feature summaries in total (21 features measured for both TE and reference-genome mapping read pairs, and each summarized by 4 values) in the final vector.

LightGBM model

When the locations of true TE insertions are known, labeled feature vectors can be used for training a machine learning model. We used the gradient-boosted decision tree classifier LightGBM (Ke et al. 2017), and trained with 500 estimators (i.e. decision trees), using objective="multiclass" as we have 3 classes in our dataset (homozygous insertion, heterozygous insertion, and no insertion). To address potential class imbalance, we set class_weight="balanced," which adjusts weights inversely proportional to class frequencies. Unless otherwise noted, other hyperparameters follow LightGBM defaults.

Breakpoint detection

After the LightGBM model described above is used for TE detection, all candidate regions classified as homozygous and heterozygous insertions are further processed to identify precise breakpoints of TE insertions. TE-mapping read pairs are checked for split reads, and if none are identified, then all read pairs mapping to the candidate region are checked for split reads. For each split read, we locate the site in the candidate region where the 3′-most site in the read maps to before the split, calling this a split-read-end. We find the top 2 sites with the largest number of split-read-ends. We use these 2 sites as the final TE breakpoints, as most TEs make staggered double-strand breaks when inserting into the genome, creating a target site duplication (TSD) at the insertion site (Craig 2002; Linheiro and Bergman 2012). If only 1 site with split reads is identified, that site is used as the insertion site. If no split reads are identified, then the center of the candidate region is used as the breakpoint.

Algorithm to detect and genotype TE insertions annotated in the reference genome

When examining short-read data from a given individual, TEforest classifies TEs annotated in the reference genome as homozygous, heterozygous, or absent in a similar manner to nonreference insertions. For this task, no detection of putative candidate regions or breakpoints is needed. User-provided coordinates of reference TEs are expanded 500 base pairs in either direction to capture information about reads surrounding the TE breakpoints. Feature vectors are calculated as previously described, and a LightGBM classifier is trained and used to label a reference TE as present or absent using the same hyperparameters used for classifying nonreference TEs.

Training and testing TEforest on high-quality *D. melanogaster* data

Creation of synthetic partially heterozygous genomes for training/testing TEforest's genotyping of nonreference TEs

A subset of the *D. melanogaster* genomes annotated with the manually curated TE library created by Rech et al. (2022) were used for training and testing the TEforest model. In Rech et al. (2022), TE coordinates from each strain were projected onto the ISO1 reference genome by aligning TE sequences and their flanking regions to ISO1. Only insertions with unambiguous, high-confidence flank mappings were retained as reliable (see their Supplementary Note 11 for details). As this annotation only contains homozygous insertions found in the long-read assembly, we selected 6 genomes with low heterozygosity (as reported in Supplementary Table 2 of Rech et al. (2022)) to minimize the risk of our model identifying false-positive insertions that are present in only a subset of flies from the inbred line used for DNA extraction and sequencing. Specifically, we used the assemblies for samples A2 (*aka* BOG1; ASM340180v1), A3 (*aka* BS1; ASM340179v1), AKA-017 (ASM2014210v1), GIM-024 (ASM2014200v1), JUT-008 (ASM2014195v1), and MUN-009 (ASM2014183v1).

To include known heterozygous insertions in model training, we created synthetic partially heterozygous genomes by combining reads from 2 genomes sequenced with the same read length. To do this, reads from both genomes were first preprocessed by *fastp* and then mapped to the reference genome (*dm6*; RefSeq GCF_000001215.4). A final heterozygous fastq file was created such that half the reads from contigs 2R and 3R were taken from each genome, while all other chromosome arms were taken from one of the genomes in the pair, with reads in each chromosome arm downsampled to the same mean coverage. Thus, most differences between the 2 lines on chr2R and chr3R would be heterozygous, while all variants on the remaining chromosome arms would be homozygous. To train on a diverse set of read lengths, we created 3 synthetic partially heterozygous genomes: 1 from a pair of genomes sequenced with 54-bp reads (A2, reads available at SRA accession SRR457711 and A3 = SRR457708), 1 for 125-bp reads (AKA-017 = SRR9951105 and GIM-024 = SRR9951106), and 1 for 151-bp reads (JUT-008 = SRR12187569 and MUN-009 = SRR12187566). All sequenced flies were female.

Training and validation dataset for nonreference insertions

TEforest was trained using nonreference insertions from contigs 3L, 3R, and X and tested on contigs 2L and 2R to ensure that no insertions used in model training were used in model testing. Additionally, we restricted our analysis to the euchromatic boundaries (defined in Fiston-Lavier et al. (2010); 2L, 530,000 to 18,870,000; 2R, 5,982,495 to 24,972,477; 3L, 750,000 to 19,026,900; 3R, 6,754,278 to 31,614,278; X, 1,325,967 to 21,338,973), filtering out heterochromatic regions as they were not annotated by Rech et al. (2022). A full description of the numbers of labeled candidate regions used for training and testing can be found in Supplementary Table 2. To construct the negative training set (the "no insertion" class), we utilized candidate regions identified by the initial TEforest search that did not overlap with the high-confidence ground-truth insertions. This approach allows the model to learn to differentiate true TEs from difficult false-positive signals found in the background, rather than simply distinguishing TEs from random genomic regions. Separate models

were trained and tested at target coverages of 5×, 10×, 20×, 30×, 40×, and 50×.

Training and validation data for reference insertions

A model to classify reference TEs as homozygous or heterozygous present or absent was trained using the same synthetic partially heterozygous genomes as the nonreference model. However, to generate more heterozygous reference TE insertions, we created 3 additional synthetic partially heterozygous genomes by combining short reads from the ISO1 reference strain with a nonreference strain at our 3 read lengths of 54 bp (A2, reads available at SRA accession SRR457711 and ISO1 = ERX645969) (these were originally 100-bp reads and were trimmed down to 54 bp using *fastp*), 125 bp (AKA-017 = SRR9951105 and ISO1 = SRR11906527), and 151 bp (JUT-008 = SRR12187569 and ISO1 = SRR29479671). Reads were downsampled to the same target coverages, and training and testing was conducted on the same contigs as the nonreference model. The numbers of labeled candidate regions used for training and testing can be found in Supplementary Table 3.

Benchmarking other TE detectors

To compare TEforest to other TE detectors, we utilized McClintock v.2.0.0, a meta-pipeline that implements popular TE detection methods in a controlled workflow (Nelson et al. 2017; Chen et al. 2023). We ran McClintock using the same dm6 reference genome described above with default settings, using PoPoolationTE (Kofler et al. 2012), PoPoolationTE2 (Kofler et al. 2016), RetroSeq (Keane et al. 2013), TEFLoN (Adrion et al. 2017), TEMP (Zhuang et al. 2014), and TEMP2 (Yu et al. 2021). These methods were chosen based on their ability to reliably produce calls for all read lengths tested and finish running within 5 days using 8 central processing unit (CPU) cores. Specifically, both versions of RelocaTE were excluded due to runtime, both versions of ngs_te_mapper and TEBreak were excluded because they produced zero calls on at least 1 read length, and TE-locate was excluded because it consistently predicted hundreds of false positives.

Methods were benchmarked for accuracy of TE detection, genotyping, and breakpoint localization. To assess the overall ability to detect TEs, calls made by a TE caller within 500 bp of a TE of the same family in the truth dataset were counted as true positives, with the rest being counted as false positives. We used this liberal distance threshold for true positives to allow small positional deviations when benchmarking, since stricter thresholds would disproportionately penalize callers with slightly less precise breakpoint localization despite correctly identifying the underlying insertion. Overlapping TEs of the same family in the truth dataset were condensed to avoid the double-counting of nested TEs; this was not required for nested TEs of different types. The truth dataset provided by Rech et al. (2022) does not provide perfectly precise breakpoints because it does not report the TSDs produced by the staggered cuts of transposase upon TE integration (Craig 2002; Linheiro and Bergman 2012); rather, it provides an approximate breakpoint near the 2 breakpoints of the TSD. For this reason, the center of 2 breakpoints predicted by each TE caller was used to calculate distances between the truth dataset breakpoint and the prediction of the TE caller. If 1 breakpoint is predicted, we simply found the distance between this breakpoint and the truth dataset breakpoint. To assess genotyping accuracy, we calculated the F_1 scores separately for homozygous and heterozygous genotypes and reported the macroaveraged F_1 score as the overall measure of performance. For example, for the homozygous F_1 score, the true positives are correctly genotyped homozygotes, false positives are heterozygotes classified as homozygotes, and false negatives are homozygotes classified as heterozygotes, and vice versa for the heterozygous F_1 score. We additionally performed a separate benchmarking of breakpoint accuracy and genotyping performance for the intersection of calls made by TEforest, TEMP2, and RetroSeq at a given coverage. These methods were chosen as they are the best performing set across all datasets, and including more callers would reduce the number of benchmarked calls.

Benchmarking allele frequency accuracy

To assess the ability of TE callers to accurately predict the allele frequencies of nonreference TEs in a population of individuals, we benchmarked the performance of all previously used TE callers in 13 genomes from the *Drosophila* Synthetic Population Resource, sequenced with 54-bp reads, that were included in our truth dataset (King et al. 2012; Chakraborty et al. 2019; Rech et al. 2022). True allele frequencies were derived from the long-read assemblies in this truth dataset; callers were evaluated on the corresponding 54-bp short-read data. The median coverage of these genomes was 46×, so the TEforest model trained with 50× coverage was used for this task. Only calls within the euchromatic regions of contigs 2L, 2R, 3L, 3R, and X were used for benchmarking. As these genomes are highly homozygous, each true-positive call (a prediction of a TE insertion of the same family within 500 bp of an annotated insertion site) was counted as a frequency of 1, with a max frequency of 13. For each insertion in the population, the true allele frequency was subtracted from the predicted allele frequency.

Benchmarking on human data

To evaluate the cross-species generalizability of TEforest and compare it against recent human-specific tools, we benchmarked performance on a human whole-genome sequencing dataset (NA12878). We utilized 30× coverage, 150-bp paired-end Illumina reads (Accession: ERR3239334) generated as part of the high-coverage 1000 Genomes Project (Byrska-Bishop et al. 2022). The human genome hg38 (RefSeq GCF_000001405.40) was used as the reference genome. To create a consensus fasta file, conserved TE consensus sequences for Alu, LINE1, and SVA families were obtained from Gardner et al. (2017).

We compared TEforest to 4 other detection methods: RetroSeq, TEMP2, DeepMEI (downloaded August 2025), and MEGAnE v1.0.l. RetroSeq and TEMP2 were executed via the McClintock pipeline as done previously. DeepMEI and MEGAnE were run as standalone tools using the BWA-mem alignments produced by McClintock as input. For MEGAnE, we performed tests using both its built-in TE library and our own user-provided library to ensure optimal performance comparisons.

The ground-truth dataset consisted of 893 validated germline insertions for NA12878 (Sudmant et al. 2015; Rishishwar et al. 2017; Yu et al. 2021), predominately composed of Alu elements (93.3%). To test TEforest, we applied the nonreference model trained on *D. melanogaster* synthetic heterozygotes (30× coverage) without any retraining on human data.

Optional retraining workflow

For reproducibility and extension to other species or datasets, we provide a Snakemake-based training workflow in the TEforest GitHub repository that automates feature-vector generation and model fitting from standard TEforest inputs plus a BED file of validated TE insertion sites annotated with zygosity. The workflow outputs a LightGBM model file that can be used directly by TEforest for inference.

Results

TEforest outperforms other TE detection methods

We developed TEforest, a machine learning tool that detects the positions of transposon insertions in a reference genome using short reads (Fig. 1 and Methods). This is accomplished by identifying read pairs where one read maps to an annotated set of TE sequences and the other to the reference genome (TE-mapping read pairs), extracting a comprehensive set of features describing read-mapping patterns in those genomic regions containing TE-mapping read pairs, and then using a LightGBM model to predict if the region contains a true TE insertion. By comprehensively incorporating all available read-mapping information from both true TE insertions and noninsertion sites, our approach attempts to capture a comprehensive spectrum of read-mapping patterns, thereby enabling the model to more effectively distinguish genuine insertions from false positives. We trained and tested TEforest using the annotated TE insertions in 6 *D. melanogaster* genomes sequenced with both long and short reads by Chakraborty et al. (2019) and Rech et al. (2022). Because these genomes were sequenced from inbred lines and we wished to train TEforest to be able to detect and genotype both homozygous and heterozygous TE insertions, we created synthetic heterozygous nonreference TE insertions by combining the reads of highly homozygous genomes for some chromosome arms, while others remained homozygous. These data were partitioned into training and testing sets as described in the Methods and in Supplementary Table 2. In addition to benchmarking TEforest on this hold-out test set, we benchmarked 6 other TE detectors available in the McClintock meta-pipeline.

Overall, TEforest demonstrated higher accuracy than competing methods across nearly all tested read lengths and coverage levels, except for the lowest-coverage 54-bp datasets (10× and

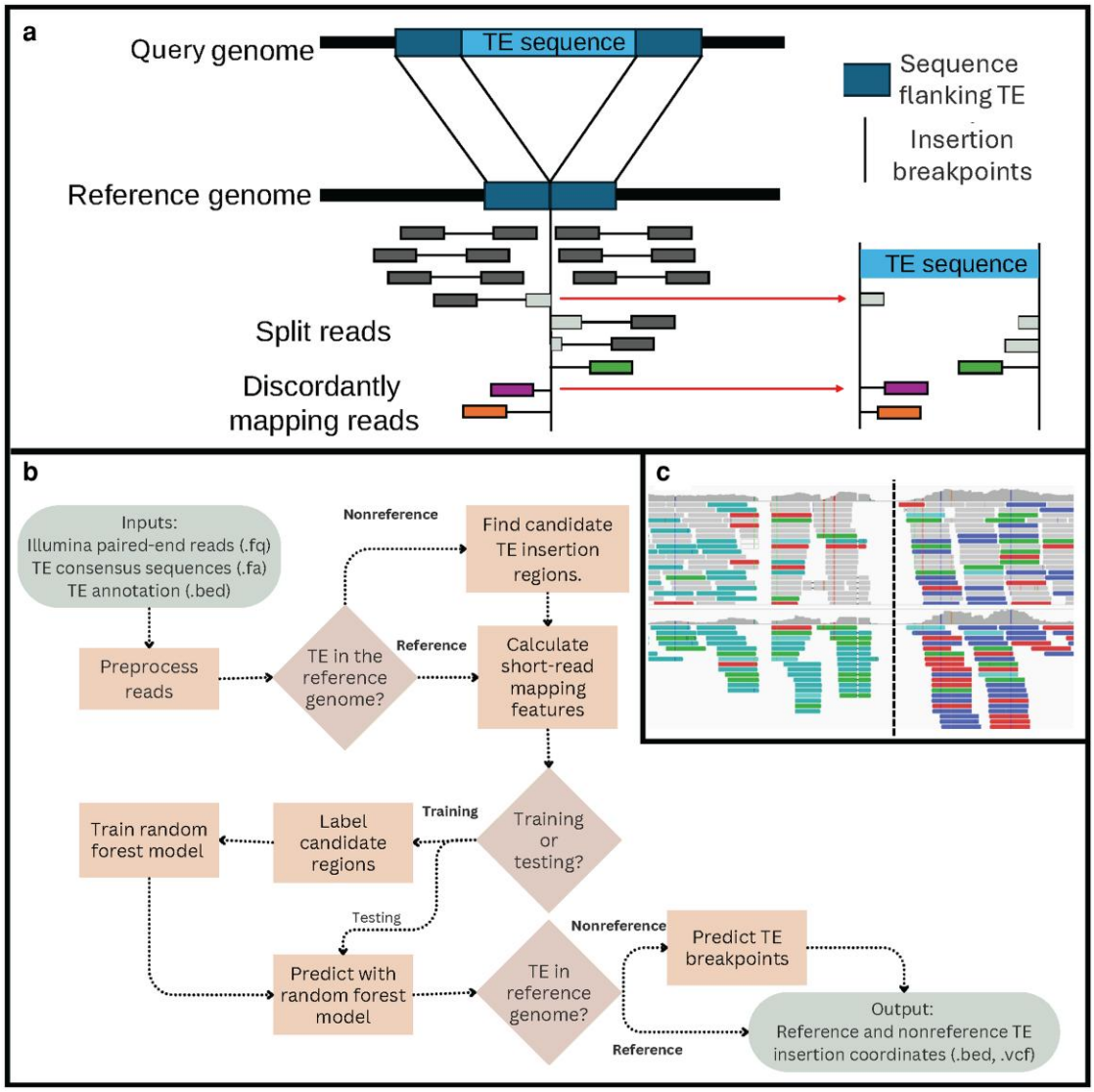

Fig. 1. a) Schematic of the read-mapping signatures used to detect nonreference TE insertions from paired-end short-read data. Short reads are aligned to the reference genome, and evidence for an insertion is provided by (i) discordantly mapping read pairs, in which 1 mate maps uniquely to the reference flanking sequence and the other maps to the TE sequence, and (ii) split reads, in which a single read spans the junction between reference flanking DNA and the TE sequence. b) Overview of the TEforest pipeline. Input and output files are shown in ovals, important branching points in the pipeline are shown in diamonds, and computational steps of the pipeline are shown in rectangles. c) Example Integrative Genomics Viewer (IGV) view of short-read alignments at a TE insertion site. Colored reads represent the local anchors of discordantly mapping read pairs whose mates map to TE-derived sequence elsewhere in the genome, either at a distant position on the same chromosome or on a different chromosome.

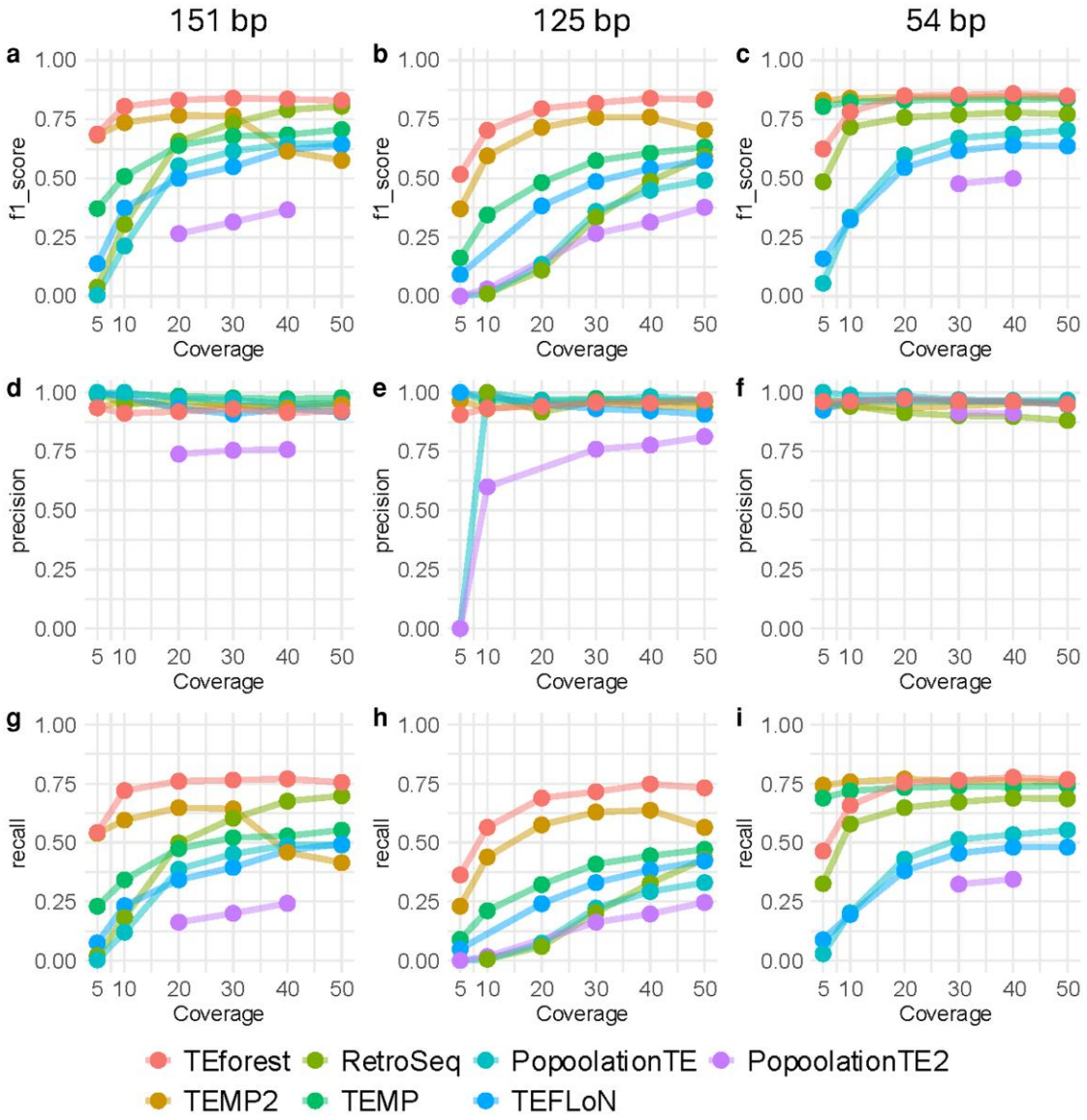

Fig. 2. The performance of TEforest compared to other short-read TE callers at detecting nonreference TE insertions annotated by long-read assemblies of *D. melanogaster* strains in 3 short-read sequencing datasets with varying read lengths and insert sizes. True positives were defined as predicting a TE insertion of the correct TE family with any prevalence within 500 bp of the true insertion site. Performance was quantified with (a, d, and g) F_1 scores, (b, e, and h) precision, and (c, f, and i) recall. The proportions of TEs that were successfully detected by TEforest (green), lost in the candidate region detection stage (red), or misclassified by the LightGBM model (red) for each read length are shown in (d). The mean breakpoint accuracy of TE callers for all (e) true-positive calls or (f) true-positive calls shared by TEforest, RetroSeq, and TEMP2 was quantified by finding the distance between the center of true-positive breakpoint ranges predicted by the TE callers and the breakpoint in the truth dataset. PopoolationTE2 failed to run at some coverages and thus is not represented for those coverages.

below; Fig. 2). TEforest's superior performance is largely due to an increase in recall, as most TE detectors, including TEforest, called very few false positives. In general, the F_1 score (the harmonic mean of recall and precision) increased with coverage for all callers (except TEMP2) due to an increased number of informative reads. Interestingly, TEforest tended to approach its highest F_1 score (~0.84) at low coverages (10× to 30×) before leveling off, while other callers continued to show slight increases up to 50× coverage. However, no method achieved recall above 77% for any coverage, indicating that a subset of TEs was undetectable with any amount of short-read data. In the case of TEforest, the majority of these false negatives (76.5% of false negatives for the 151-bp read dataset with 50× coverage) were lost in the candidate region detection stage instead of being mislabeled by the LightGBM model (Supplementary Fig. 1), which could be due to a lack of supporting reads or proximity to an annotated reference TE resulting in the region being filtered out (Methods).

The other TE callers tested exhibited more variable performance than TEforest across different read lengths and coverages. When tested with 151-bp reads, TEMP2 had the next-best F_1 score compared to TEforest from 5× to 30× coverage (e.g. 0.68 at 5× coverage and 0.77 at 30× coverage vs. 0.69 at 5× coverage and 0.84 at 30× coverage for TEforest), before seeing a rapid decline in recall at 40× (leading to a F_1 score of 0.62 for TEMP2 vs 0.84 for TEforest) and 50× coverage (F_1 of 0.58 for TEMP2 vs 0.83 for TEforest). RetroSeq continued to steadily increase in recall, eventually approaching but not exceeding the performance of TEforest at 50× coverage ($F_1 = 0.81$).

When tested with 125-bp reads, TEMP2 was consistently the next-best caller behind TEforest (TEforest's $F_1 = 0.83$ at 50×), though it still exhibited a decline in performance when increasing coverage to 50× ($F_1 = 0.76$ at 40× and 0.70 at 50×; Fig. 2), although this was not as severe as in the 151-bp dataset. In contrast to the 151-bp dataset, RetroSeq's performance lagged behind the other TE callers until 50× coverage ($F_1 = 0.49$ at 40× and 0.59 at 50×) and did not approach the F_1 score of TEforest. Interestingly, the median insert size for this dataset was 208, meaning that the paired-end reads typically had no gap between them when aligned to the genome. As RetroSeq only utilizes information from discordant paired-end reads rather than information about split reads for TE detection, it may struggle to detect TEs in this context. The performance gap between TEforest and other callers narrowed significantly when identifying TEs with shorter reads. In contrast to the 151- and 125-bp datasets, when tested on the 54-bp dataset, F_1 scores were approximately equal among TEforest, TEMP, and TEMP2 for coverage values between 20 and 50× (~0.85; Fig. 2). This trend highlights that while TEforest maintains consistent performance regardless of read length, methods such as TEMP and TEMP2 see significant performance gains as read lengths decrease. Interestingly, in contrast to results with longer reads, TEMP and TEMP2 did not experience a noticeable decline in performance at lower coverage with the 54-bp read dataset ($F_1 = 0.80$ and 0.83, respectively, at 5× coverage), while TEforest experienced a qualitatively similar dip in performance as it and other methods did for the datasets with 125 and 151 bp (F_1 score dropping to 0.63 at 5×), We also note that in the 54-bp dataset, TEMP2 did not exhibit any of the decline in performance at higher coverages that it did in the 151- and 125-bp datasets.

We also evaluated the computational efficiency of each tool with the Linux GNU time utility using 8 CPUs on an Intel Xeon Gold 6140 CPU (2.30 GHz) running RHEL 9.6 by running each tool 3× on the dataset with 151-bp reads at 30× coverage. While TEforest required the longest runtime (~160 min, 1.5× longer than the next slowest tools TEMP, PoPoolationTE1 and 2), its peak memory usage (~6.5 GB) remained comparable to most other callers, with only TEFLoN requiring significantly more memory (~15 GB) (Supplementary Fig. 2).

TEforest is robust to read lengths not in its training set, while TE callers show unexpected performance gains with shorter reads

To evaluate whether TEforest's performance declines on read lengths not represented in training, we trimmed 151- and 125-bp reads from our synthetic genomes to 100 bp at the 3′ end using *fastp* (Chen 2023) and downsampled them to 30× coverage for direct comparison with the previously evaluated 30× model. For reads trimmed from 125 to 100 bp, TEforest slightly increased in F_1 score of from 0.82 to 0.83 despite a reduction in read length (Supplementary Fig. 3). The recall for TEforest improved slightly from 0.72 to 0.75, while precision decreased to from 0.96 to 0.92. These results suggest that TEforest's feature vectors are resilient to variations in read length not contained in its training dataset. Some TE callers detected TEs much more accurately with these trimmed reads. RetroSeq exhibited the largest improvement, with its F_1 score increasing from 0.33 to 0.68, while TEMP2 had the smallest improvement, from 0.76 to 0.80. Despite these gains, none of the callers outperformed TEforest. The improvements in other callers' performance were driven by higher recall, though this came at the cost of slightly reduced precision in all cases. For the 151-bp reads, TEforest exhibited a slight decline in F_1 score after trimming down to 100 bp, dropping from 0.84 to 0.80 (Supplementary Fig. 4). Recall decreased from 0.77 to 0.73, while precision declined from 0.93 to 0.89. In contrast, the performance of other TE callers improved, though not as dramatically as the 125-bp case. PopoolationTE2 showed the largest increase, with its F_1 score rising from 0.32 to 0.46. Notably, RetroSeq slightly surpassed TEforest's performance, achieving an F_1 score of 0.81, up from 0.74.

These unexpected increases in the performance of other TE callers may be attributed to their reliance on discordant and split reads that can be mapped to TE sequences. While split reads are a valuable resource for precisely detecting the breakpoints of TE insertions, it is more challenging to correctly map each piece of the split read compared to mapping a discordant read. When reads spanning a TE insertion are trimmed, the length of the effectively unsequenced portion of the fragment is increased, making it more likely that the 2 reads will be on either side of the breakpoint. As a result, some non-TE-mapping split reads may instead fully map to TE sequences after trimming, increasing support for insertion calls for some tools. This explanation is supported by the larger performance gains seen when trimming the 125-bp dataset than when trimming the 151-bp dataset, as the former had an estimated mean insert size of 208 bp (meaning that prior to trimming there was often no space between the 2 paired reads), while the latter had a mean insert size of 436 bp (meaning that prior to trimming there was already an average 134-bp space between paired reads). Importantly, TEforest incorporates split-read information directly into its feature vectors, regardless of whether the read mapper correctly aligns the split piece back to the TE consensus sequence. This capability of TEforest is particularly desirable, as it allows researchers to utilize the full range of information present in their dataset without needing to manipulate or preprocess the data extensively to achieve optimal performance.

TEforest nonreference calls are close to the true insertion breakpoints

We next evaluated breakpoint accuracy by measuring the mean absolute distance between predicted and true insertion sites. For 151-bp reads, TEforest had a mean breakpoint error below 25 bp at all coverages, improving to ~10 bp at 50×, and generally performed better than or comparably to other callers (Supplementary Fig. 5a). The full distribution of predicted breakpoint distances from the true breakpoint shows that TEforest, TEMP2, and TEFLoN predict ~50% of calls 0 or 1 bp away from the annotated breakpoint, although TEMP2 has an excess of calls ≥ 25 bp away from the annotated breakpoint, causing TEMP2 to be 2× further than TEforest from the true breakpoint on average (Supplementary Fig. 6). For 125-bp reads, breakpoint accuracy improved with coverage but was broadly similar across methods (Supplementary Fig. 5b). For 54-bp reads, TEforest was less accurate than competing methods, likely because shorter reads provide less split-read information for breakpoint resolution (Supplementary Fig. 5c). Because breakpoint comparisons can be biased by differences in recall, we repeated the analysis using only insertions detected by TEforest, TEMP2, and RetroSeq. In this shared set, TEforest predicted breakpoints more accurately than TEMP2 and RetroSeq for 151- and 125-bp reads, but still less accurate for 54-bp reads at ≥30× coverage (Supplementary Fig. 5d to f).

TEforest accurately genotypes nonreference TE insertions

We also assessed the ability of TEforest and other callers to accurately genotype known TE insertions. Other callers estimate the

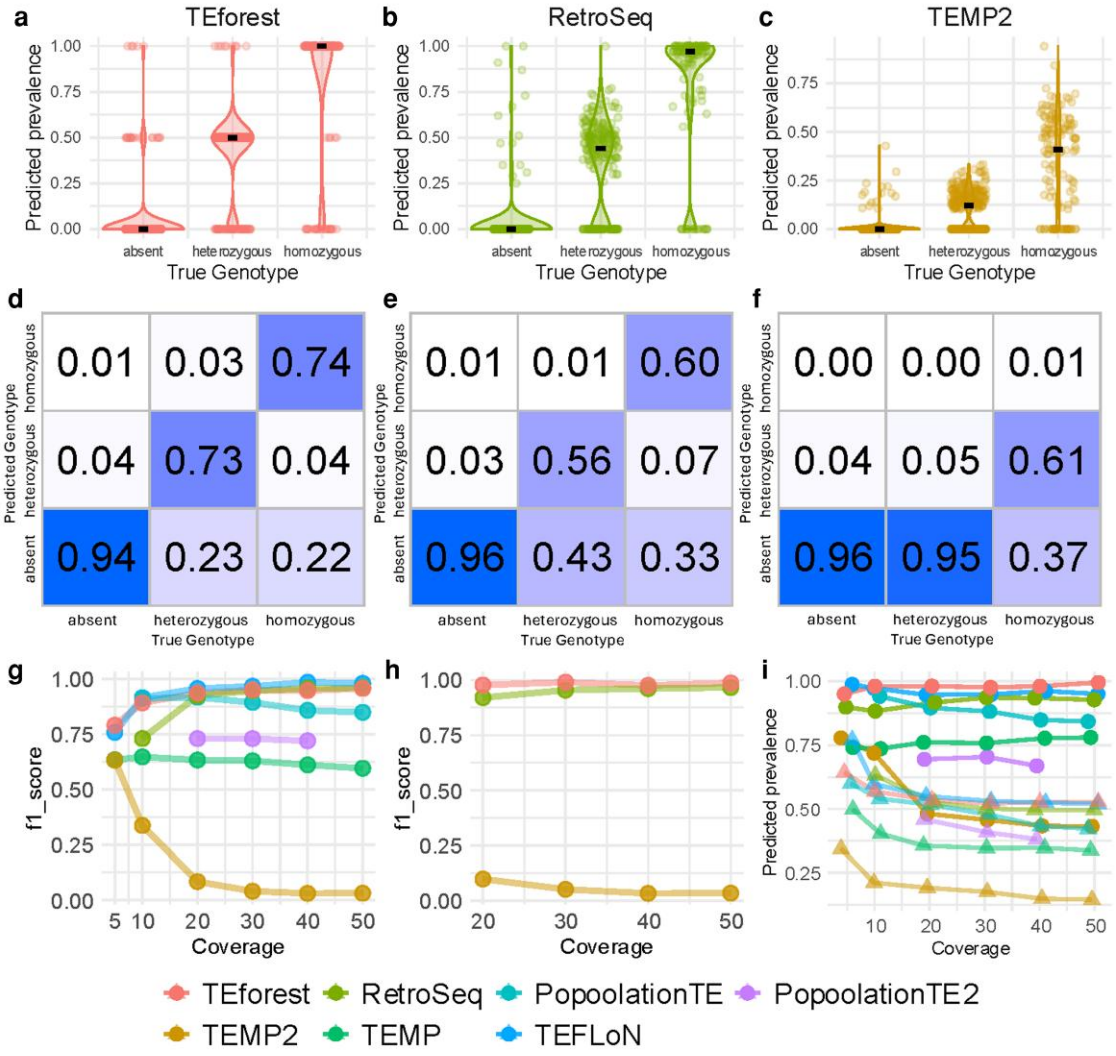

Fig. 3. The performance of TEforest compared to other short-read TE callers at genotyping nonreference TE insertions. Short reads used for detection of TEs were 151 bp with ~436-bp insert sizes. Prevalence predictions for insertions that were homozygous, heterozygous, or absent were calculated for (a) TEforest, (b) RetroSeq, and (c) TEMP2. The absent class consisted of (i) a random sample of true-negative loci from the TEforest candidate regions, matched in number to the true positives (and shared across all methods), together with (ii) false-positive calls from each method. True-negative loci not called by a method were assigned predicted prevalence 0, whereas false-positive calls retained their predicted prevalence values. The median prevalence is shown as a black line. d to f) The accuracy of these prevalence predictions is quantified in confusion matrices, where heterozygous predictions are defined as prevalence predictions between 0.25 and 0.75 and homozygous predictions are above 0.75. For (a to f), the methods were evaluated at a sequencing coverage of 30×. The genotyping F_1 score for (g) all true-positive calls or (h) true-positive calls shared by TEforest, RetroSeq, and TEMP2 was quantified using true-positive predictions of each caller. The mean for the predicted prevalence of true-positive homozygous (circles) and heterozygous (triangles) insertions are shown in (i), calculated using only the prevalences of true-positive predictions of each caller (false positives are excluded).

fraction of genomes in the sample from which DNA was collected that contain an insertion, yielding a prediction that can range anywhere from 0 to 1. While some tools refer to this as the TE insertion's "frequency," to avoid conflation with population frequency, we refer to this as the insertion's "prevalence." In contrast, since we designed TEforest to classify insertions as heterozygotes or homozygotes, it makes discrete predictions of 0 (homozygous absent), 0.5 (heterozygous), or 1 (homozygous present). Thus, to allow for a comparison of genotyping accuracy among methods, we transformed the predictions made by the other methods into distinct genotypes by labeling estimates below 0.25 as homozygous absent, between 0.25 and 0.75 as heterozygotes, and above 0.75 as homozygous present.

In the dataset with 151-bp reads at 30× coverage, the predicted prevalences of true-positive predictions were similar to the true values for TEforest (predicted mean prevalence of 0.98 for homozygous insertions and 0.52 for heterozygous insertions) and RetroSeq (homozygous mean prevalence = 0.94; heterozygous mean = 0.50), while TEMP2 systematically underestimated the prevalence (homozygous mean = 0.46; heterozygous mean = 0.18), despite having a high F_1 score of 0.77 for detecting insertions at this coverage (Fig. 3a to f). The average prevalence predicted by TEforest for homozygous true-positive insertions (that is, the average of its predicted genotypes) remained close to 1 across all coverages, while it tended to misclassify heterozygotes as homozygotes at lower coverages and make more accurate predictions at higher coverages (Fig. 3i). Other callers such as RetroSeq, TEFLoN, and PopoolationTE also had reasonable average prevalences, while TEMP and TEMP2 underpredicted the true prevalence of insertions. We also calculated a genotyping F_1 score, where we computed F_1 scores separately for homozygous and heterozygous genotypes and reported the macroaveraged F_1 score as

the overall measure of performance. TEforest displayed a competitive genotyping F_1 score relative to other across callers across coverages when calculated for all of its calls (Fig. 3g) or when comparing to the intersection of TE insertions detected by TEforest, RetroSeq, and TEMP2 (Fig. 3h). Despite recovering fewer known TEs, PopoolationTE and TEFLoN performed well at the genotyping task for those TEs they were able to detect (Fig. 3g). To further investigate the sources of error among callers, we calculated detection recall stratified by the true zygosity, where a site was considered detected if it was classified as either heterozygous or homozygous (i.e. not absent). TEforest exhibited consistent, high sensitivity across genotypes. For example, in the dataset with 151-bp reads at 30× coverage (Fig. 3d to f), TEforest successfully detected 78% of homozygous insertions and 77% of heterozygous insertions. In contrast, other methods showed a severe reduction in sensitivity for heterozygous sites. RetroSeq detected 67% of homozygotes but only 57% of heterozygotes, while TEMP2 detected 63% of homozygotes but failed to detect 95% of heterozygous insertions (heterozygote recall = 0.05). This indicates that the lower overall performance of competing methods is at least partially explained by an inability to distinguish heterozygous TE signals from background, whereas TEforest remains robust regardless of zygosity.

For 125-bp reads with short insert sizes, patterns of genotyping were largely the same as for the 151-bp dataset, with TEFLoN, PopoolationTE, and TEforest performing well (Supplementary Fig. 7). For example, at 30× coverage, TEforest's mean prevalence was 0.98 for homozygotes and 0.51 for heterozygotes. TEMP and TEMP2 also underpredicted the prevalence of insertions at this read length (e.g. for TEMP2 at 30× coverage, homozygous mean prevalence = 0.50; heterozygous mean = 0.18). RetroSeq, which also performed poorly at detecting known TE insertions, overpredicted the prevalence of heterozygous insertions for this dataset across all coverages (at 30× coverage, homozygous mean = 1.00; heterozygous mean = 0.77; Supplementary Fig. 7i). Genotyping performance for all callers was generally higher for the 54-bp reads than other read lengths (Supplementary Fig. 8), which is to be expected as more reads are sequenced per genome to reach a given coverage, and the space between the 2 reads on a sequenced fragment is larger. TEMP and TEMP2 still underpredicted the prevalence of TE insertions, though this bias was less severe than for the longer read lengths (Supplementary Fig. 8c and i). Specifically, at 30× coverage, the mean prevalence for homozygotes was 0.99, 0.98, and 0.94 for TEforest, RetroSeq, and TEMP2, respectively, while the mean prevalence for heterozygotes was 0.51, 0.54, and 0.39.

TEforest false negatives are associated with truncated TEs, nested TEs, and ground-truth discrepancies

Despite its high recall relative to other methods, TEforest still failed to detect a considerable fraction of true-positive nonreference insertions (24.4% of the true-positive calls in genome with 151 bp reads at 50× coverage). To better understand TEforest false negatives, we examined their genomic and annotation context in the 151-bp, 50× dataset. Undetected insertions were enriched for truncated (Supplementary Figs. 9 and 10) and nested TEs (Supplementary Table 4), especially among highly fragmented families (Supplementary Table 5), and manual inspection also suggested that a subset of apparent false negatives reflected discrepancies between the long-read-derived truth set and the short-read validation data rather than clear failures of TEforest (Supplementary Figs. 11 to 14 and Table 6). We also found that high levels of ambiguous bases in a small number of TE consensus sequences may impair read mapping for some families (Supplementary Table 7). Together, these analyses indicate that many residual false negatives arise from insertions that are inherently difficult to detect with short reads or limitations of the benchmark itself, rather than from a weakness of the specific method. These analyses are described in more detail in Supplementary Note 1.

Combining TE insertion detectors improves performance

Previous benchmarking studies have demonstrated that individual TE detection tools often detect different sets of false positives (Nelson et al. 2017; Vendrell-Mir et al. 2019; Xu et al. 2023), similar to more general structural variant analyses for which merging call sets can improve detection (Jeffares et al. 2017). To determine whether TEforest false negatives were identified by other methods and whether its improved performance stemmed from recovering insertions missed by all other methods, we evaluated the combined performance of various TE detectors in the genome with 151-bp reads at 50× coverage. TEforest and RetroSeq identified an equal number of unique true positives (3 insertions each). Of the 401 total true-positive insertions in the tested chromosome arms, the vast majority of those successfully detected were found by multiple tools. Specifically, ~97% of the 316 true-positive insertions detected by any caller were identified by at least 1 additional caller (Supplementary Fig. 15a). Notably, of all true insertions detected by at least 1 other method, only ~4.4% were not recovered by TEforest. Although each caller also produced unique false positives, 22% of false positives were shared by at least 2 callers (11 out of 49 total false-positive insertions; Supplementary Fig. 15b). Since recall was a bigger limiting factor than precision in the F_1 score, unions of call sets generally improved recall and F_1 relative to single callers (Supplementary Figs. 9b and 15b). However, TEforest alone outperformed most multicaller combinations (Supplementary Fig. 9b; $F_1 = 0.837$; recall = 0.753; precision = 0.941). The top-performing multicaller combinations yielded very similar overall performance. The highest-scoring combination consisted of PopoolationTE, RetroSeq, and TEMP ($F_1 = 0.851$; recall = 0.778; precision = 0.940), while the closely following second-best combination included TEforest alongside PopoolationTE and TEMP ($F_1 = 0.848$; recall = 0.781; precision = 0.929). These results show that combining callers can improve performance, although the best combination may vary across datasets, read lengths, and coverage levels.

TEforest uses biologically relevant features to make its predictions

We analyzed feature importances in TEforest's nonreference model trained at 50× coverage to identify the features contributing most to TEforest's predictions (Supplementary Table 8). Feature importance was quantified using total gain, defined as the cumulative reduction in the loss function attributable to all splits involving a given feature across the model (Ke et al. 2017). Features with higher gain therefore contributed more substantially to improving model fit. The top 10 features were primarily related to discordantly mapping read pairs, split reads, and proper pairing of reads, indicating that TEforest relies on alignment signatures similar to those used by earlier TE detection methods. In many cases, the standard deviation or IQR of a feature was more informative than central tendency metrics, such as the mean or median. Two highly ranked examples are shown in Supplementary Fig. 16: variation in split-read counts ("Cigar4

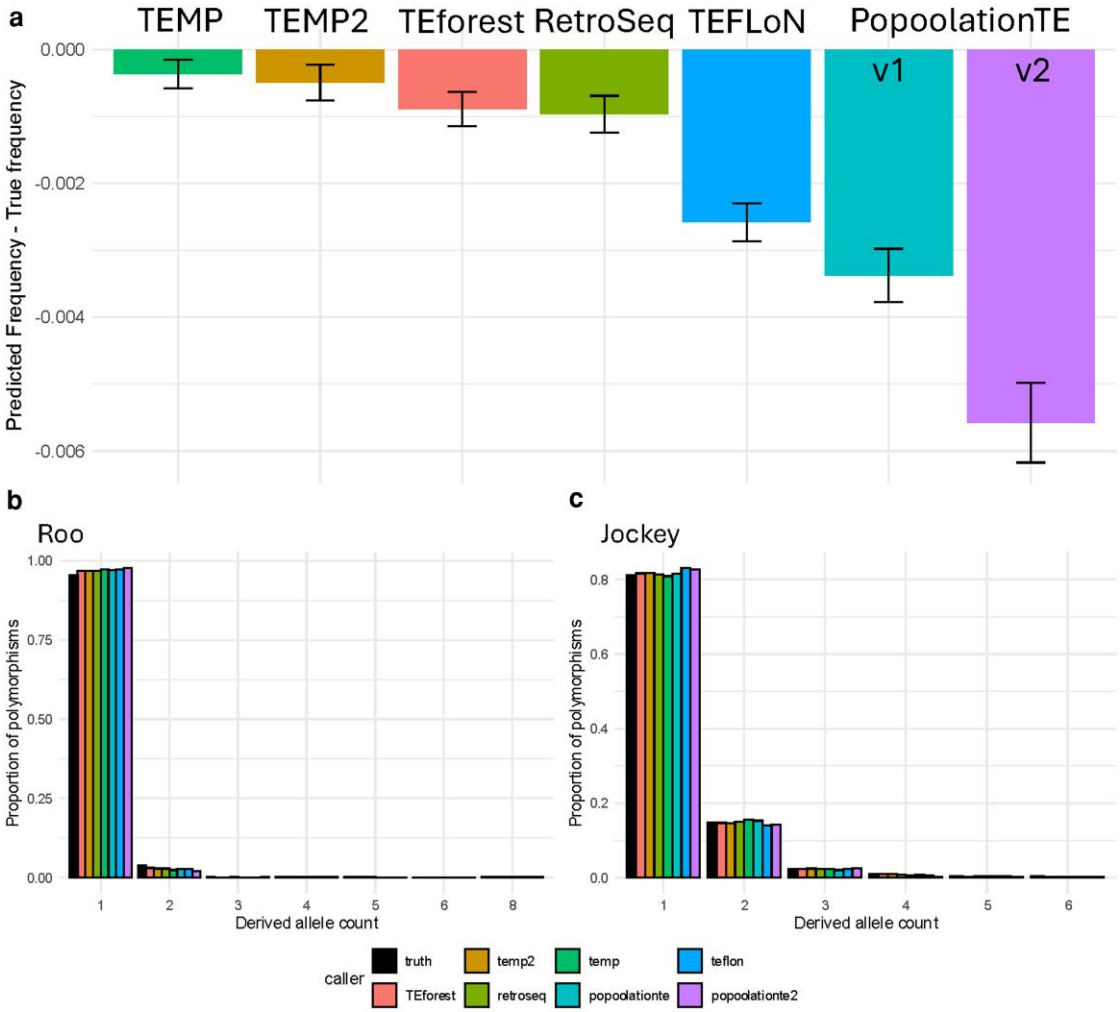

Fig. 4. a) Performance of TEforest compared to other TE callers at correctly predicting the allele frequencies of nonreference TE insertions in 13 fully homozygous individuals sequenced with 54-bp reads from the *Drosophila* Synthetic Population Resource. The *y* axis represents the average difference in allele frequency predicted by each caller to the true allele frequency at that locus. Error bars represent the standard error for each caller. The SFS of (b) Roo and (c) jockey TE insertions, which had the highest number of TE insertions in these individuals.

sd") appeared to help distinguish heterozygous from homozygous insertions, whereas the interquartile range of TE-specific discordantly mapping read counts ("TE-specific Discordant Read IQR") appeared to be more useful for identifying true-negative candidate regions.

Allele frequency prediction: can we use TEforest for population genetics?

To characterize the impact of any class of mutation, it is essential not only to identify these mutations but also to genotype them accurately enough to measure their frequency in a population. For example, the distribution of allele frequencies in a population, also called the site frequency spectrum (SFS), is a useful tool for understanding the population genetic forces governing a class of mutations (Williamson et al. 2005; Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Emerson et al. 2008). To accurately measure allele frequencies in a population, a TE caller will need to successfully detect and genotype the TE in all individuals carrying the allele (unless there are multiple detection/genotyping errors that happen to offset one another). For example, false-positive predictions could skew the SFS toward rare mutations, potentially inflating estimates of the strength of natural selection acting against new mutations (Emerson et al. 2008; Johri et al. 2022).

To assess whether the genotyping accuracy of TE detection methods examined here is sufficient to produce accurate estimates of the SFS of TE insertions, we ran all previously tested TE detection methods in 13 genomes, sequenced with 54-bp reads, from the *Drosophila* Synthetic Population Resource included in our truth dataset (King et al. 2012; Chakraborty et al. 2019; Rech et al. 2022). We assessed the allele frequency prediction accuracy for each nonreference insertion in the population by subtracting the true TE frequency from the number of true-positive predictions (Fig. 4a). Due to the tendency toward false negatives instead of false positives, all TE callers underpredicted the true TE frequency on average. TEMP, TEMP2, RetroSeq, and TEforest predicted frequencies more accurately than other callers, reflecting the previous benchmarking on single genomes. TEFLoN, PopoolationTE, and PopoolationTE2 struggled at this task due to many false-negative predictions. When viewing the entire SFS for Roo (Fig. 4b) and Copia (Fig. 4c) insertions, which were the 2 TEs with the highest number of nonreference copies in this dataset, we found that most callers predict the proportion of alleles in each frequency category fairly accurately. Perhaps in part

because this dataset is composed of a global sample of individuals rather than a local population, the vast majority of TE alleles were rare in the sample, limiting our ability to assess allele frequency prediction for high-frequency alleles. However, combined with our previous benchmarking results, we find that TEforest is uniquely capable of both recovering TE insertions and obtaining accurate allele frequency estimates for downstream population genetic inference.

TEforest outperforms other methods in reference TE detection

We also trained a model to detect and genotype TEs present in the reference genome, calculating the same feature vector as the non-reference model for each reference TE region. Because reference TEs' locations are predefined, this model does not require candidate region or breakpoint identification steps. TEforest showed the highest accuracy for reference TE detection across all tested read lengths and coverages (Supplementary Fig. 17a to c). For example, at 151 bp and 50× coverage, F_1 scores were 0.92 for TEforest, 0.86 for TEMP2, 0.84 for TEFLoN, and 0.31 for PopoolationTE2. TEforest also achieved the highest recall while maintaining high precision (recall = 0.95; precision = 0.90), compared with TEMP2 (0.88, 0.86) and TEFLoN (0.81, 0.89). TEforest, along with TEMP2, maintained a high F_1 score across all tested coverages (e.g. for 151-bp reads at 5× coverage $F_1 = 0.91$ and 0.87 for TEforest and TEMP2, respectively), whereas TEFLoN and PopoolationTE2 had lower recall at 20× coverage and below. Performance also varied little across read lengths: at 50× coverage, TEforest reached $F_1 = 0.93$ for both 125 and 54-bp reads, while TEMP2 reached 0.89 and 0.87, respectively (Supplementary Fig. 18a to c).

In addition, we tested the genotyping accuracy of TEforest relative to other TE calling methods (Supplementary Fig. 18d to f). It is important to note that TEMP2, the next best tool for detecting reference TEs, has no ability to genotype reference TEs. TEFLoN slightly outperformed TEforest in its ability to genotype true-positive reference TE insertions when tested on the datasets with 151- and 125-bp reads (for example, at 151 bp and 50×, $F_1 =$ 0.67 for TEFLoN vs 0.64 for TEforest), whereas TEforest performed slightly better in the dataset with 54-bp reads (0.69 for TEforest vs 0.66 for TEFLoN). The main source of error in reference TE detection was annotating heterozygous insertions as homozygous, leading the mean prevalence estimates for heterozygotes to be higher than 0.5 for TEforest and TEFLoN (0.80 and 0.75, respectively, for 151-bp reads at 50× coverage), with prevalence predictions slightly improving with increased coverage (Supplementary Fig. 17d to f). In contrast, the mean prevalence for true homozygotes was close to 1.0 for TEforest (0.99 for 151-bp reads at 50× coverage) and slightly lower for TEFLoN (0.96 for 151-bp reads at 50× coverage). PopoolationTE2 consistently underestimated prevalence for homozygous reference insertions, with this bias worsening at higher coverage, while its heterozygote estimates were initially inflated but approached 0.5 as coverage increased.

TEforest generalizes effectively to human data without retraining

To assess whether the feature sets learned from *D. melanogaster* function in other species, we applied the *Drosophila*-trained TEforest model to the human NA12878 genome. Despite the vast evolutionary distance and differences in TE landscape, TEforest achieved the highest F_1 scores for the detection of *Alu* and *SVA* insertions among all tested tools (Supplementary Fig. 19). For *Alu* elements, which comprise the vast majority of the ground-truth set (833/893 insertions), TEforest achieved an F_1 score of 0.74, outperforming RetroSeq (0.72), MEGAnE with its built-in TE library (0.70) or our custom TE library (0.65), TEMP2 (0.69), and DeepMEI (0.66). Notably, TEforest exhibited the highest precision for *Alu* detection (0.77), significantly reducing false discoveries compared to DeepMEI (0.53) and MEGAnE (0.55). Performance on *LINE1* (46/893 TEs) and *SVA* elements (14/893 TEs) was lower across all callers, primarily driven by low precision, indicating an excess of false positives or an incomplete annotation in the ground-truth set. For *SVA* elements, TEforest achieved the highest F_1 score (0.39), followed by TEMP2 (0.30), RetroSeq (0.26), DeepMEI (0.25), MEGAnE with its built-in TE library (0.24), and MEGAnE with our custom TE library (0.21). For LINE1 elements, TEMP2 achieved the highest F1 score (0.38), followed by RetroSeq (0.32), MEGAnE with our custom TE library (0.31), TEforest (0.29), MEGAnE with its built-in TE library (0.27), and DeepMEI (0.21).

Interestingly, while TEMP2 and DeepMEI achieved F_1 scores similar to previous benchmarks using human data, RetroSeq demonstrated unexpected competitiveness on this dataset (e.g. *Alu* $F_1 = 0.72$). Notably, RetroSeq outperformed newer methods, such as TEMP2 and MEGAnE, as well as the neural-network-based DeepMEI. This finding stands in contrast to the human benchmarks reported in the original publications for these tools, where RetroSeq was typically outperformed. This may be due to the specific sequencing parameters of this dataset (30× coverage, 150-bp reads with a 433-bp insert size), which closely match the ideal read lengths for RetroSeq identified here (Fig. 2), unlike the benchmarks of the other papers, which used different read lengths and coverages. Overall, these results demonstrate that the short-read mapping signatures captured by TEforest are biologically robust and transferable across species boundaries.

Discussion

Long-read sequencing technologies are becoming increasingly affordable and useful for analyzing TEs (Ewing et al. 2020; Zhou et al. 2020; Chu et al. 2021; Han et al. 2022; Rech et al. 2022; Mohamed et al. 2023), but short-read sequencing remains the more accessible option for many researchers, especially for large-scale population studies. Moreover, the vast majority of population genomic datasets that are currently available to date consist of short-read sequences. To integrate recent advancements in long-read sequencing with existing population-scale short-read sequencing datasets, we developed TEforest—a machine learning tool for detecting and genotyping TE insertions using short-read data. By incorporating high-quality annotations from long-read *D. melanogaster* genome assemblies in the training process, TEforest leverages the precision of long-read sequencing while taking advantage of the widespread availability of short-read data. Our comprehensive benchmarking, one of the most extensive real-data evaluations to date involving 6 different TE detection tools, demonstrates that TEforest not only outperforms existing methods but also maintains more stable performance across various read lengths and coverages. While other tools exhibit fluctuating performance depending on sequencing parameters, such as a surprising increase in performance with shorter read lengths, TEforest consistently delivers accurate detection and genotyping of TE insertions.

Our benchmarking assessments highlight the advantages of using real data over simulated data. While simulated datasets allow controlled testing of algorithms, they currently do not capture the full complexity of real genomic data, including sequencing errors, structural variations near TE insertion breakpoints or within TE sequences, and the complexity of nested TE insertions. Early

testing on real data also shaped decisions made during the development of our algorithm, such as accounting for the fact that many TE insertion breakpoints have small gaps between the 2 breakpoints rather than the classic TSD signature (see the example insertion in Fig. 1c). Such details might have been overlooked if we developed the algorithm using a simulated dataset.

In our study, we observed that in some cases, PoPoolationTE and PoPoolationTE2 did not perform as well as other methods both in simulated datasets (also reported in Chen et al. (2023) and in our real datasets) (see Vendrell-Mir et al. (2019) for a counterexample). Conversely, TEMP2 and RetroSeq, which have shown strong performance in other studies, also performed reasonably well in our benchmarks.

However, our results also suggest that the combination of TE callers that performs the best on a given dataset will vary depending on the insert size, read length, sequencing coverage, and other species-specific parameters and thus may be difficult to choose in practice without prior benchmarking. The benchmarks of previous studies have shown that TE detection tools often perform variably across different species due to differences in genome structure and TE content. For instance, tools, such as MELT (Gardner et al. 2017), originally developed for use in the human genome performed worse when applied to *Drosophila* genomes due to the mislabeling of related TE families, while TEMP2 and RetroSeq performed well in both contexts (Yu et al. 2021). Similarly, TEPID performed well when detecting simulated *Arabidopsis* insertions, but declined in performance when applied to *Drosophila* (Verneret et al. 2025). A potential solution to the variable performance of different methods depending on context is to use a combination of methods to make more confident predictions. Similar to Xu et al. (2023), we find that certain combinations of TE detection methods can increase performance by increasing recall without a severe decline in precision. We also note that when evaluating different combinations of tools, we simply took the union of all TE insertions detected by a given set of callers, and it is possible that more complex rules for combining TE call sets from different methods could be more effective (e.g. requiring that all calls produced by a method with lower precision be recovered by at least 1 additional caller). In addition, one could optimize a model or model ensemble to maximize precision at the cost of recall or vice versa depending on the use case.

The human-focused tools XTea (Chu et al. 2021) and DeepMEI (Xu et al. 2023) also use machine learning in the context of nonreference TE detection, while other recent methods, such as MEGAnE (Kojima et al. 2023), rely on advanced heuristic approaches. XTea employs random forest with a less comprehensive feature vector for short-read genotyping, while DeepMEI utilizes a convolutional neural network (CNN) for TE detection and genotyping. When we benchmarked these tools on 1 human dataset, we found that the *Drosophila*-trained TEforest model generalized effectively without retraining, outperforming both DeepMEI and MEGAnE in detecting *Alu* and SVA insertions. This demonstrates that the feature vectors extracted by TEforest capture structural variation signatures that are robust across species. While TEforest generalized well to a human dataset without retraining, we also provide an automated Snakemake workflow in our GitHub repository to retrain TEforest for other species or datasets given validated TE insertion sites. Interestingly, we also observed that RetroSeq performed competitively on this human dataset, outperforming the newer neural-network-based methods. This stands in contrast to the benchmarking reported by Xu et al. (2023), in which DeepMEI outperformed RetroSeq. This discrepancy underscores the variability of TE caller performance across datasets and sequencing parameters, suggesting that continued, comprehensive benchmarking across different coverages, read lengths, and insert sizes will be essential as new datasets and tools emerge.

The success of short-read TE detection approaches, such as TEforest, is fundamentally dependent on the availability of high-quality reference genomes from which comprehensive TE annotations can be produced and for which short-read sequencing data can be obtained. Thankfully, these resources are readily available in our tested species *D. melanogaster* (Kaminker et al. 2002; Rech et al. 2022). The performance of our approach of training on TE annotations and short-read mapping patterns from such assemblies would degrade when using poorly assembled or annotated reference genomes or species with an incomplete library of TE consensus sequences. Recent methodological advances, such as GraffiTE (Groza et al. 2024), have also sought to bridge the gap between long-read resolution and short-read accessibility by constructing pangenome graphs that incorporate TE structural variants. While both GraffiTE and TEforest leverage high-quality long-read data to improve short-read genotyping, they address different challenges. Pangenome approaches excel at characterizing common variants and improving read mapping near common structural variants and in complex, repetitive regions, such as heterochromatin. However, machine learning approaches, such as TEforest, may be more effective for detecting the rare or de novo insertions typical of *D. melanogaster* populations (Fig. 4), which are unlikely to be represented in a pangenome graph constructed from a limited number of individuals. As pangenome alignment tools mature, future extensions of TEforest could potentially combine these strengths by extracting feature vectors from graph-based alignments to call novel insertions. Bearing this in mind, along with the often-unpredictable performance of TE detection tools when tested in a new species for which high-quality training/benchmarking data are not available, we recommend careful evaluation of TE detection tools when analyzing TEs in nonmodel species, perhaps by simulating TE insertions as realistically as possible for testing (Chen et al. 2023; Verneret et al. 2025). Despite these challenges, we expect the performance and utility of short-read TE detection tools to increase as genome assemblies become more complete (Li and Durbin 2024) and tools to automatically curate TE consensus sequences and annotate them in a reference genome become more advanced (Baril et al. 2024; Orozco-Arias et al. 2024).

Our study and others (e.g. Xu et al. 2023) have found that shorter, fragmented TE insertions are more challenging to detect than full-length insertions. As shorter TE copies are often inactivated or made reliant upon the transposition machinery encoded by other TE copies, it could be useful to extend TEforest to classify insertions as "full-length" vs "fragmented" based on where informative reads map along the TE consensus. Such an approach may detect truncations at the TE ends (via shifts in breakpoint-proximal read-mapping patterns), but its utility would be limited by short-read length: Short reads typically provide strong evidence near the insertion junctions yet offer little coverage of internal TE sequence, making internal deletions difficult to observe. For example, with the longest paired-end reads used here, informative read evidence generally extends only a few hundred base pairs into the inserted TE, whereas some TE families exceed 10 kb in length. Another potential extension of TEforest involves detecting low-prevalence or somatic TE insertions, which are typically more challenging to detect due to the low number of informative reads per insertion (Yu et al. 2021); training data for such a method could be obtained by combining reads from genomes known to have different genotypes for a TE insertion to produce a pooled sample where the TE insertion is present at the desired prevalence. A machine

learning algorithm, such as TEforest, could be useful in distinguishing true insertions from false positives, provided that appropriate training data is available.

TEforest represents a significant step forward in the detection and genotyping of TE insertions from short-read sequencing data. By combining machine learning techniques with comprehensive feature extraction from read alignments, it consistently outperforms existing methods. As genomic resources and annotations continue to improve, approaches, such as TEforest—which leverage high-quality ground-truth TE insertion datasets to enhance the utility of less costly sequencing data—will become increasingly robust. In turn, this will enable broader and more in-depth studies on the evolutionary and functional impact of TEs.

Data availability

The code used to train and run TEforest, along with the trained models and code used for benchmarking, is available at https://github.com/SchriderLab/TEforest.git (figshare DOI: 10.6084/m9.figshare.28278599). The MCTE library of consensus TE sequences and annotations are available at the DIGITAL.CSIC repository and can be accessed with the DOIs 10.20350/digitalCSIC/13765 and 10.20350/digitalCSIC/13894, respectively. The short-read sequences used for this study are available in the NCBI database under the BioProject accessions PRJNA559813 and SRP011971.

Supplemental material available at GENETICS online.

Acknowledgments

We thank Marta Coronado Zamora and Josefa González for their assistance with the *Drosophila* truth dataset. We thank Tianxiong Yu for the assistance in obtaining the human truth dataset. We thank 3 anonymous reviewers for their helpful suggestions that improved the manuscript. The research in this study was conducted using computational resources provided by ITS Research Computing at the University of North Carolina at Chapel Hill.

Funding

AD received support from the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R35GM154969 and T32GM067553, and DRS received support from the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138286.

Conflicts of interest

None declared.

Author contributions

A.D.: conceptualization, formal analysis, software, visualization, writing—original draft, and writing—review & editing. L.S.W.: conceptualization, formal analysis, software, and writing—review & editing. R.Z.: resources, software, and writing—review & editing. J.J.E.: conceptualization, resources, and writing—review & editing. D.R.S.: conceptualization, formal analysis, visualization, writing—review & editing, and funding acquisition.

Literature cited

Adrion JR *et al.* 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. Genome Biol Evol. 9:1329–1340. https://doi.org/10.1093/gbe/evx050.

Baril T, Galbraith J, Hayward A. 2024. Earl grey: a fully automated user-friendly transposable element annotation and analysis pipeline. Mol Biol Evol. 41:msae068. https://doi.org/10.1093/molbev/msae068.

Bourque G *et al.* 2018. Ten things you should know about transposable elements. Genome Biol. 19:199. https://doi.org/10.1186/s13059-018-1577-z.

Byrska-Bishop M *et al.* 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. 185:3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22:1503–1517. https://doi.org/10.1111/mec.12170.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 10:4872. https://doi.org/10.1038/s41467-019-12884-1.

Charlesworth B, Langley CH. 1989. The population genetics of *Drosophila* transposable elements. Annu Rev Genet. 23:251–287. https://doi.org/10.1146/annurev.ge.23.120189.001343.

Chen J *et al.* 2023. Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast. Mob DNA. 14:8. https://doi.org/10.1186/s13100-023-00296-4.

Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2:e107. https://doi.org/10.1002/imt2.107.

Chu C *et al.* 2021. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat Commun. 12:3836. https://doi.org/10.1038/s41467-021-24041-8.

Craig NL. 2002. Mobile DNA: an introduction. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington (DC): ASM Press. p. 1–11. https://doi.org/10.1128/9781555817954.ch1.

Danecek P *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10:giab008. https://doi.org/10.1093/gigascience/giab008.

Drongitis D, Aniello F, Fucci L, Donizetti A. 2019. Roles of transposable elements in the different layers of gene expression regulation. Int J Mol Sci. 20:5755. https://doi.org/10.3390/ijms20225755.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science. 320:1629–1631. https://doi.org/10.1126/science.1158078.

Ewing AD *et al.* 2020. Nanopore sequencing enables comprehensive transposable element epigenomic profiling. Mol Cell. 80:915–928.e5. https://doi.org/10.1016/j.molcel.2020.10.024.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 173:891–900. https://doi.org/10.1534/genetics.106.057570.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 9:397–405. https://doi.org/10.1038/nrg2337.

Finnegan DJ. 1992. Transposable elements. Curr Opin Genet Dev. 2:861–867. https://doi.org/10.1016/S0959-437X(05)80108-X.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. Gene. 463:18–20. https://doi.org/10.1016/j.gene.2010.04.015.

Gardner EJ *et al.* 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 27:1916–1929. https://doi.org/10.1101/gr.218032.116.

Groza C et al. 2024. A unified framework to analyze transposable element insertion polymorphisms using graph genomes. Nat Commun. 15:8915. https://doi.org/10.1038/s41467-024-53294-2.

Han S et al. 2022. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. Nucleic Acids Res. 50:e124. https://doi.org/10.1093/nar/gkac794.

Hill T, Unckless RL. 2019. A deep learning approach for detecting copy number variation in next-generation sequencing data. G3 (Bethesda). 9:3575–3582. https://doi.org/10.1534/g3.119.400596.

Hoyt SJ et al. 2022. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. Science. 376: eabk3112. https://doi.org/10.1126/science.abk3112.

Jeffares DC et al. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 8:14061. https://doi.org/10.1038/ncomms14061.

Johri P et al. 2022. Recommendations for improving statistical inference in population genomics. PLoS Biol. 20:e3001669. https://doi.org/10.1371/journal.pbio.3001669.

Kaminker JS et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3:research0084.1. https://doi.org/10.1186/gb-2002-3-12-research0084.

Ke G et al. 2017. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 30:3146–3154. https://doi.org/10.5555/3294996.3295074.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 29:389–390. https://doi.org/10.1093/bioinformatics/bts697.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 177:2251–2261. https://doi.org/10.1534/genetics.107.080663.

King EG et al. 2012. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. Genome Res. 22: 1558–1566. https://doi.org/10.1101/gr.134031.111.

Kirov I 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. https://doi.org/10.3390/plants10122681.

Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 8:e1002487. https://doi.org/10.1371/journal.pgen.1002487.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using pool-seq. Mol Biol Evol. 33:2759–2764. https://doi.org/10.1093/molbev/msw137.

Kojima S et al. 2023. Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. Nat Genet. 55:939–951. https://doi.org/10.1038/s41588-023-01390-2.

Lawrence M et al. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol. 9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118.

Lee YCG. 2015. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. PLoS Genet. 11:e1005269. https://doi.org/10.1371/journal.pgen.1005269.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [preprint]. arXiv, arXiv:1303.3997. https://doi.org/10.48550/arXiv.1303.3997.

Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. Nat Rev Genet. 25:658–670. https://doi.org/10.1038/s41576-024-00718-w.

Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. PLoS One. 7:e30008. https://doi.org/10.1371/journal.pone.0030008.

Lynch M. 2007. The origins of genome architecture. Sinauer Associates.

Makałowski WGotea VPande A, Makałowska I. 2019. Transposable elements: classification, identification, and their use as a tool for comparative genomics. In: Anisimova M, editor. Evolutionary genomics. Humana Press Inc.; Vol. 1910. p. 177–207. https://doi.org/10.1007/978-1-4939-9074-0_6.

Mohamed M et al. 2023. TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. Genome Biol. 24:63. https://doi.org/10.1186/s13059-023-02911-2.

Montgomery E, Charlesworth B, Langley CH. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet Res. 49:31–41. https://doi.org/10.1017/S0016672300026707.

Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. Genetics. 129:1085–1098. https://doi.org/10.1093/genetics/129.4.1085.

Nelson MG, Linheiro RS, Bergman CM. 2017. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3 (Bethesda). 7:2763. https://doi.org/10.1534/g3.117.043893.

Orozco-Arias S, Sierra P, Durbin R, González J. 2024. MCHelper automatically curates transposable element libraries across eukaryotic species. Genome Res. 34:2256–2268. https://doi.org/10.1101/gr.278821.123.

Rech GE et al. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. Nat Commun. 13:1948. https://doi.org/10.1038/s41467-022-29518-8.

Rishishwar L, Mariño-Ramírez L, Jordan IK. 2017. Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinform. 18:908–918. https://doi.org/10.1093/bib/bbw072.

Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. Mol Ecol. 28:1537–1549. https://doi.org/10.1111/mec.14794.

Shen W, Sipos B, Zhao L. 2024. SeqKit2: a Swiss army knife for sequence and alignment processing. iMeta. 3:e191. https://doi.org/10.1002/imt2.191.

Sudmant PH et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature. 526:75–81. https://doi.org/10.1038/nature15394.

Vasimuddin MMisra SLi H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium, Rio de Janeiro, Brazil. p 314–324 https://doi.org/10.1109/IPDPS.2019.00041 2019.

Vendrell-Mir P et al. 2019. A benchmark of transposon insertion detection tools using real data. Mob DNA. 10:53. https://doi.org/10.1186/s13100-019-0197-9.

Verneret M et al. 2025. Particular sequence characteristics induce bias in the detection of polymorphic transposable element insertions. Peer Community J. 5:e63. https://doi.org/10.24072/pcjournal.570.

Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. Annu Rev Genet. 54:539–561. https://doi.org/10.1146/annurev-genet-040620-022145.

Williamson SH *et al.* 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 102:7882–7887. https://doi.org/10.1073/pnas.0502300102.

Xu X *et al.* 2023. Identification of mobile element insertion from whole genome sequencing data using deep neural network model [preprint]. bioRxiv 531451. https://doi.org/10.1101/2023.03.07.531451.

Yu T *et al.* 2021. A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. Nucleic Acids Res. 49:e44. https://doi.org/10.1093/nar/gkab010.

Zhou W *et al.* 2020. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. 48:1146–1163. https://doi.org/10.1093/nar/gkz1173.

Zhuang J, Wang J, Theurkauf W, Weng Z. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. 42:6826–6838. https://doi.org/10.1093/nar/gku323.

Editor: K. Lohse