**S1. Parameters recovered\* with regular expression (RegEx) pattern-matching\*\* in unaligned genomes.**

**Motif description Coding strand Non-coding Reverse Complement**

CpG site CG CG

high affinity E2 site ACC[GT]A\w{4}GGT ACC\w{4}T[AC]GGT

non-canonical E2 site ACAC\w{5}GGT ACC\w{5}GTGT

APOBEC3 site TC[AT] [AT]GA

TLR9 stimulatory [ACT][AG][AT]CGTT AACG[AT][TC][TCG]

[AG]{2}CG[CT]{2} [GA]{2}CG[TC}{2}

TLR9 suppressing (CC\w[GTA]{2}\w{2}GGG CCC\w{2}[CTA]{2}\wGG

[AT][GT]{2}[ACT]GGGG CCCC[AGT]CA{2}[AT]

TTAGGG CCCTAA

G quadruplexes ([G]{3,5}\w{1,12}){3}[G]{3,5} ([C]{3,5}\w{1,12}){3}[C]{3,5}

G duplexes GGG\w{1,12}GGG CCC\w{1,12}CCC

\* additional parameters that were not found with regular expression pattern matching included base composition, palindromes, inverted repeats, large duplicated regions in forward and reverse orientation, and the proportion of APOBEC sites on either strand (in light of the single-stranded DNA dependency of APOBEC3 enzymes).

\*\* relevant RegEx rules:

[GT] = a single nucleotide that is either G or T

\w = any single letter

{4} = 4 multiples of the immediately preceding portion of the expression

{3,7} = 3, 4, 5, 6 or 7 multiples of the immediately preceding portion of expression

Because two APOBEC3 sites can overlap (i.e., TCTCT,TCTCA) full discovery employed the iterative '?=TC[AT]' expression.

For example:

(G{3,7}\w{3,7}){2} matches any pattern that satisfies 3 or more G in a row followed by 3 or more nucleotides in a row, repeated twice. Including but not limited to:

GGGNNNGGGNNN

…

GGGNNNGGGGGGGNNNN

…

GGGGGGGNNNNNNNGGGGGGGNNNNNNN