SUPPLEMENTARY FILE

Supplementary Note 1. Sources of TEforest false negatives

Truncated and nested insertions

Despite its high recall relative to other methods, TEforest still failed to detect a considerable fraction of true positive non-reference insertions (24.4% of the true positive calls in genome with 151 bp reads at 50X coverage). We hypothesized that insertions of TEs that were shorter or more fragmented relative to the consensus TE sequence, nested inside of other TE insertions, and/or located near other structural variants would be more difficult to detect than full-length TE insertions in non-repetitive regions. Focusing on TEforest false negatives for the genome sequenced with 151 bp reads at 50X coverage, we found that insertions that were not detected in the candidate region detection stage of the algorithm were significantly shorter in total length (Figure 4A) compared to their consensus TE sequence (SFigure 9). Lengths of TE insertions were taken from Supplementary Table 10 from Rech *et al.* (2022). This length bias largely explains the poor performance on specific families (STable 1); for instance, families with zero detected insertions, such as *1731* and *NOF*, consisted almost entirely of highly truncated fragments. Similarly, ~20% of false negatives belonged to the *INE-1* and *Gypsy-2-Dsim* families. *INE-1* is a TE family that has been thought to be inactive for millions of years (Kapitonov and Jurka 2003; Singh and Petrov 2004), and thus most copies are expected to be old and fragmented, so low performance on this family is not surprising and perhaps not as relevant to efforts to detect genuine new TE insertion polymorphisms. In addition, TEs nested with copies from different TE families were much more difficult to detect, with 32/41 TEs nested with different TE families not identified by TEforest, with 22 of these lost in the candidate region detection stage (STable 2). TEs nested with copies of the same TE family were easier to detect (though not as easy as unnested TEs), with ~17% undetected. Extreme nesting scenarios proved particularly challenging; notably, the four undetected *Kepler* insertions represent a single complex event consisting of four distinct TEs nested at the same genomic coordinate. These issues are not limited to TEforest: for example, RetroSeq, like TEforest, failed to detect any of the 31 total insertions of the particularly challenging *INE-1*, *Gypsy-2-Dsim*, *NOF*, *1731*, and *Kepler* families. We also note that detection sensitivity will be inherently limited in genomic regions with low mapping quality, such as dense heterochromatin, in which Rech *et al.* (2022)

did not attempt to detect TEs. In these contexts, the repetitive nature of the sequence can interfere with the unique alignment of short reads, effectively masking the split and discordant read signatures required by TEforest and other tools that use short reads.

Ground-truth discrepancies

Finally, we investigated whether a subset of false negatives could be attributed to discrepancies between the ground-truth annotations (derived from long-read assemblies) and the short-read data used for testing. Recent benchmarking of the Rech *et al.* (2022) dataset identified a subset of false positive annotations resulting from errors mapping TE insertions in the non-reference genome ground truth dataset to the reference genome (Groza et al. 2024). False positives in the assembly would be expected to result in false negatives in the short-read data. We note that any such errors would be removed from the positive set of our training data because of the filtering during our candidate region detection step; the impact of any false TE insertions in our truth data is thus limited to our hold-out evaluation sets. To assess if such discrepancies contributed to TEforest's false negatives, we manually inspected the alignment of short and long reads to the long-read assembly for the A1 strain (STable 3). We examined 15 randomly selected false negatives and found that 29% (4/14; one insertion inconclusive) lacked short-read support entirely (SFigure 10), while 14% (2/14) appeared to be polymorphic insertions present in a subset of the flies sequenced for the long-read assembly but absent in the short-read data (SFigure 11). In contrast, 15 randomly selected true positive TEs were all supported by both data types (SFigure 12). Combined with the evidence in Groza *et al.* (2024), this analysis suggests that, while overall the Rech *et al.* (2022) truth dataset is reliable, a portion of the reported false negatives are likely due to biological divergence between the sequenced individuals or potential errors in mapping assemblies to each other, rather than a failure of TEforest. Consequently, the performance metrics reported here are likely to represent a conservative estimate of the true recall of TE detection methods.

Effects of ambiguous bases in TE consensus sequences

We next assessed whether ambiguous bases in the TE consensus library could reduce the sensitivity of TEforest by limiting the mapping of TE-associated reads during the consensus-alignment step. To do so, we quantified the fraction of ambiguous bases in each consensus

sequence and found that the vast majority contained little to no ambiguity; however, a small subset of families showed elevated IUPAC content (STable 4), including *Kepler* (~21% IUPAC bases) and *INE-1* (~9%), with several additional families exceeding ~1–5%. To evaluate the practical impact of this ambiguity on read recruitment, we aligned all short reads from the DSPR strain A1 (54-bp Illumina reads) directly to the TE consensus library and examined coverage across the consensus sequences. For *Kepler*, which has the highest ambiguity content, we observed that the highly ambiguous portion of the consensus had little to no read coverage, consistent with reduced alignment sensitivity in that region (SFigure 13). Notably, TEforest failed to detect all four *Kepler* insertions in A1 (a single complex nested event), suggesting that extensive ambiguity in the consensus sequence can contribute to false negatives for certain TE families.

REFERENCES

Groza C et al. 2024. A unified framework to analyze transposable element insertion polymorphisms using graph genomes. Nat Commun. 15(1):8915. https://doi.org/10.1038/s41467-024-53294-2

Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA. 100(11):6569–6574. https://doi.org/10.1073/pnas.0732024100

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(1):1948. https://doi.org/10.1038/s41467-022-29518-8

Singh ND, Petrov DA. 2004. Rapid sequence turnover at an intergenic locus in *Drosophila*. Mol Biol Evol. 21(4):670–680. https://doi.org/10.1093/molbev/msh060

SUPPLEMENTARY FIGURES AND TABLES

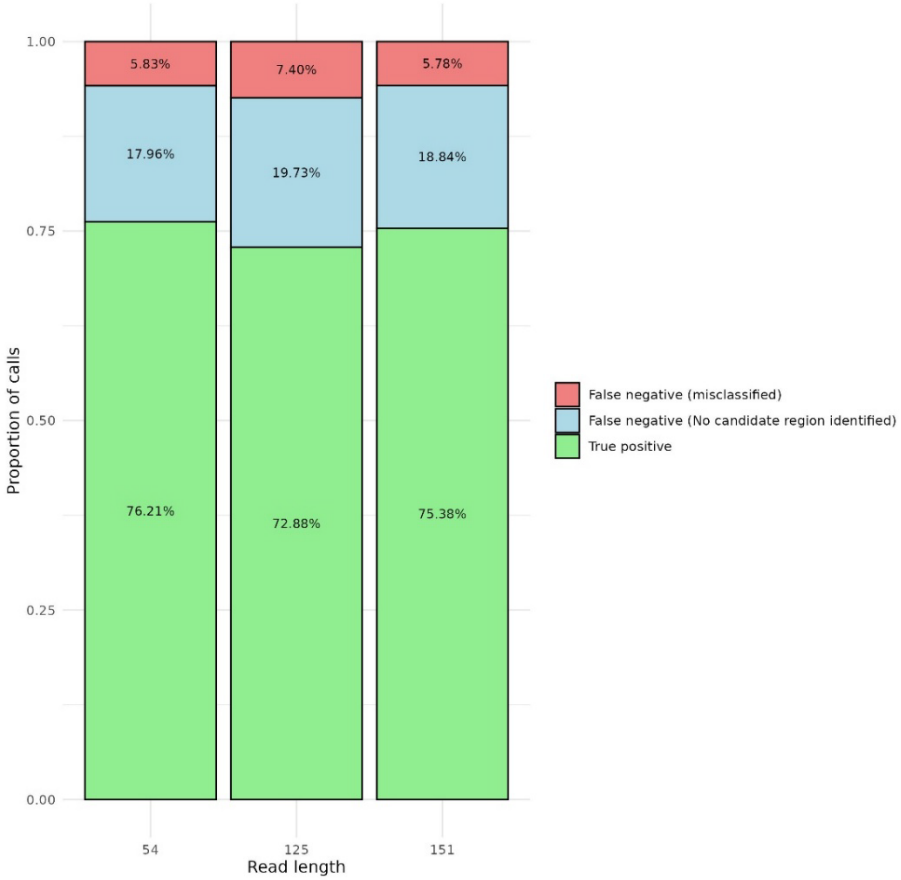


Figure S1: The proportions of TEs that were successfully detected by TEforest, lost in candidate region detection stage, or misclassified by the LightGBM model for each read length at 50X coverage.

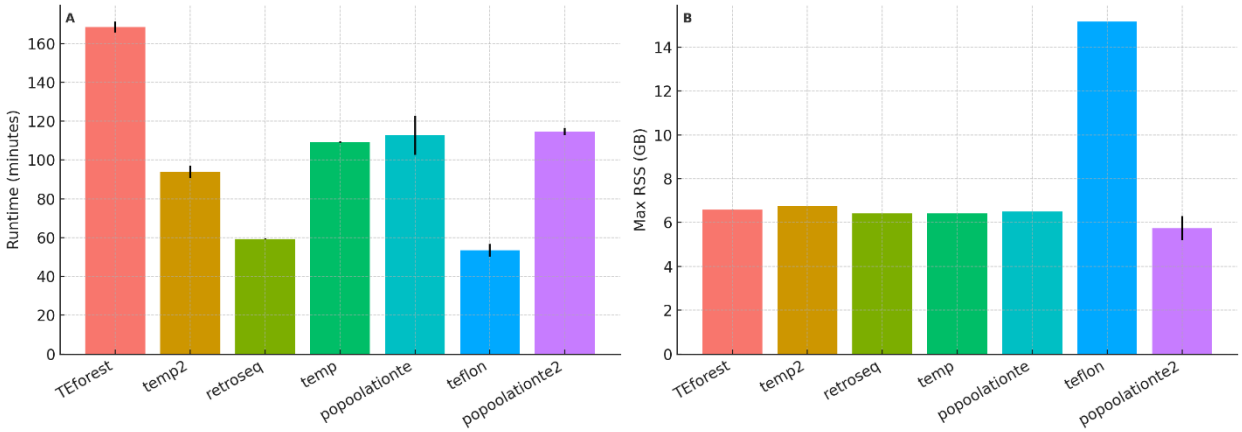


Figure S2: Computational performance of TE detection tools. (A) Mean runtime (wall-clock time in minutes) and (B) peak memory usage (Maximum Resident Set Size in GB) for TEforest and competing TE callers. Benchmarks were performed using the 151 bp read length dataset at 30X coverage. Each tool was run in triplicate on a single computational node (Intel Xeon Gold 6140 CPU @ 2.30GHz, Red Hat Enterprise Linux 9.6). Error bars represent the standard deviation across three independent replicates.

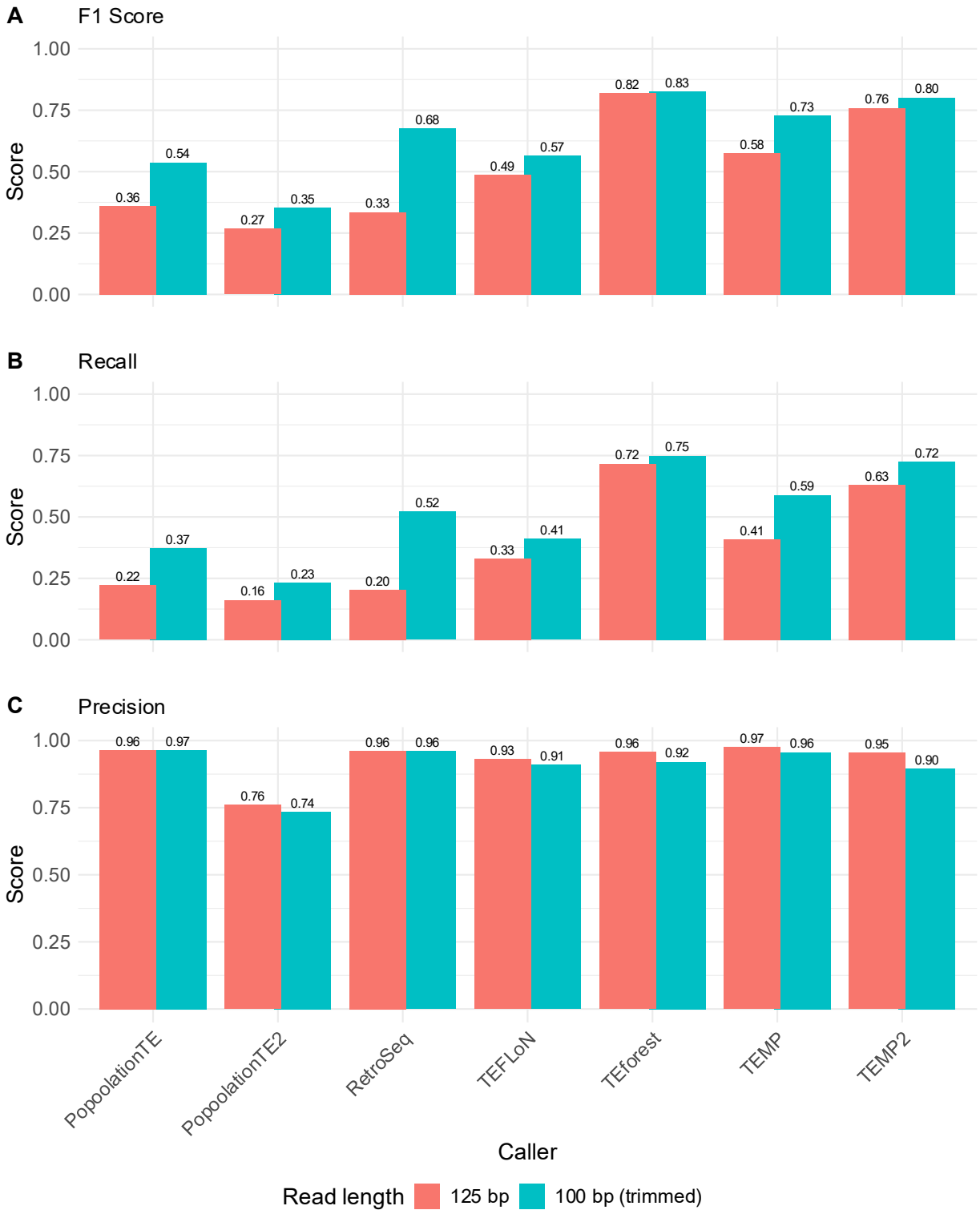


Figure S3: Effects of trimming reads on the (**A**) F1 score, (**B**) recall, and (**C**) precision of TEforest and other short-read TE callers at detecting non-reference TE insertions. Reads were originally 125 bp with ~208 bp insert sizes, then shortened to 100 bp with the same insert size by trimming at the 3' end of the reads. Both datasets were downsampled to 30X coverage.

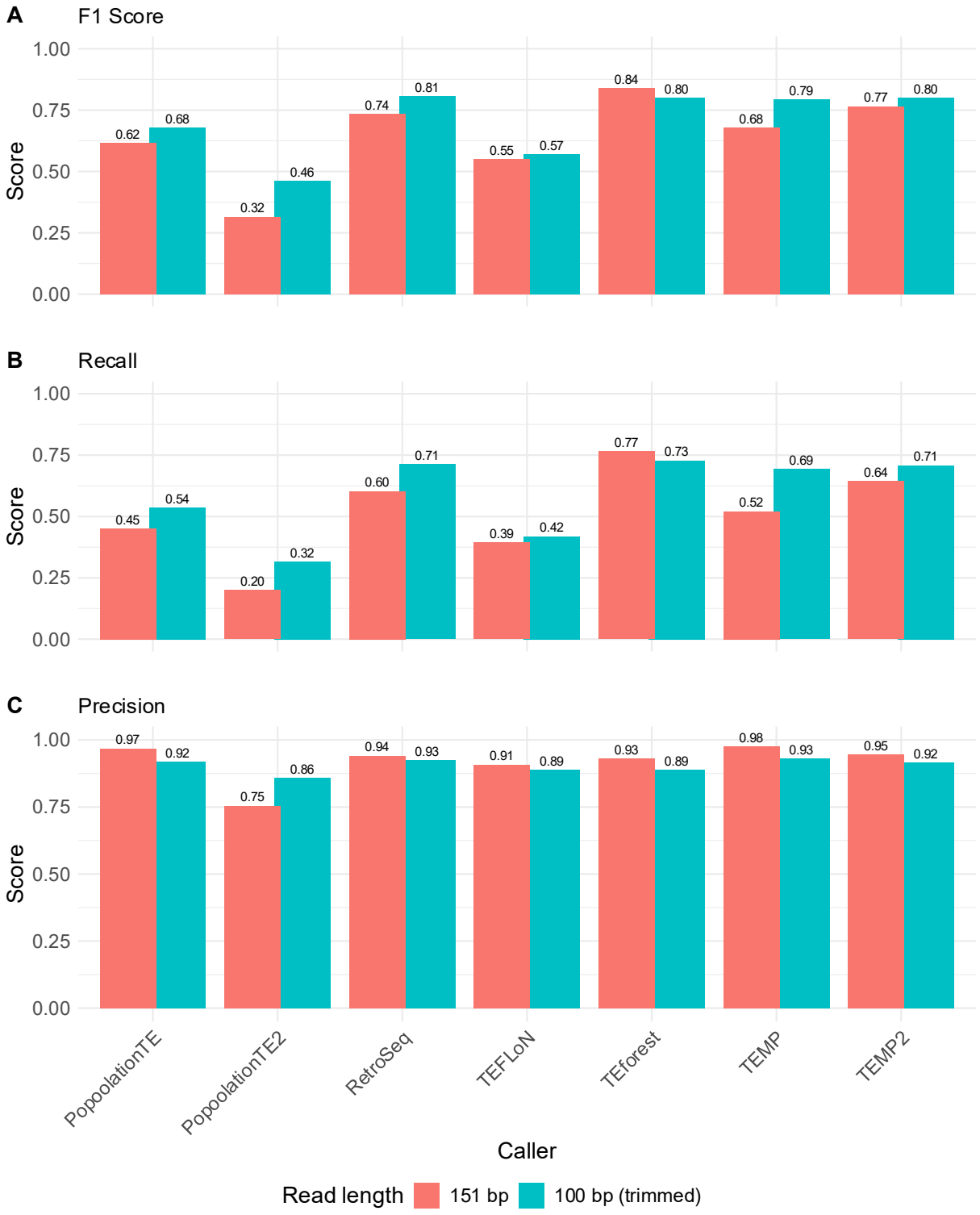


Figure S4: Effects of trimming reads on the (**A**) F1 score, (**B**) recall, and (**C**) precision of TEforest and other short-read TE callers at detecting non-reference TE insertions. Reads were originally 151 bp with ~436 bp insert sizes, then shortened to 100 bp with the same insert size by trimming at the 3' end of the reads. Both datasets were downsampled to 30X coverage.

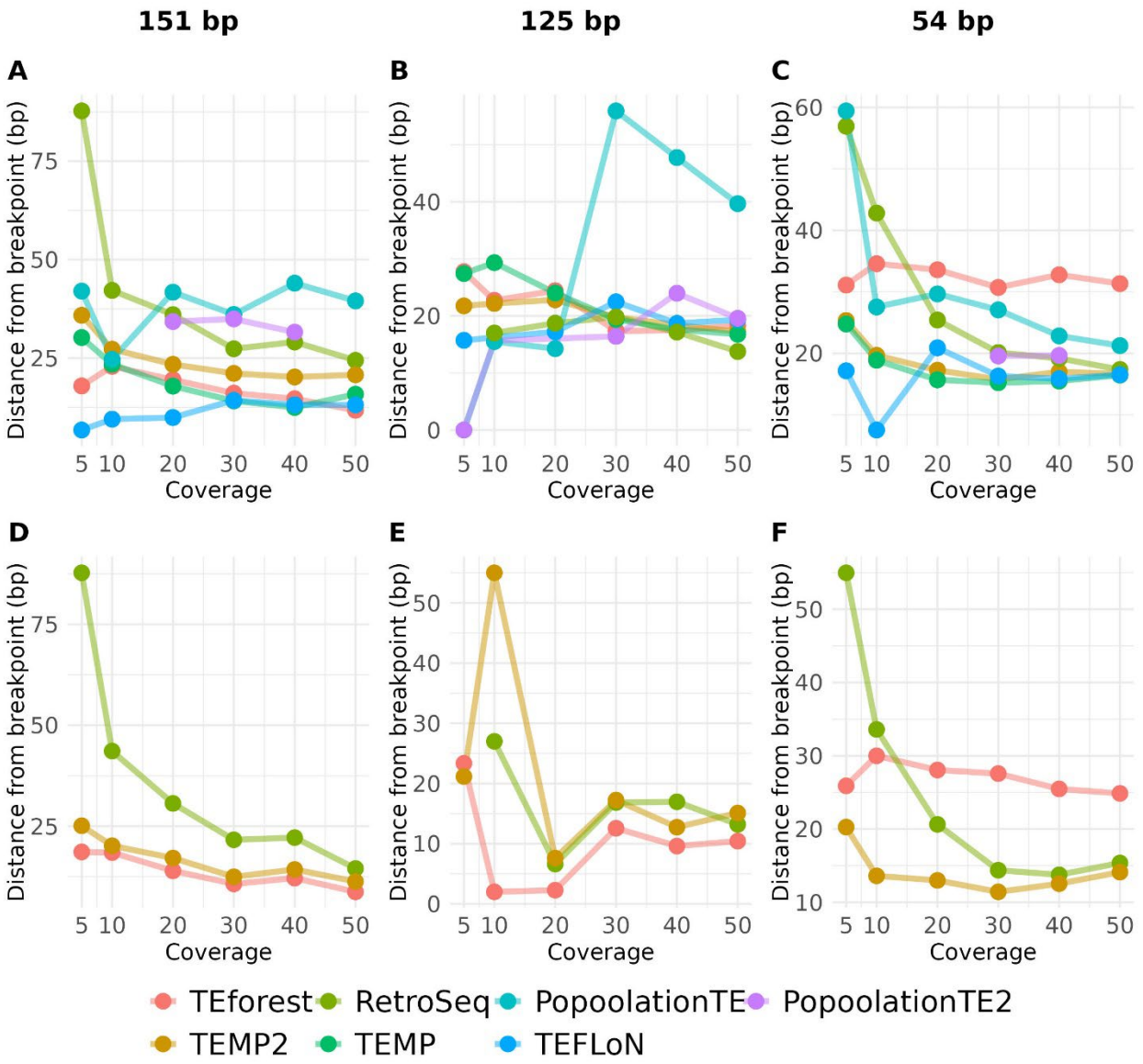


Figure S5: Breakpoint accuracy of TEforest and other short-read TE callers for true positive TE insertions annotated from long-read assemblies of *D. melanogaster* strains. The mean breakpoint error of TE callers for (**A-C**) all true positive calls or (**D-F**) true positive calls shared by TEforest, 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.

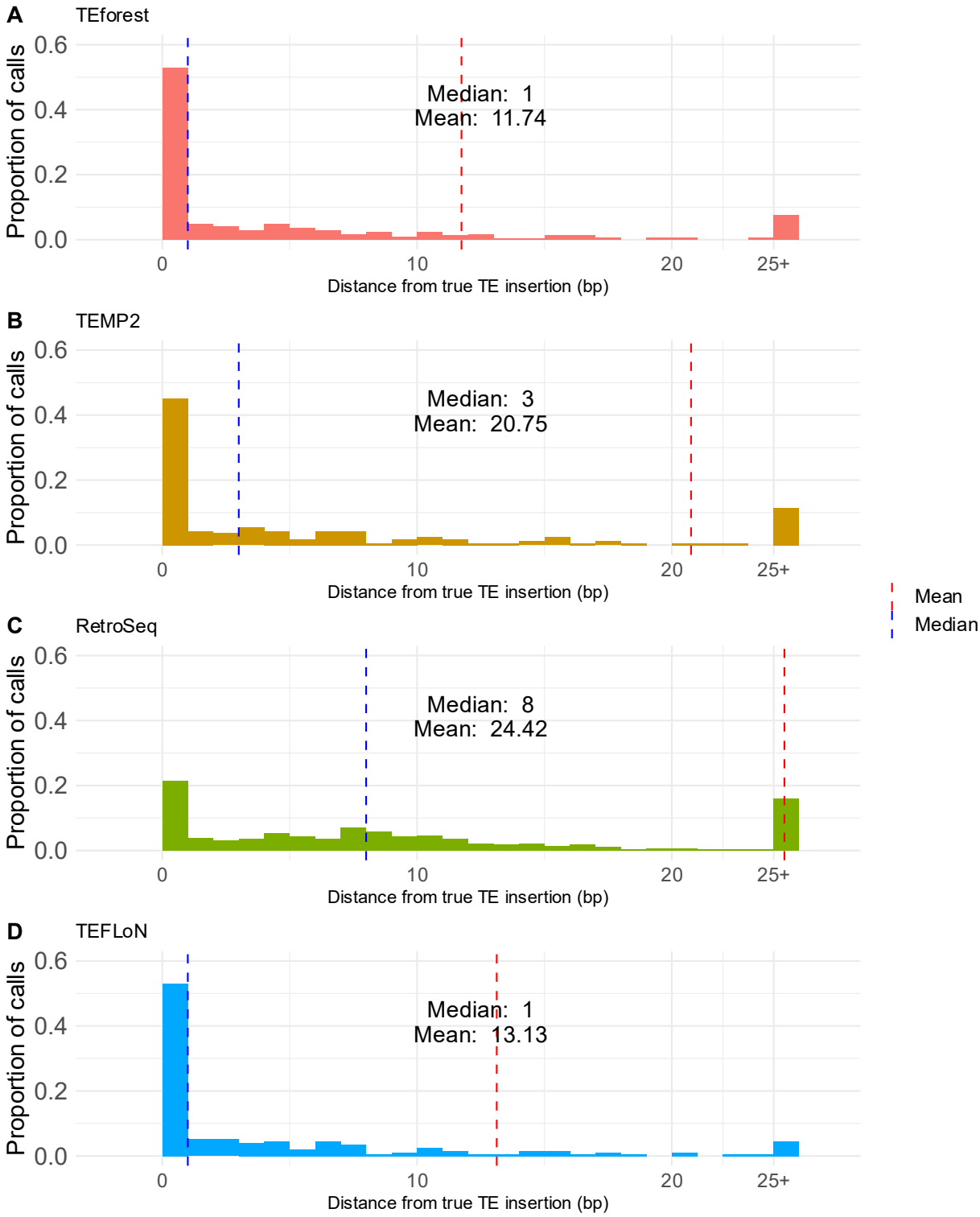


Figure S6: The normalized distribution of breakpoint accuracy for all true positive calls for the 151 bp dataset from **(A)** TEforest, **(B)** TEMP2, **(C)** RetroSeq, and **(D)** TEFLoN was quantified by finding the distance between the center of true positive breakpoint ranges predicted by the TE callers and the breakpoint in the truth dataset. The last bin represents the proportion of calls >=25 bp away from the annotated breakpoint.

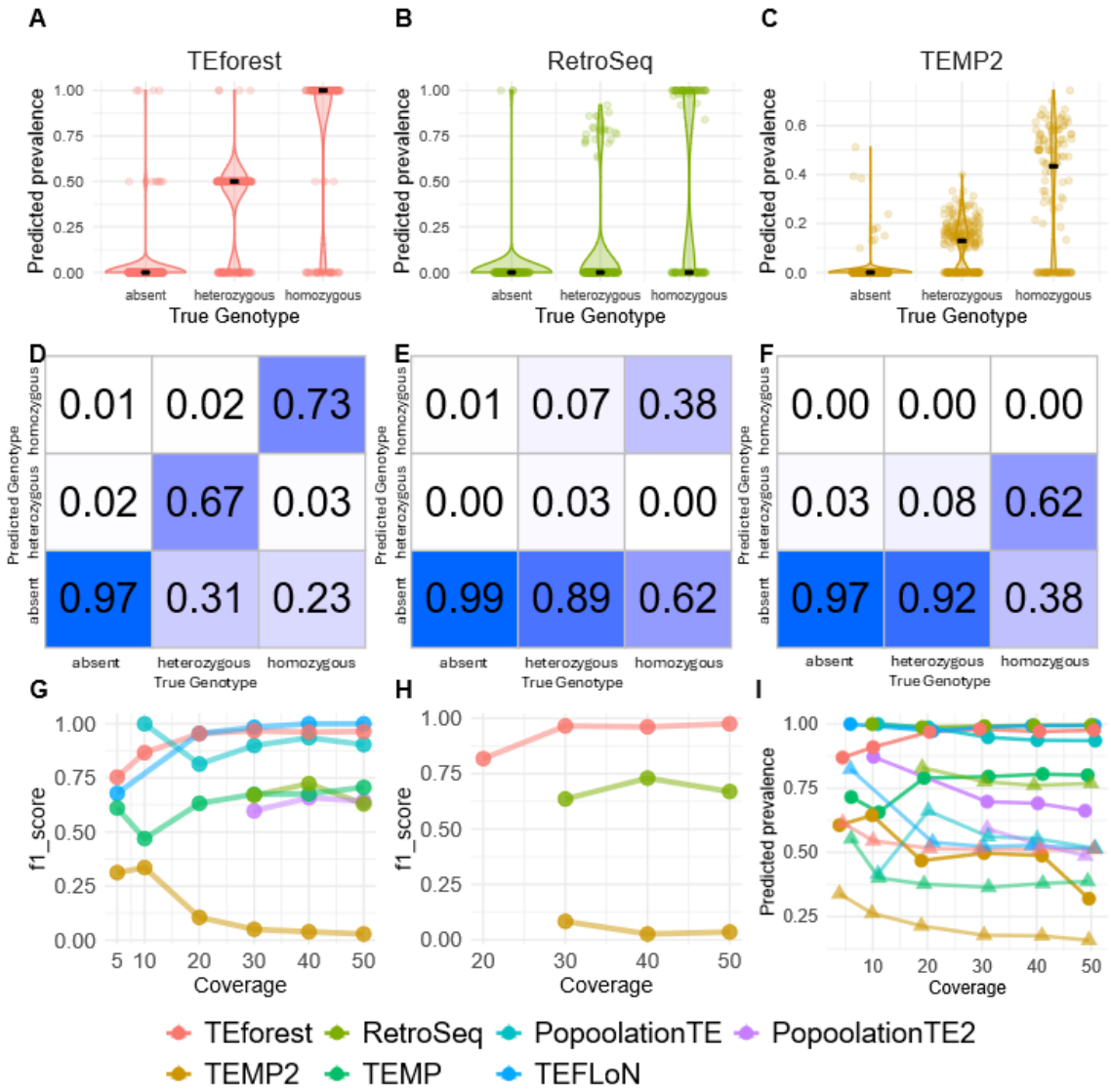


Figure S7: The performance of TEforest compared to other short-read TE callers at genotyping nonreference TE insertions in the 125 bp dataset. The mean insert size of this dataset was ~208 bp. Prevalence predictions for insertions that were homozygous, heterozygous, or absent were calculated for (**A**) TEforest, (**B**) RetroSeq, and (**C**) TEMP2. The absent class consisted of a random sample (equal to the number of true positives in the dataset) of true-negative candidate regions from the TEforest candidate regions together with 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-F**) Confusion matrices showing genotyping accuracy, where heterozygous predictions are defined as frequency predictions between 0.25 and 0.75 and homozygous predictions are above 0.75. For panels **A-F**, the methods were evaluated at a sequencing coverage of 30X. The genotyping F1 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 panel **I**, calculated using only the prevalences of true positive predictions of each caller (false positives are excluded).

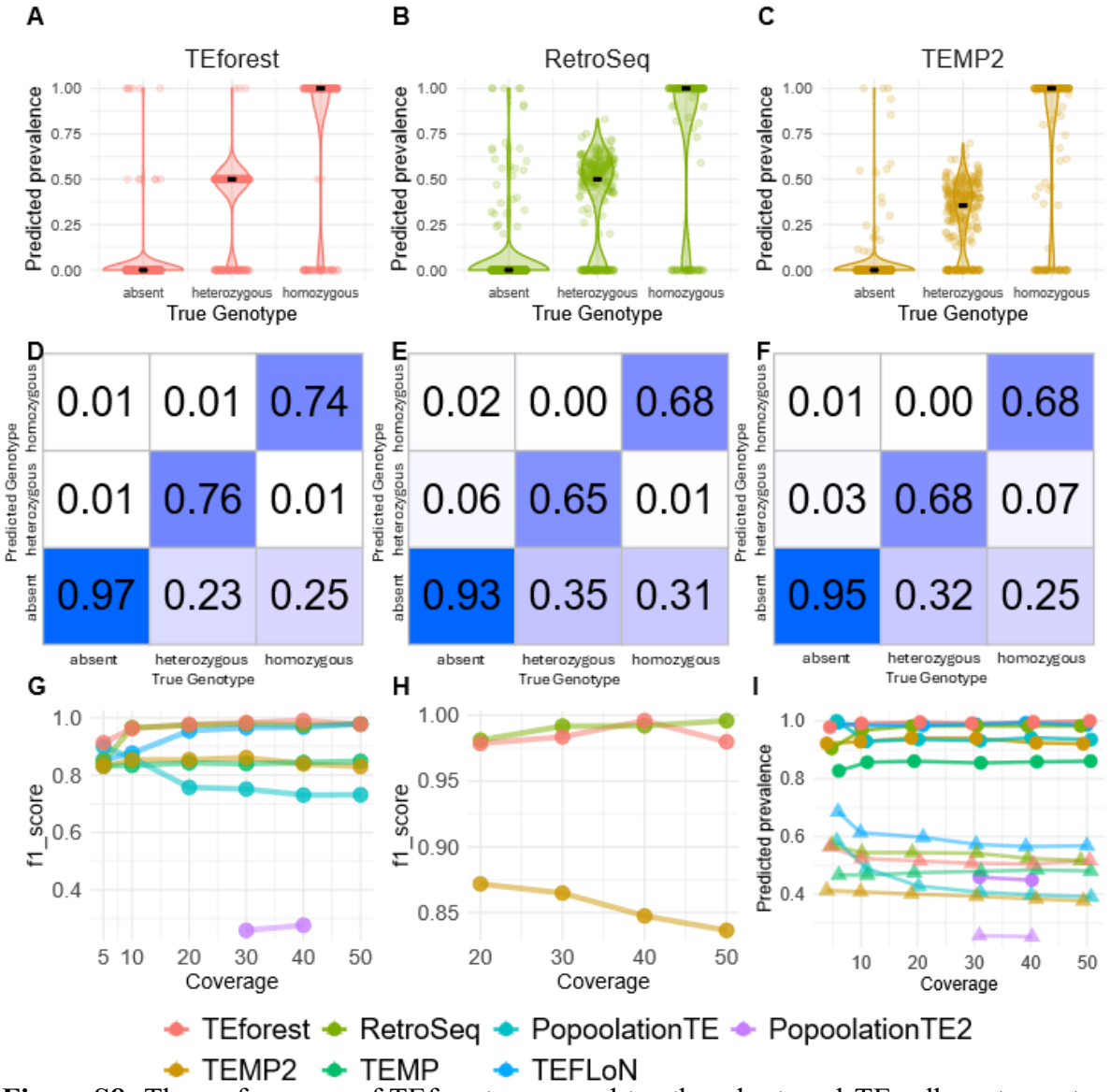


Figure S8: The performance of TEforest compared to other short-read TE callers at genotyping nonreference TE insertions in the 54 bp dataset. The mean insert size of this dataset was ~287 bp. Prevalence predictions for insertions that were homozygous, heterozygous, or absent were calculated for (**A**) TEforest, (**B**) RetroSeq, and (**C**) TEMP2. The absent class was defined in the same manner as Figure S7, as were the predicted prevalences for true negatives and false positives. (**D-F**) Confusion matrices showing genotyping accuracy, with prevalence predictions converted to discrete genotypes in the same manner as Figure S7. For panels **A-F**, methods evaluated at 30X coverage. The genotyping F1 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).

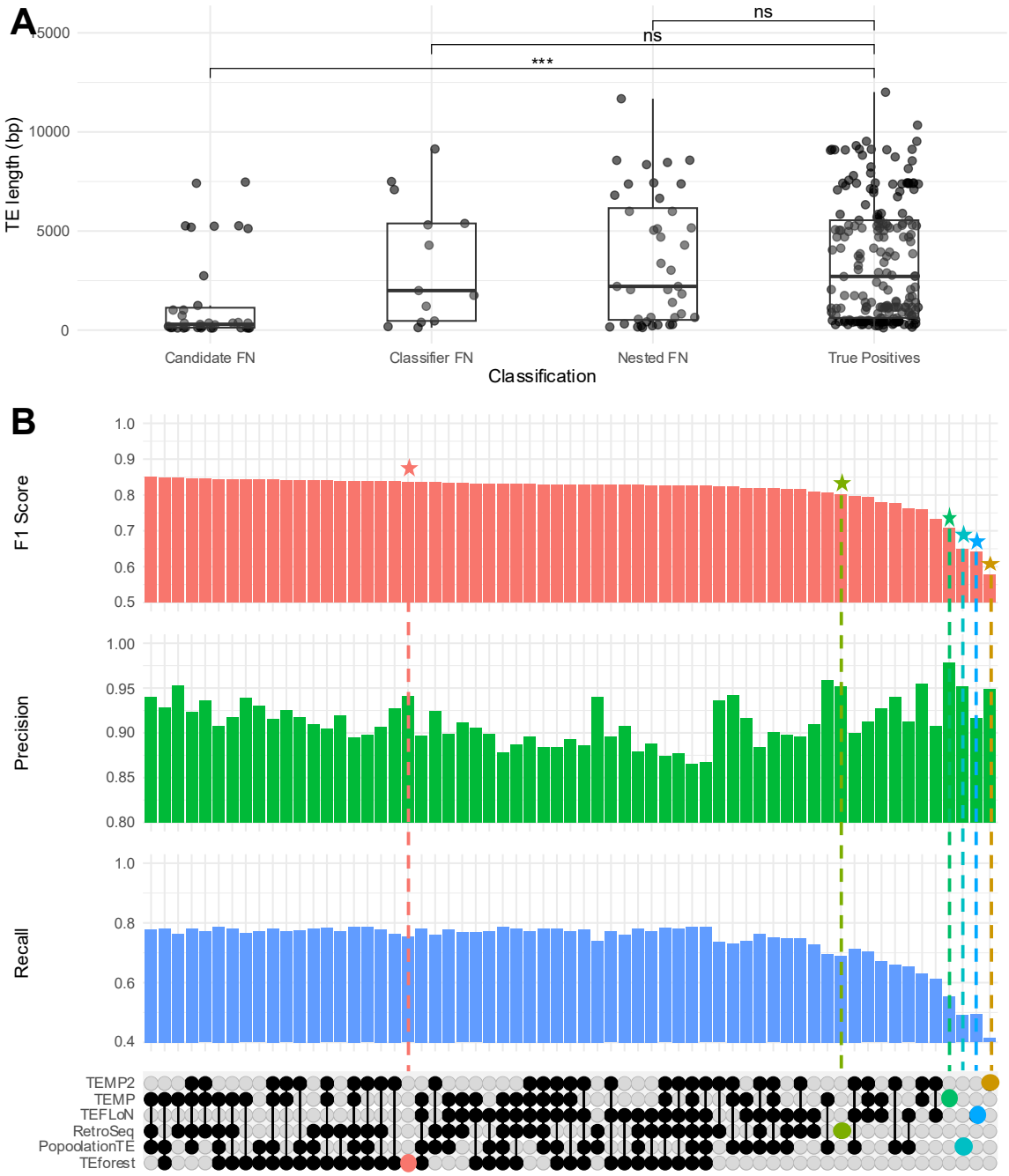


Figure S9: (A) Lengths of non-reference TEs insertions that were in the genome with 151 bp reads with ~436 bp insert sizes, divided into classes depending on whether they were unnested and not detected in the candidate region identification stage (Candidate FN), unnested and mislabeled as absences by the LightGBM classifier (Classifier FN), nested and lost in either the candidate regions or classifier step (Nested FN), or successfully detected by TEforest (True Positives). Overhead bars represent the results of pairwise Wilcox rank-sum tests. **(B)** Upset plot representing the results of combining different TE callers on the F_1 Score, Precision, and Recall for detecting non-reference TEs. Colored stars mark results when each caller is run alone (not combined).

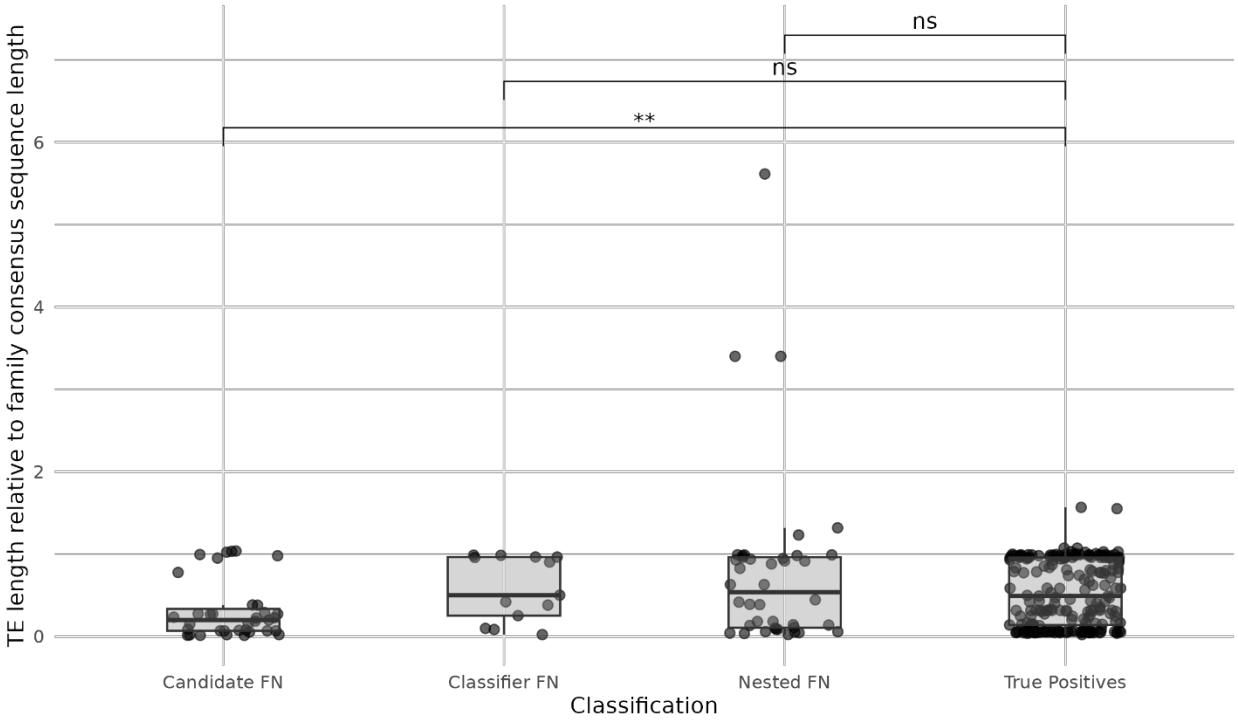
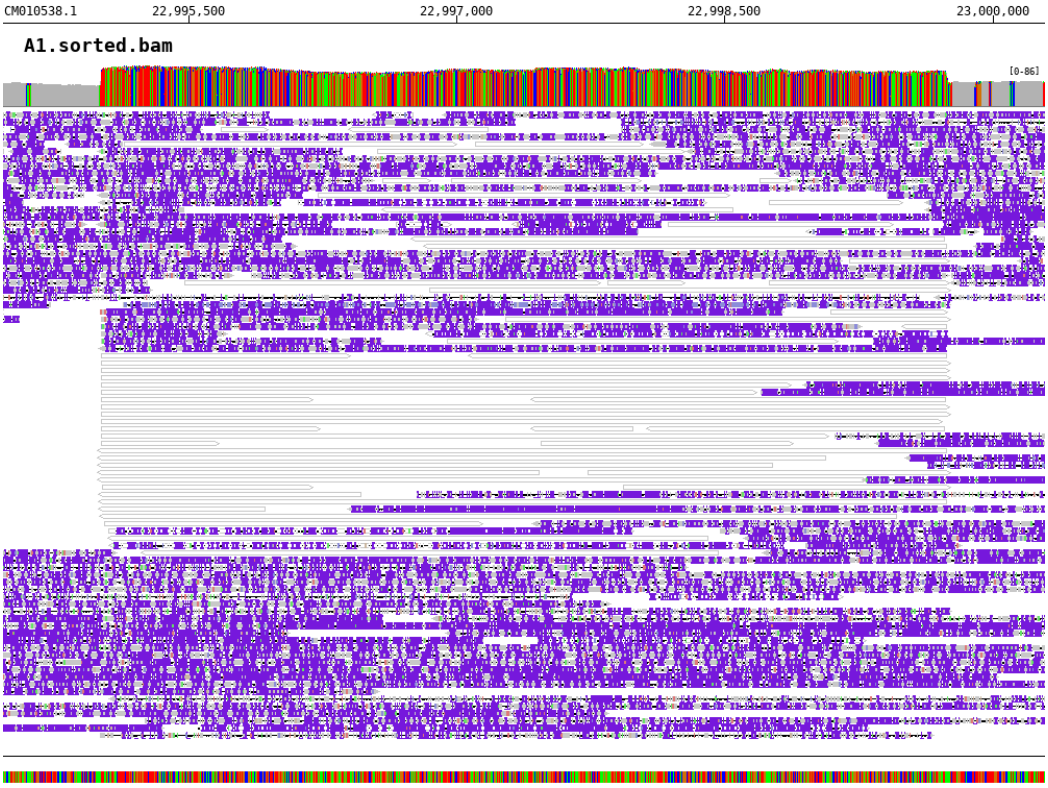


Figure S10: Ratio of TE sequence length to the consensus length of that TE family of non-reference TEs insertions that were in the genome with 151 bp with ~436 bp insert sizes, divided into classes depending on whether they were unnested and not detected in the candidate region identification stage (Candidate FN), unnested and mislabeled as absences by the random forest classifier (Classifier FN), nested and lost in either the candidate regions or classifier step (Nested FN), or successfully detected by TEforest (True Positives). Overhead bars represent the results of pairwise Wilcox rank-sum tests.


CM010538.1
22,995,500
22,997,000
22,998,500
23,000,000
A1.sorted.bam
[0-86]

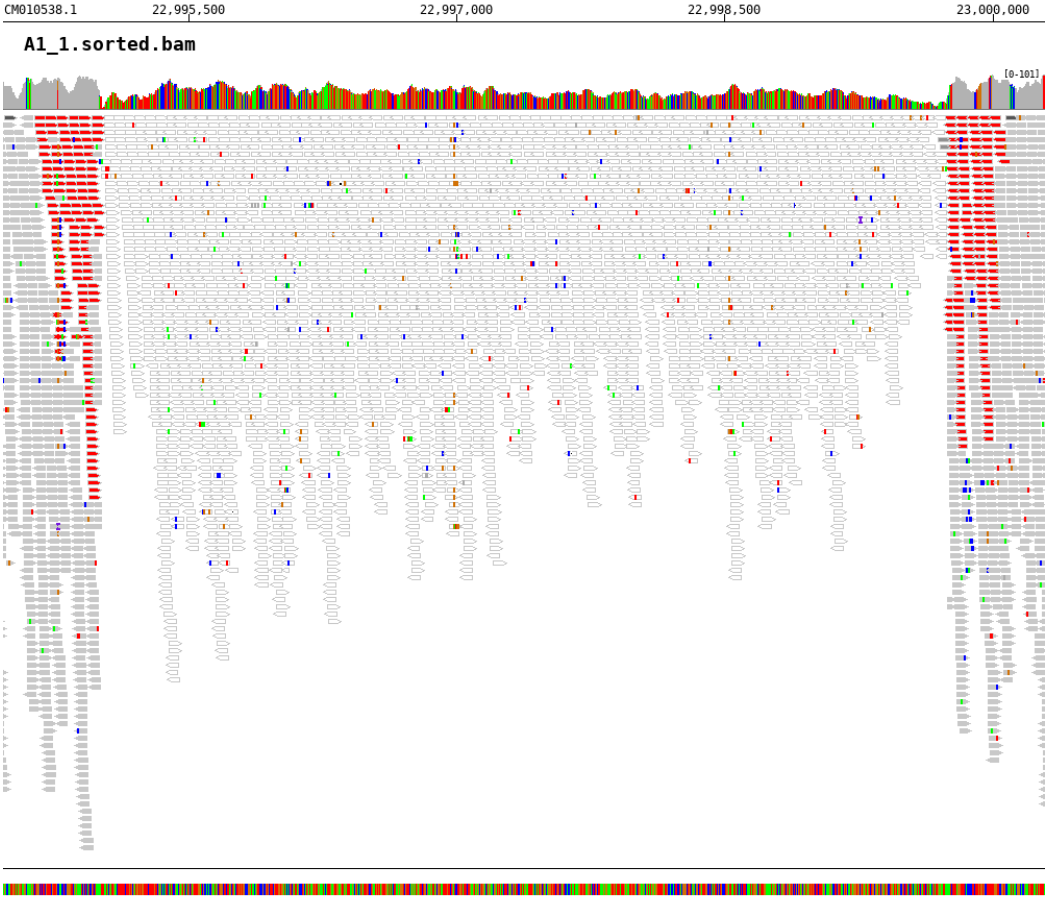

CM010538.1
22,995,500
22,997,000
22,998,500
23,000,000
A1_1.sorted.bam
[0-101]

Figure S11: Example alignment 1 of short and long reads from the sequenced sample A1 to the A1 long-read assembly. This IGV image represents evidence for a Doc insertion at 3R:22494133-22494133 (reference genome coordinates), which was not detected by TEforest. The image displays the region where the TE is annotated in the A1 genome +/- 500 bp. Short reads show split alignments and discordant read-pair evidence at the breakpoint. Red read pairs in IGV denote mates mapping farther apart than expected based on the insert-size distribution; here, because the mates align on opposite sides of the annotated TE, this pattern is consistent with a deletion across the annotated insertion interval. Gray reads represent non-discordant alignments, whereas lighter/transparent white reads indicate low mapping quality, due to reads mapping to multiple locations in the genome (in this case, likely to other TE copies). In the long-read panel, several non-repetitive long reads span the breakpoint interval continuously, supporting the TE insertion.

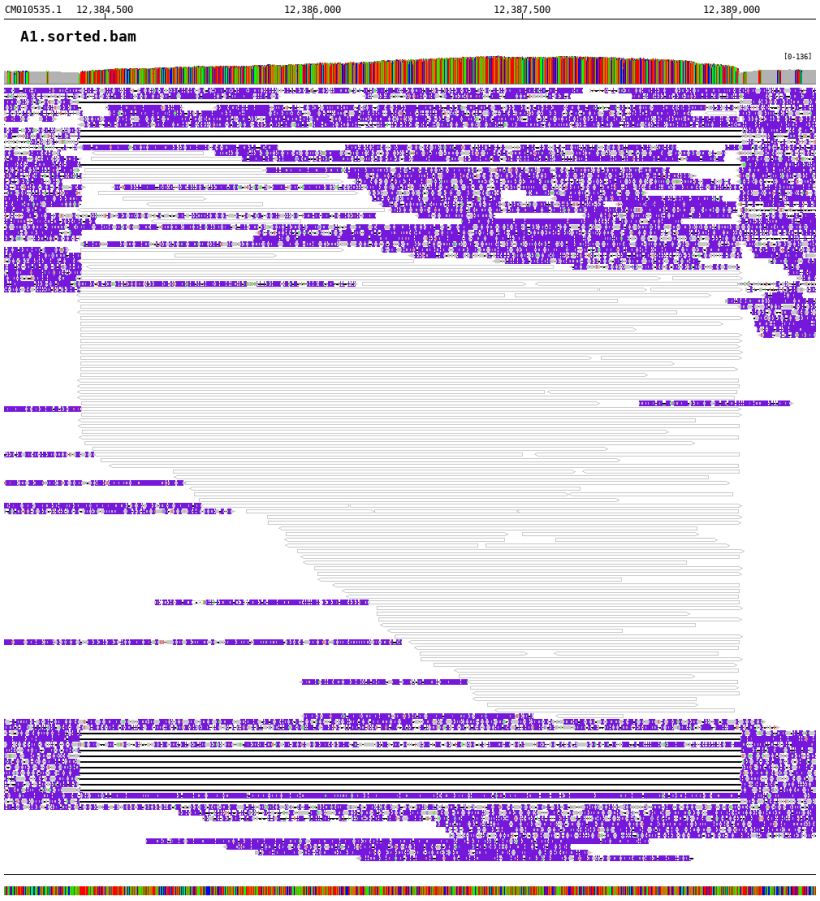
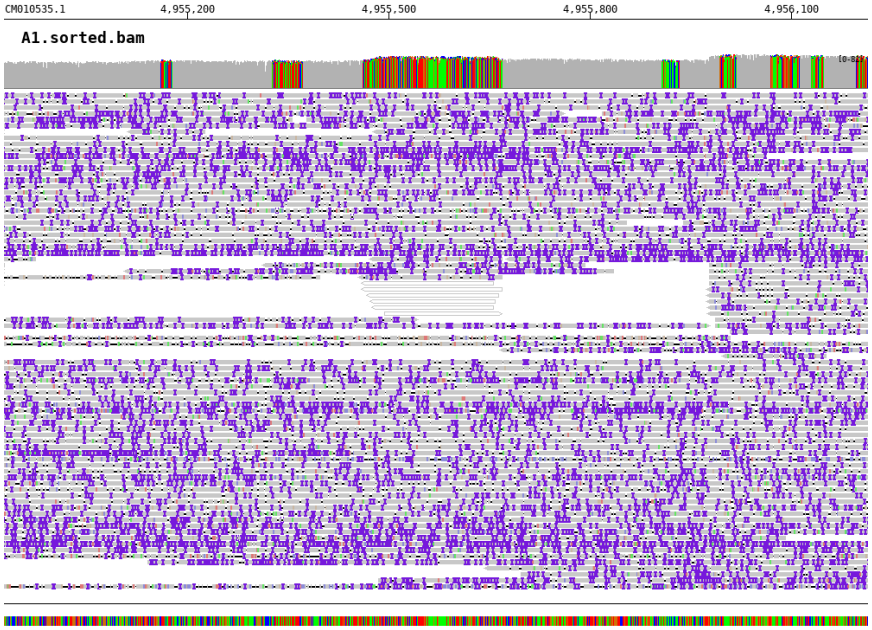

CM010535.1 12,384,500
12,386,000
12,387,500
12,389,000
A1.sorted.bam
[0-136]
CM010535.1 12,384,500
12,386,000
12,387,500
12,389,000
A1_1.sorted.bam
[0-127]

Figure S12: Example alignment 2 of short and long reads from the sequenced sample A1 to the A1 long read assembly. This IGV image represents evidence for a Doc insertion at 2L:12359453-12359453 (reference genome coordinates) which was not detected by TEforest. The image displays the region where the TE is annotated in the A1 genome +/- 500 bp. Short reads are split or map discordantly across the breakpoints, and long reads show evidence for a polymorphic insertion in the sequenced sample. Short reads show split alignments and discordant read-pair evidence at the breakpoint. Red read pairs in IGV denote mates mapping farther apart than expected based on the insert-size distribution; here, because the mates align on opposite sides of the annotated TE, this pattern is consistent with a deletion across the annotated insertion interval. Gray reads represent non-discordant alignments, whereas lighter/transparent white reads indicate low mapping quality, due to reads mapping to multiple locations in the genome (in this case, likely to other TE copies). In the long-read panel, some reads map continuously across the interval, whereas others contain a long black gap in the alignment, indicating a deletion relative to the assembly across the annotated TE interval. The presence of both spanning reads and reads with an internal deletion is consistent with a polymorphic insertion in the sequenced sample.


CM010535.1
4,955,200
4,955,500
4,955,800
4,956,100
A1.sorted.bam
[0-166]


CM010535.1
4,955,200
4,955,500
4,955,800
4,956,100
A1_1.sorted.bam
[0-166]

Figure S13: Example alignment 3 of short and long reads from the sequenced sample A1 to the A1 long read assembly. This IGV image represents evidence for a pogo insertion at 2L:4951112-4951112 (reference genome coordinates) which was successfully detected by TEforest. The image displays the region where the TE is annotated in the A1 genome +/- 500 bp. Short and long reads map without disruption across the breakpoints, with no discordantly mapping short reads or abnormally split long reads.

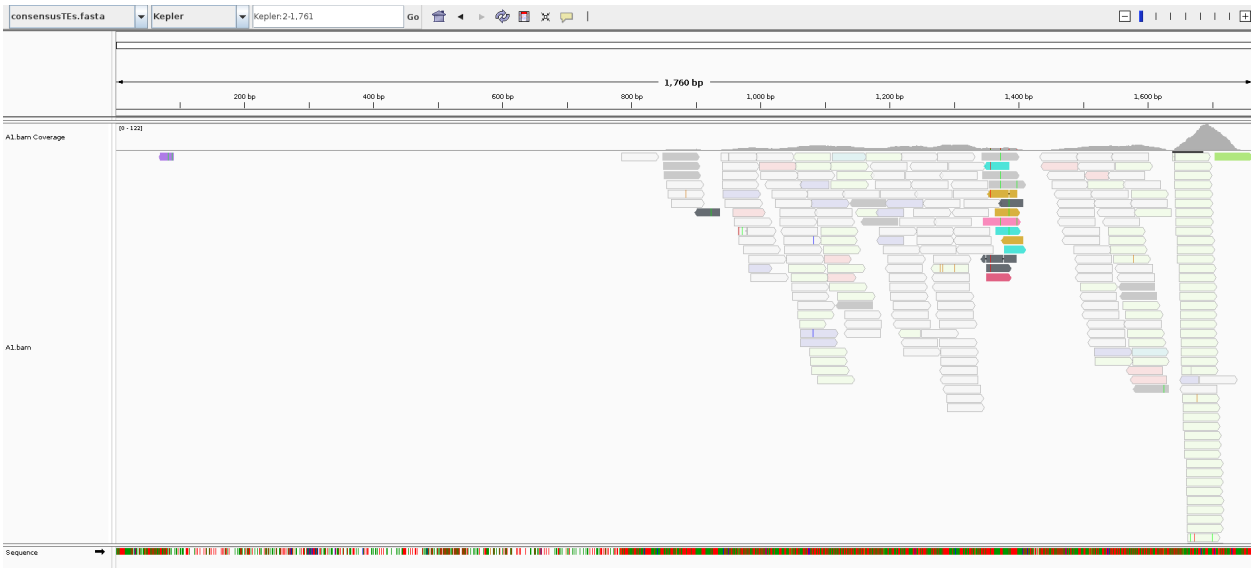


Figure S14: IGV image of the alignment of all sequenced reads from the DSPR genome A1 (54 bp reads; King et al. 2012) to the Kepler TE consensus sequence.

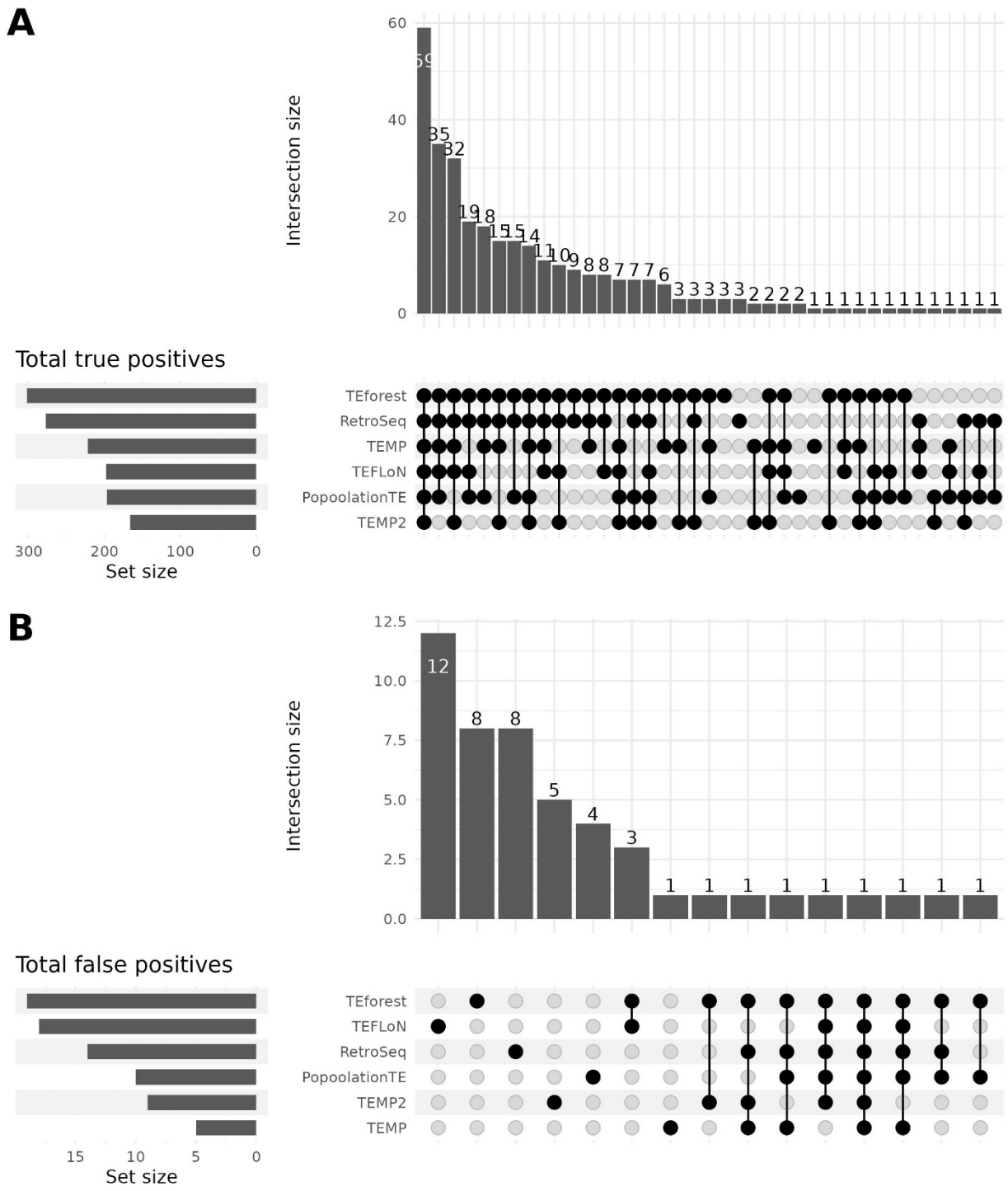


Figure S15: (A) Upset plot representing the number of unique true positives for different combinations of TE callers, for the dataset with 151 bp reads at 50X coverage. True positives were defined as a prediction made within 500 bp of an annotation within the truth dataset. Nested TEs from the same family were only counted once. **(B)** Upset plot representing the number of unique false positives for different combinations of TE callers.

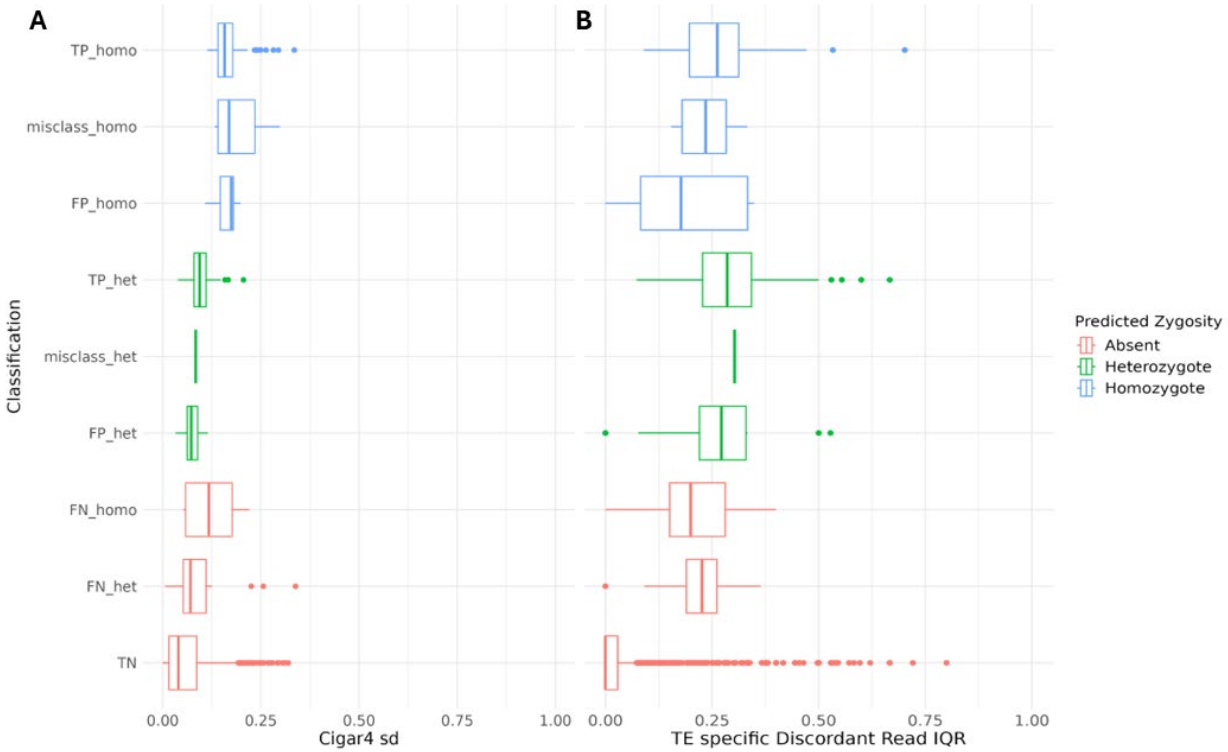


Figure S16: Distributions of two high-importance TEforest features—**(A)** Cigar4 (split/soft-clipped read signal) standard deviation and **(B)** TE-specific discordant read signal interquartile range (IQR)—shown across prediction/ground-truth categories that summarize the relationship between true and predicted zygosity. In TEforest, each feature is first computed per base across the full candidate region, then summarized into a single value per region; thus, Cigar4 SD quantifies how spatially variable the split-read/soft-clipping signal is across the candidate region. Likewise, the TE-specific discordant read IQR quantifies the spread of the per-base discordant-pair signal computed from the TE-family–specific BAM (i.e., using only TE-associated read pairs), capturing how concentrated versus diffuse discordant evidence is across the region. Categories include true positives (TP_homo, TP_het), true negatives (TN), candidate regions without true insertions that were predicted as insertions (FP_homo, FP_het), false negatives (FN_homo, FN_het), and zygosity misclassifications (misclass_homo, misclass_het; e.g., true homozygotes predicted as heterozygotes or vice versa). Boxplots show the distribution of the per-region feature summary values for each category, with outliers plotted as points.

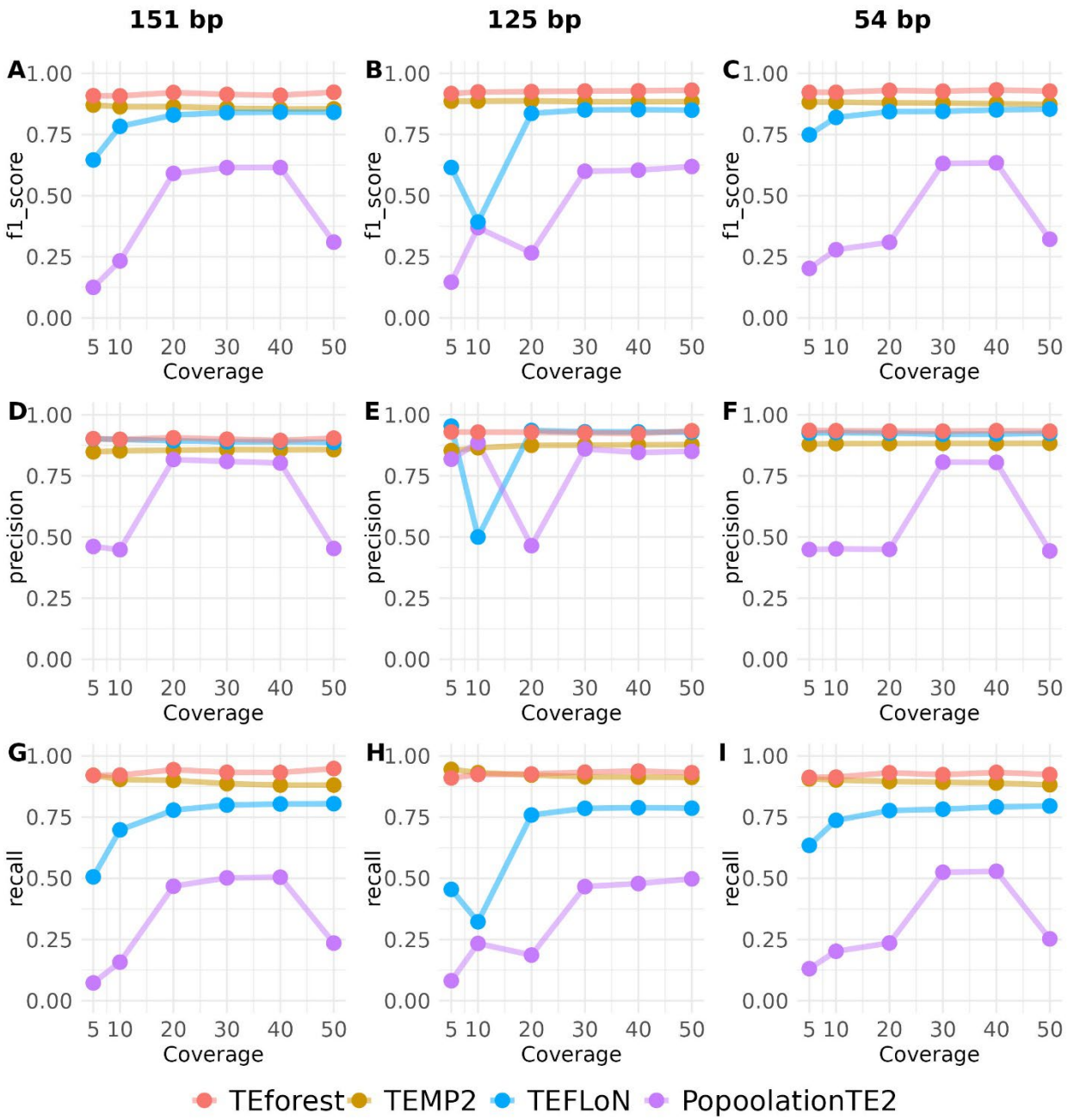


Figure S17: The performance of TEforest compared to other short-read TE callers at detecting reference TE insertions annotated by long read assemblies of *D. melanogaster* strains. Short reads used for detection of TEs were (**A**) 151 bp with ~208 bp insert sizes. Performance was quantified with F1 scores, precision, and recall, (**B**) 125 bp with ~208 bp insert sizes, and (**C**) 54 bp with ~436 bp insert sizes.

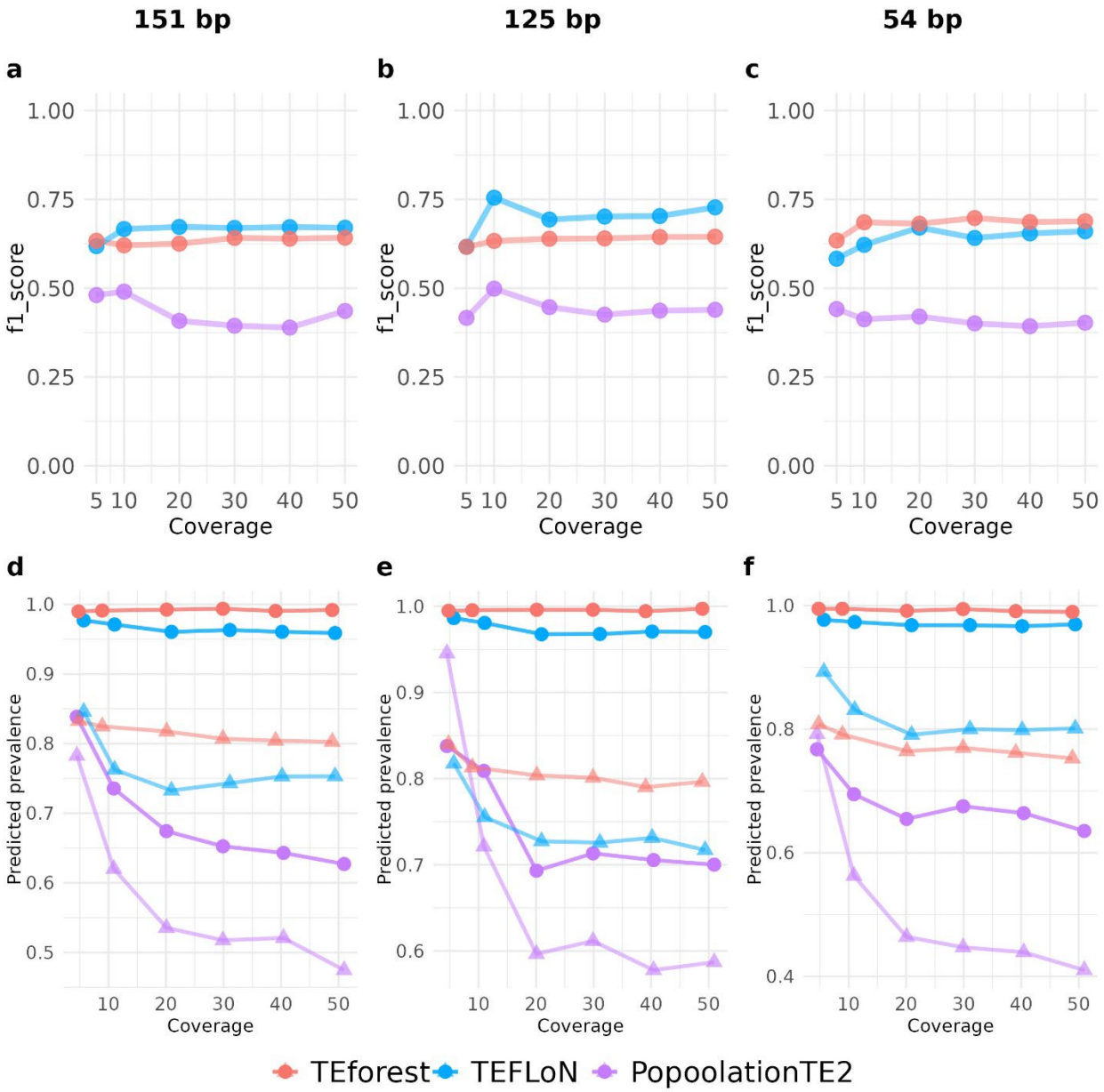


Figure S18: The performance of TEforest compared to other short-read TE callers at genotyping reference TE insertions. (**a-c**) The genotyping F1 score for all true positive calls was quantified using true positive predictions of each caller. The mean for the predicted prevalence of true positive homozygous (circles) and heterozygous (triangles) insertions are shown in panels **d-f**, calculated using only the prevalences of true positive predictions of each caller (false positives are excluded).

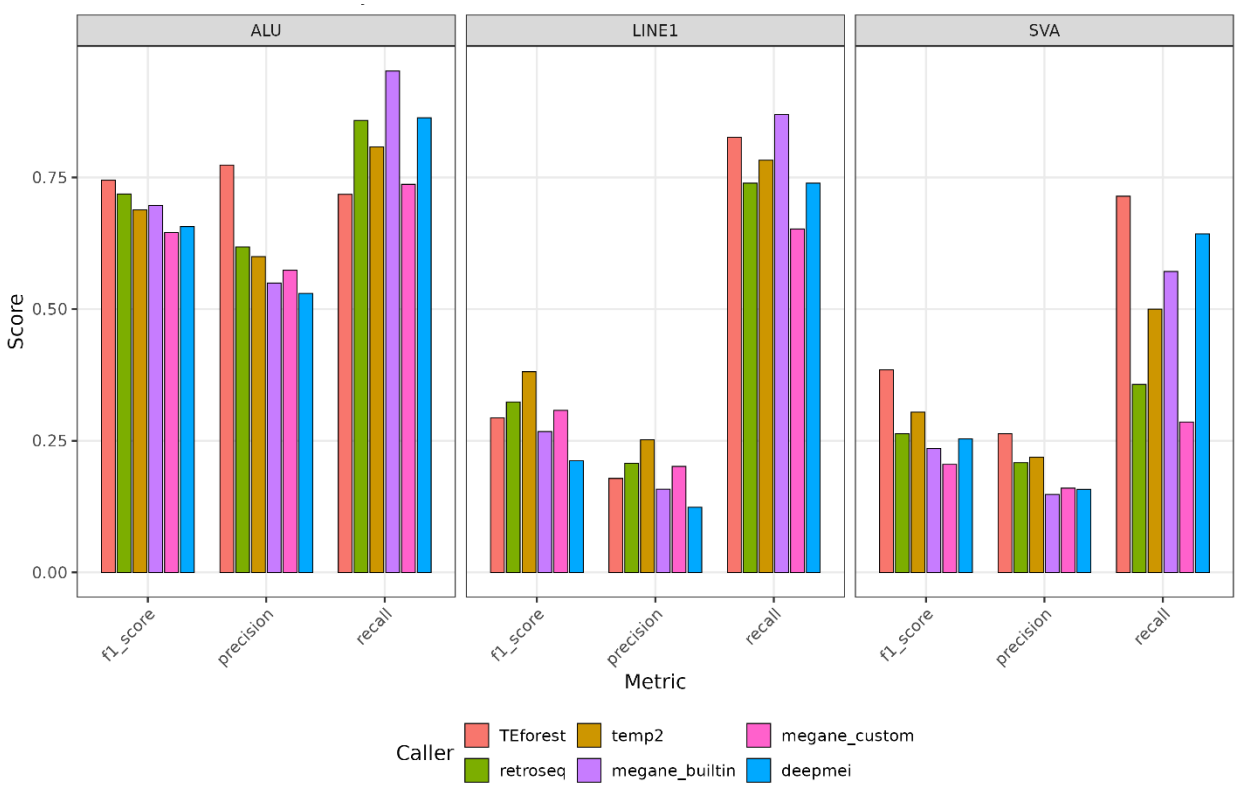


Figure S19: The performance of TEforest on human genome NA12878 using 893 validated germline insertions (Stewart et al. 2011; Yu et al. 2021). The F1 score, precision, and recall for detecting non-reference *Alu*, *LINE1*, and *SVA* insertions. Benchmarking was performed using a 30X coverage whole-genome sequencing dataset aligned to the hg38 reference genome. TEforest was applied using the non-reference model trained on *Drosophila melanogaster* synthetic heterozygotes without species-specific retraining. MEGAnE was benchmarked using both its built-in and the TE consensus library used for the other TE callers.

Table S1: Description of features used to describe the alignment of short reads to the reference genome. Each feature is first calculated for each base pair of each read and summed for each basepair before summary statistics (mean, median, sd, and IQR) are calculated for the feature across the region. Features are calculated for each specific candidate region using the full BAM alignment as well as a BAM file containing only reads or their mate aligning to the TE of interest.

Feature	Description (T=1, F=0)
Cigar1	Base matches reference.
Cigar2	Deletion has occurred in the read.
Cigar3	Insertion in one or more bases in the read, preceding this base.
Cigar4	Soft clip in the read at this position.
Cigar5	Hard clip in the read at this position.
Paired	Read has a pair.
Proper Pair	Both reads in the pair are aligned with correct orientation and distance.
Is Read1 Unmapped	First read in the pair is unmapped.
Is Read2 Unmapped	Second read in the pair is unmapped.
Is Read1 Rev Comp	First read in the pair is reverse complemented.
Is Read2 Rev Comp	Second read in the pair is reverse complemented.
Is First Read	Is the first read in the pair.
Is Second Read	Is the second read in the pair.
Split	Read is part of a split alignment.
Long Insert	Template length of this read is in the top 1% of template lengths in this region.
Short Insert	Template length of this read is in the bottom 2% of template lengths in this region.
Parallel Read	Both read and its mate align in the same direction.
Everted Read	Both read and its mate align in opposite directions.
Discordant Read	Mate of the read is not mapped to the same chromosome.
Template Length	Distance between the two ends in the read pair on the reference genome.
Quality	Quality score of the alignment at this position.

Table S2: Numbers of homozygous, heterozygous, and false positive candidate regions used as input for training and testing the non-reference insertion model (50X coverage).

Training/testing	Read length (bp)	Insert size (bp)	# homozygotes	# heterozygotes	# candidate regions without true positives
Training	54	287	199	244	2867
Training	125	208	193	264	16947
Training	151	436	200	304	3684
Testing	54	287	120	218	955
Testing	125	208	108	185	6616
Testing	151	436	106	217	1420

Table S3: Numbers of candidate regions with a TE present or absent used as input for training and testing the reference insertion model (50X coverage).

Training/testing	Read length (bp)	Insert size (bp)	# homozygotes	# heterozygotes	# candidate regions without true positives
Training	54	287	2031	260	595
Training	125	208	2065	247	583
Training	151	436	1961	268	655
Testing	54	287	888	207	301
Testing	125	208	908	204	302
Testing	151	436	808	269	327

Table S4: Numbers of true positive and false negative TEs, sorted by whether they are nested or not. Since short-read TE callers cannot distinguish multiple copies of the same TE family at one site, true positives and false negatives represent the number of nested TE groups rather than the total copy number.

	Non-nested TE	TE nested with different family	Group of TEs nested with same family
True positives	303	9	38
False negatives	48	32	8
False negatives lost during candidate region detection	35	22	4

Table S5: TE family-specific numbers of false negative and true positive calls, along with the proportion of the false negatives that consists of that TE family.

TE	Total False Negatives	Candidate Region False Negatives	Fraction of all False Negatives	Total True positives
INE_1	11	10	0.13	0
Gypsy_2_Dsim	7	7	0.08	0
FB4	6	1	0.07	14
NOF	5	4	0.06	0
1731	4	4	0.05	0
blood	4	2	0.05	5
F_element	4	1	0.05	14
Kepler	4	4	0.05	0
BS	3	2	0.03	3
Copia	3	2	0.03	13
roo	3	2	0.03	46
1360	2	0	0.02	0
297	2	1	0.02	15
Copia2	2	2	0.02	0
gypsy12	2	1	0.02	1
HMS_Beagle	2	1	0.02	2
HMS_Beagle2	2	1	0.02	0
Max_element	2	2	0.02	1
P_element	2	2	0.02	31
17_6	1	0	0.01	2
accord	1	1	0.01	8
Blastopia	1	1	0.01	8
Circe	1	1	0.01	0
G4	1	1	0.01	0
G6	1	0	0.01	1
GATE	1	1	0.01	0
Gypsy_24_Dya	1	0	0.01	4
gypsy8	1	1	0.01	0
I_element	1	0	0.01	7
Invader3	1	1	0.01	0
mdg1	1	1	0.01	9
NewFam14	1	1	0.01	0
Quasimodo	1	0	0.01	3
Rt1b	1	1	0.01	4
S_element	1	0	0.01	1
Stalker2	1	1	0.01	1
Transpac	1	1	0.01	6
3S18	0	0	0	4
412	0	0	0	9

Bari1	0	0	0	2
BS2	0	0	0	5
Burdock	0	0	0	4
diver	0	0	0	1
Doc	0	0	0	9
Doc6	0	0	0	5
G_element	0	0	0	1
gypsy1	0	0	0	3
gypsy6	0	0	0	1
H	0	0	0	25
hopper	0	0	0	1
Idefix	0	0	0	1
Ivk	0	0	0	5
jockey	0	0	0	44
mdg3	0	0	0	3
Nomad	0	0	0	4
pogo	0	0	0	17
Rt1a	0	0	0	1
Stalker4	0	0	0	1
Tirant	0	0	0	5

Table S6: Summary of manual validation of TE insertions in the truth dataset. Short- and long-read sequences from DSPR genome A1 were aligned to the Chakraborty et al. (2019) assembly, and the regions surrounding 30 annotated insertions (15 detected by TEforest and 15 missed by TEforest) were visually inspected in BamSnap using the insertion coordinates reported for this assembly in Rech et al. (2019). Each site was assigned a manual validation code based on read-level evidence: T (supported/true insertion) indicates clear support for the annotated TE at that locus (e.g., TE-associated discordant/split-read signatures in short reads and/or long reads consistent with an insertion at the breakpoint); F (no support/likely absent) indicates evidence for the *absence* of the annotated TE, such as read pairs spanning the locus without TE-associated signatures (including spanning pairs with increased insert size across the breakpoint and/or split reads consistent with the non-insertion allele in short reads) and long reads aligning continuously across the locus without an inserted TE; P (polymorphic/mixed) indicates that the long-read data contain both TE-present and TE-absent allele evidence at the same locus (e.g., some long reads support a clean deletion/non-insertion while others support TE presence), consistent with heterozygosity or within-line polymorphism; and I (inconclusive) indicates that presence/absence could not be confidently assessed due to ambiguous mapping, typically when mapping quality was 0 across one or both breakpoint regions.

TEforest prediction	Insertion ID (chr_start_stop_TE)	Short-read support	Long-read support
True positive	2L_4951112_4951112_pogo	T	T
True positive	2L_10181634_10181634_roo	T	T
True positive	2L_12845323_12845325_Transpac	T	T
True positive	2L_13154224_13154226_roo	T	T
True positive	2R_9271128_9271128_roo	T	T
True positive	2R_17979247_17979249_Blastopia	T	T
True positive	3L_14781802_14781802_roo	T	T
True positive	3L_18320212_18320212_Doc	T	T
True positive	3R_9830000_9830000_roo	T	T
True positive	3R_13253984_13253984_Rt1b	T	T
True positive	3R_17612152_17612152_roo	T	T
True positive	3R_19493130_19493130_roo	T	T
True positive	3R_28477291_28477291_roo	T	T
True positive	X_3212704_3212704_Nomad	T	T
True positive	X_4249096_4249096_blood	T	T
False negative	2L_12359453_12359453_Doc	F	P
False negative	2L_16859593_16859595_BS2	T	T
False negative	3L_11581060_11581060_jockey	T	T
False negative	3R_12498613_12498616_Micropia	T	T
False negative	3R_22494133_22494133_Doc	F	T
False negative	X_7081917_7081917_Blastopia	I	I
False negative	3R_15232056_15232056_pogo	T	T
False negative	2R_6006492_6006492_Doc	F	T
False negative	2L_15902599_15902599_Ivk	T	T
False negative	2R_9138738_9138738_pogo	T	T

False negative	2R_10022986_10022993_NOF	T	T
False negative	2L_4891846_4891846_Doc	F	P
False negative	2R_6372367_6372369_roo	T	T
False negative	X_8348807_8348807_roo	T	T
False negative	2R_6404778_6404781_1360	T	T

Table S7: TE families with the highest percentages of IUPAC bases, along with their copy number in the euchromatin of the *dm6* reference genome.

TE family	Consensus sequence length	iupac bases count	iupac bases %	Number of copies in reference sequence
Kepler	1760	378	21.477	42
INE-1	683	64	9.37	561
gypsy10	7396	375	5.07	1
diver2	5242	228	4.349	2
H	3009	114	3.789	40
HeT-A	6659	218	3.274	3
FB4	3490	111	3.181	23
Tc3	4007	91	2.271	4
G3	4611	53	1.149	0
Fw2	3967	45	1.134	0

Table S8: Feature importance metrics for the top non-reference model features. The table lists the top ten features ranked by Total Gain. This metric quantifies the total reduction in the model's objective function (loss) contributed by splits on a given feature across all trees in the LightGBM model.

Feature	Total Gain
TE specific Discordant Read sd	77831.24
Cigar4 sd	70616.584
Proper Pair sd	28262.405
Cigar4 mean	10713.418
TE specific Discordant Read IQR	9681.501
Proper Pair mean	6732.586
TE specific Proper Pair mean	3405.193
Short Insert sd	3300.483
TE specific Insert Size median	3249.76
Insert Size median	3212.889