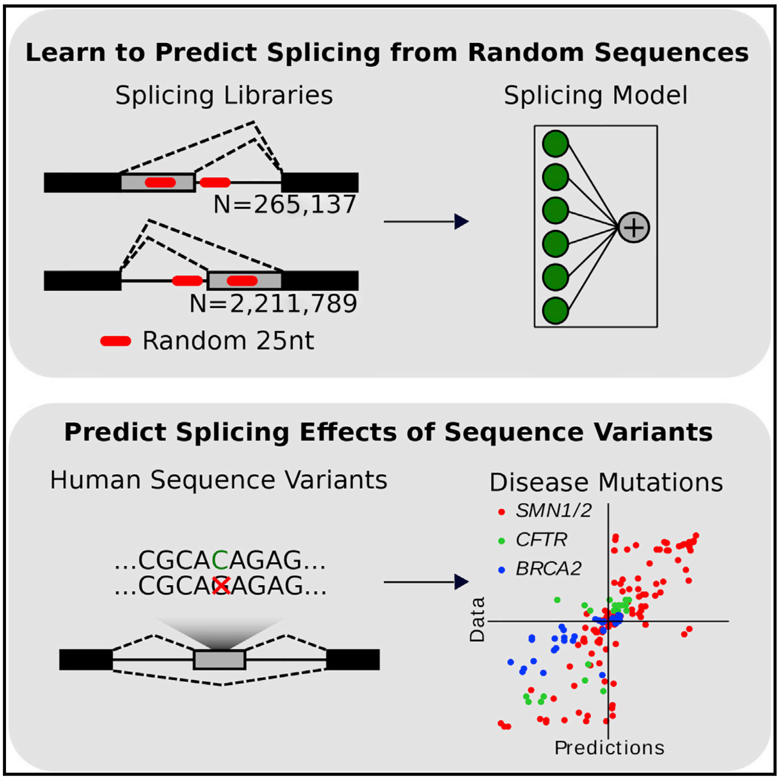
Cell

Article

Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences

Graphical Abstract



Highlights

- Measured splicing patterns of nearly 2M synthetic alternatively spliced mini-genes
- *cis*-regulatory elements primarily act additively rather than cooperatively
- Model trained only on synthetic data predicts effects of human SNPs on isoform ratios
- Model of alternative 5′ and 3′ splicing predicts effect of SNPs in skipped exons

Authors

Alexander B. Rosenberg, Rupali P. Patwardhan, Jay Shendure, Georg Seelig

Correspondence

gseelig@uw.edu

In Brief

A combination of synthetic biology and machine learning approaches identifies universal rules of RNA splicing and enables the accurate prediction of the effects of disease-related human SNPs on isoform levels.

Accession Numbers

GSE74070

Rosenberg et al., 2015, Cell *163*, 698–711
October 22, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.09.054

Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences

Alexander B. Rosenberg,[1] Rupali P. Patwardhan,[2] Jay Shendure,[2] and Georg Seelig[1,3,*]
[1]Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
*Correspondence: gseelig@uw.edu

http://dx.doi.org/10.1016/j.cell.2015.09.054

SUMMARY

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.

INTRODUCTION

Alternative splicing is a major source of proteome diversity in eukaryotes (Nilsen and Graveley, 2010). Regulation of alternative splicing is vital to cellular processes that depend on the precise ratios of isoforms. For example, mutations that lead to even subtle changes in the ratio of *MAPT* isoforms 3R and 4R cause an inherited form of dementia (Garcia-Blanco et al., 2004). While new sequencing technologies have enabled the comprehensive cataloging of human genetic variation, the functional consequences of these variants on even molecular phenotypes, such as alternative splicing, remain poorly predictable.

Experimentally testing the consequence of every possible genetic variant on endogenous alternative splicing is impractical, motivating the development of predictive models of the "splicing code." The core splicing signals—5′ splice donor, 3′ splice acceptor, branchpoint, and polypyrimidine tract—form the basis of the splicing code; they are required for recognition of intron-exon boundaries and for correct intron removal by the splicing machinery. Computational methods have been developed to score the likelihood of splicing at different splice donor and acceptor sequences (Yeo and Burge, 2004). Splice regulatory elements (SREs)—sequence motifs in exons or introns shown to regulate splicing—form the next level of regulatory information. SREs typically regulate alternative splicing by binding *trans*-acting splice factor proteins (Ule et al., 2006; Wang et al., 2013). Depending on their position and mode of action, SREs are classified as exonic splice enhancers (ESEs), exonic splice silencers (ESSs), intronic splice enhancers (ISEs), or intronic splice silencers (ISSs). Examples of SREs have been identified computationally by analyzing motif enrichment near splice sites (Castle et al., 2008; Fairbrother et al., 2002; Zhang and Chasin, 2004) or sequence conservation between species (Goren et al., 2006). Recently, a deep neural network was trained on exon skipping events in the genome to generate a comprehensive model of the splicing code that can be used to predict exon inclusion percentages (Xiong et al., 2014). Despite this progress, current models of alternative splicing do not perform well enough to be used in clinical genetics (e.g., to reclassify "variants of uncertain significance"), and many machine learning strategies result in "black boxes" that limit mechanistic insight.

We hypothesized that a model of alternative splicing learned from very large libraries of synthetic sequences could outperform models trained only on the genome. Current technology makes it possible to create and test gene libraries with millions of synthetic sequences—orders of magnitude more than the number of alternative splice events in the human genome. In other applications of machine learning, such as computer vision, predictive power has increased greatly with access to larger datasets (Le et al., 2012). Previous work supports the idea that synthetic gene libraries with extensive and targeted variation can provide mechanistic insight into biological phenomena. In vivo (Culler et al., 2010; Wang et al., 2012) and in vitro (Yu et al., 2008) randomized selections have identified potential SREs. Massively parallel reporter assays (MPRAs) that combine next-generation sequencing with extensive variation have been applied to study transcription (Melnikov et al., 2012; Patwardhan et al., 2012; Patwardhan et al., 2009; Sharon et al., 2012; Smith et al., 2013; White et al., 2013), translation (Noderer et al., 2014), mRNA stability (Oikonomou et al., 2014), and even alternative

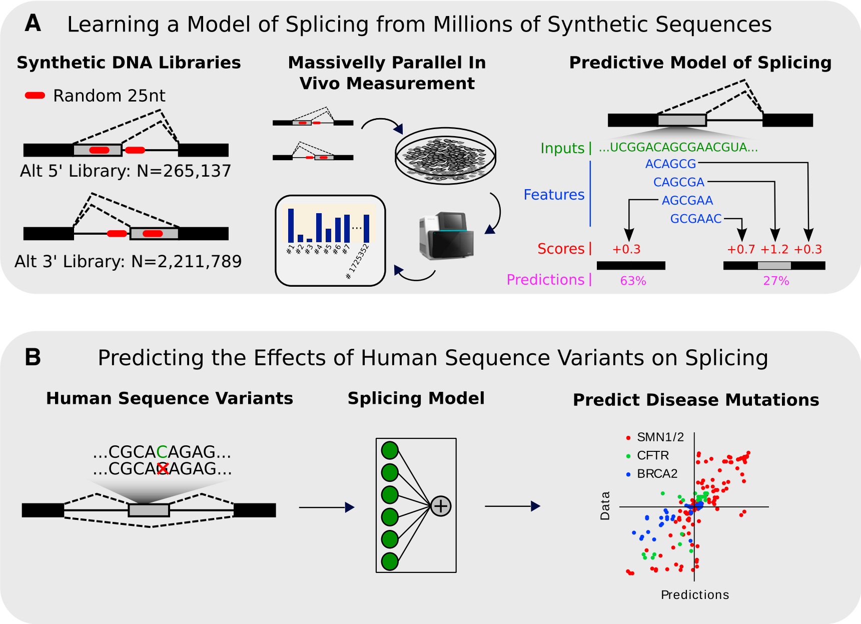

A
Learning a Model of Splicing from Millions of Synthetic Sequences
Synthetic DNA Libraries
Random 25nt
Alt 5' Library: N=265,137
Alt 3' Library: N=2,211,789
Massivelly Parallel In Vivo Measurement
#1 #2 #3 #4 #5 #6 #7 # 1725352
Predictive Model of Splicing
Inputs
...UCGGACAGCGAACGUA...
Features
ACAGCG
CAGCGA
AGCGAA
GCGAAC
Scores
+0.3
+0.7 +1.2 +0.3
Predictions
63%
27%
B
Predicting the Effects of Human Sequence Variants on Splicing
Human Sequence Variants
...CGCACAGAG...
...CGCAXAGAG...
Splicing Model
Predict Disease Mutations
SMN1/2
CFTR
BRCA2
Data
Predictions

Figure 1. A Predictive Model of Alternative Splicing Learned from Millions of Synthetic Sequences

(A) Two libraries with either alternative 5′ or 3′ splice sites were constructed with two 25-nt randomized regions. The library was transfected into human cells, and massively parallel measurement of isoform ratios was performed with RNA-seq. These two datasets were used to learn a predictive model of alternative splicing. The model takes a sequence as input, which is then converted to 6-mer features. A score for each 6-mer is learned and then used to predict the fractional usage of each splice site.

(B) When human sequence variants are fed to the model as inputs, the model makes more accurate predictions than the current state of the art algorithms.

splicing (Ke et al., 2011). However, MPRA studies to date have overwhelmingly focused on measuring the consequences of variants in endogenous sequences (e.g., saturation mutagenesis) or on validating predicted activities (e.g., enhancers predicted by the ENCODE project). There are thus far few, if any, examples of predictive biological models learned entirely on MPRA data.

To test whether it is possible to learn predictive biological models from synthetic data alone, we developed an MPRA that measures alternative splice site selection in a highly complex library of “degenerate introns” (Figure 1A). We added degenerate regions into an otherwise fixed sequence context, ensuring that any differences in gene expression can be causally attributed to the degenerate region. We created two libraries, one with alternative 5′ splice donors consisting of 265,137 members and one with alternative 3′ splice acceptors containing 2,211,739 members. We transfected these libraries to human cells, performed RT-PCR and RNA sequencing (RNA-seq) to quantitatively measure isoform ratio for all mini-genes and used the results to learn a predictive model of alternative splicing. To assess the quality of the resulting model, we predicted the effects of human sequence variants on isoform levels and compared our results to available experimental data (Figure 1B). We tested variants in alternative 5′ splicing events, both within the alternative splice donors themselves and within the alternative exon. Although our MPRA did not include a skipped exon library, our model also predicted with high accuracy the effect of sequence variants in skipped exons.

RESULTS

Molecular Phenotyping of Millions of Alternatively Spliced Mini-Genes Containing Random Sequences

We chose to study both alternative 5′ and alternative 3′ splice site selection. In the case of alternative 5′ splicing, we first generated a complex library by introducing 2 × 25 nt fully degenerate regions into a single-intron plasmid mini-gene (Figure 2A). Specifically, the intron was designed with two competing splice donors separated by 44 nt; one degenerate region was inserted between the splice donors and the other downstream of the second donor. Neither degenerate sequence overlapped a splice donor. The mini-genes contained an additional degenerate 20 nt barcode in the 3′ UTR. This barcode was used to create a look-up table linking barcodes and intronic sequences. Thus, even when both degenerate regions were spliced out, their sequences could be recovered from the barcode sequence (Figure 2A). To maximize intron sequence variability, we constructed and sequenced a complex library of 265,137 such mini-genes. Thus, over 13 Mb of unique intronic sequence are represented within the degenerate regions of this library (265,137 × 50 nt).

In the case of alternative 3′ splicing, we inserted 2 × 25 nt fully degenerate regions into a single-intron system designed to have two alternative 3′ splice sites (Figure 2C). The degenerate regions did not overlap either splice acceptor, but the upstream degenerate region did overlap the typical position of the first splice acceptor’s branchpoint (−44:−19 relative to SA_1). Similarly to the alternative 5′ library, we included an additional degenerate 20-nt barcode in the 3′ UTR. The alternative 3′ library contained 2.2 million unique mini-genes encompassing over 110 Mb of unique sequence variation (2,211,739 × 50 nt).

We transfected the pooled libraries of plasmids into HEK293 cells and then quantified isoform ratios with targeted RNA-seq. To identify both the isoform and originating plasmid of each mRNA, we used paired-end sequencing with one read across the exon junction and the other read across the 3′ UTR barcode (Figures 2A and 2C). We used 13 million reads for the alternative 5′ library and 5.4 million reads for the alternative 3′ library. We were then able to calculate the isoform ratios for each mini-gene in each library. We averaged 50.0 reads per mini-gene in the 5′ library with reads mapping to 265,044/265,137 (99.96%) of all mini-genes. On the other hand, in the 3′ library we averaged only 2.47 reads per mini-gene with reads mapping to 1,686,096/2,211,739 (76.23%) of all mini-genes.

Degenerate Sequences in Both Libraries Strongly Influence Isoform Ratios

In the alternative 5′ library, isoforms were present from several different splicing events. The most upstream splice donor (SD_1) was used on average 22.4% of the time, while SD_2 was used 50.0% of the time (Figure 2B). The remaining transcripts were spliced at new splice donors inserted into the randomized regions (11.3%), a cryptic splice donor site (SD_{CRYPT}) 35 nt downstream of SD_2 (7.9%), or not spliced at all (8.4%). However, as evidenced by the broad distributions of usage at each SD (Figure 2B), the degenerate regions had a strong influence on splice site selection. For instance, although 49.7% of mini-genes spliced at SD_1 with less than 5% frequency, 7,705 mini-genes (2.9%) spliced at SD_1 with over 95% frequency.

In the alternative 3′ library, we also found isoforms from different splicing events, although splice site usage was less evenly balanced than in the 5′ library. SA_1 was used an average of 3.3% of the time, while SA_2 was used 89.2% of the time (Figure 2D). In this library, new splice sites in the randomized regions were only used with 0.3% frequency, probably reflecting the larger information footprint of splice acceptors (>20 nt) compared to splice donors (9 nt), which makes the occurrence of new sites within the degenerate regions less likely. Similarly to the 5′ library, we inadvertently inserted a cryptic splice acceptor 16 nt upstream of SA_2 that was used with 4.6% frequency. Many other cryptic splice sites were used with very low frequency (1×10^{-7} to 5×10^{-3}) accounting for a total of 2.3% of transcripts. In contrast with the alternative 5′ library, only 0.3% of transcripts were unspliced. Although SA_2 was the dominant splice site, 0.7% of the 1.2 M of mini-genes represented by multiple reads spliced 100% at SA_1.

With so many transcripts in each library splicing at new splice sites, we asked whether we could rediscover the known motifs for splice donors and splice acceptors from the de novo sites alone. When we plotted the relative frequencies of each base at each position for new splice donors (Figure 2E) and new splice acceptors (Figure 2F), both splice site motifs were nearly identical to the expected motifs for splice donors and splice acceptors. More specifically, the splice donors contained the canonical GT at the +1:+2 positions, while the splice acceptors contain a clear polypyrimidine tract (T and C rich), followed by N[CT]AGG. The ability to fully rediscover canonical signals for splice donors and splice acceptors demonstrates the rich type of information contained in each dataset.

We also asked whether translation might affect the mRNA stability in our libraries. Sequencing of the alternative 5′ library yielded fewer median reads on mRNA from mini-genes that were primarily spliced out of frame than in frame (Figure S1A). However, when the mini-genes contained a premature stop codon, the median number of reads per mRNA was similar for all three reading frames (Figure S1B). These results indicate that a large string of amino acids translated out of frame will destabilize the mRNA, likely through the no-go decay pathway (Doma and Parker, 2006; Shoemaker et al., 2010) as ribosomes stall due to protein misfolding. We also find evidence of nonsense-mediated decay, but only if the premature stop codon occurred >40 nt upstream of the splice donor. This is consistent with previous studies on nonsense-mediated decay that suggest the premature stop codon must occur >50 nt upstream of the last exon junction (Lewis et al., 2003).

Splicing Is More Likely to Occur at Upstream Splice Donors

From an analysis of the new splice sites, we found strong evidence that upstream splice donors were favored over downstream splice donors; new splice donors inserted in the first degenerate region were 4.1 times more likely to be used than new splice donors inserted into the second degenerate region (region 1: 849,666 spliced reads; region 2: 208,396 spliced reads). Furthermore, the effect of position of splice donors within each degenerate region was significant ($p < 0.005$; Figure S1C). The number of spliced reads at a new splice site decayed exponentially with the distance from SD_1 (Figure 2G). Splicing has been shown to be co-transcriptional, and spliceosome components can begin to assemble at a 5′ splice donor before downstream alternative slice sites are transcribed (Listerman et al., 2006), suggesting a potential mechanistic explanation for the observed effect. This strong bias for upstream splice donors is consistent with the typically short length of exons in the human genome (Burge and Karlin, 1997).

Splicing Is Less Likely to Occur at Splice Acceptors with Distal Branchpoints

Large-scale mapping of human branchpoints with RNA-seq found that 90% of mapped branchpoints occur between 19–37 nt upstream of the splice acceptor (Mercer et al., 2015). However, it remains unclear just how detrimental a distal branchpoint is toward efficient splicing. Consensus branchpoints (CU[AG]A[CU]) occur over 10,000 times at every position between 40 to 19 nt upstream of SA_1 in our dataset, allowing us to answer this question. We found that mini-genes with a consensus branchpoint sequence 19 nt upstream of SA_1 were approximately six times more likely to be spliced at SA_1 relative to those with a branchpoint 40 nt upstream of SA_1 (Figure 2H). One explanation for this phenomenon could be that distal branchpoints are more likely to contain another AG between the branchpoint and SA_1 that could be used as an alternative splice acceptor. However, we observed a strong distance dependence on branchpoint position for sequences both with and without an AG between the branchpoint and SA_1 (Figure S1D). This result suggests that mechanism by which distal branchpoints reduce splicing efficiency is primarily due to the increased distance between the branchpoint and the splice acceptor and/or polypyrimidine tract.

Sequence Motifs in Alternative Exons Have a Stronger Regulatory Role than Intronic Sequences

Next, we asked how short sequence motifs affect splice site selection in different contexts. We chose to analyze the effects of 6-mer because each possible 6-mer occurs within an average of 1,294 mini-genes for the alternative 5′ library, and 8,232 mini-genes for the alternative 3′ library. Furthermore, most known RNA binding proteins (RBPs) are reported to bind sequences between 4–8 nt (Lunde et al., 2007). In order to estimate the effect of

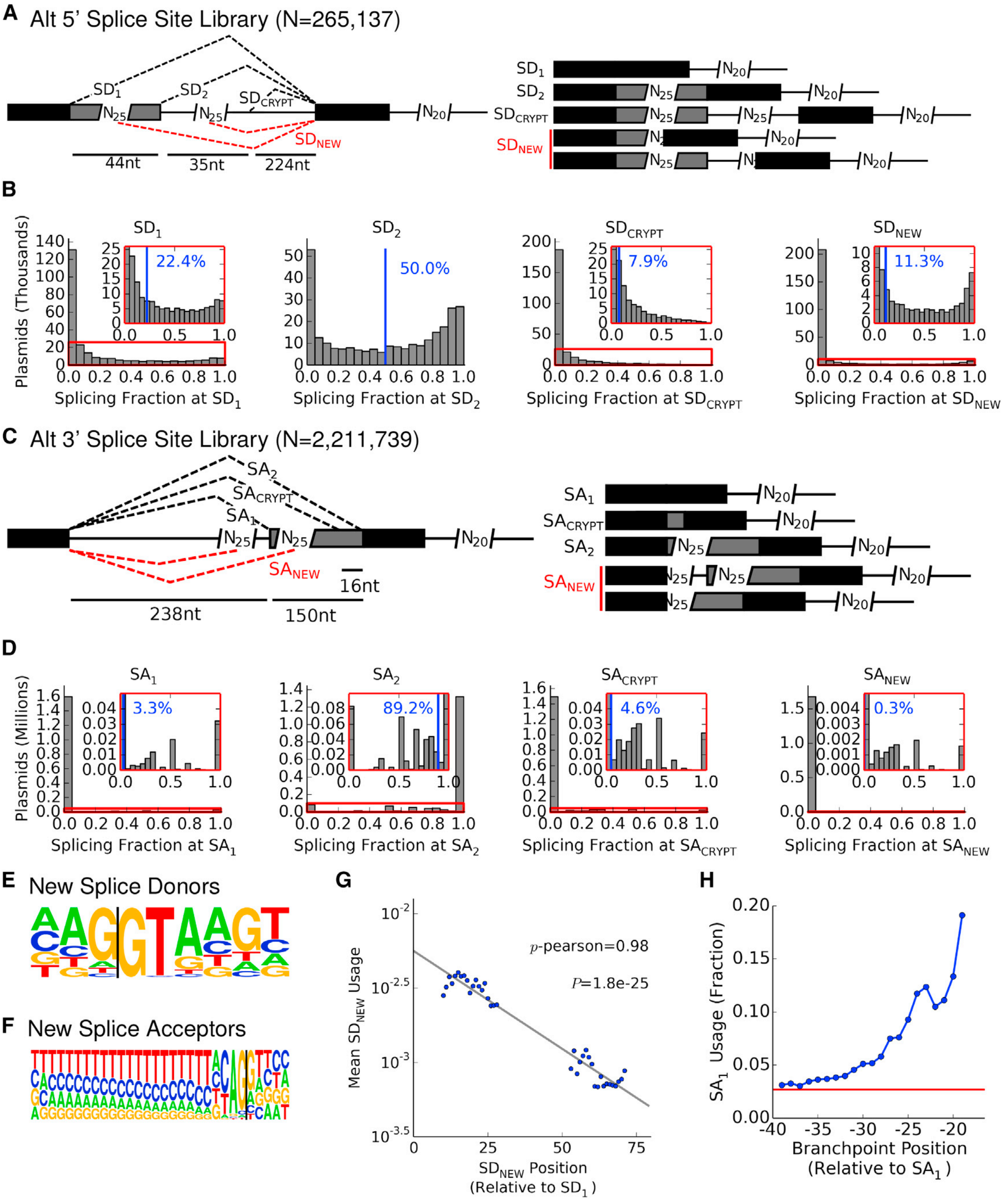


Figure 2. Splice Site Selection in Two Million Alternative 5′ and 3′ Spliced Sequences

(A) A schematic of the alternative 5′ library. Spliced reads map to SD_1, SD_2, and a cryptic splice site (SD_{CRYPT}), as well as new splice donors (SD_{NEW}) created in the degenerate regions.

(B) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.

(C) A schematic of the alternative 3′ library. Spliced reads map to SA_1, SA_2, and a cryptic splice site (SA_{CRYPT}), as well as new splice donors (SD_{NEW}) created in the degenerate regions.

(D) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.

(E) The splice donor motif recovered from the new splice alternative 5′ library matches the previously known human splice donor site.

(F) The splice acceptor motif recovered from the new splice alternative 5′ library matches the previously known human splice acceptor site.

(legend continued on next page)

each possible 6-mer in each region, we calculated splice site usage for the subset of mini-genes containing the 6-mer and for the much larger subset not containing the motif. We then asked to what extent the odds of splicing at a splice site changed in the presence of the motif relative to the control set. To quantify this "effect size," we used the $\log_2$ odds ratio with and without the 6-mer present (Supplemental Experimental Procedures). For example, we found that mini-genes containing the 6-mer GTGGGG in the first degenerate region of the 5′ library were spliced at SD_2 only 19.0% of the time, while RNA derived from mini-genes not containing this motif spliced at SD_2 50.2% of the time, resulting in an effect size of −2.1 (Figures S2A–S2D). In other words, the odds of splicing at SD_2 are 4.29 ($2^{2.1}$) times lower in the presence of GTGGGG compared to its absence.

In Figure 3A, we plot the empirically measured effect sizes of all hexamers in the first degenerate region on the relative usage of SD_2 and SD_1, with 95% confidence intervals. The strongest enhancers located in the alternative exon (included when splicing occurs at SD_2, but excluded when splicing occurs at SD_1) increased the odds of splicing at SD_2 4.38-fold, while the strongest silencers decreased the odds 16-fold. Approximately 15% of 6-mer have been previously identified as SREs (Culler et al., 2010; Fairbrother et al., 2002; Wang et al., 2004, 2012) (622/4,096), but here 82.9% of 6-mer (3,396/4,096) exhibited a significant effect on isoform selection (95% confidence interval does not contain zero effect size). Intriguingly, the cumulative effects of previously identified SREs accounted for only 20% of the cumulative effects of all possible 6-mer. The strongest silencers were G rich, consistent with known binding sites for hnRNPs (Martinez-Contreras et al., 2006). On the other hand, some of the strongest enhancers for SD_2 appear to act by generating secondary structure around SD_1: the 6-mer perfectly complementary to part of SD_1 (−3 to +8) were all in the top 6% of SD_2 enhancers (percentiles: 97.77, 99.75, 99.97, 94.23, 94.79, and 98.92).

We then looked at the effects of 6-mers in the second degenerate region (3′ to SD_2). Unlike the first degenerate region, which is located within the alternative exon region, the second degenerate region is intronic to both SD_1 and SD_2. We found that the effect sizes were much smaller than in the first degenerate region (Figure 3B). The strongest enhancer and silencer of SD_2, respectively, only changed the odds of splicing at SD_2 relative to SD_1 1.95-fold and 1.48-fold. Furthermore, only 36.7% of 6-mer (1,505/4,096) had a statistically significant effect.

We performed a similar analysis for each degenerate region on the usage of SA_1 in the alternative 3′ library (Figures 3C and 3D). Again, we found that motifs in the alternative exon (3′ of SA_1, but 5′ of SA_2) had strong effect sizes (statistically significant 6-mer effect sizes: 3,500/4,096, 85.4%; strongest enhancer: 3.84-fold increase in odds of splicing at SD_2; strongest silencer: 9.87-fold decrease in odds of splicing at SD_2). Unlike in the alternative 5′ library, we found that motifs in the intronic degenerate region (5′ of SA_1 and SA_2) also had quite strong effects (statistically significant 6-mer effect sizes: 3,248/4,096, 79.3%; strongest enhancer: 3.45-fold increase in odds-ratio; strongest silencer: 4.63-fold decrease in odds-ratio), although still generally smaller in magnitude than the downstream alternative exon region. When we looked at the strongest 6-mer enhancers of SA_1 in this intronic region, we found they all fit the consensus branchpoint sequence CU[AG]A[CU] (Figure 3D).

The Same Sequence Motifs Regulate Alternative Exon Inclusion Independent of the Type of Alternative Splicing

Surprisingly, we found that the effect sizes of 6-mers occurring within the alternative exon regions were extremely similar between the alternative 5′ and 3′ libraries (Figure 3E; R^2 = 0.68). We looked at several motifs known to bind splice factors or that have previously been identified as ESEs/ESSs (G-run, SRSF1, hnRNPA1, hnRNPH2) and found the effect sizes to be highly correlated. In both libraries, GGGGGG was the strongest exonic silencer (5′ library: 16.0-fold change in odds ratio; 3′ library: 9.87-fold reduction in odds ratio).

We also compared the effect sizes of intronic 6-mers (second randomized region in the alt. 5 library; first randomized region in the alt. 3′ library) between the two libraries. We found a significant, but weaker, correlation between the 6-mer scores (R^2 = 0.27; Figure S2E). The first randomized region in the alternative 3′ library overlaps the expected location of the SA_1 branchpoint, which may reduce the effect size correlation. However, the weaker correlation can also be explained by the fact that the effect sizes of intronic 6-mer were much smaller in magnitude compared to 6-mer within the alternative exon regions.

Sequence Motifs Regulate Exon Inclusion Additively Rather than Cooperatively

Although previous studies have observed co-occurrence of conserved sequence motifs around splice sites (Barash et al., 2010), it remains unclear whether such motifs act cooperatively or additively and independently of one another to regulate alternative splicing. In an additive and independent model of regulation, the joint effect size of multiple motifs should simply equal the sum of the individual effect sizes (Figure 4A). To assess this, we examined the joint effect sizes of pairs of 4-mers on alternative exon-inclusion levels in both the 5′ and 3′ libraries. We chose 4-mers because pairs of 4-mers occur sufficiently often within each randomized region to allow for robust effect size measurements (alt. 5′ library: 692 mini-genes/4-mer pair; alt. 5′ library: 4,399 mini-genes/4-mer pair).

We first calculated the individual effect size of all 4-mers on exon inclusion in the 5′ library. We then calculated the joint effect size of every possible pair of non-overlapping 4-mers. Surprisingly, we found that combinatorial effects were extremely well

(G) The number of spliced reads at each position within the randomized regions shows a strong position dependency. Splicing is more likely to occur at an upstream (5′) splice donor than at a downstream (3′) splice donor. The gray line is a fit that shows the linear relationship between the location of splice donor and the log read count at that location.

(H) Mini-genes with a consensus branchpoint near SA_1 are much more likely to use SA_1 than mini-genes with a distal branchpoint. The red line indicates the SA_1 usage, when there is no consensus branchpoint.

See also Figure S1.

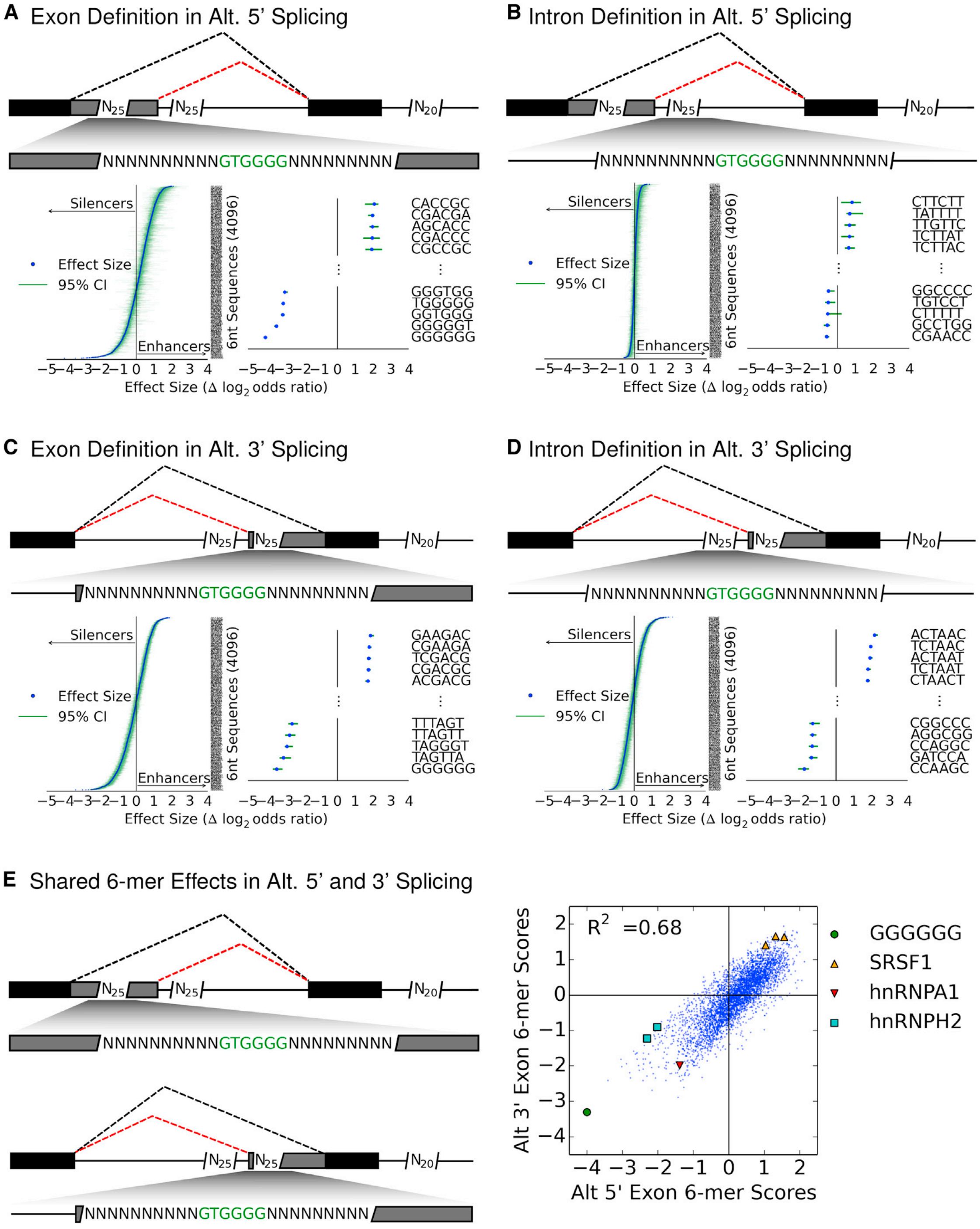


Figure 3. Measured Effect Sizes of Individual 6-mer in Each Degenerate Region

(A–D) To measure how sequence motifs alter the relative use of SD_2/SD_1 or SA_1/SA_2, we calculated effect sizes for every 6-mer (n = 4,096) within each degenerate region in both libraries. We defined effect sizes as the log odds ratio of SD_2 or SA_1 usage between mini-genes with/without the 6-mer of interest. The 6-mer are ranked by estimated effect size and plotted with 95% confidence intervals generated by bootstrapping with replacement. (A) Alternative exon region in 5′ library. (B) Intronic region in 5′ library. (C) Alternative exon region in 3′ library. (D) Intronic region in 3′ library.

(E) The 6-mer scores in the alternative exon region in both the 5′ and 3′ libraries (A and C) are highly similar, suggesting alternative splicing in both libraries is regulated by the same mechanism.

See also Figure S2.

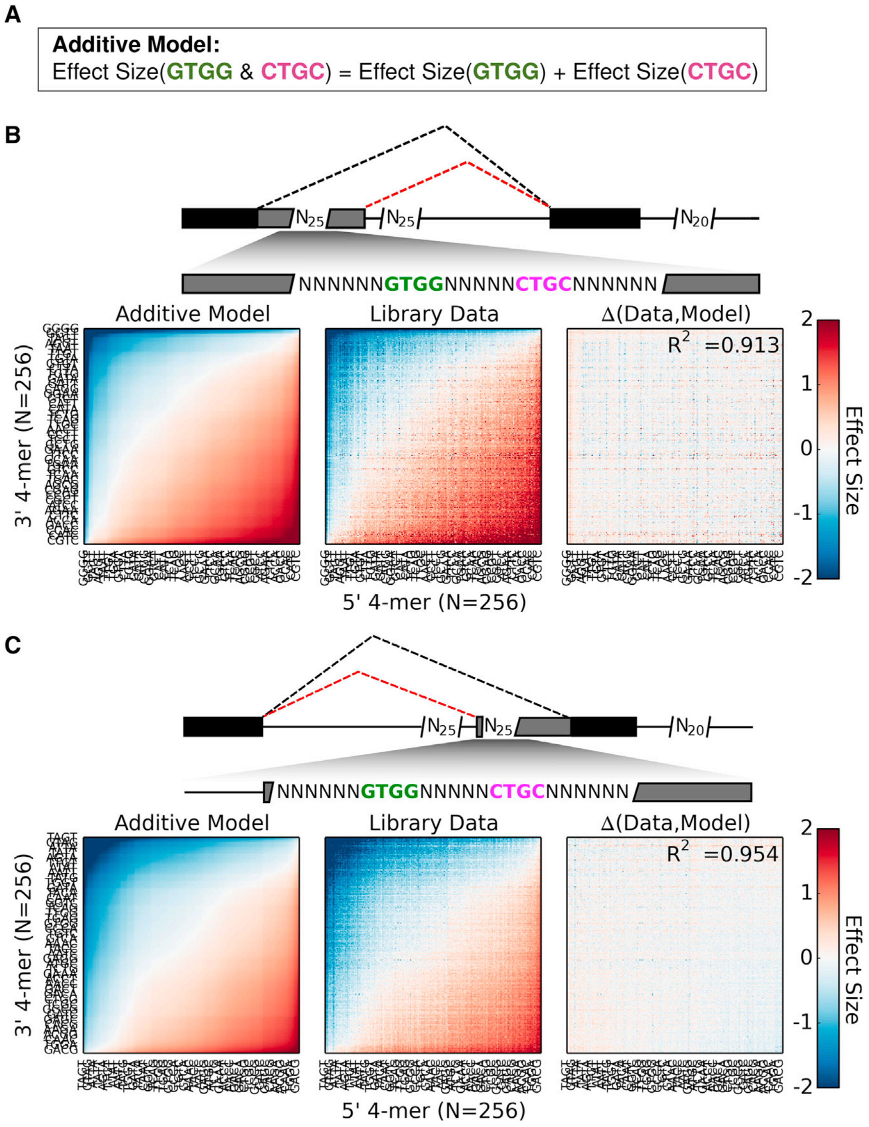


Figure 4. Combinatorial Regulation of Alternative Splicing Is Additive

(A) An additive model of alternative splicing regulation: the joint effect size of two 4-mer is equal to the sum of the individual 4-mer effects.

(B) Using an additive model, the predicted combinatorial effect size of every pair of 4-mer (n = 65,536) is plotted on the left. Each pixel corresponds to a pair of 4-mer with the 5′ 4-mer on the x axis and the 3′ 4-mer on the y axis. The measured combinatorial effect sizes from the 5′ library data are plotted in the middle. The residuals between the additive model and the observed data are plotted on the right. The additive model explains >90% of the combinatorial effect sizes (R^2 = 0.913).

(C) The same analysis is repeated for the alternative 3′ library. In this library the additive model explains over 95% of the combinatorial effect sizes (R^2 = 0.954).

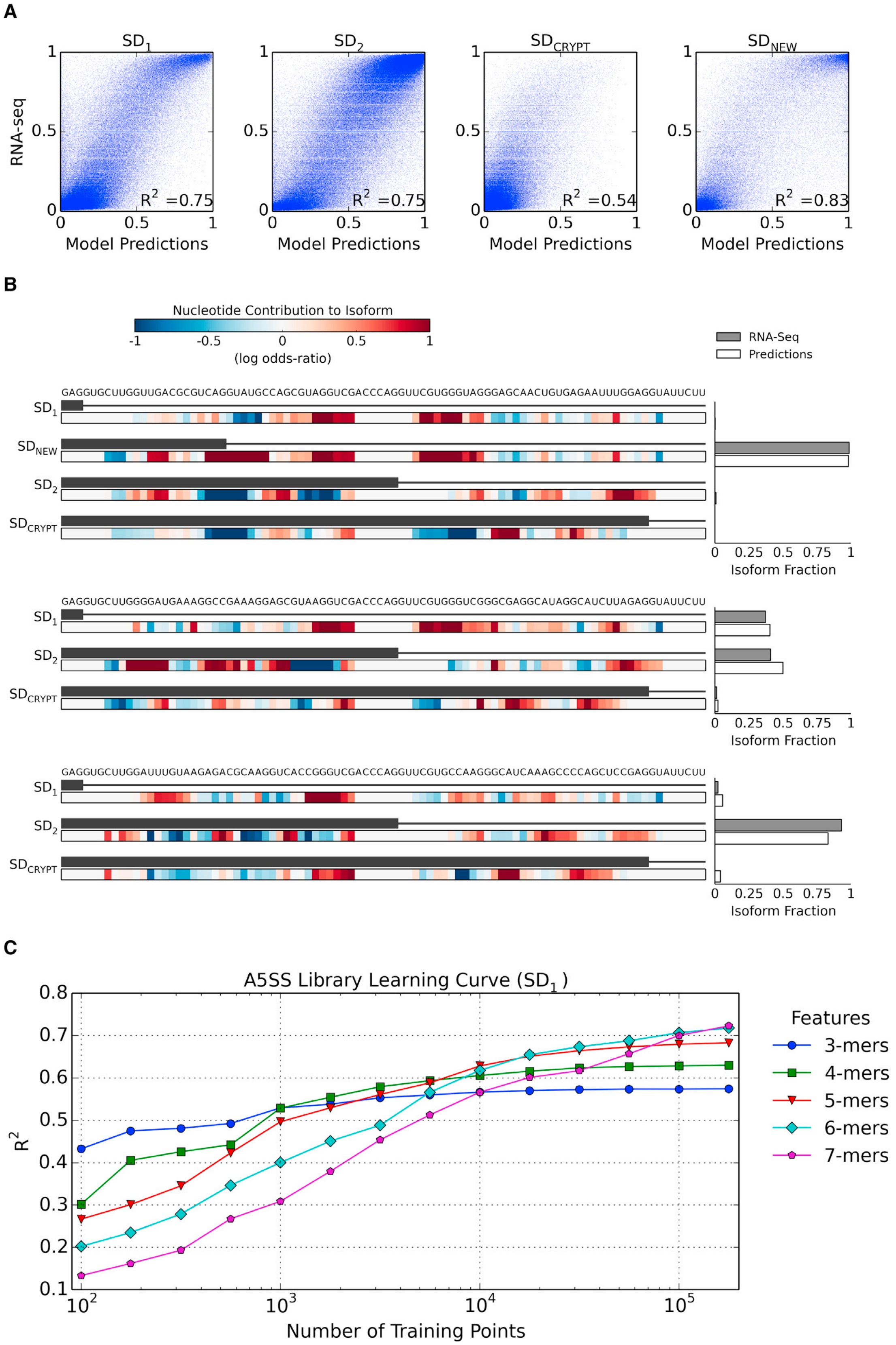
See also Figure S3.

captured by the sum of the 4-mer's individual effect sizes (R^2 = 0.913; Figure 4B). We did the same analysis for 4-mers located in the second degenerate region of the 3′ library. Here, the linear model fit the experimental data even better (R^2 = 0.954; Figure 4C). Thus, while specific instances of cooperative sequence interactions have been well documented (Huelga et al., 2012; Oberstrass et al., 2005), our results suggest the majority of motifs primarily exert their influence on exon inclusion independently of the surrounding motifs.

Predicting Isoform Ratios in Alternative Splicing from Sequence Information

We then turned to the task of learning a model of alternative splicing to predict isoform levels from sequence information. Because combinatorial regulation of alternative splicing was accurately captured by an additive model, we postulated that an additive model with short sequences as input features would perform well for prediction. Using both the 5′ and 3′ libraries, we trained a joint model of alternative exon definition in which a score is learned for each of the 4,096 possible 6-mers (Figure S3A). The scores learned here are similar to the previously calculated effect sizes, but rather than measuring the effects of a single 6-mer one at a time, we learned all the scores together through regression. Given the large number of new splice donors appearing within the 5′ library, we also chose to train a model of the splice donor site itself (Figure S3B). When we tested the splice donor model using cross validation, we found it accurately predicted the fraction of reads mapping to the three original splice donors, accounting for up to 75% of observed isoform variability (R^2: SD_1 = 0.75, SD_2 = 0.75, SD_{CRYPT} = 0.54; Figure 5A). It also proved accurate in predicting the position and fraction of reads mapping to newly created splice donor sites within the degenerate regions (R^2: 0.83; Figures 5A and 5B).

A fundamental advantage of testing synthetic sequences is the ability to learn from larger datasets than were previously available. As an attempt to quantify this advantage, we calculated learning curves on a simple model predicting usage of SD_1 in the alternative 5′ library. We split our data into training and test sets (90%/10% split) and trained models using subsets of the training data (between 100 to 177,827 training points). We also trained separate models using 3-mers, 4-mers, 5-mers, 6-mers, or 7-mers. With limited data (1,000 or fewer training points), the simplest model (3-mers) made the most accurate predictions, while the 7-mer model made the least accurate predictions, with the other models ordering between (Figure 5C). However, with the largest training subset (177,827 points), the results were reversed with the 7-mer model achieving the highest accuracy. Based on the slopes of



(legend on next page)

the learning curves, the 3-mer to 5-mer models would not benefit significantly from more data points (> 177,827), but the 6-mer, and especially the 7-mer, models seem likely to achieve significantly higher prediction accuracy with larger training sets. These results highlight the intuitive point that richer feature sets can improve predictions accuracy, but require more data to properly train.

Predicting the Effects of Human Genomic SNPs on Alternative Isoform Ratios

Next, we asked whether we could apply our model (HAL [hexamer additive linear])—developed entirely in the context of synthetic mini-genes—to predict changes in alternative splice donor usage caused by common polymorphisms in human genomes. As a first test case, we focused on 5′ alternative splicing. Combining DNA and RNA sequencing data, respectively, from the 1000 Genomes Project (Abecasis et al., 2012) and GEUVADIS consortium (Lappalainen et al., 2013), we calculated the percent of splicing at the downstream alternative splice donor (percent spliced in [PSI]) of wild-type genotypes for 8,546 5′ alternative splicing events using the MISO software package (Katz et al., 2010). We separately calculated mean isoform levels for genotypes heterozygous or homozygous for a single SNP in the region between the two competing splice donors or within the splice donors themselves (Table S1).

We began by investigating whether the model of the actual 9-nt splice donor sequence—again learned completely from our synthetic mini-genes—could accurately predict the effects of SNPs occurring within splice donor sequences. We also compared our prediction accuracy to a leading splice donor prediction tool trained directly from splice donor usage in the human genome (MaxEnt) (Yeo and Burge, 2004). Among heterozygous SNPs in alternative splice donors occurring in multiple individuals, we found that 93 of 199 SNPs altered PSI by >5% (Figures 6A and 6B). Within this set, HAL predicted the direction of change with 87.1% accuracy (81/93; binomial $p = 9.83 \times 10^{-14}$), while MaxEnt predicted the direction of change with 81.7% accuracy (76/93; binomial $p = 4.45 \times 10^{-10}$). Among the 35 homozygous SNPs in splice donors that alter PSI by >5%, our model predicted every SNP correctly, while MaxEnt made two mistakes (HAL: 35/35, binomial $p = 5.82 \times 10^{-14}$; MaxEnt: 33/35, binomial $p = 3.67 \times 10^{-10}$). For the set of SNPs within splice donors, our model explained 59.3% of the observed heterozygous effects ($R^2 = 0.593$, $p = 6.38 \times 10^{-8}$) and 67.7% of the observed homozygous effects ($R^2 = 0.677$, $p = 4.65 \times 10^{-24}$). This is a substantial improvement over MaxEnt, which accounted for 39.8% of the observed heterozygous effects ($R^2 = 0.398$, $p = 1.22 \times 10^{-11}$) and 41.1% of the observed homozygous effects ($R^2 = 0.411$, $p = 3.3 \times 10^{-5}$). Even when we extended our analysis to all SNPs (including those with less than 5% change in PSI), we found HAL substantially outperformed MaxEnt (HAL: $R^2 = 0.48$; MaxEnt: $R^2 = 0.22$; Figure S4A).

We then applied the model to predict the effects of human genomic SNPs in the alternative exon region between, but not overlapping, splice donors. Because most SNPs not occurring in actual splice sites are likely to only have modest effects, we restricted our analysis to SNPs with at least ten homozygous wild-type or ten heterozygous samples expressing the relevant mRNA. Moreover, we focused on SNPs that resulted in a change in the PSI of at least 5% to minimize the impact of measurement noise on the validation dataset; 43/344 heterozygous and 20/131 homozygous SNPs altered the PSI by >5% (Figure 6C). HAL correctly predicted the direction of change for 37/43 heterozygous and 17/20 homozygous SNPs (p: heterozygous = 1.63×10^{-6}, homozygous = 2.58×10^{-3}, combined = 6.11×10^{-9}). Furthermore, our model explained around half of the total observed effects of these SNPs (heterozygous: $R^2 = 0.570$, $p = 9.23 \times 10^{-9}$; homozygous: $R^2 = 0.442$, $p = 1.39 \times 10^{-3}$). Thus, our model not only outperformed the state of the art splice donor algorithm (MaxEnt) at predicting the effects of SNPs within splice donors but also successfully predicted the effects of SNPs within the alternative exon region, which to our knowledge, no other tool can do.

Predicting Alternative 5′ Isoform Levels from Sequence Information

To further assess the accuracy of our splice donor model, we predicted the isoform ratios in 6,152 alternative 5′ splicing events expressed in lymphoblastoid cell lines and compared our results to four other splice donor prediction algorithms. Our splice donor model substantially outperformed all of the other algorithms (Figure S5; Table S2). Interestingly, all of the models (including ours) performed better on events with shorter alternative exon regions (i.e., the region between splice donors). In these events, there is less space for regulation between the splice donors, possibly simplifying the prediction task.

Predicting the Effects of Variants on Exon Skipping in Mendelian Diseases

The most common form of alternative splicing is neither alternative 5′ or 3′ splicing, but exon skipping. Exon skipping is a highly regulated form of alternative splicing in human cells, and misregulation of cassette exon splicing can cause disease (Garcia-Blanco et al., 2004) and cancer (Kim et al., 2008). Given the relatively more complex structure of skipped exons, it might on first sight seem unlikely that a model trained only on 5′and 3′ alternative splicing should be able to predict levels of exon inclusion. However, we hypothesized that the similarity between

Figure 5. A Model Accurately Predicts Alternative 5′ Splicing and the Location of New Splice Donors

(A) For each splice donor (SD_1, SD_2, SD_{CRYPT}), model predictions are plotted against the observed splice site usage fraction. Each point represents a single test plasmid. The results are also plotted for all new splice sites (SD_{NEW}).

(B) The prediction results for three different mini-genes are shown with the associated nucleotide scores for each isoform. Each nucleotide score is calculated by averaging the model weights of all 6-mer overlapping the nucleotide. In the first example mini-gene, HAL predicts the usage and position of a new splice donor, which is confirmed by RNA-seq.

(C) A learning curve was generated for different models that predict the fraction of splicing at SD_1. The simplest model (3-mer features) performed the best with small training sets (<1,000 data points), but with more data points, richer feature sets offer better performance.

See also Figure S5.

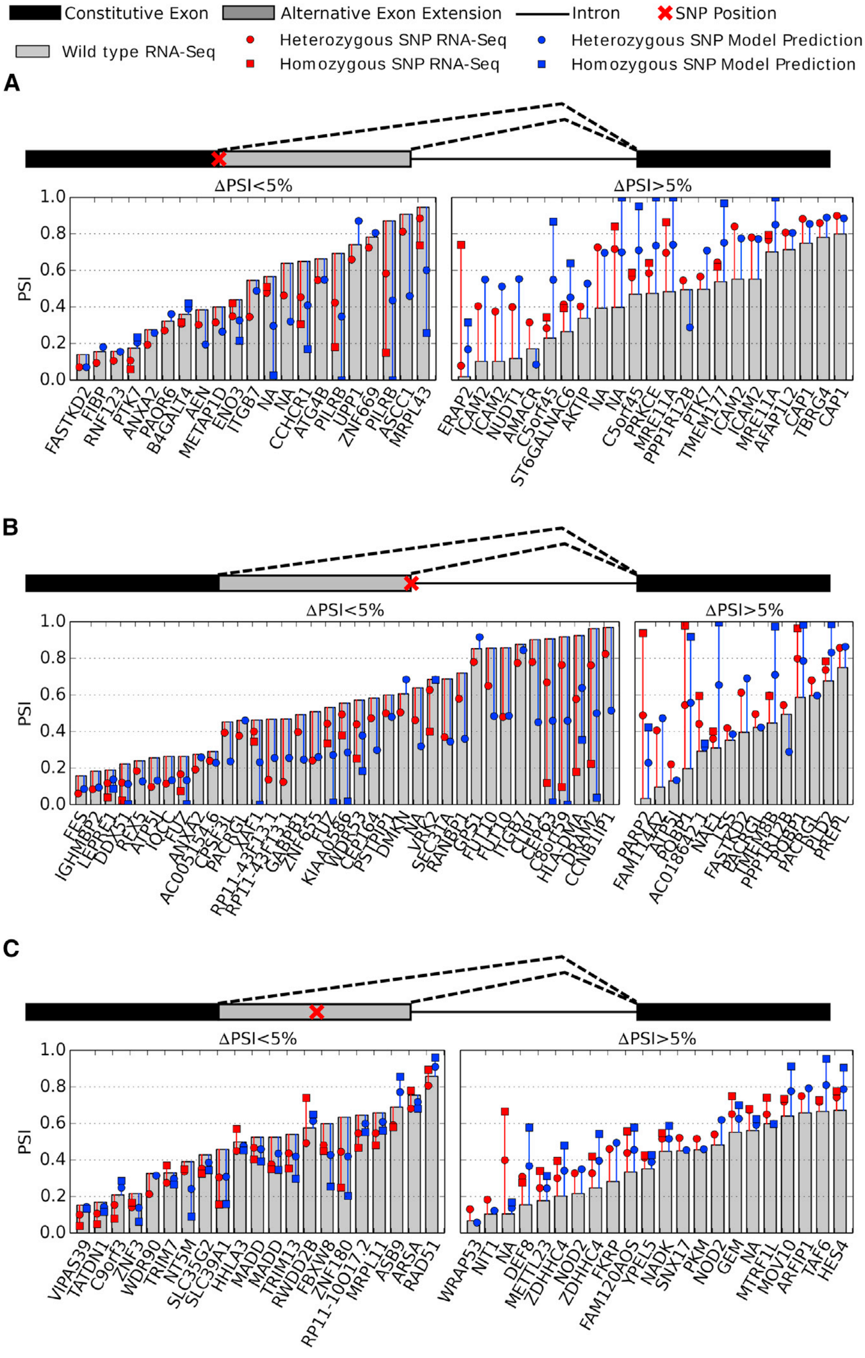


Figure 6. Splicing Model Identifies the Functional Effect of SNPs on Alternative Splicing

(A–C) Model predictions are plotted with the PSI measured from RNA-seq for SNPs occurring in the upstream splice donor (A), the downstream splice donor (B), and between the competing splice donors (C) that alter the measured PSI by greater than 5%. The observed PSI from RNA-seq for the wild-type genotype (gray bar) and genotypes containing the SNP (red) are plotted together with the model prediction (blue). The model accurately predicts the direction of change of the heterozygous SNPs in splice donors with 87.1% accuracy (81/93; binomial p = 9.83 × 10^{-14}) and the heterozygous SNPs between splice donors with 86.0% accuracy (37/43; binomial p = 8.18 × 10^{-7}). See also Figure S4 and Tables S1 and S2.

the sequence determinants of alternative exons in alternative 5′ and 3′ splicing might extend to exon skipping as well. If this were the case, we would expect our model to accurately predict the effects of exonic sequence variants on skipped exon-inclusion levels, even though it was never trained directly on any exon skipping data. We tested this hypothesis in the context of mutations in several distinct genes that are known to cause Mendelian disease by promoting exon skipping (Figure 7A; Table S3).

First, we compared model predictions to experimental data for the *SMN1* and *SMN2* genes, whose misregulation can lead to spinal muscular atrophy. Our model correctly predicted increased or decreased exon 7 inclusion in 205/229 (89.5%; Figure 7D) variants with experimental data. In Figure 7B, we compare predictions (increased or decreased exon inclusion) to experimental data. To make the plot more readable, we only included a single SNP at each position. Our model accurately predicts increased/decreased exon inclusion for 20/22 of the plotted SNPs. On just the variants with quantitative data (n = 131), our model explained 65% of the observed variance (R^2 = 0.65; Figure 7E). The *SMN1/2* variants that we tested included SNPs, indels, and combinations of up to 30 nt changes.

We then tested our model on variants in *CFTR*, whose misregulation can lead to cystic fibrosis. Our model correctly predicted increased/decreased exon 12 inclusion in 19/22 variants (Figure 7D). When we only looked at the SNP with the largest effect at each position, our model accurately predicted increased/decreased exon inclusion for 11/12 SNPs (Figure 7C). Among all the *CFTR* variants, our model explained 60% of the observed variance (Figure 7E; R^2 = 0.60).

Next, we tested our model predictions on variants in exon 7 of the *BRCA2* gene, a tumor suppressor responsible for DNA damage repair. Mutations in *BRCA2* affecting the ability of the protein to repair DNA lead to such an increased risk of ovarian and breast cancer that patients with these mutations may choose to have prophylactic surgery. However, the effect of many variants on alternative splicing and hence protein function remain unknown, forcing patients and doctors to make clinical decisions with limited information. The ability to identify deleterious variants computationally can provide valuable information to

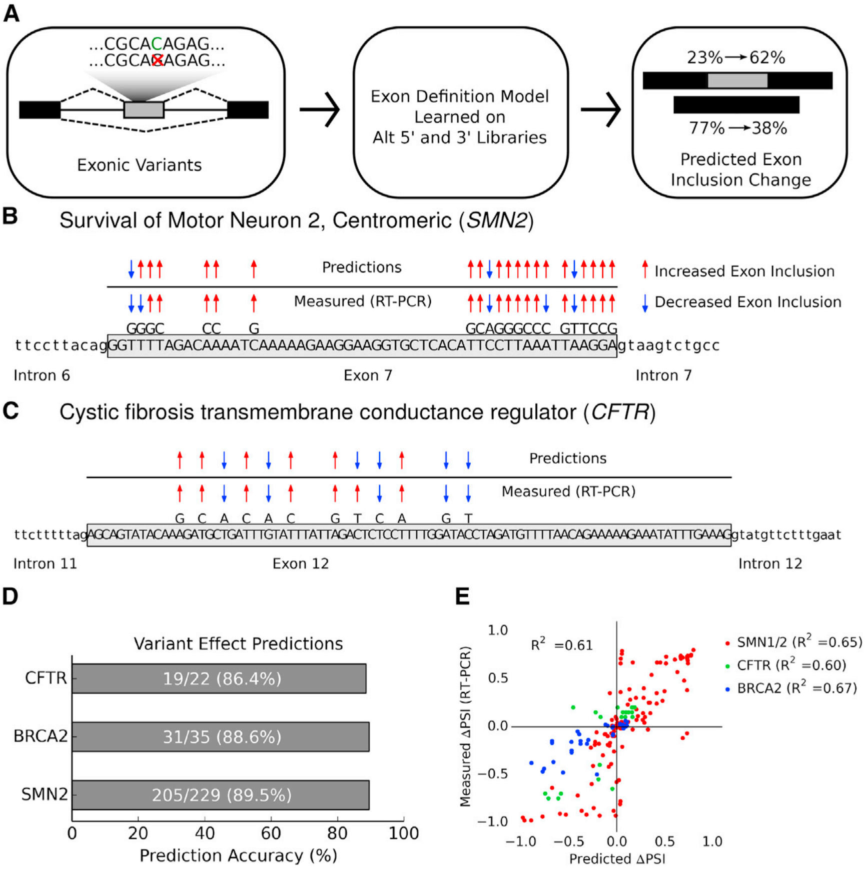


Figure 7. Predicting the Effects of Exonic Variants on Exon Skipping

(A) The inputs to the splicing model can include SNPs, indels, or complex variants within the alternative exon. The splicing model then predicts the exon inclusion levels with the variant present.

(B) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 7 of SMN2. For positions with data for multiple SNPs, the SNP with the largest measured change in PSI was plotted. The model accurately predicted the directional change in PSI (increased exon inclusion/exclusion) for 20/22 SNPs plotted.

(C) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 12 of CFTR. The model accurately predicted the directional change in PSI for 11/12 SNPs plotted.

(D) The prediction accuracy for variants in SMN2, CFTR, and BRCA2 ranged from 86% to 90%.

(E) The change in PSI is plotted for every variant with RT-PCR data. The model explains over 60% of the effects of SNPs for variants each gene tested (SMN1/2, CFTR, and BRCA2).

See also Figure S6 and Table S3.

patients with these variants of unknown significance. Our model correctly predicted increased/decreased exon 7 inclusion for 31/35 variants that experimentally altered inclusion levels (Figure 7D). The model correctly predicted 19/22 of the SNPs with the largest effect at each position within the exon (Figure S6B). Among all the *BRCA2* variants, our model explained 67% of the observed variance (R^2 = 0.67; Figure 7E).

We then compared our results to SPANR (Xiong et al., 2014)—the current state of the art in predicting the effects of SNPs on exon skipping. SPANR consists of a Bayesian deep learning algorithm trained on exon skipping events in the human genome with 1,393 carefully hand-selected features. As of this paper, SPANR only supports predictions of SNPs, so we were not able to compare our predictions on more complex variants. However, for SNPS in *SMN1/2*, *CFTR*, and *BRCA2*, we found that HAL accounted for three times more of the observed effects than SPANR (HAL: R^2 = 0.51; SPANR: R^2 = 0.17; Figure S6A). We made HAL publicly available at http://splicing.cs.washington.edu. All of the code to reproduce this study is publicly available at https://github.com/Alex-Rosenberg/cell-2015.

DISCUSSION

We present a framework based on massively parallel analysis of synthetic sequences to dramatically improve our understanding of alternative splicing and the ability to predict the impact of natural human genetic variation. Our model accurately predicts the effects of sequence variants on alternative 5′ splicing that occur both within the alternative exon and in the competing splice donors. Even more importantly, our model learned regulatory rules about alternative splicing that generalized to exon skipping—a completely different form of alternative splicing than those on which the model was trained.

Our results suggest that a common regulatory mechanism is shared between all major forms of alternative splicing. Additional evidence for such a common mode of regulation comes from previous smaller-scale studies of ESEs or ESSs that have shown similar effects across different forms of alternative splicing (Wang et al., 2006, 2012). It is unlikely that this shared form of regulation occurs during splice site recognition; any exonic splice regulatory element that alters splice donor or splice acceptor recognition should have different effects in alternative 5′ and 3′ splicing events. It is more likely that alternative exon inclusion is modulated during exon definition, that is the pairing of splice site across exons, which often precedes the eventual pairing of splice donors and acceptors across introns (Robberson et al., 1990).

Furthermore, our data also suggest that the exon-defining interactions between the upstream splice acceptor and downstream splice donor are regulated additively. In both alternative 5′ and 3′ splicing, we found the joint effect size of multiple 4-mer to be highly correlated with the sum of the individual 4-mer effects. This result may indicate that each sequence motif can contribute additively to stabilizing the splice acceptor-splice donor interaction, likely through the *trans*-factors that bind these sites. However, the true mechanistic basis for this additivity will require further investigation. Although, there is evidence supporting specific examples of functional interactions between *cis*-splicing regulatory elements (Oberstrass et al., 2005), our results indicate that these examples are likely uncommon.

A potential limitation of our approach is that mRNAs are transcribed from plasmids rather than directly from the genome, especially considering evidence suggesting that chromatin can influence alternative splicing (Luco et al., 2010). However, advances in high-throughput genome editing may make it possible to perturb the genome in a massively parallel fashion, which will enable extensions of our approach to probe the effects of chromatin on alternative splicing. In fact, recent work demonstrated that small-scale genomic libraries could be created through insertion of degenerate sequences directly into an alternatively spliced gene locus (Findlay et al., 2014). Moreover, our current work focused on mini-genes with short alternative exons, and more work will be necessary to understand to which extent our results generalize to other gene architectures. However, human exons are typically short (an average 147 bp for internal exons) (IHGSC et al., 2001), and, moreover, analysis of sequence conservation suggests that most sequence determinants of alternative splicing can be found within a few hundred nucleotides of intron-exon junctions. It is important to emphasize that our approach uncovers only *cis*-regulatory rules. Complementary experiments that connect this *cis*-grammar to a repertoire of *trans*-acting splice factor proteins are necessary to fully understand the mechanisms underlying the regulation of alternative splicing.

We have demonstrated that learning the sequence determinants of gene regulation from large libraries of synthetic sequences can be used as a complementary approach to learning directly from the human genome. We assayed over two million alternatively spliced constructs, nearly two orders of magnitude more events than the 38,000 that are present in the human genome (Wang et al., 2008), containing over 100 Mb of synthetic sequence. Our improved understanding of alternative splicing and performance in predicting the effects of genetic variants is not a result of more sophisticated machine learning algorithms but simply the result of learning from a larger and more reliable dataset. We anticipate that this general approach will be useful for advancing our biological understanding of diverse forms of gene regulation, such as transcription, translation, and polyadenylation.

EXPERIMENTAL PROCEDURES

Cloning of Degenerate Libraries

The libraries were assembled with PCR and standard Gibson assembly (Gibson et al., 2009) using degenerate oligonucleotides (IDTDNA). First Citrine was split into two exons, and the first exon of the Citrine gene was altered to remove any potential splice donors, without altering the amino acid sequence. The introns with degenerate sequences were inserted between the two exons of Citrine. The barcode sequence was inserted into the 3′ UTR of Citrine.

Cell Culture and Transfection

HEK293 cells were cultured in in DMEM (Cellgro) plus 10% FBS and L-glutamine/penicillin/streptomycin on coated plates. Plates were coated for 24 hr with 8 ml of 100× diluted extracellular matrix gel (Sigma-Aldrich) before HEK293 cells were added to the plates. For transfection of a complex pool of plasmids, 1.2 million cells were seeded in a 10-cm dish 24 hr before transfection. We mixed 10 μg of the plasmid library in 1 ml of Opti-MEM Reduced Serum Medium (Life Technologies) with 30 μl of Lipofectamine LTX and 10 μl of Plus Reagent (Life Technologies), before transfecting into the 10-cm dish. The DMEM was replaced 5 hr after transfection.

Isolation of RNA and Generation of cDNA

Total RNA was extracted using RNeasy (QIAGEN) kits 24 hr after transfection. The optional on column DNaseI digest was performed with the RNase-Free DNase Set (QIAGEN). Total RNA quality and purity was tested by measuring the A260/A280 ratio on a NanoDrop 1000 Spectrophotometer and, in some cases, by measuring the ratio of the 18S and 28S rRNA bands on a native 1% agarose gel. mRNA was separated from 35–48 μg total RNA using polyA Spin mRNA Isolation Kits (New England Biolabs). Isolated mRNA was again digested by DNaseI for 30 min using the Turbo DNA-free Kit (Ambion). cDNA was then synthesized from 109–374 ng mRNA using MultiScribe Reverse Transcriptase (Ambion) and Oligo d(T)16 primers (Ambion). cDNA synthesis was performed by holding reactions at 25°C for 10 min, 42°C for 110 min, and 85°C for 5 min. The quality of cDNA and presence of DNA contamination were checked through qPCR: Citrine, mCherry, and TBP were compared using cDNA, no reverse transcription controls (NRTC), and a no template control (NTC). The results indicated that there was no plasmid or genomic DNA carryover into the cDNA reactions.

Generation of Illumina Flow Cell Compatible PCR Products from RNA and DNA Library

The resultant cDNA was then amplified by PCR to generate products compatible with the Illumina HiSeq2000 Flow Cell. PCR reactions were performed in 100 μl with 2x Phusion HF Master Mix (New England Biolabs), 50 pmol forward primer, and 50 pmol reverse primer with sample specific barcodes and 20% of each cDNA reaction. Cycling was done on a BioRad T100 Thermal Cycler with the following protocol: 98°C for 5 min, then seven cycles of 98°C for 10 s, 67.5°C for 15 s, 72°C for 30 s, and a final extension step at 72°C for 5 min. The necessary number of cycles was determined for each sample by first running qPCR reactions with EvaGreen in a Biorad CFX and determining when fluorescence began to plateau. Following PCR, 10% of the products were run on a 2% agarose gel to determine if the expected bands were present. The remainder of the PCR products was purified using the QIAquick PCR Purification Kit (QIAGEN) and eluted into 30 μl of EB. Concentrations, as well as A260/280 and A260/230 ratios, were measured on a NanoDrop 1000 Spectrophotometer.

Illumina-compatible PCR products were also generated from the DNA plasmid library with the same protocol as above, except the cDNA template was replaced with 10 ng of plasmid library DNA and the PCR reaction was performed with 20 cycles.

Sequencing Plasmid Library and RT-PCR Products

Both the RT-PCR products and plasmid library PCR products were sequenced on either an Illumina HiSeq2000 or Illumina MiSeq with paired end reads. The forward read crossed the post-splicing exon-exon junction and the reverse read covered the 3′ UTR barcode. A 6-nt index read was used to sequence the sample barcode to determine if the read came from a DNA library or a cDNA library.

Associating Degenerate Intronic Regions with 3′ UTR Barcode Tags

Using the sequencing results of the DNA plasmid library, we first counted the number of reads for every observed barcode and calculated an average Phred quality score for each position. We discarded any barcode tags with less than two reads or less than an average Phred score of 20 at any position. We then mapped each remaining tag to the associated degenerate sequence with the most reads. If each degenerate sequence had a single read, we chose the sequence with the highest minimum Phred score.

Measuring Isoform Fractions from Sequencing Results

For every read on an RT-PCR product, we recorded the splicing position (or lack of splicing) by aligning the read to the unspliced plasmid. Using the associated barcode read, we were then able to tally the number of reads splicing at each position for every plasmid in our library. With respect to the alternative 5′ library, only reads that mapped to a splice donor with GT or GC in the +1 to +2 intronic positions were counted.

ACCESSION NUMBERS

The accession number for the raw and processed sequencing data reported in this paper is GEO: GSE74070.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.09.054.

AUTHOR CONTRIBUTIONS

A.B.R. designed and performed experiments, analyzed data, built and tested the splicing model, wrote the manuscript, and developed the web tool; R.P.P. designed experiments; and J.S. and G.S. designed the experiments and wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by a National Science Foundation (NSF) Career Award (0954566) and a Burroughs Wellcome Career Award at the Scientific Interface (to G.S.).

Received: July 14, 2015
Revised: August 28, 2015
Accepted: September 21, 2015
Published: October 22, 2015

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., and Johnson, J.M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. *40*, 1416–1425.

Culler, S.J., Hoff, K.G., Voelker, R.B., Berglund, J.A., and Smolke, C.D. (2010). Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. Nucleic Acids Res. *38*, 5152–5165.

Doma, M.K., and Parker, R. (2006). Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature *440*, 561–564.

Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Garcia-Blanco, M.A., Baraniak, A.P., and Lasda, E.L. (2004). Alternative splicing in disease and therapy. Nat. Biotechnol. *22*, 535–546.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. Mol. Cell *22*, 769–781.

Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. *1*, 167–178.

International Human Genome Sequencing Consortium, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Kim, E., Goren, A., and Ast, G. (2008). Insights into the connection between cancer and alternative splicing. Trends Genet. *24*, 7–10.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Le, Q.V., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., and Ng, A.Y. (2012). Building high-level features using large scale unsupervised learning. In Proceedings of International Conference in Machine Learning.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA *100*, 189–192.

Listerman, I., Sapra, A.K., and Neugebauer, K.M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat. Struct. Mol. Biol. *13*, 815–822.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Martinez-Contreras, R., Fisette, J.-F., Nasim, F.U., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol. *4*, e21.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human splicing branchpoints. Genome Res. *25*, 290–303.

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Noderer, W.L., Ross, J.F., Aparna, B., Alexander, J.D.A., Jiajing, Z., Paul, A.K., and Clifford, L.W. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol. Syst. Biol. *10*, 1–14.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., and Allain, F.H. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science *309*, 2054–2057.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell Rep. *7*, 281–292.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Shoemaker, C.J., Eyler, D.E., and Green, R. (2010). Dom34:Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. Science *330*, 369–372.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580–586.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wang, Z., Xiao, X., Van Nostrand, E., and Burge, C.B. (2006). General and specific functions of exonic splicing silencers in splicing control. Mol. Cell *23*, 61–70.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat. Struct. Mol. Biol. *19*, 1044–1052.

Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat. Struct. Mol. Biol. *20*, 36–45.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. USA *110*, 11952–11957.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2014). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, no. 6218.

Yeo, G., and Burge, C. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yu, Y., Maroney, P.A., Denker, J.A., Zhang, X.H., Dybkov, O., Lührmann, R., Jankowsky, E., Chasin, L.A., and Nilsen, T.W. (2008). Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. Cell *135*, 1224–1236.

Zhang, X.H., and Chasin, L.A. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. *18*, 1241–1250.