

B)

Vector B
Shared exon counts

Vector A Shared exon counts	1	2	3	4	5
1	892,250	378,742	317,543	325,946	576,377
2		609,149	224,971	231,320	404,272
3			517,062	194,795	337,672
4				516,352	381,412
5					931,546

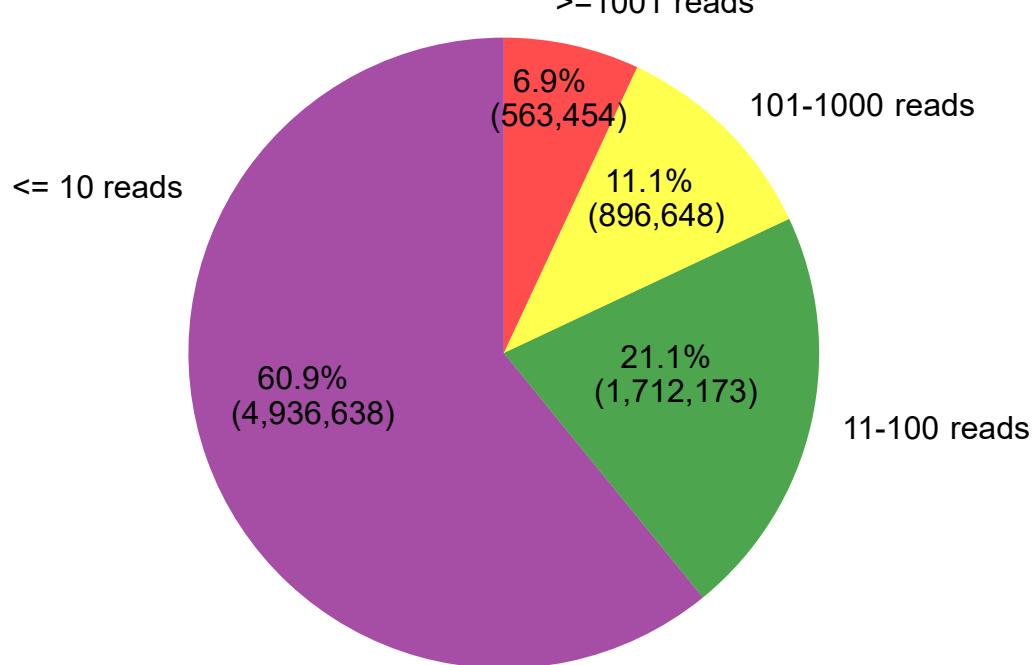
Figure S1

Figure S1. Exon trapping vector design and exon capture.

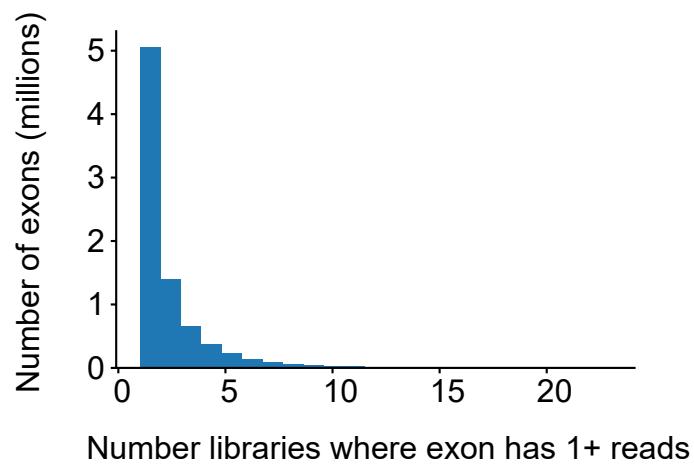
A) Diagram depicting exon trapping vectors. Vector 1-3 were made by gene synthesis and gateway cloning. Vectors 4 and 5 are derivatives of vector 1, with a 5'SS mutation to weaken the associated MaxEntScan score or a mutation to scramble a putative RBFOX binding site, respectively.

B) Matrix depicting exon detection in each of the vectors and overlap between vectors. The diagonal shows how many of the 1.25 million exons were detected at least once in at least one library generated from each of the vectors. Off-diagonal entries indicate how many exons were detected in both libraries (row and column).

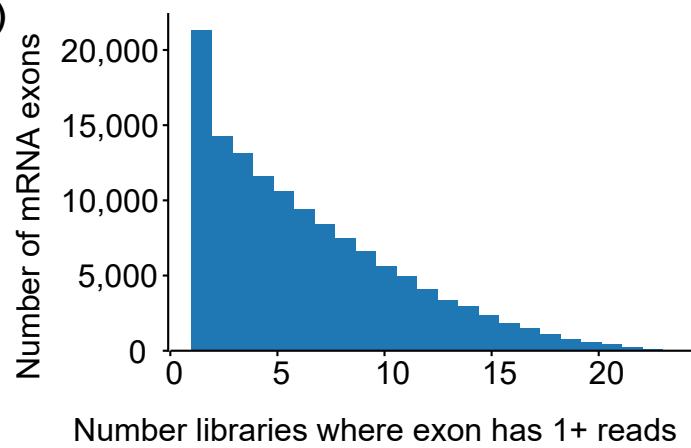
A)



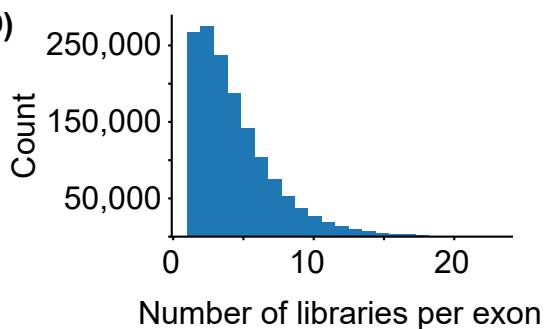
B)



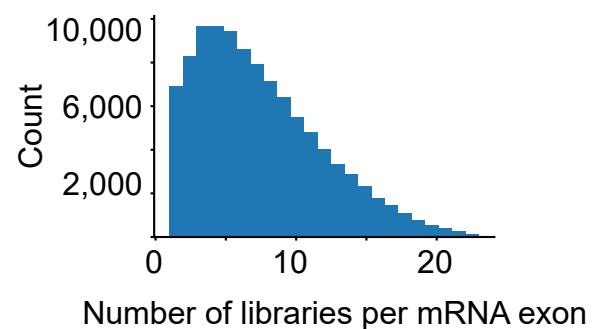
C)



D)



E)



F)

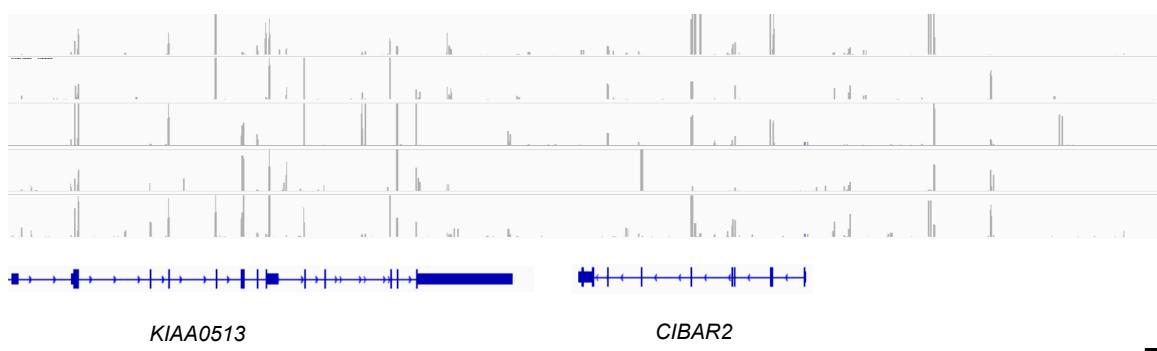


Figure S2

Figure S2. Exon sequencing reads.

- A) Pie chart depicting the fraction and number of sequenced exons at different read count thresholds.
- B) Histogram of the number of exons found dependent on the number of sequencing libraries each exon is found within for all exons.
- C) Histogram of the number of exons found dependent on the number of sequencing libraries each exon is found within for mRNA exons.
- D) Same as B) except for exons with sequencing read counts of 100 or more.
- E) Same as C) except for exons with sequencing read counts of 100 or more.
- F) Mapped sequencing read counts (combined forward and reverse strand pileups) for the different vector backbones for a region containing *KIAA0513* and a portion of *CIBAR2* (display region coordinates: Chr16:85,062,938-85,134,585). The zoom-in region corresponds to exon 7 of *CIBAR2*.

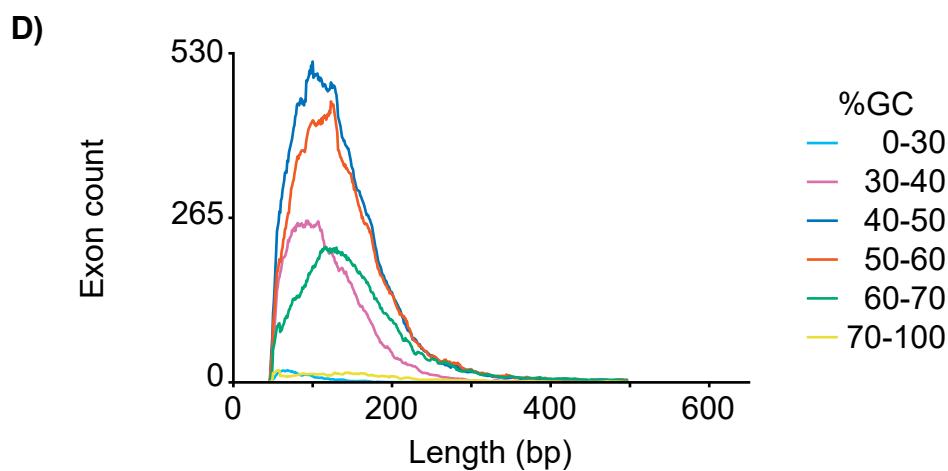
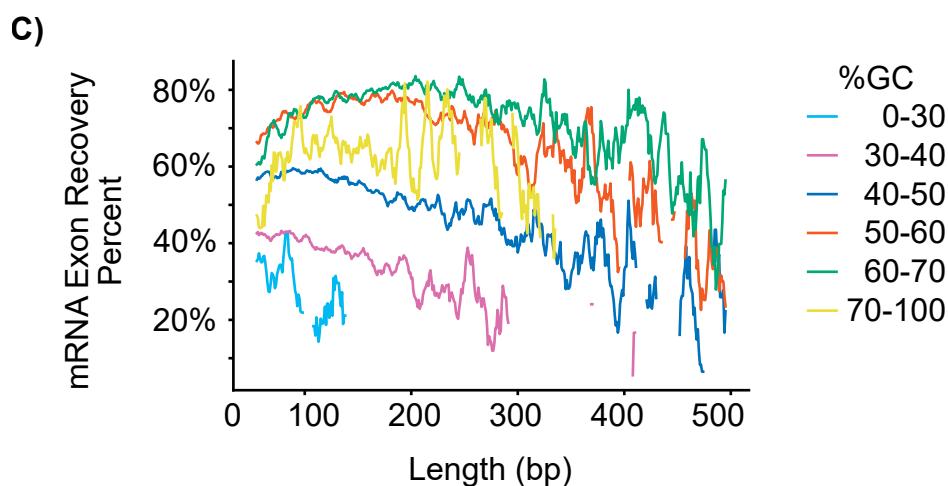
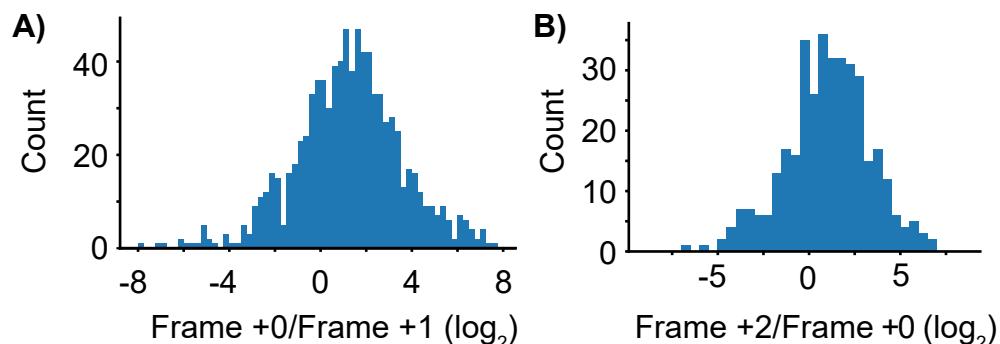


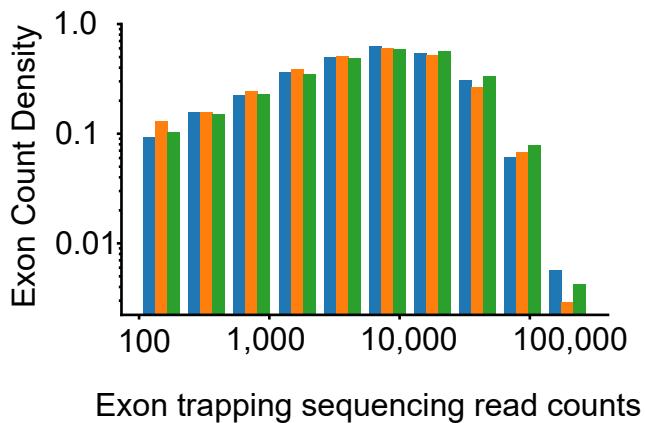
Figure S3

Figure S3. Exon recovery.

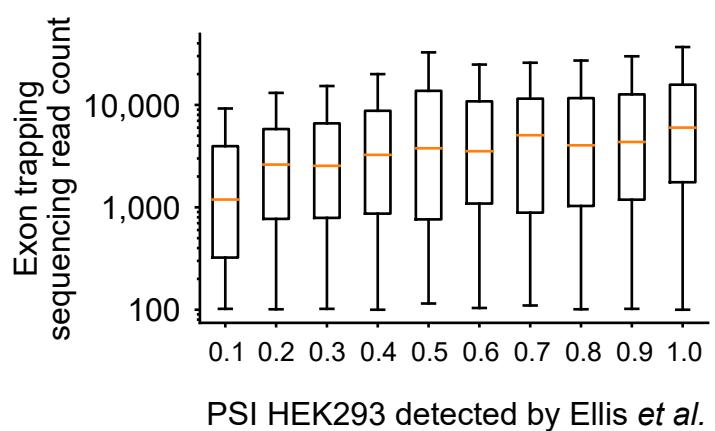
- A) Histogram depicting ratio of exon sequencing read counts for different reading frames depending on the presence or absence of stop codons. Exons have no stop codon in frame +0 and at least one stop codon in frame +1.
- B) Histogram depicting ratio of exon sequencing read counts for different reading frames depending on the presence or absence of stop codons. Exons have no stop codon in frame +2 and at least one stop codon in frame +0.
- C) Line plots depict the distribution mRNA exon recovery percent as a function of exon length, subset by exon GC content. Plots have a 9 bp smoothing window applied.
- D) Line plots depict the distribution mRNA exon lengths, subset by exon GC content. Plots have a 9 bp smoothing window applied.

A)

- Housekeeping Genes Hounkpe *et al.*
- Top Express Genes Nieborak *et al.*
- All other Genes



B)



C)

- Found in exon trapping data
- Missed in exon trapping data

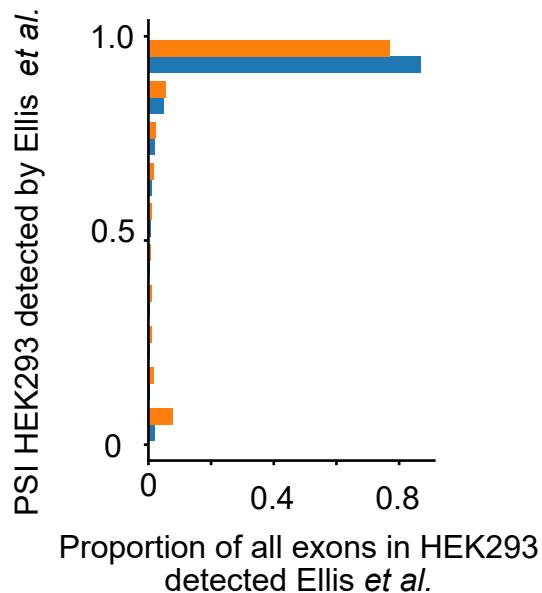


Figure S4

Figure S4. Exon trapping exons HEK293 cell line properties.

- A) Bar plots depicting the sequencing read counts of exon trapping exons from three groups of transcribed HEK293 transcripts: housekeeping genes, remaining top 20% HEK293 expression, and remaining bottom 80% expression genes.
- B) Boxplot series depicting exon trapping sequencing read count as a function of exon PSI value ranges associated with those exons, for genes expressed in HEK293.
- C) Horizontal bar series depicting the proportion of exon belonging to different PSI values in expressed HEK293 genes for two categories of exons: exons found by exon.

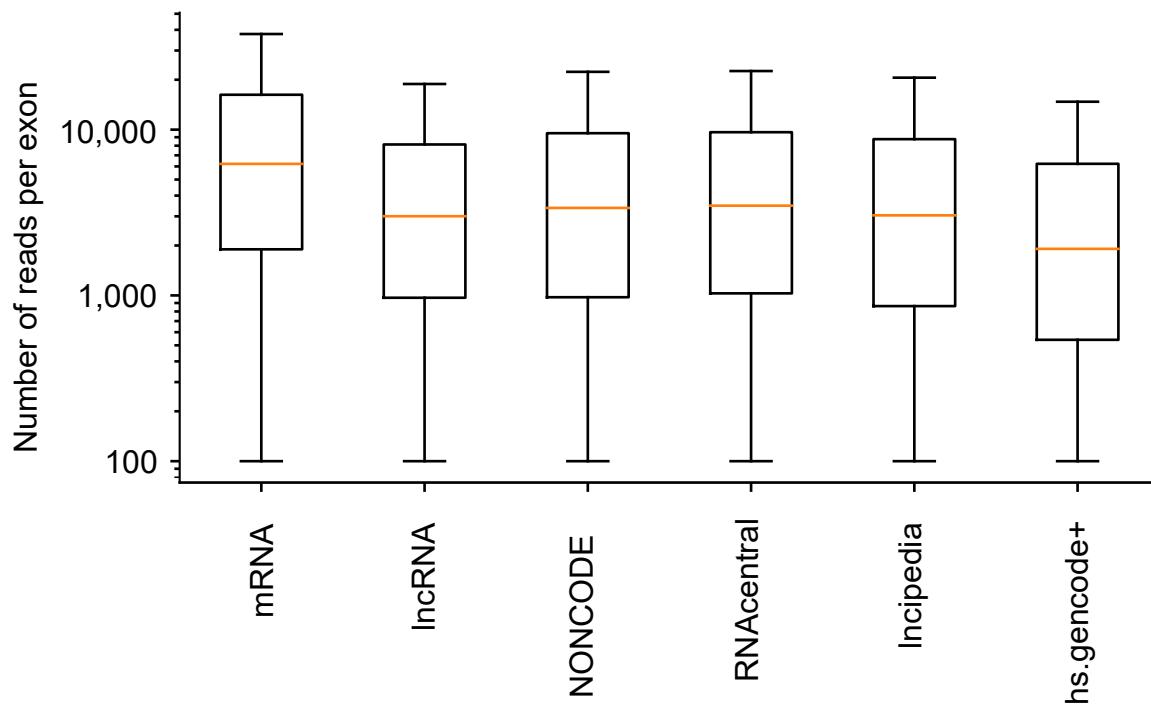


Figure S5

Figure S5. Exon recovery of different lncRNA annotations.

Boxplots showing the number of exon trapping sequencing reads for exons of mRNA, GENCODE lncRNA and exons of different lncRNA annotations with the exons also contained in GENCODE lncRNA removed.

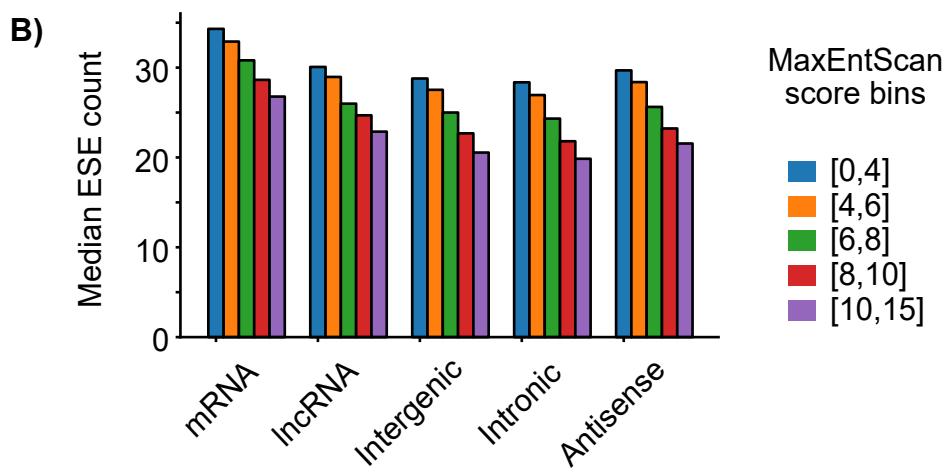
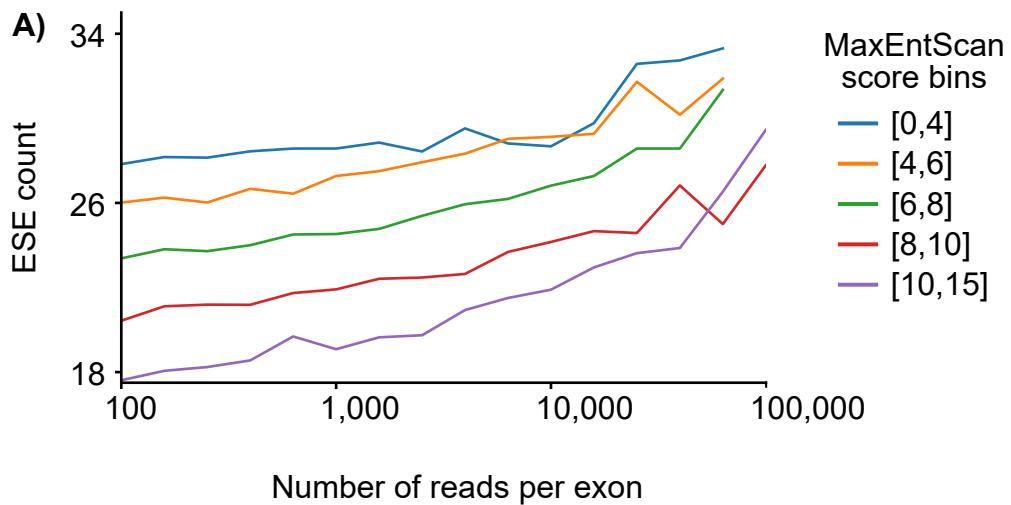


Figure S6

Figure S6. Exons splicing enhancer frequencies for different splice site strengths.

- A) Line plots showing binned exon read counts versus corresponding median exon ESE counts for intergenic exons, selecting groups defined by both splice site scores being in the specified range. Exons are binned by their sequencing read counts using logarithmic bins ranging from 2 to 5 with 0.2 steps.
- B) Bar plots of the median number of splicing enhancers for different exon genomic region categories with exons subdivided into different splice site bins. Exons in splice site bins are required to have both MaxEntScan scores (3'SS and 5'SS) within the indicated range.

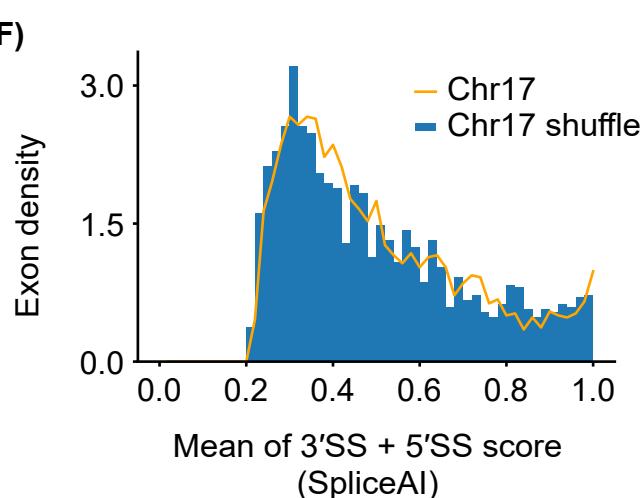
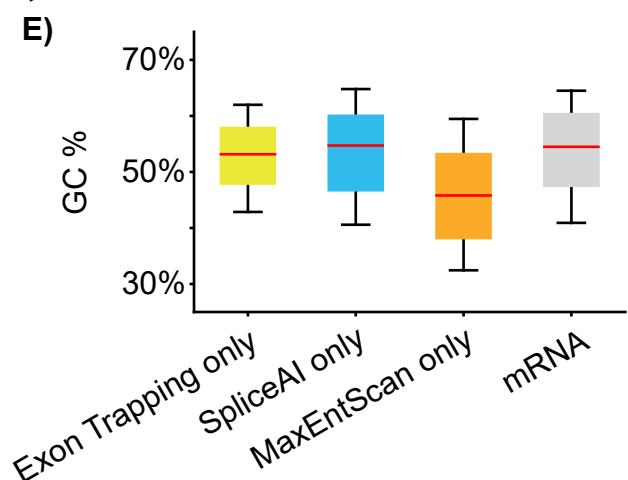
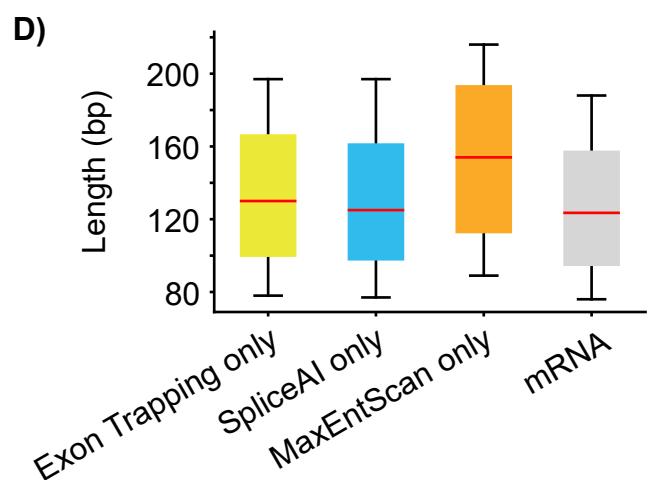
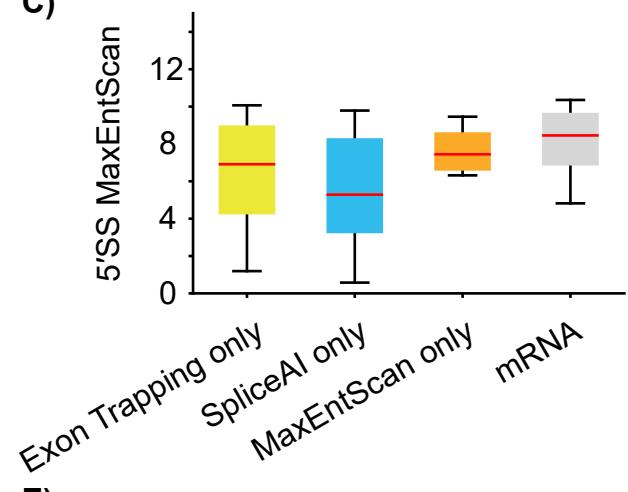
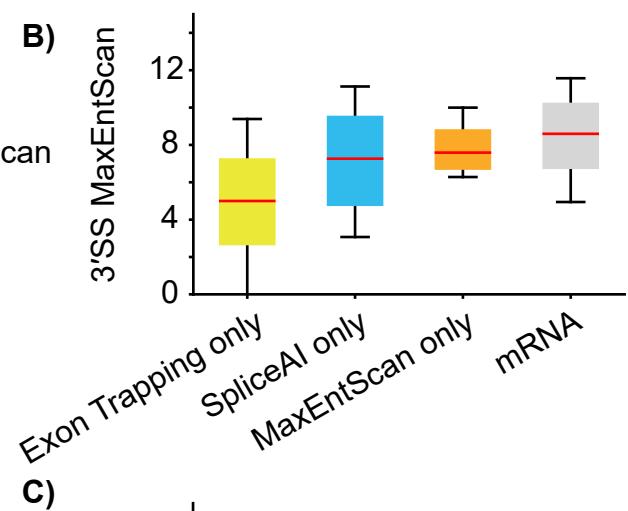
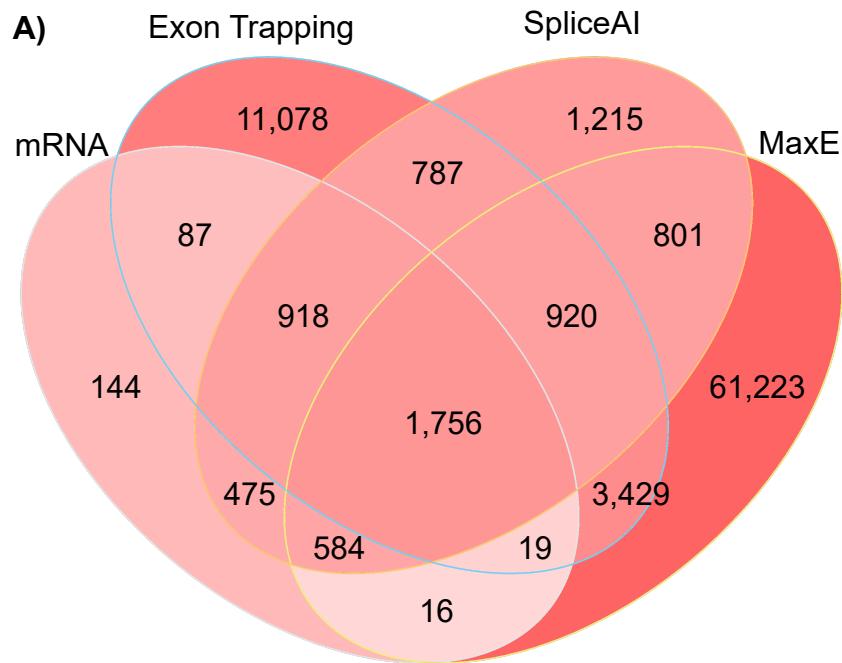


Figure S7

Figure S7. Dissection of overlap and properties of exons detected using different approaches.

- A) Venn diagram depicting Chromosome 17 forward strand exons found by different exon calling approaches and that share the same 3'SS, but not necessarily 5'SS (annotated mRNA & lncRNA internal exons, exons found by exon trapping, MaxEntScan based exon finding approach, and SpliceAI based exon finding approach). Venn overlaps colored red with intensity determined by $(\log_{10} \# \text{exons})/8$.
- B) Boxplot of MaxEntScan 3'SS scores for exons belonging only to the indicated exon finding approach.
- C) Boxplot of MaxEntScan 5'SS scores for exons belonging only to the indicated exon finding approach.
- D) Boxplot of exon lengths for exons belonging only to the indicated exon finding approach.
- E) Boxplot of GC content for exons belonging only to the indicated exon finding approach.
- F) Histogram showing exon density of mean of 3'SS + corresponding 5'SS SpliceAI score for each unannotated exon found by SpliceAI on Chromosome 17 forward strand (line) and for SpliceAI found exons on a shuffled Chromosome 17 forward strand (bars) for those exons found with SpliceAI score > 0.2 . Histograms are area normalized to 1 and depict exon density.

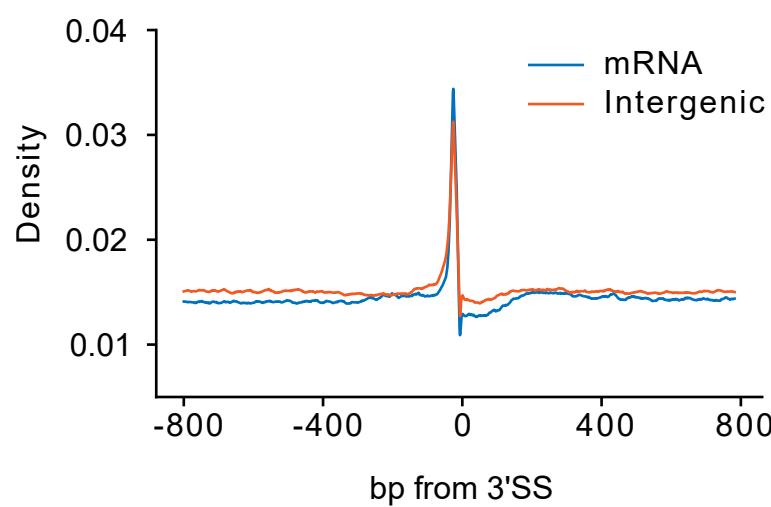
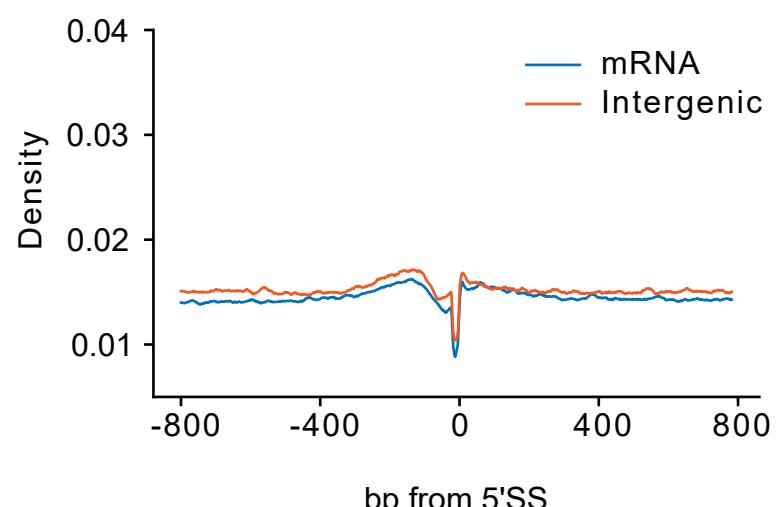
A)**B)**

Figure S8

Figure S8. Intronic splice site silencer frequencies around splice sites.

- A) Line plots depicting the frequency of ISS sequences found at a given bp distance from the 3'SS (left) or the 5'SS (right) for mRNA or intergenic exons. Plots are smoothed with a 19 bp moving average. ISS sequences from (Wen et al. 2010).
- B) Line plots depicting the frequency of ISS sequences found at a given bp distance from the 3'SS (left) or the 5'SS (right) for exons unique to different exon finder approaches with the graph of mRNA exon also shown. Plots are smoothed with a 19 bp moving average. ISS sequences from (Wen et al. 2010).