Additional File 1. Supplementary Figures

AluMine: alignment-free method for the discovery of polymorphic Alu element insertions

Tarmo Puurand, Viktoria Kukuškina, Fanny-Dhelia Pajuste and Maido Remm*

Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

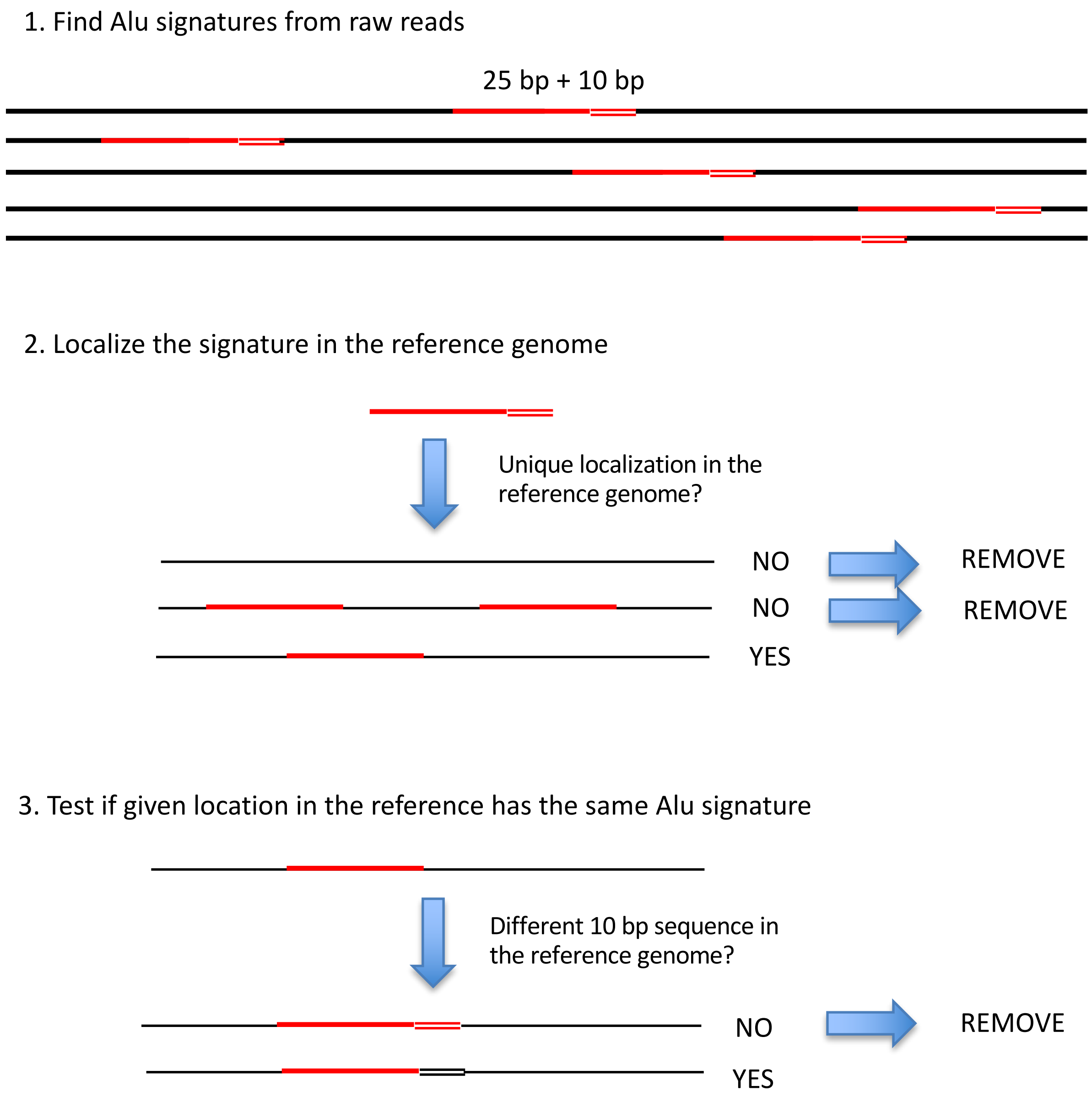
1. Find Alu signatures from raw reads
25 bp + 10 bp
2. Localize the signature in the reference genome
Unique localization in the reference genome?
NO
REMOVE
NO
REMOVE
YES
3. Test if given location in the reference has the same Alu signature
Different 10 bp sequence in the reference genome?
NO
REMOVE
YES

Figure S1. Main steps of the REF– discovery pipeline.

Step 1. Search for 10 bp Alu signature sequences in raw reads from sequenced individuals. Extract 25 bp sequence from the 5′-flanking region of the signature sequence and add it to the 10 bp signature. Remove the candidate if the frequency of the resulting 35-mer in a given individual is <5 or >100.

Step 2. Use the 25 bp region to determine the location of the Alu element in the reference genome using gtester4. Remove the candidate if its location is not detectable in the reference genome (Alu elements from heterochromatin). Remove the candidate if the 25 bp sequence is present in multiple locations in the reference genome (Alu elements from repeated regions).

Step 3. Extract the 10 bp sequence from the reference genome. Compare it with the Alu signature. Remove the candidate if the reference genome already contains an Alu element in this position (fixed Alu elements).

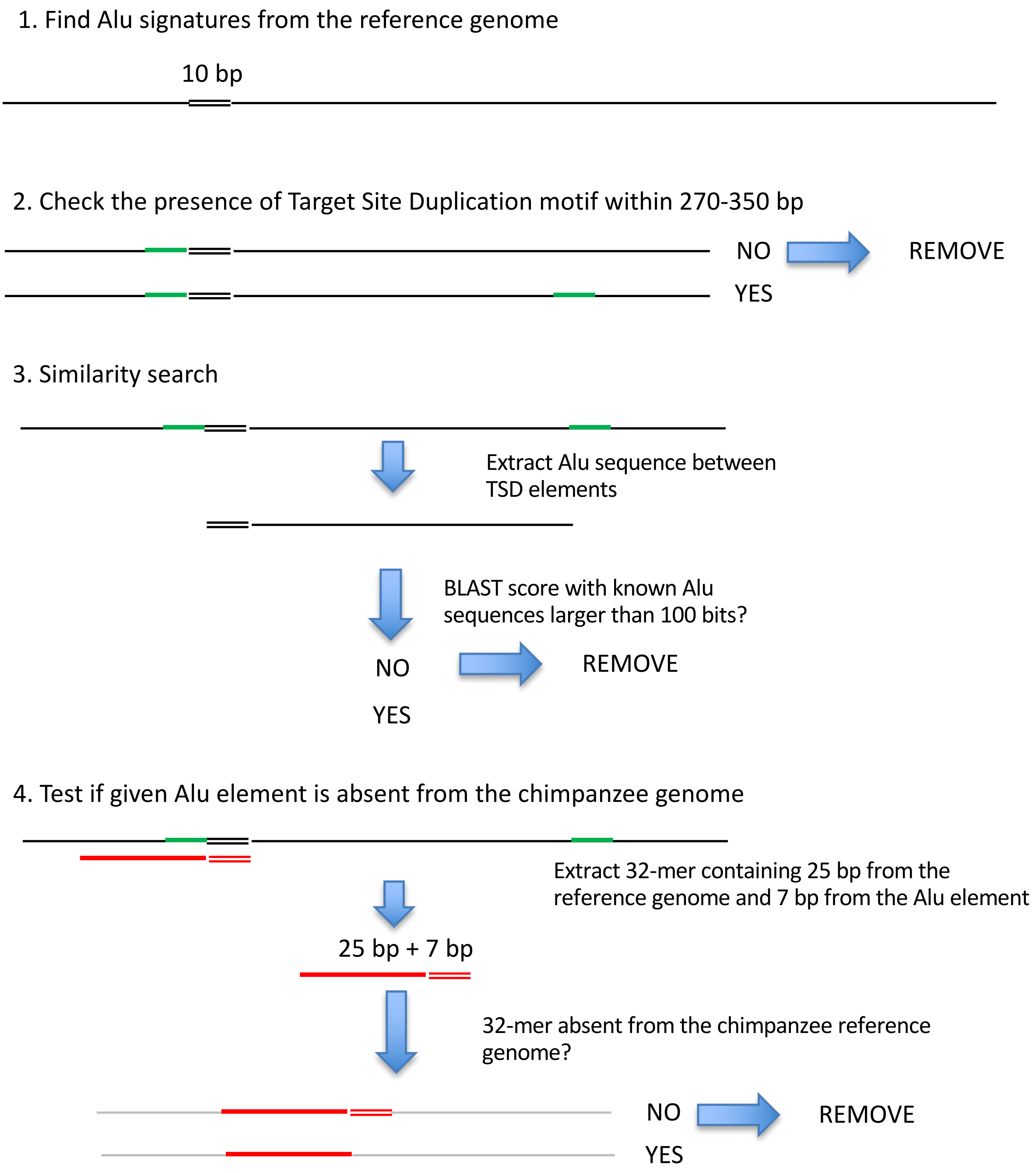
1. Find Alu signatures from the reference genome
10 bp
2. Check the presence of Target Site Duplication motif within 270-350 bp
NO
REMOVE
YES
3. Similarity search
Extract Alu sequence between TSD elements
BLAST score with known Alu sequences larger than 100 bits?
NO
REMOVE
YES
4. Test if given Alu element is absent from the chimpanzee genome
Extract 32-mer containing 25 bp from the reference genome and 7 bp from the Alu element
25 bp + 7 bp
32-mer absent from the chimpanzee reference genome?
NO
REMOVE
YES

Figure S2. Main steps of the REF+ discovery pipeline.
Step 1. Find all 10 bp Alu signature sequences, allowing up to 1 mismatch with the reference genome.
Step 2. Identify the 5 bp target site duplication (TSD) sequence at the 5' end of the Alu signature sequence. Search for identical 5 bp TSD sequences at the 3' end of the Alu element. Remove the candidate if the 3' end TSD is not detected within 270 - 350 bp of the start of the Alu signature sequence.
Step 3. Test the similarity between detected Alu elements with known Alu elements.
Step 4. Generate REF+ *k*-mers (25 nt from the genome and 7 nt from the Alu sequence) for each candidate. Count the frequencies of these *k*-mers in the chimpanzee genome, allowing 2 mismatches. Remove the candidate if the 32-mer was detected at least once in the chimpanzee genome.