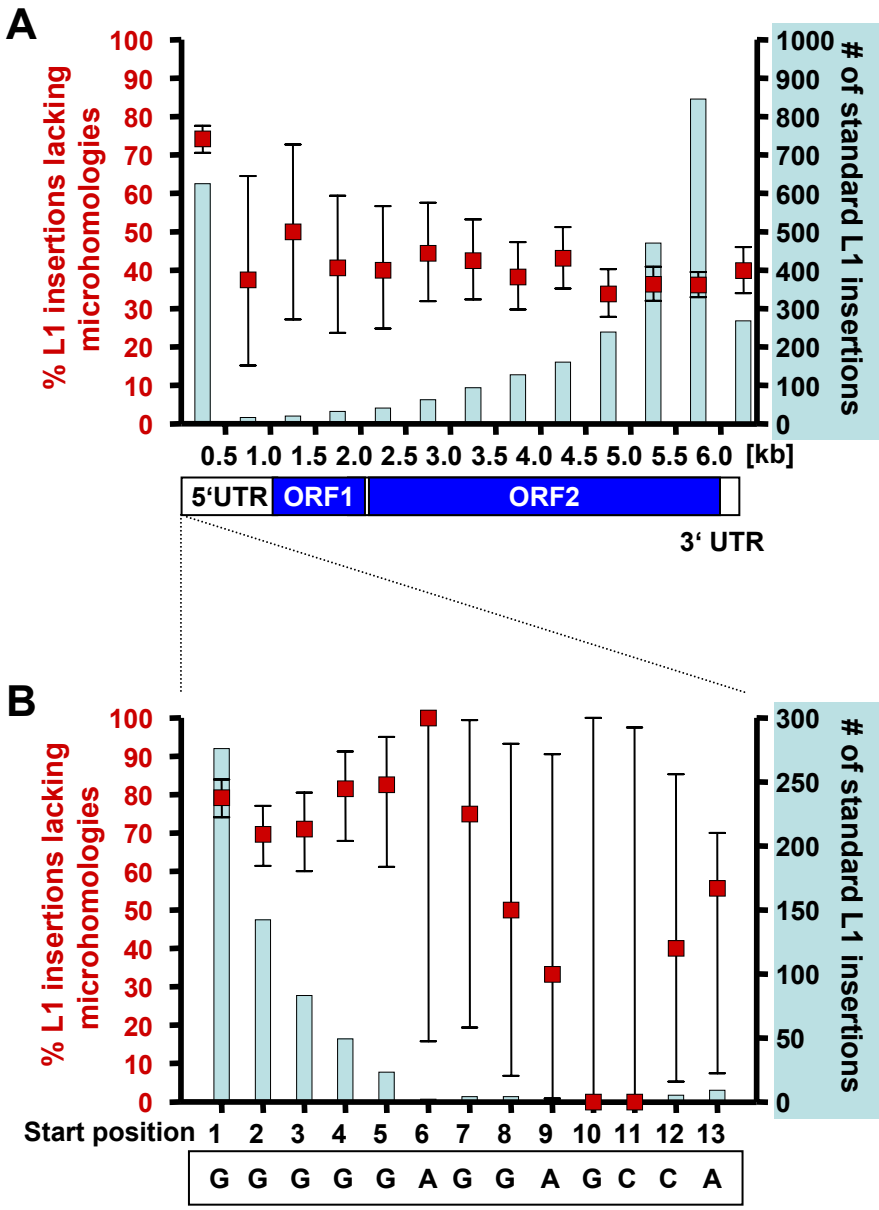

Supplementary Figure 3: Correlation between L1 insertion length and occurrence of microhomologies.
(A) L1 start positions of the 3000 canonical endogenous L1 insertions identified by TSDfinder were grouped into 500-nucleotide bins. The schematic of L1 along the x axis indicates the L1 start position represented by each bin. Blue bars represent the number of L1 insertions (y axis on the right hand side) as a function of the length of the 5'-truncation. Red squares denote the relative frequency of L1 insertions without any microhomologies at the 5' junction (y axis on the left hand side). Error bars indicate 95% Clopper-Pearson confidence intervals. Consistent with previous reports (Szak et al. 2002; Symer et al. 2002) we found a bimodal distribution of L1 lengths. While the majority of L1 elements was 5' truncated with increasing copy numbers at shorter lengths, a significant fraction consisted of full-length elements. As expected for full-length elements, 75% of the integrants in the first bin did not display microhomologies. In all other bins, independent of the degree of truncation, only 30-50% of integrants (vs. 50-56% expected) did not share nucleotides with their TSD.
(B) Detailed analysis of the correlation between insertion length and frequency of microhomologies in L1 insertions starting at position +1 to +13. The consensus sequence of the 5' end of a full-length L1 element is given below the corresponding nucleotide position (x axis). 70 to 100% of L1 insertions that start within the first six nucleotides of the L1 full-length consensus sequence lack microhomologies, indicating that these elements have some preference for a mechanism not requiring the formation of microcomplementarities. In the case of L1 insertions starting from position 6 to 9, the fraction of elements without any microhomologies decreases from >70% to ~40%, although the statistical uncertainty is rather high due to the small copy numbers of these elements in the human genome. However, the observed trend is consistent with our definition of full-length elements starting within the first 5 nucleotides of the L1.3 reference sequence.