Learning the sequence determinants of alternative splicing from millions of random sequences

Most human transcripts are alternatively spliced, and many disease-causing mutations affect RNA splicing. Toward better modeling the sequence determinants of alternative splicing, we measured the splicing patterns of over two million (M) synthetic mini-genes, which include degenerate subsequences totaling over 100 M bases of variation. The massive size of these training data allowed us to improve upon current models of splicing, as well as to gain new mechanistic insights. Our results show that the vast majority of hexamer sequence motifs measurably influence splice site selection when positioned within alternative exons, with multiple motifs acting additively rather than cooperatively. Intriguingly, motifs that enhance (suppress) exon inclusion in alternative 5′ splicing also enhance (suppress) exon inclusion in alternative 3′ or cassette exon splicing, suggesting a universal mechanism for alternative exon recognition. Finally, our empirically trained models are highly predictive of the effects of naturally occurring variants on alternative splicing in vivo.