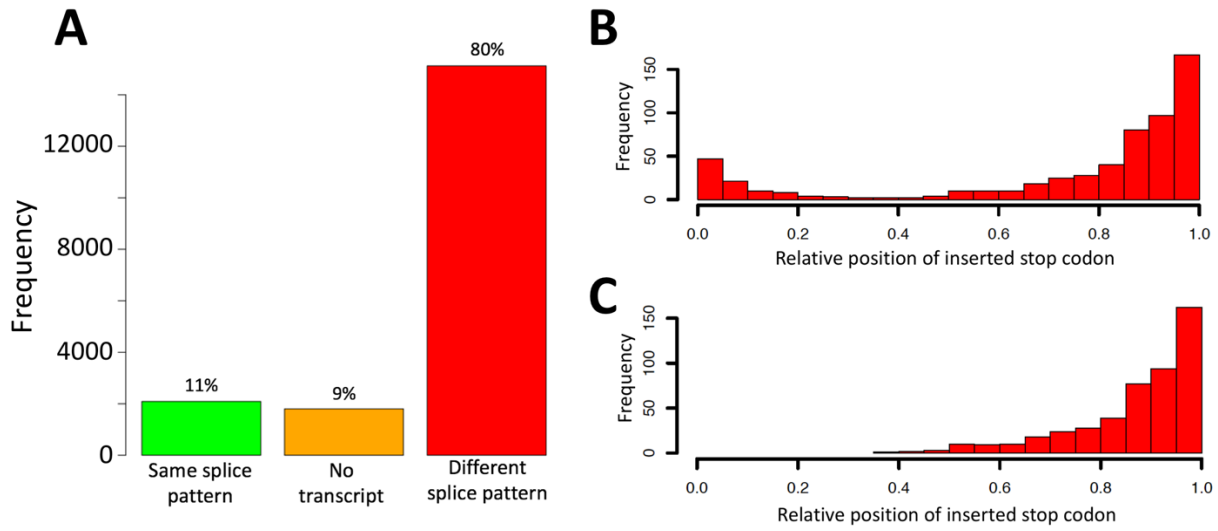
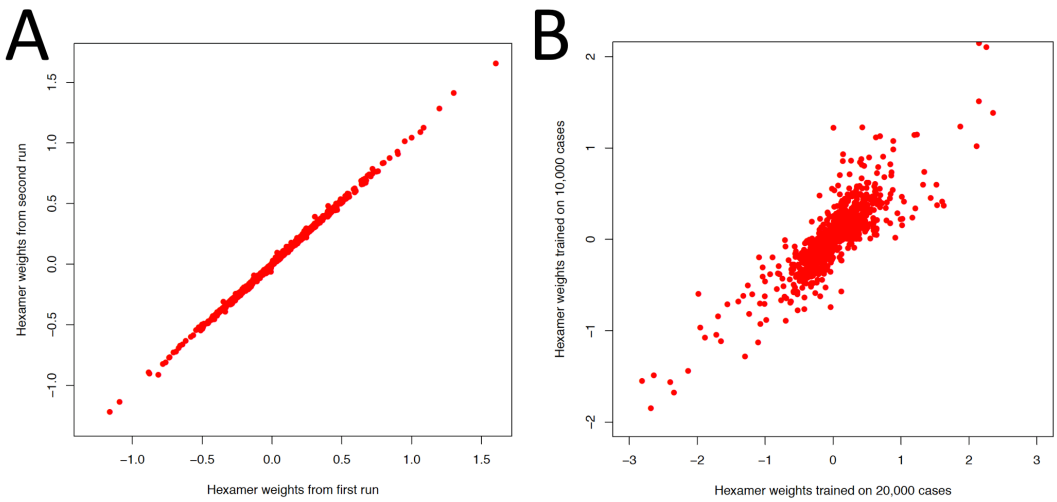
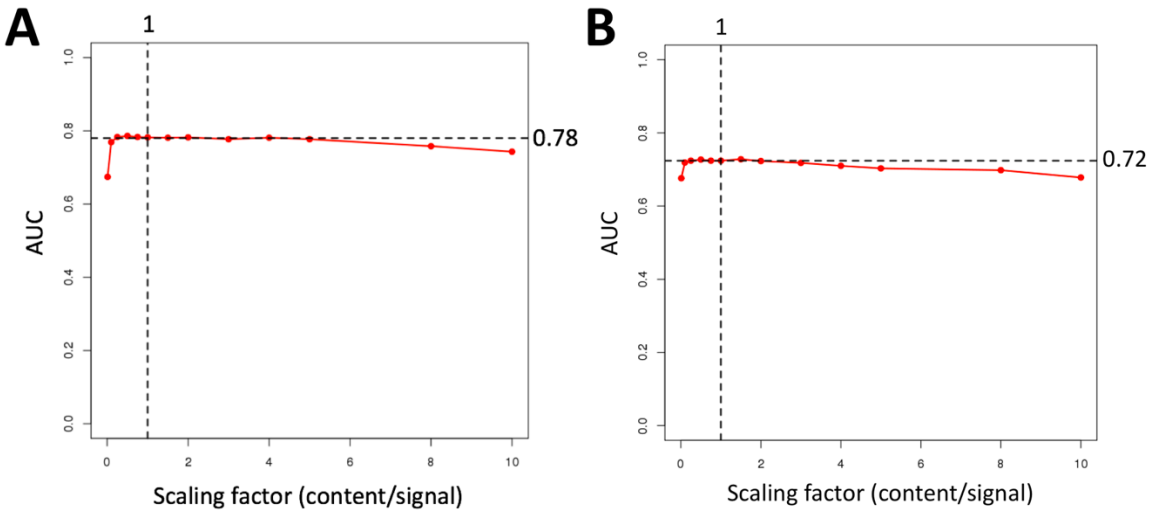
Supplementary Figures for Majoros et al., "Predicting Gene Structure Changes Resulting from Genetic Variants via Exon Definition Features."



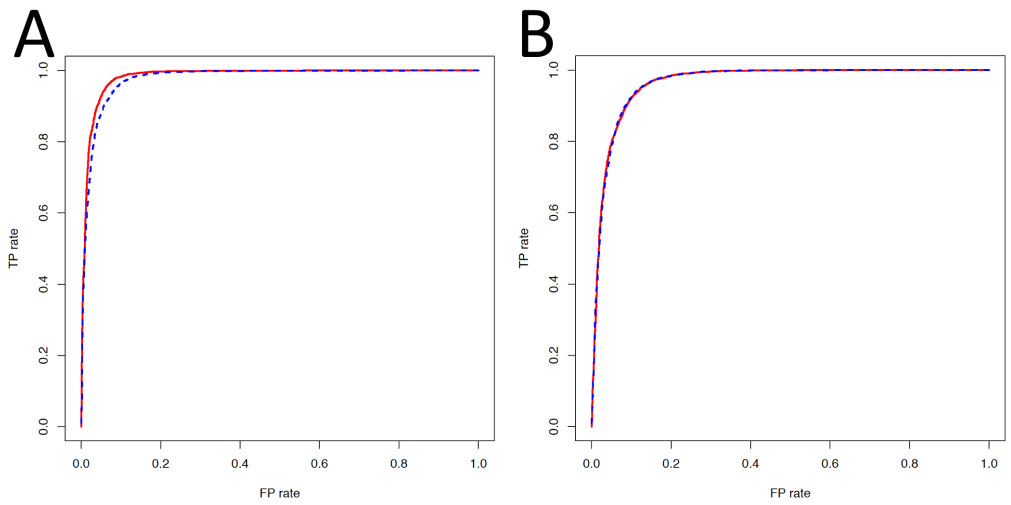
Supplementary Fig. S1: **(A)** Results of running an HMM gene finder on 19,000 broken genes. The gene finder was run on each gene, then a stop codon was inserted in a random location in the CDS without creating a splice site, and the gene finder was run on the modified sequence. In 11% of cases the gene finder predicted the same splice pattern on both the original sequence and the sequence modified to contain a premature stop. In 9% of cases, the gene finder predicted that no gene was present after the stop was inserted. In the remaining 80% of cases, the gene finder predicted a different splice pattern after the stop codon was inserted. **(B)** Relative position of inserted stop codon (relative to the spliced transcript) in cases in which the gene finder predicted the same splice pattern. There was a strong enrichment for stop codons inserted near the end of the coding segment, in which only the terminal portion of the protein would be affected, as well as a weaker enrichment near the beginning of the coding segment, in which the gene finder was able to find another start codon in the same reading frame that avoided the inserted stop codon. **(C)** Relative position of inserted stop codon in cases in which both the splice pattern did not change and the start codon was not changed.



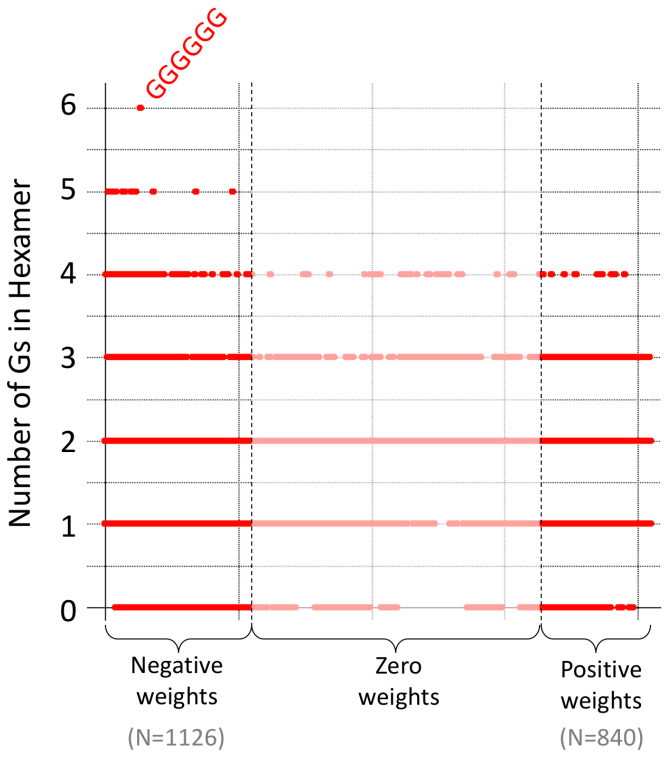
Supplementary Fig. S2: Reproducibility of logistic regression training for content sensors. **(A)** Hexamer weights trained on 10,000 exon-intron pairs, versus weights for the same hexamers from an independent logistic regression applied to the same training cases. **(B)** Hexamer weights estimated from 20,000 training cases (x-axis) and weights for the same hexamers estimated from a subset of 10,000 training cases (y-axis).



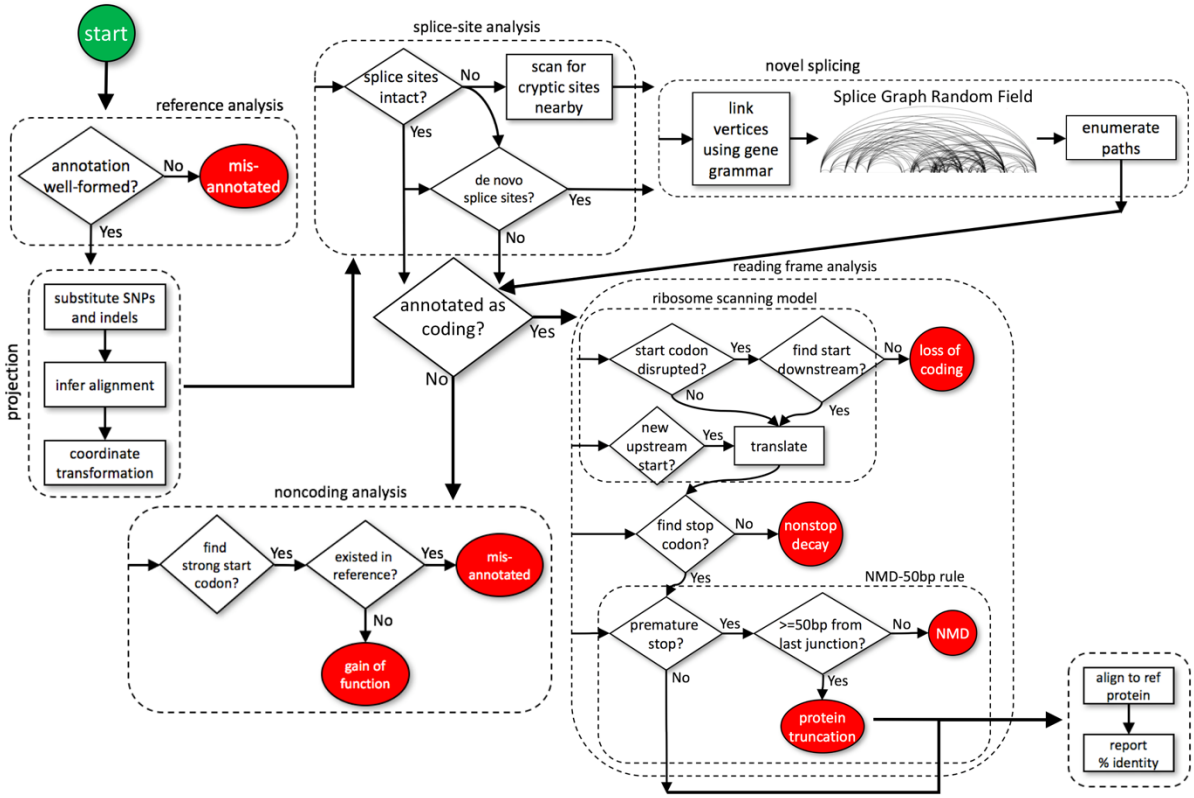
Supplementary Fig. S3: **(A)** Effect of scaling factor $r_{content/signal}$ on AUC for SGRF model applied to Thousand Genomes individual HG00096, for the human logistic content sensor. **(B)** AUC versus scaling factor for the SGRF using the minigene content sensor.



Supplementary Fig. S4: Predictive accuracy of logistic signal sensors versus probabilistic weight matrices (PWMs). **(A)** AUC for classification of annotated donor splice sites versus decoy sites, using logistic signal sensor (red) and PWM (blue). **(B)** AUC for classification of annotated acceptor splice sites versus decoy sites, using logistic signal sensor (red) and PWM (blue).



Supplementary Fig. S5: Number of Gs in 4096 hexamers (y-axis) as a function of logistic weights; points are sorted along the x-axis by weight.



Supplementary Fig. S6: Overview of the ACE+ framework. ACE+ constructs personal genomes from a phased VCF file and projects reference annotations onto the alternate sequence. The SGRF scores predicted alterations to splicing patterns. Each predicted pattern is evaluated as to its potential for multiple forms of loss of function, including NMD, nonstop decay, loss of coding potential, and protein change.