Supplementary Material

LoRTE is a Python 2.7 program composed of two main modules (Supplementary Fig. S1) that only required BLAST+ suite as a dependency:

1) The first module is designed to verify the presence/absence in the PacBio reads of a list of annotated TEs in the reference genome. Briefly, the program acquires the flanking sequences of each TEs and align them on the reference genomes using MEGABLAST (Johnson, et al., 2008). The length of the flanking sequences is specified by the user (default=200bp). At this stage, a filter verifies if the TE is correctly annotated and if the flanking sequences map uniquely on the genome. TE wrongly annotated or located in region too much enriched in repeats are categorized as “irresolvable locus” in the final output file. The remaining 3' and 5' flanking sequences are aligned on the PacBio read using MEGABLAST. All the sequences located between a 3' and 5' flanking sequences in the same orientation, and in a specified window size in the PacBio reads are extracted. These extracted sequences are then searched with BLASTN against the TE consensus sequences. For a given locus if the sequence matches on the TE consensuses, the TE is considered as “present” in the read. Sequences <50nt without any match on the TE consensuses correspond to a deletion (“absent”). “Polymorphic locus” corresponds to a situation in which a given TE is “absent” in some reads and “present” in some others. Finally some locus are characterized as “ambiguous” if the extracted sequences between the 3' and the 5' flanking are >50nt but do not match with a TE consensus sequences. This latter case may correspond to partially deleted TEs.

2) The second step aims to identify new TE insertions present in the reads but absent in the reference genome. The program removes from the PacBio reads all the TE sequences identified by the first module. Then, the TE consensuses are aligned using BLASTN on the reads to identify all the remaining TEs. The flanking 3' and 5' ends of these putative new TE insertions are extracted and searched using MEGABLAST on the reference genome. All the sequences between a 3' and 5' ends, in the same orientation, and in a specified window size are extracted and the program verifies if they match with a TE consensus using BLASTN. If the extracted sequences are <50nt and do not resemble to a given consensus the program considers these cases as new insertions. Finally, all the reads testifying for a new insertion for the same locus are clustered together.

To assess the performance and accuracy, we have tested LoRTE on two *Drosophila melanogaster* datasets: (i) synthetic PacBio reads generated by random cutting of the reference genome (release 5) in segments of 3 to 30kb in length. Benchmark of the program are monitored by random insertion of 250 TEs and random deletion of 100 TEs in the reference genome before its segmentation (ii) genuine PacBio reads of adult males of the ISO1 strains (same stock used in the official reference

assembly) with a sequencing depth of 90x. In order to identify false positives, LoRTE predictions are then compared with the genome assembly of the PacBio reads. Reads and the assembly are available at https://github.com/PacificBiosciences/DevNet/wiki/Drosophila-sequence-and-assembly . To test the impact of the coverage on the performance of LoRTE we have sub-sampled the datasets to a lower coverages (from 1x to 40x). For these experiments, we have used a list of 4239 annotated TEs and corresponding TE consensuses obtain from FlyBase (Attrill, et al., 2016) and RepBase (Bao, et al., 2015).

Input and raw output files used in this study are available at http://www.egce.cnrs-gif.fr/?p=6422

References

Attrill, H., *et al.* (2016) FlyBase: establishing a Gene Group resource for Drosophila melanogaster, *Nucleic Acids Res*, **44**, D786-792.
Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes, *Mob DNA*, **6**, 11.
Johnson, M., *et al.* (2008) NCBI BLAST: a better web interface, *Nucleic Acids Res*, **36**, W5-9.

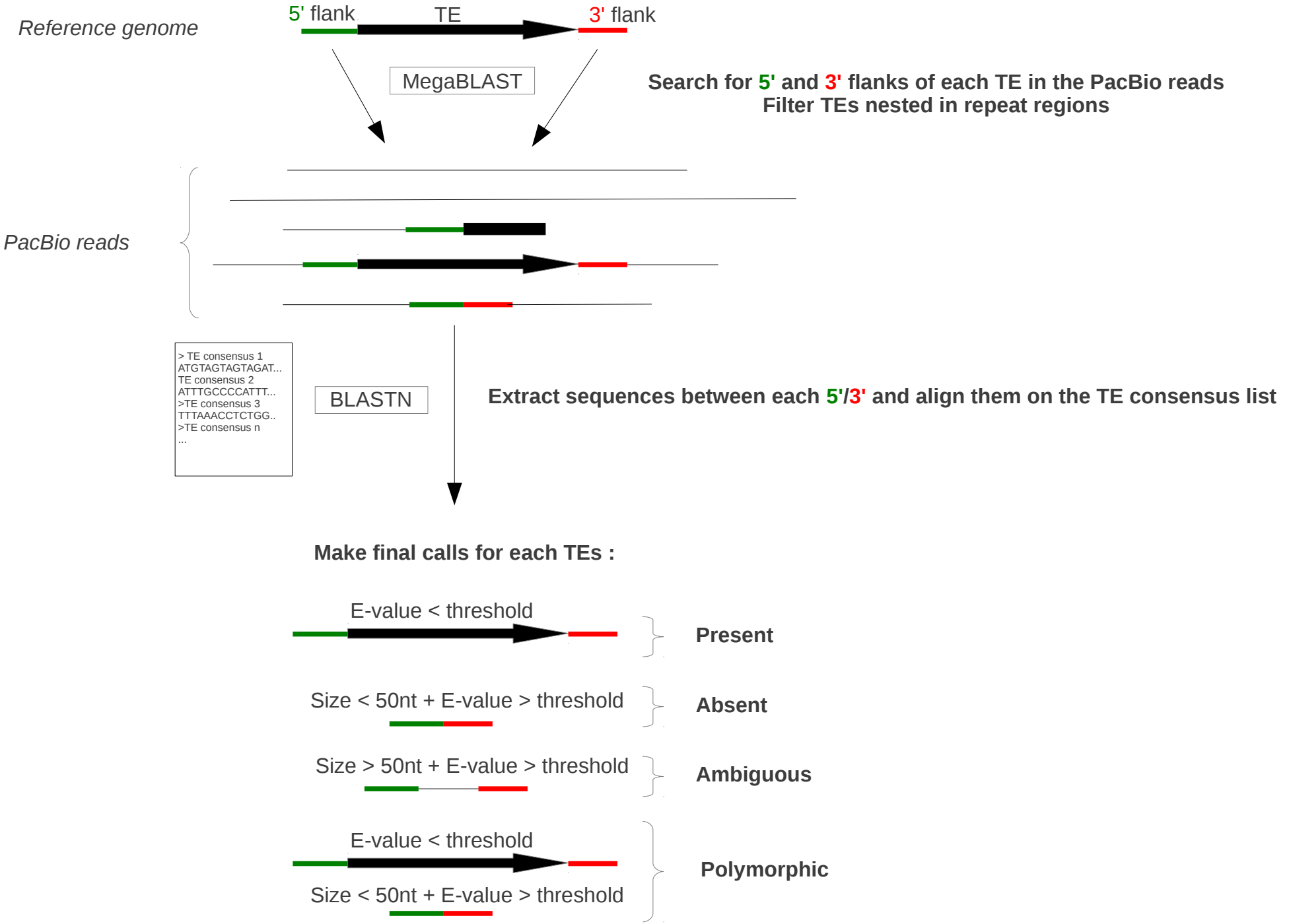
Supplementary Figure Legends

Supplementary Fig. S1: Simplified workflow of the two modules composing LoRTE

Supplementary Fig. S2: Family distribution of the new TE insertion and deletion found in the *Drosophila melanogaster* PacBio reads and absent in the reference genome.

A. Presence/Absence module

Figure S1



B. New insertion module

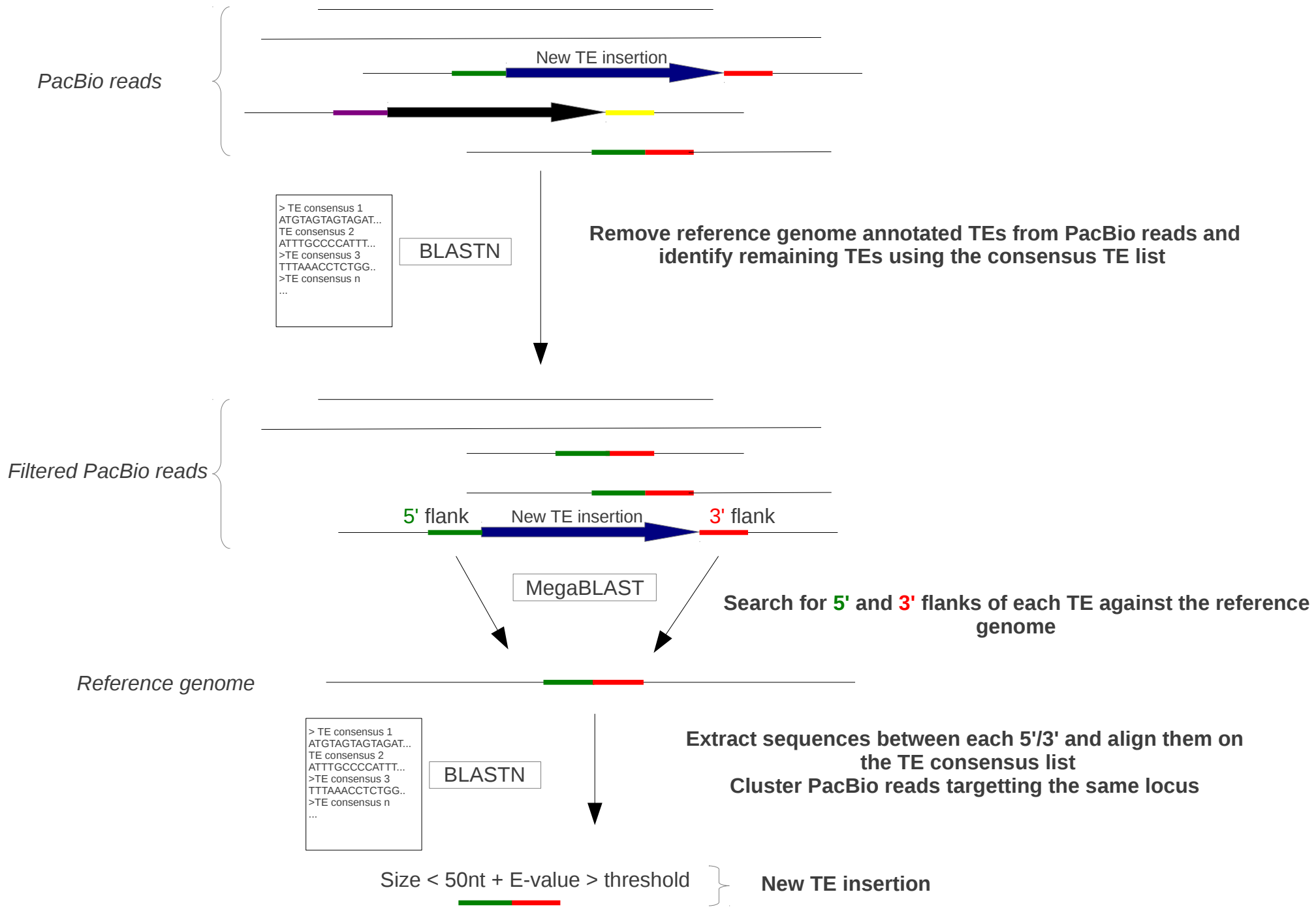
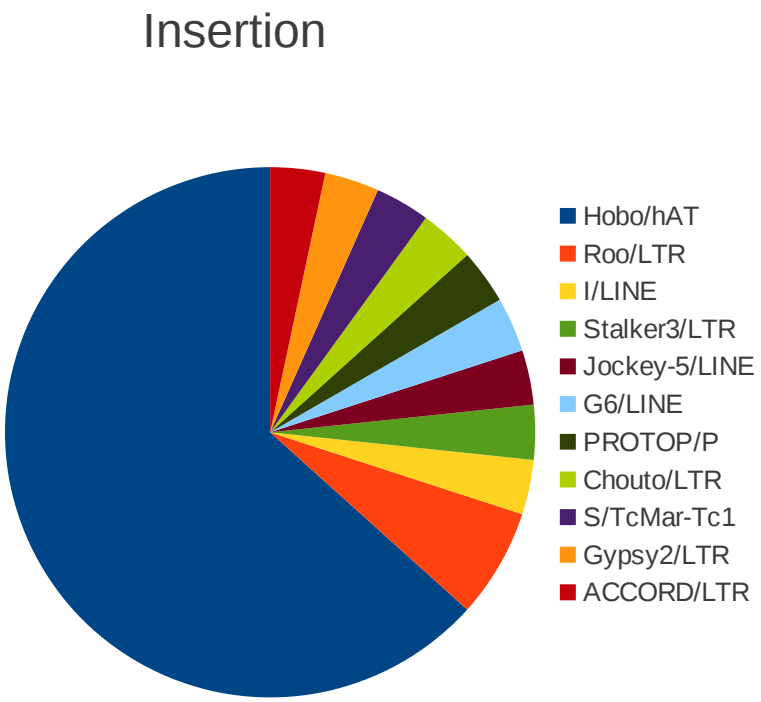
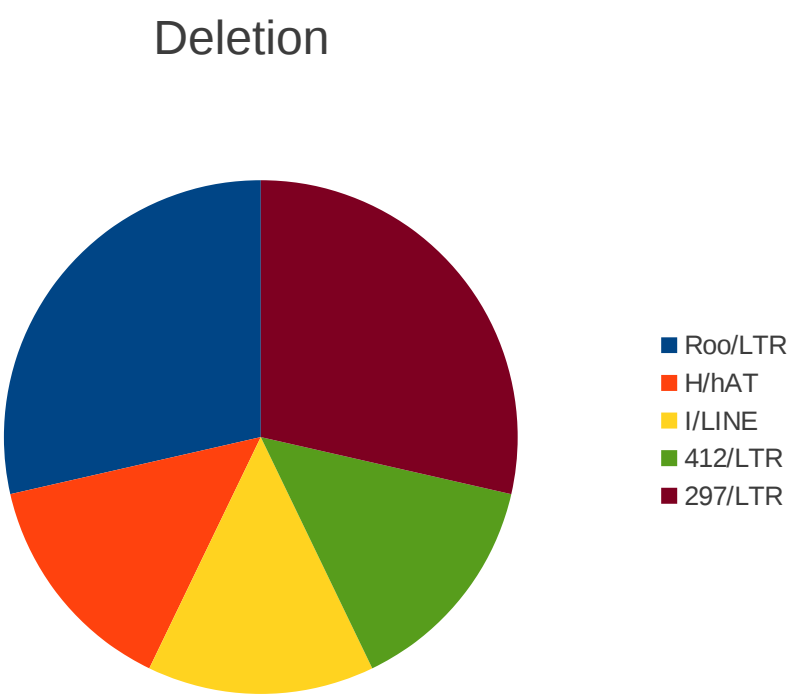


Figure S2


Insertion
Hobo/hAT
Roo/LTR
I/LINE
Stalker3/LTR
Jockey-5/LINE
G6/LINE
PROTOP/P
Chouto/LTR
S/TcMar-Tc1
Gypsy2/LTR
ACCORD/LTR


Deletion
Roo/LTR
H/hAT
I/LINE
412/LTR
297/LTR