



Figure S2 **A**) Distribution of the CRE 8-mer TGACGTCA in human promoters. **B**) LF for CRE 8-mer with insert length ranging from 0-bps to 30-bps (TGAC-N₀₋₃₀-GTCA). **C**) Distribution of the ETS TFBS (CCGGAAGT) in promoters counting occurrence in 20 bp bins. **D**) LF for ETS 8-mer (CCGG-N₀₋₃₀-AAGT) with insert length ranging from 0-bps to 30-bps. The 8-mer CCGGAAGT preferentially localizes in promoters when the two 4-mers are abutted. High LF values are also observed at additional insert

length, including CCGG-N₆-AAGT, which represents the ETS \leftrightarrow ETS motif and CCGG-N₂₃-AAGT. **E)** Distribution of the split 8-mer CGCC-N₂₄-TATA in promoters. **F)** LF for CGCC-N₀₋₃₀-TATA with insert length ranging from 0-bps to 30-bps. **G)** Distribution of the split 8-mer CAAT-N₂₄-CCGC in human promoters. The most localizing motif is the split 9-mer CCAAT-N₂₄-CCGC marked in red. **H)** LF for CAAT-N₀₋₃₀-CCGC with insert length from 0-bps to 30-bps. **I)** Distribution of the split 8-mer GGGA-N₄-TAGT and more localizing split 10-mer GGGA-N₂-TGTAGT. **J)** Occurrences of split 8-mer GGGA-N₀₋₄₀-TAGT with insert length ranging from 0-bps to 40-bps. The reverse order of the two 4-mers TAGT-N₀₋₄₀-GGGA is shown with negative values.