Supplementary Figures

Figure S1. Overall nucleotide content of flank and core exonic regions.

Figure S2. SSM density in 5' versus 3' exon flanks. A) TCGA exome data. B) ICGC whole genome sequencing data.

Figure S3. Scatter plot of SSM density in flank versus core regions in individual tumours. For each tumour, levels of SSM in flank and core regions were computed. Green line represents y=x line and blue line represents least square fit ("lm" in R) with grey band representing 1 SEM.

Figure S4. Comparison of SSM density in flank region adjacent to canonical splice sites compared to exonic cores. We repeated the analysis shown in Figure 1a and 1b but did not include canonical splice sites (i.e. the 3 nucleotides from the ends of each exon) for both TCGA exomic data (a) and ICGC WGS data (b). Statistical significance was computed using Wilcoxon Signed Rank test. The y-axis unit is SSM rate per bp per tumour. Error bars, 95% confidence interval computed by bootstrapping.

Figure S5. Comparison of SSM density between flanks and cores for 96 trisnps. For this analysis we considered tumour samples from top 12 highly mutated cancer types (SKCM, UCEC, STAD, COAD, LUSC, BLCA, LUAD, READ, ESCA, PAAD, CESC, HNSC). For each SSM in the flank or the core region, we identified it’s a nucleotide before and after (i.e. trisnp). For each tumour that had atleast 5 trisnps, we quantified the frequency of each of the 96 possible trisnps. The y-axis unit is frequency of each trisnp per tumour. Error bars, 95% confidence interval computed by bootstrapping.

Figure S6. SSM density in Alternate versus Constitutive exon flanks in which CG dinucleotides are masked out. Similar to Figure S2 except that flanks were partitioned into Alternative or Constitutive using classification given in HexEvent database.

Figure S7. Overall nucleotide content of ESE and non-ESE exonic regions. ESE motifs from the INT3 database were identified in each exon as described in methods.

Figure S8. Comparison of SSM density between ESE and non-ESE for 96 trisnps. Similar to Figure S5, for this analysis we considered tumour samples from top 12 highly mutated cancer types (SKCM, UCEC, STAD, COAD, LUSC, BLCA, LUAD, READ, ESCA, PAAD, CESC, HNSC). For each SSM in the ESE or the non-ESE region, we identified it’s a nucleotide before and after (i.e. trisnp). For each tumour that had atleast 5 trisnps, we quantified the frequency of each of the 96 possible trisnps. The y-axis unit is frequency of each trisnp per tumour. Error bars, 95% confidence interval computed by bootstrapping.
Supplementary Figure S1

Proportion

0.2 0.3 0.4 0.5 0.6

A+T G+C A C G T

outer
middle

Nucleotide
Figure S2

(a) Exon flank (5') vs Exon flank (3')

(b) Exon flank (5') vs Exon flank (3')

SSM per bp per tumour

-dots- Exon flank (5')

-dots- Exon flank (3')

P-values:
- P=0.021
- P=0.088
- P=0.22

1 \times 10^{-8}
1 \times 10^{-9}
1 \times 10^{-7}
1 \times 10^{-6}
1 \times 10^{-5}

ACCBLCABRCACESCCOADDLBCESCAHNSCLUADLUSCPAADREADSKCMSTADUCEC

Exon flank (5') Exon flank (3')

Supplementary Figure S2
Figure S3

Number of SSM in exon flanks vs. Number of SSM in exon cores for different cancer types:

- ACC
- BLCA
- BRCA
- CESC
- COAD
- DLBC
- ESCA
- HNSC
- LUAD
- LUSC
- PAAD
- READ
- SKCM
- STAD
- UCEC

Linear fit: $y = mx + b$

Line $x = y$
Figure S4

(a) Flank_excluding_canonical_splice_sites vs. Core

(b) SSM per bp per tumour

ACCB | BLCA | BRCA | CESC | COAD | DLBC | ESCA | HNSC | LUAD | LUSC | PAAD | READ | SKCM | STAD | UCEC

P-values:
P = 0.0026
P = 0.028
P = 0.16
P = 0.0027
P = 0.0047
P = 0.00013
P = 0.82
P = 3.3e−05
P = 0.29
P = 0.0075
P = 0.00049
P = 2.5e−10
P = 0.0039
P = 0.00091
P = 0.4
P = 1.9e−08
P = 3.2e−11
P = 1.5e−08
Figure S5

Flank Core

SSM per bp per tumour

Supplementary Figure S5
Figure S6

Supplementary Figure S6

SSM per bp per tumour

- Flank (CG masked)
- Core (CG masked)

P=0.00068
P=0.04
Figure S7

A+T  G+C  A  C  G  T

Proportion

Nucleotide