



Figure S2 Classifying the repeat motif length of STRs. (A,B) Example of a locus targeted for cleaning (misfit <5%). This is locus is clearly a tri locus with some noise, which could be cause by stuttering (or a point mutation). A: The relative overlap $\sum_i f_i f_{i+L} / \sum_i f_i^2$, where f_i is the frequency of allele i in the sample, and L is the repeat unit length. Repeat length three has the best fit (yellow circle). B: The frequencies of alleles with different values of the remainder g . (C,D) Example of a locus targeted for removal. Panels C and D corresponds to panels A and B, respectively. This could be either a mix of di- and tetra-nucleotide repeats, or two shifted tetra-nucleotide sequences.