Supplementary for "Statistical Methods for Identifying Sequence Motifs Affecting Point Mutations"

Yicheng Zhu, Teresa Neeman, Von Bing Yap & Gavin A Huttley

November 15, 2016

Position(s)	**Deviance**	**df**	**p-value**
-2	1574.2	3	0.0
-1	18674.9	3	0.0
+1	346848.0	3	0.0
+2	2174.5	3	0.0
(-2, -1)	1603.1	9	0.0
(-2, +1)	555.5	9	0.0
(-2, +2)	352.7	9	1.7×10^{-70}
(-1, +1)	2341.3	9	0.0
(-1, +2)	315.1	9	1.6×10^{-62}
(+1, +2)	1965.0	9	0.0
(-2, -1, +1)	939.7	27	0.0
(-2, -1, +2)	523.0	27	2.7×10^{-93}
(-2, +1, +2)	264.6	27	7.3×10^{-41}
(-1, +1, +2)	467.8	27	6.5×10^{-82}
(-2, -1, +1, +2)	273.9	81	9.1×10^{-23}

Table S1: Log-linear analysis of C→T autosomal intergenic mutations. Position(s) are relative to the index position (see manuscript Figure 1). Deviance is from the log-linear model, with df degrees-of-freedom and corresponding p-value obtained from the χ^2 distribution. p-values listed as 0.0 are below the limit of detection.

Position(s)	Deviance	df	p-value
-2	26528.3	3	0.0
-1	20038.7	3	0.0
+1	57037.8	3	0.0
+2	1802.0	3	0.0
(-2, -1)	9058.8	9	0.0
(-2, +1)	3615.8	9	0.0
(-2, +2)	701.1	9	0.0
(-1, +1)	3233.2	9	0.0
(-1, +2)	1516.8	9	0.0
(+1, +2)	2329.1	9	0.0
(-2, -1, +1)	2018.3	27	0.0
(-2, -1, +2)	561.1	27	0.0
(-2, +1, +2)	362.2	27	2.4×10^{-60}
(-1, +1, +2)	1191.2	27	0.0
(-2, -1, +1, +2)	426.5	81	2.3×10^{-48}

Table S2: Log-linear analysis of A→G autosomal intergenic mutations. Position(s) are relative to the index position (see manuscript Figure 1). Deviance is from the log-linear model, with df degrees-of-freedom and corresponding p-value obtained from the χ^2 distribution. p-values listed as 0.0 are below the limit of detection.

Direction	$RE_{max}(1)$	RE Dist.	p-val Dist.
A→C	0.0042	4	10
A→G	0.0186	2	10
A→T	0.0093	3	10
C→A	0.0093	4	10
C→G	0.0057	3	10
C→T	0.0861	1	10
G→A	0.0860	1	10
G→C	0.0054	3	10
G→T	0.0091	4	10
T→A	0.0095	3	10
T→C	0.0190	2	10
T→G	0.0039	4	10

Table S3: The most distant positions from the mutation with RE(1) $\geq$ 10% of $RE_{max}(1)$. RE(1) is the first order RE for the position, and $RE_{max}(1)$ the largest RE from a first order effect for the surveyed positions. RE Dist. is the absolute value of the relative position based on the RE value. p-val Dist. is the corresponding distance based on the p-value. The maximum possible distance is 10. Only point mutations significant after correcting for 20 tests using the Holm-Šidäk procedure were considered.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)
A→C	1.6×10^{-5}	+1	-	-
A→G	4.2×10^{-5}	+1	2.6×10^{-5}	(-2, -1)
A→T	9.3×10^{-5}	-1	1.5×10^{-5}	(-2, +2)
C→A	2.7×10^{-5}	-1	3.4×10^{-5}	(-1, +1)
C→G	3.8×10^{-5}	-1	1.5×10^{-5}	(-2, -1)
C→T	3.2×10^{-5}	+1	1.2×10^{-5}	(-1, +1)

Table S4: Neighbour associations with point mutations differ between autosomal intronic and intergenic point mutations. As there was no significant strand asymmetry detected for either sequence class, only + strand effects are shown. Only point mutations with at least one significant test after correcting for 15 tests using the Holm-Šidäk procedure are shown. Non-significant results are indicated by '-'.

Direction	**RE**$_{max}$(1)	**Pos.**(1)
A→C	1.7×10^{-5}	+1
A→G	6.3×10^{-6}	+1
C→G	1.4×10^{-5}	+1
C→T	5.0×10^{-6}	+1
G→A	6.2×10^{-6}	+1
T→A	1.6×10^{-5}	+2
T→C	8.3×10^{-6}	-1
T→G	2.1×10^{-5}	-1

Table S5: Significant differences in neighbour associations between intergenic autosomal and X-chromosomal point mutations. RE$_{max}$(1) the largest RE from a first order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown.

Direction	**RET**
G→A	-0.0032
A→G	-0.0031
C→T	-0.0031
T→C	-0.0026
C→G	-0.0019
G→C	-0.0017
T→G	-0.0011
A→C	-0.0007
T→A	0.0038
A→T	0.0039
G→T	0.0051
C→A	0.0052

Table S6: Significant differences in mutation spectra between autosomal intergenic and intronic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. Only RET from the intergenic group are shown. A positive (negative) RET indicates a excess (deficit) of that mutation in the intergenic group. All tests returned p-values that were below the limit of detection and thus were statistically significant after correcting for 4 tests using the Holm-Šidäk procedure.

Direction	RET
T→A	-0.0004
C→A	-0.0003
G→T	-0.0003
A→T	-0.0002
A→C	-0.0002
T→G	-0.0001
G→C	-0.0000
C→G	0.0000
C→T	0.0003
G→A	0.0003
A→G	0.0004
T→C	0.0005

Table S7: Significant differences in spectra between autosomal and X-chromosomal intergenic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. p-value is from the corresponding hypothesis test. Only RET from the autosomal group are shown. A positive (negative) RET indicates a excess (deficit) of that mutation in autosomes. All tests returned p-values that were $\leq 4.7e^{-9}$ and thus were statistically significant after correcting for 4 tests using the Holm-Šidäk procedure.

Direction	RET
T→G	-0.0001
A→C	-0.0001
G→T	-0.0001
C→A	-0.0001
G→C	-0.0001
A→T	-0.0001
T→A	-0.0000
C→G	0.0000
C→T	0.0001
G→A	0.0002
T→C	0.0002
A→G	0.0002

Table S8: Significant differences in spectra between autosomal and X-chromosomal intronic point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. p-value is from the corresponding hypothesis test. Only RET from the autosomal group are shown. A positive (negative) RET indicates a excess (deficit) of that mutation in autosomes. All tests returned p-values that were $\leq 8.6e^{-5}$ and thus were statistically significant after correcting for 4 tests using the Holm-Šidäk procedure.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0132	-1	0.0093	(-1, +1)	0.0039	(-2, -1, +1)
A→G	0.0134	-1	0.0164	(-1, +1)	0.0032	(-2, -1, +1)
A→T	0.0116	-1	0.0030	(-2, +1)	0.0027	(-2, -1, +1)
C→A	0.0276	-1	0.0076	(-1, +1)	0.0029	(-1, +1, +2)
C→G	0.0259	+1	0.0028	(-1, +1)	0.0025	(-2, -1, +1)
C→T	0.0840	-1	0.0110	(-1, +1)	0.0006	(-2, -1, +1)

Table S9: Test of strand symmetric neighbourhood associations for malignant melanoma point mutations. RE$_{max}$(#) is the maximum RE for order # and Pos.(#) the corresponding position(s). Only effects significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown.

Direction	**RET**	**p-value**
A→C	-0.0025	0.1650
C→G	-0.0024	8.0×10^{-50}
C→A	-0.0020	8.0×10^{-50}
A→T	0.0007	0.1650
A→G	0.0018	0.1650
C→T	0.0048	8.0×10^{-50}

Table S10: Differences in spectra between strands for malignant melanoma point mutations. Separate log-linear models were used for the + strand starting bases A and C. RET is the RE term for that row mutation direction. Only RET from the + strand are shown. A positive (negative) RET indicates a excess (deficit) of that mutation on the + strand. p-value is from the corresponding hypothesis test. Only mutations from C were significant after correcting for 2 tests using the Holm-Šidäk procedure.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0034	-1	0.0016	(+1, +2)	0.0012	(-2, -1, +1)
A→G	0.0205	+1	0.0042	(-2, -1)	0.0007	(-2, -1, +1)
A→T	0.0089	+1	0.0051	(-1, +1)	0.0025	(-1, +1, +2)
C→A	0.0092	+1	0.0035	(-1, +1)	0.0012	(-1, +1, +2)
C→G	0.0049	+1	0.0022	(+1, +2)	0.0008	(-1, +1, +2)
C→T	0.0924	+1	0.0004	(+1, +2)	0.0002	(-2, -1, +1)

Table S11: Neighbour associations with point mutations within autosomal introns. RE$_{max}$(1) the largest RE from a first order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown.

Direction	**RE**$_{max}$(1)	**Pos.**(1)	**RE**$_{max}$(2)	**Pos.**(2)	**RE**$_{max}$(3)	**Pos.**(3)
A→C	0.0033	-1	0.0009	(-1, +1)	-	-
A→G	0.0023	+1	0.0005	(-1, +1)	0.0004	(-1, +1, +2)
A→T	0.0071	-1	0.0023	(+1, +2)	-	-
C→A	0.0065	-1	0.0013	(-2, -1)	0.0007	(-2, +1, +2)
C→G	0.0007	+1	0.0004	(+1, +2)	0.0006	(-2, -1, +2)
C→T	0.0268	-1	0.0030	(-1, +1)	0.0002	(-2, -1, +1)
G→A	0.0275	+1	0.0017	(-1, +1)	0.0002	(-1, +1, +2)
G→C	0.0008	-1	0.0004	(+1, +2)	0.0006	(-2, -1, +2)
G→T	0.0056	+1	0.0011	(+1, +2)	0.0007	(-1, +1, +2)
T→A	0.0080	+1	0.0018	(-2, -1)	0.0018	(-1, +1, +2)
T→C	0.0023	-1	0.0014	(-1, +1)	0.0005	(-1, +1, +2)
T→G	0.0014	+1	0.0015	(-1, +1)	0.0013	(-2, +1, +2)

Table S12: Significant differences in the association of neighbours on exonic point mutations between germline and malignant melanoma. RE$_{max}$(1) the largest RE from a first order test and Pos.(1) is the corresponding position. Only mutations significant after correcting for the 15 different tests using the Holm-Šidäk procedure are shown. Non-significant results are indicated by '-'.

Direction	**RET**
T→C	-0.0332
A→G	-0.0327
C→G	-0.0109
C→A	-0.0092
G→C	-0.0091
G→T	-0.0080
A→C	0.0045
T→G	0.0061
G→A	0.0263
C→T	0.0346
T→A	0.0624
A→T	0.0624

Table S13: Significant differences in spectra between germline exon and malignant melanoma point mutations. Separate log-linear models were used for each starting base (X in X→Y). RET is the RE term for that row mutation direction. Only RET from the melanoma group are shown. A positive (negative) RET indicates a excess (deficit) of that mutation in the melanoma group. All tests returned p-values that were below the limit of detection and thus were statistically significant after correcting for 4 tests using the Holm-Šidäk procedure.

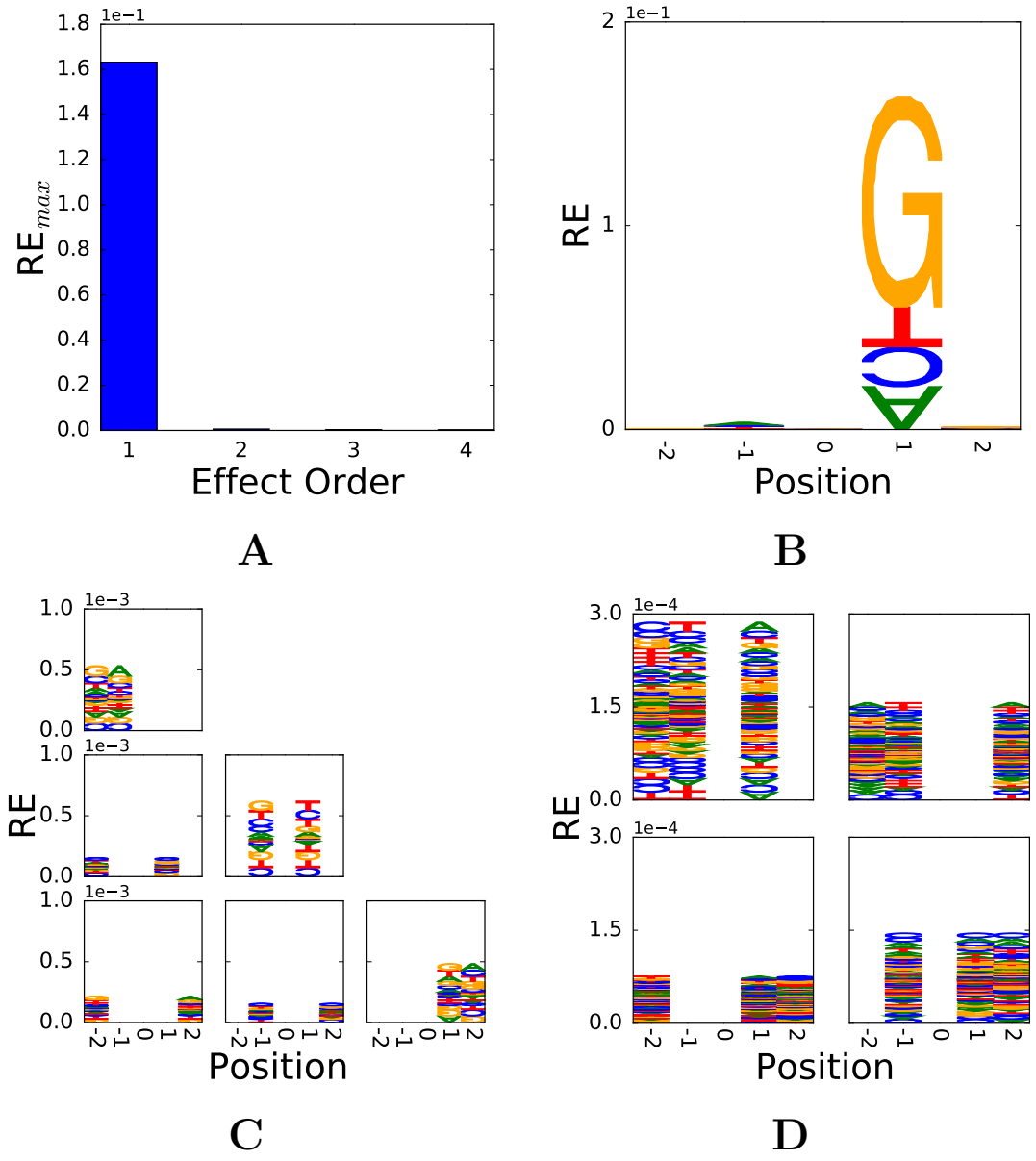


Figure S1: Flanking influences on C→T mutation in autosomal exon sequences. A First order effects are the dominant neighbourhood influence, RE_{max} (y-axis) is the maximum RE from the possible evaluations for a motif length (x-axis), B Single position effects, C Two-way effects, and D Three-way effects.

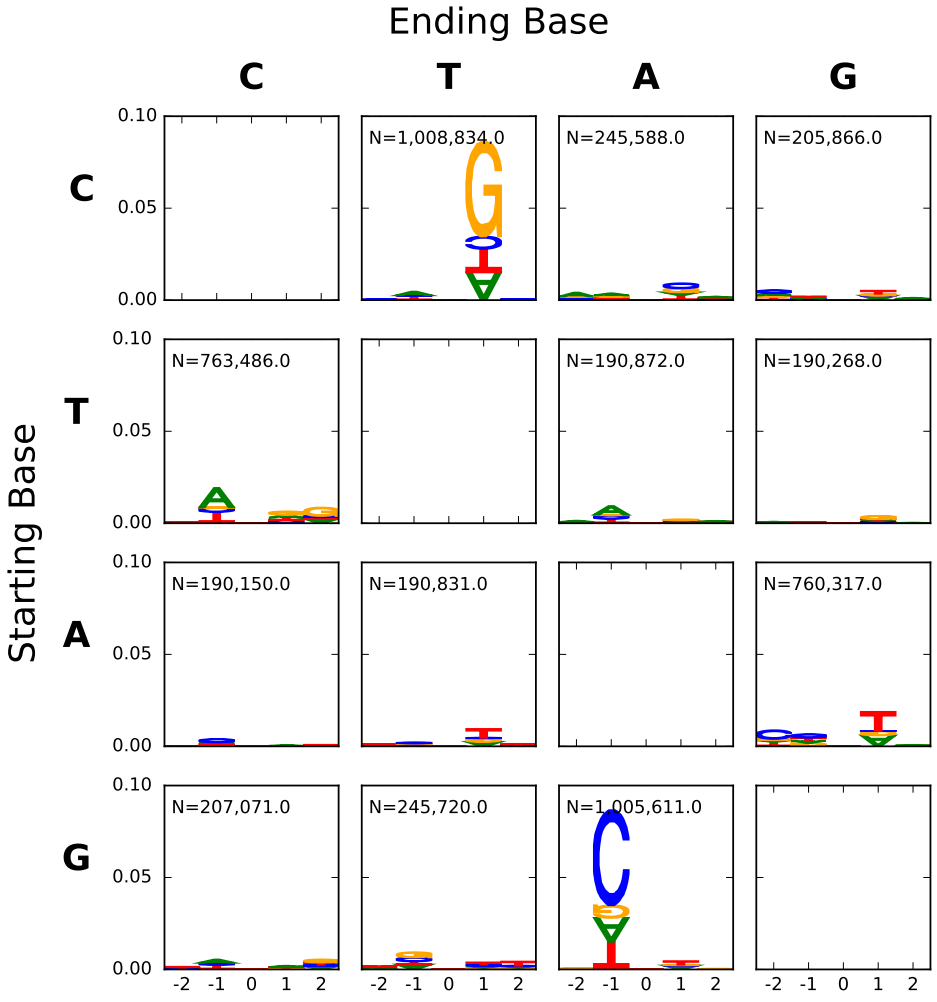


Figure S2: A panel of all 12 point mutations from autosomal intergenic germline mutations. Text in each panel indicates the number of genetic variants analysed.

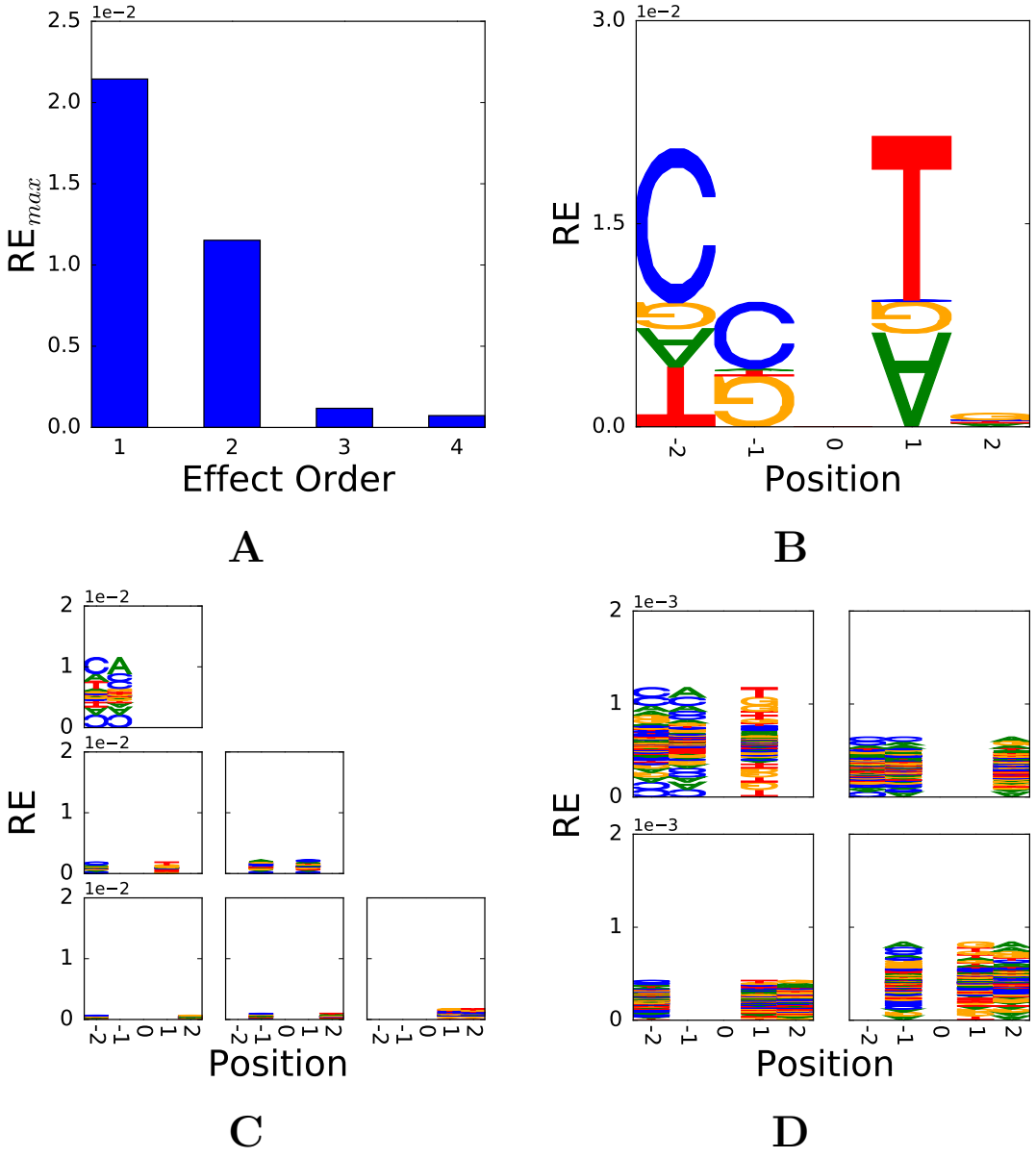
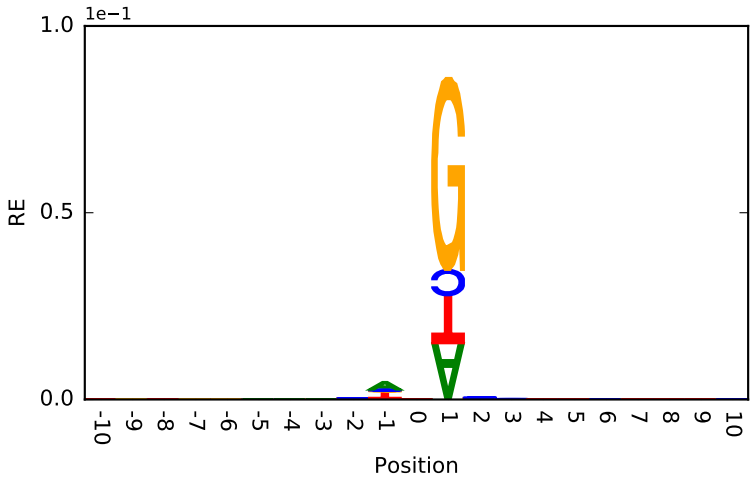
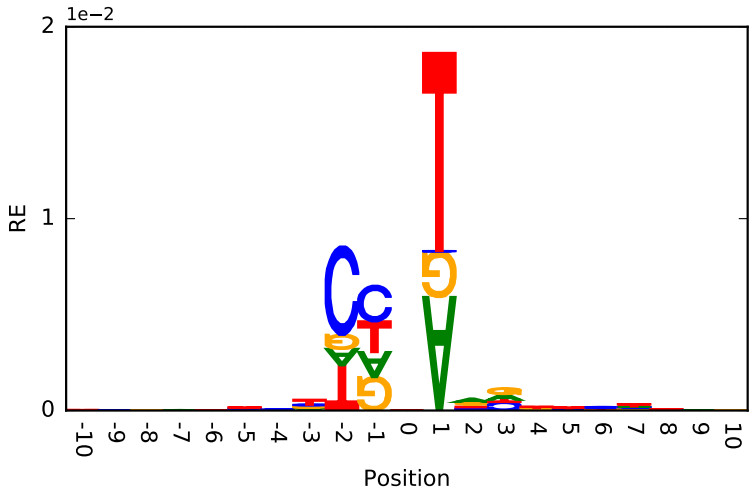


Figure S3: Flanking influences on A→G mutation in autosomal exon sequences. A First order effects are the dominant neighbourhood influence, B Single position effects, C Two-way effects, and D Three-way effects



(A)



(B)

Figure S4: The extent of neighbourhood effects on autosomal intergenic mutations. A C→T, B) A→G.

Mutation motifs estimated using the whole genome as the reference

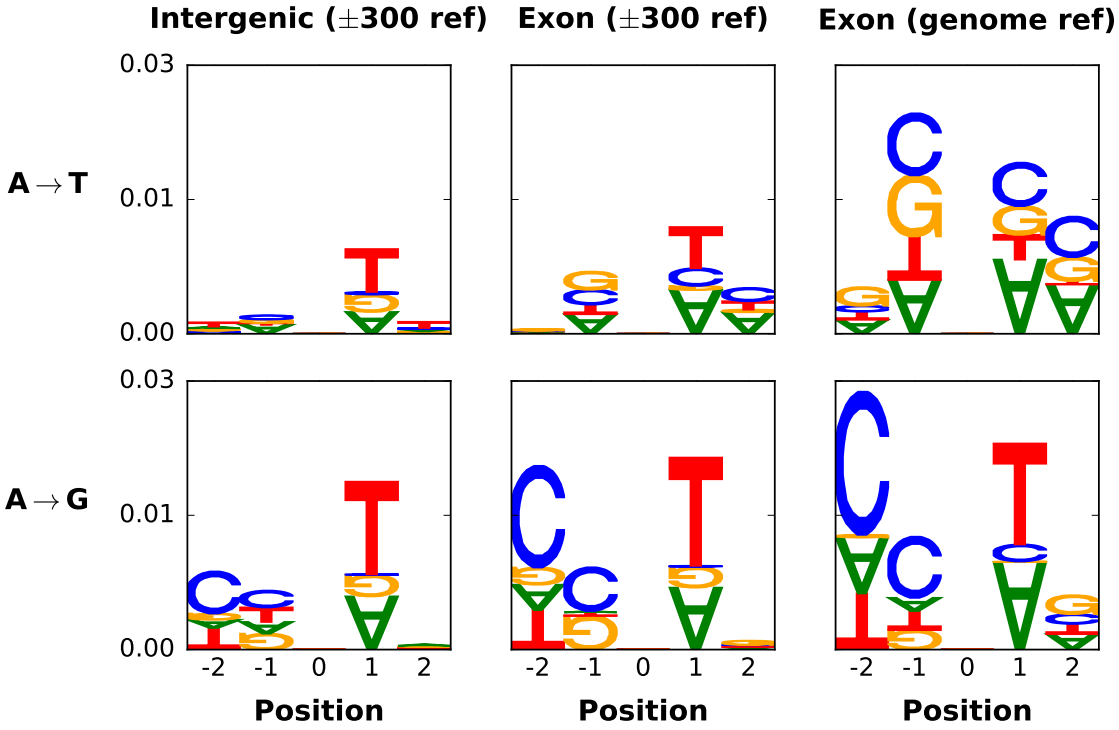


Figure S5: Using the genome as the reference introduces bias. "±300 ref" uses reference bases selected at random within ±300bp of the mutated base. "genome ref" uses reference bases selected at random from the entire human genome. Only autosomal mutations were used for the analysis.