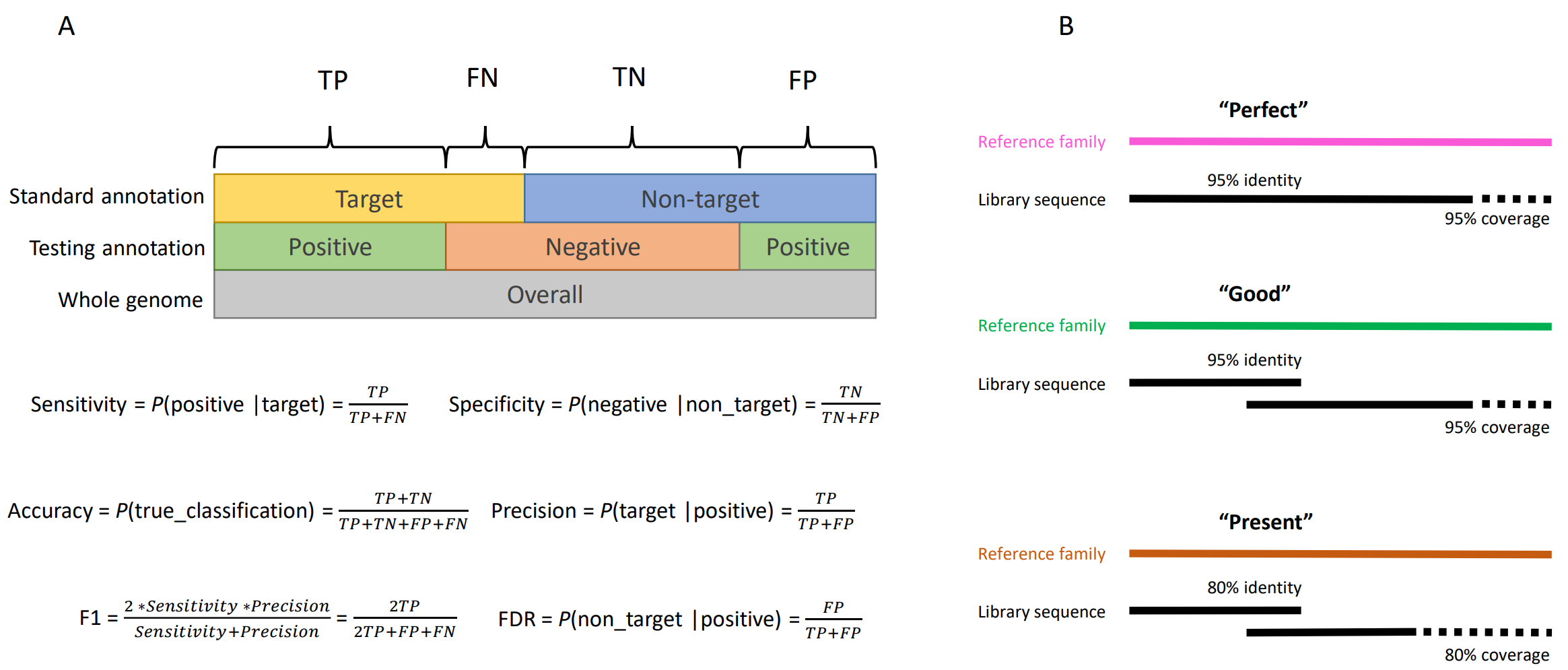
## Additional File 1. Benchmarking Methods

### Setting up benchmarking methods for TE library evaluation

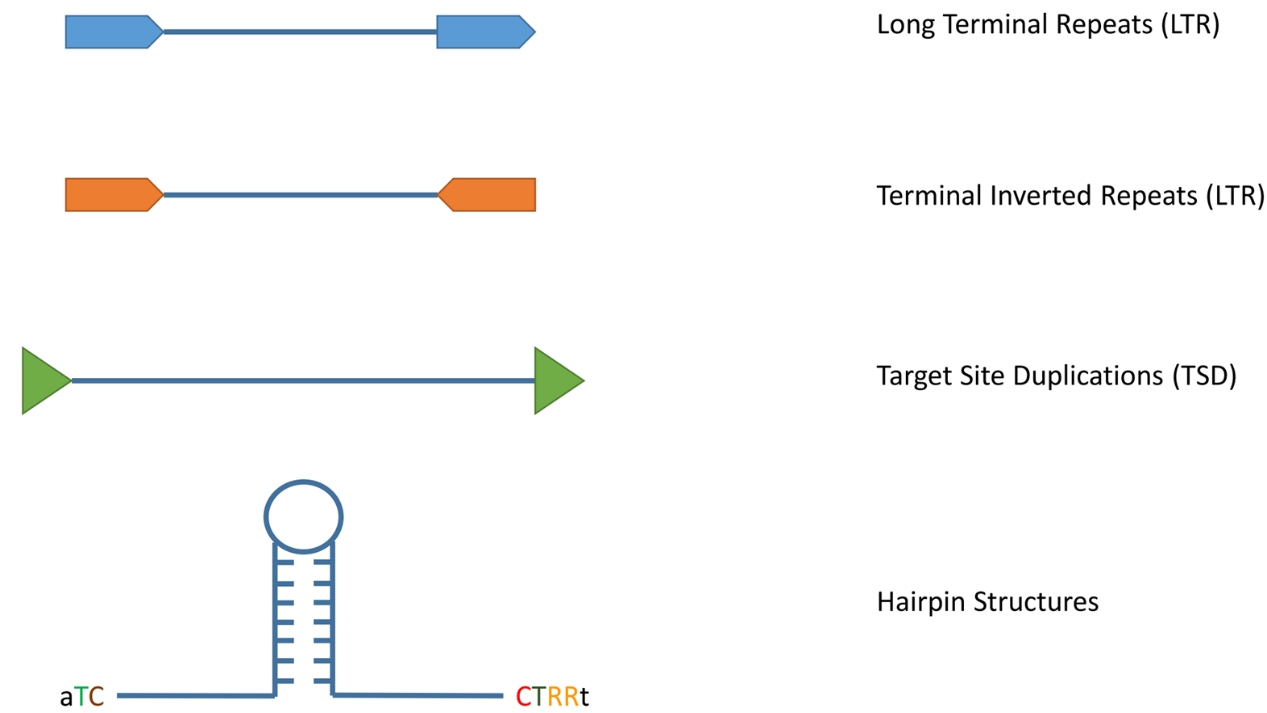
To fairly and comprehensively measure the quality of TE libraries generated by different TE identification tools, we use the benchmarking methods released in two recent studies, EDTA and RepeatModeler2. For convenience, we hereafter refer to the benchmarking methods of EDTA and RepeatModeler2 as BM\_EDTA and BM\_RM2, respectively.

As shown in Fig. S1(A), the BM\_EDTA evaluates the performance of various tools by annotating the genome with the gold standard TE library and the tested TE library generated by these tools. Based on the total number of genomic DNA bases, six metrics, including sensitivity, specificity, accuracy, precision, FDR, and F1, are used to characterize the annotation performance of the tested library. The BM\_EDTA can display detailed metrics, including the rates of false positives, which are common in many TE identification methods. However, it cannot reflect the integrity of the TE models. All general repeat identification programs, even those with many fragments and unclear boundaries, still performed well in benchmarking. For example, while a 1 kbp intact TE sequence and ten 100 bp fragments may obtain the same performance, the former is obviously more valuable in terms of TE integrity and biological significance.

As shown in Fig. S1(B), the BM\_RM2 aligns the tested TE library with the gold standard library and divides the gold standard sequences into four levels: “Perfect”, “Good”, “Present”, and “Not found”. “Perfect” families are those for which one sequence in the tested library matches with >95% sequence similarity and >95% length coverage to a family consensus in the gold standard library. “Good” families are those in which multiple overlapping sequences in the tested library match with >95% similarity and >95% coverage to the curated consensus. A family is considered “present” if one or multiple library sequences align with >80% similarity and >80% coverage to the reference consensus sequence. Below these thresholds, a family is considered “not found”. The BM\_RM2 takes the integrity of the sequence into consideration. Intact TE models usually get a perfect level, while fragments can only get a good, present, or even not found level. However, it cannot display the rate of false positives in the tested TE library. By combining the two complementary benchmarking methods, we can accurately evaluate the integrity of TE models and the rate of false positives in the whole TE library.



**Fig. S1.** Schematic representation of benchmarking methods. A. EDTA benchmarking methods. B. RepeatModeler2 benchmarking methods.



**Fig. S2.** Examples of commonly used TE structural features for detection. The LTR, TIR, TSD, and hairpin structures are used to identify LTR elements, class II elements, and Helitrons, respectively.