Title: Large-scale screening of rare genetic variants in humans reveals frequent splicing disruptions

Authors:

Rocky Cheung[1†], Kimberly D. Insigne[2†], David Yao[3], Christina P. Burghard[2], Eric M. Jones[1], Daniel B. Goodman[4], Sriram Kosuri[1,5*]

Affiliations:

[1] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA

[2] Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095, USA

[3] Genetics Graduate Program, Stanford University, Stanford, CA 94035, USA

[4] Department of Microbiology and Immunology, University of California, San Francisco, CA 94143, USA

[5] UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed. Tel: +1 310 825 8931; Email: sri@ucla.edu

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract:

Pre-mRNA splicing is an important mechanism by which genetic variation influences complex traits. We developed a Multiplexed Functional Assay of Splicing using Sort-seq (MFASS) that allows us to quantify exon inclusion in large libraries of human exons and surrounding intronic contexts. We used MFASS to explore >10,000 designed mutations intended to alter regulatory elements that govern splicing. Many classes of mutations led to large-effect splicing disruptions including mutations far from canonical splice sites, and these effects were not easily predicted. We assayed 29,531 extant variants in the Exome Aggregation Consortium, and found that >1000 variants (3.6%) within or adjacent to 2393 assayed human exons led to almost complete loss of exon recognition. While most variants at the canonical splice site disrupt splicing, they represent <20% of splice-disrupting variants overall because genetic variation elsewhere dominates. Our results indicate that loss of exon recognition caused by rare genetic variation may play a larger role in trait diversity than previously appreciated, and that MFASS may provide a scalable way to functionally test such variants.

Main Text:

Any individual's genome contains ~4-5 million deviations from the reference human genome, almost all of which are very rare[1]. How this collection of differences give rise to trait diversity and disease susceptibility is a central question in human genetics. Recent genetic studies implicate pre-mRNA splicing as a major and underappreciated means through which variation imparts functional consequences[2–5]. However, genetic variation is depleted at the major splicing recognition sites[2,6]. If genetic variation is having major impacts on splicing, how does it impart its effects if not through the major sites known to affect splicing?

In humans, genetic and biochemical studies show that exons are first recognized in a process called exon definition, and then introns between them are removed[7–11]. The major exon recognition elements, including the splice donor, acceptor, branchpoint and polypyrimidine tract, taken together are too degenerate alone to discriminate true exons from those not utilized *in vivo*[12–14]. Numerous computational, *in vitro*, and genetic studies have shown that other *cis*-regulatory elements are required to distinguish false exons from included ones[12,13,15]. These sequences are short motifs that are broadly classified as exonic splicing enhancers (ESEs) and suppressors (ESSs) as well as their intronic counterparts[16,17] (ISEs & ISSs). Machine learning methods use these and other genomic features trained against genome-wide RNA sequencing datasets to build predictive models of splicing regulation[7–9]. However, the predictive power of these models may come almost entirely from sequence conservation rather than the mechanistic understanding of splicing[18,19]. These models predict that human genetic variation, and especially rare variation, often disrupt sequence features required for proper exon recognition, but it is difficult to verify the accuracy of these predictions at large scales[7].

Several groups have developed massively parallel reporter assays of splicing[8,14,20,21]. Most of these assays look at a small set of exons and mutate them to understand which elements are important for splicing. Importantly, these methods have allowed us to better quantify how individual ESEs and ESSs combine to contribute to exon recognition in a small number of exon contexts, and can be used to build more general predictive models for exon splicing. Recently, a survey of disease variants within a much broader set of human exons found that ~10% of these variants had exon recognition defects[20]. Despite the recent progress, there are still several limitations inherent to these large-scale approaches. First, these reporters often assay exons in the contexts of short background intronic sequences, which have been shown to impact exon skipping and intron retention[22]. Second, most previous studies use transient transfections that do not reflect physiological chromatin contexts[23] and are usually highly overexpressed, which can lead to saturation of the splicing machinery[24,25]. Finally, most of these assays cannot screen both intronic and exonic changes simultaneously.

Here we develop a novel multiplexed assay that overcomes many of these shortcomings called MFASS (Multiplexed Functional Assay of Splicing by Sort-seq) that builds upon several previous approaches (Fig. 1A). MFASS allows testing of tens of thousands of chemically-synthesized exons and surrounding introns in the context of a reporter with long constant introns, stably integrated at single copy at a precise genomic locus with high efficiency (Supp. Fig. 1). Briefly, we split a GFP coding sequence with a constant intron backbone, with a downstream mCherry fluorescent marker to act as a control. Thus, the ratio of green to red fluorescence is a direct measure of exon inclusion. This is reminiscent of past approaches[13,14] but optimized for large libraries[26], readout by next-generation sequencing, and optimized to study exon definition[13] (Supp. Fig. 2). The library of exons and surrounding native intronic sequences is cloned into this constant intron backbone. We then integrated the plasmid library into an engineered serine integrase-based landing pad at the AAVS1 locus in HEK293T cells, ensuring only one integrant per cell, similar to recently published high-efficiency integration methods[26,27] (Supp. Fig. 1, 3). We sorted the integrated cell library into bins based on the GFP:mCherry ratio, followed by DNA-Seq of the integrated library (similar to past Sort-seq approaches[28–31]) to build a quantitative measure of exon inclusion level of any designed sequence.

We first designed, built and assayed a library to explore how Splicing Regulatory Elements (SRE) individually govern exon recognition across a randomly-chosen library of 205 natural human exons and surrounding intronic sequences (Figure 2A). We used fluorescence-activated cell sorting (FACS) to sort our pooled sequence library of splicing reporters into three bins (GFP_{neg}, GFP_{int} and GFP_{+}). We expanded these sorted bins over several passages and observed that the sorted populations remained stable (Fig. 1B). We also performed bulk RT-PCR for each bin and found that the observed RNA splicing efficiencies corresponded almost directly with observed fluorescence of the bins (Fig. 1C, Supp. Fig. 4). In addition, we constructed individual reporters corresponding to individual library sequences, and evaluated both fluorescence and RNA splicing under transient expression and site-specific genome integration (Supp. Fig. 5). While level of exon inclusion as measured by RT-PCR is consistent between transient and stable expression, reporter fluorescence in stably integrated constructs is more consistent with RT-PCR results because the transient transfections included signals at very high gene dosage (Supp. Fig. 4, 5).

For our SRE library studies, we first tested a variety of short constant intron contexts, but found that these resulted ~10-fold lower expression indicative of intron retention (Supp. Fig. 6), which is usually a rarer event in higher eukaryotes that contain longer introns[32]. We chose two longer intronic backbones (~300-600 bp) shown previously to not suffer from such intronic retention (*C. griseus* DHFR and human SMN1 intron backbones), and found that the longer intron lengths improved both expression and assay reproducibility[33,34]. Exon inclusion metrics obtained from both of these intron contexts were highly reproducible between biological replicates (Fig. 1E) (r = 0.94, $p < 10^{-16}$, DHFR intron backbone, and r = 0.89, $p < 10^{-16}$, SMN1 intron backbone). Exon inclusion level for the entire library also correlates highly across DHFR and SMN1 constant intron contexts (Fig. 1E) (r = 0.85, $p < 10^{-16}$), indicating our reporter assay is robust across broader intron contexts. Notably, most library sequences are represented predominantly in one exclusive bin showing either complete exon inclusion or skipping (Fig. 1D), consistent with bimodality in splicing behavior in our flow cytometry readout (Fig. 1B) and in single cells[35–37]. For all subsequent analyses, we only include constructs with Δinclusion index that agree within 0.30 for both biological replicates and across intron backbones.

We designed the SRE library using a software tool that we developed, Splicemod, that can iteratively mutate specific classes of regulatory elements that govern splicing without unintentionally creating new ones (Fig. 2A; Supp. Table 2). As expected, reducing the strength of the splice acceptor (SA) and splice donor (SD) adversely affects exon inclusion (Fig. 2B). We observe a significant correlation between decreased MaxEnt[38] score (relative to wild-type) and Δinclusion index for both SA (r = 0.33, $p < 10^{-16}$) and SD (r = 0.36, $p < 10^{-16}$) (Fig. 2B). The change in score for both SA and SD combined explains 14% of the variation in Δinclusion index (multiple linear regression, $p < 10^{-16}$). Variants designed to mutate SA and/or SD but retain comparable strength (i.e. same MaxEnt score) show that while the majority (79.2%, 236/298) shows little change relative to wild-type (-0.20 ≤ Δinclusion index ≤ 0.20), 16% (48/298) of variants exhibit large effects with Δinclusion index ≤ -0.50 (Splice-Disrupting Variants, SDVs). Taken together, while MaxEnt scores do correlate with function, there seems to be a context dependence that is not accounted for in the score alone.

Perturbations to ESEs result in a significant decrease in exon inclusion compared to random exonic changes (Mann-Whitney *U* test, $p < 10^{-16}$), while weakening or destroying ESSs results in a small but significant increase in exon inclusion (Mann-Whitney *U* test, $p = 1.33 \times 10^{-4}$). Interestingly, disrupting only the strongest ESE results in a significant decrease in Δinclusion index (Mann-Whitney *U* test, $p = 2.42 \times 10^{-7}$). We calculated an average exon hexamer score for each sequence using the HAL model, which is learned from synthetic mini-genes focused on alternative 5' and 3' splicing[8] (Fig. 2C). We quantified the change in average exon hexamer score as the difference relative to the wild-type (Δaverage exon hexamer score) and found a correlation with Δinclusion index ($r = 0.26$, $p < 10^{-16}$) and a significant difference between mutants that increase or decrease the average score (two-tailed Student's *t* test, $p < 10^{-16}$). Compared to random intronic changes, we found that weakening or destroying intronic motifs does not have an overall significant effect on exon inclusion (Mann-Whitney *U* test), although 9.4% (63/672) of these mutants are SDVs. Additionally, we designed mutations that disrupt 53 RNA-binding protein (RBP) motifs and found small changes in Δinclusion index relative to random mutations (Mann-Whitney *U* test, $p = 2.08 \times 10^{-4}$ (intronic), $p = 3.80 \times 10^{-2}$ (exonic)), with 14.1% (48/341) being SDVs. We synthesized 109 dbSNP mutations but do not observe significant changes in Δinclusion index (as compared to random changes) for either exonic or intronic single nucleotide polymorphisms (SNPs)[39] (Mann-Whitney *U* test).

Given the appreciable proportions of SDVs across many classes of elements, we sought to examine the extent to which rare human variants act as SDVs. We first examined a larger library of 4660 natural human exons and found that 2902 exons (62.2%) have an inclusion index of ≥ 0.80 in our assay (Fig. 3A). Based on these human sequences, we designed and synthesized all possible exonic and intronic single nucleotide variants (SNVs) from the Exome Aggregation Consortium[2] (ExAC, v0.3.1) (Fig. 3B), which represents a rich resource of genetic diversity from 60,706 individuals. We were able to quantify the effects of 29,531 SNVs across 2393 reference sequences, which is more than half (54.7%, 29,531/54,021) of those found in the ExAC for these exons (Fig. 3B). We evaluated all SNVs in the DHFR intron backbone, because the backbone provided more replicable data in the SRE datasets. We also only report data for variants with calculated Δinclusion index within 0.20 between biological replicates to be more conservative with potential SDVs (r = 0.80, $p < 10^{-16}$) (Fig. 1E; Supp. Fig. 11). We also included four control sets: (1) random nucleotides, (2) a previously tested set of skipped exons in the SRE library, (3) systematic mutations of both the splice donor and acceptor of wild-type sequences, (4) and two reporter constructs that split at distinct positions of GFP to assess how reading frame affects exon inclusion. 100% of random sequences (n = 27), 98.6% of skipped exons (n = 95), and 97.3% of broken SD/SA sequences (n = 1391) demonstrate exon skipping (inclusion index < 0.50) (Supp. Fig. 10). Moreover, Δinclusion indices across two separate reporter constructs located in different parts of GFP and in different frames demonstrate robust correlation (r = 0.95, $p < 10^{-16}$, Supp. Fig. 11).

Overall, we found that 3.6% (1050/29,531) of ExAC SNVs leads to large-effect splicing disruptions in exon recognition, and are spread broadly across human exon backgrounds (Fig. 3B). The annotations in ExAC use the Variant Effect Predictor classification[40], and we find that

67.8% of splice site SNVs (2 bp of intron adjacent to exon) are SDVs (Fig. 3D). Note that in our assays, alternative 5' and 3' splice site usage will be called as false negatives and thus we may be missing other potential SDVs. Variants in the broader splice region category, which includes variants located 2 bp into the exon and 8 bp into the intron (excluding splice sites), only disrupt splicing 8.5% of the time. Synonymous, non-synonymous, and further intronic SNVs disrupt splicing more rarely at 3.0%, 3.1%, and 1.5% respectively. The increased sensitivity at splice site locations mirror added evolutionary constraints at these sites (Fig. 3C). However, SNVs at splice sites are rare in our library and also for all ExAC variants as a whole (Fig. 3C, Supp. Fig. 12), and the larger number of SNVs in other regions makes up for their reduced sensitivity (Fig. 3D). Notably, SNVs at splice sites only constitute 17% of the SDVs revealed by our assay, whereas intron variants, which are the least sensitive to genetic variation, contribute 19% of the SDVs (Fig. 3D). Overall, we observe almost equal contributions from intronic (53%) and exonic (47%) SDVs.

Evolutionary conservation does correlate with whether an SNV will be an SDV, and this is most clearly seen within introns, which are enriched for highly conserved SDVs (Fig. 4A) (two-sided Fisher's exact test, $p < 10^{-16}$). However, this conservation has limited predictive power, as there are more lowly conserved intronic SDVs than highly conserved ones especially for upstream intronic regions, while there are few poorly conserved exonic sites (Fig. 4B). Looking at gene level population genetic constraints, for exons within those genes that are predicted to be intolerant to loss-of-function (pLI ≥ 0.9), we observe significantly fewer SDVs (Fig. 4C) (two-sided Fisher's exact test, $p = 2.67 \times 10^{-12}$). Finally, while a vast majority of SDVs are rare, the proportion of SNVs that are SDVs is significantly different across ExAC allele frequency bins

(p = 1.12 x 10^{-3}, chi-squared test) ranging from extremely rare variants (singletons) to more common variants with allele frequency of ≥ 0.1% (Fig. 4D).

We compared multiple prediction algorithms to our human variant dataset, some designed specifically for splicing (SPANR[7] and HAL[8]) and others to predict the impact of non-coding genetic variation (CADD[41], DANN[42], FATHMM-MKL[43], fitCons[44], and LINSIGHT[45]) (Fig. 4E, Supp. Fig. 13). Overall, we find that the two algorithms specifically designed for and trained on splicing data perform the best, mostly due to their ability to distinguish exonic SDVs (HAL only predicts exonic SNVs). Most of the models that use conservation and other functional attributes perform equally well on intronic SNVs. In particular, SPANR works best overall largely due to its increased ability to differentiate exonic SDVs (Fig. 4E, right; Supp. Fig. 13). At equivalent effect size (>50%), SPANR achieves 44.5% precision, though only 11.8% of the SDVs are called. However, SPANR is trained on bulk RNA-Seq data, and thus effect sizes can be skewed. As we lower the threshold for calling an SDV (i.e., the predicted effect size of an SNV), SPANR can achieve 14.9% precision at 50% recall level (of the SDVs called). For the other prediction algorithms, precision is below 10% at most appreciable recall levels.

As with other functional approaches, our assay has several limitations which must be considered[46]. *First*, we only perform this assay in a single cell type (HEK293T), and thus there might be trans-factors that mitigate or exacerbate splicing[47]. Using MFASS in other cell types will be important to understand the scope of these effects. *Second*, the tested regions are surrounded by non-native intron sequence that might affect the propensity of variants that affect splicing[48]. *Third*, because MFASS depends upon FACS, our limit of detection can only reliably observe large effect sizes. For calling SDVs this is tolerable, and it seems likely that only large-effect changes will translate across cell types. However, small-effect changes might be important both functionally and for constraining predictive models. *Fourth*, MFASS as designed can only observe full exon skipping events. Even though these events dominate a majority of splicing perturbations, other types of splicing disruptions, including alternative 3' and 5' splice-site usage, are likely to be false negatives from MFASS. Other multiplexed splicing assays that use barcoded RNAs can alleviate such issues, but are currently limited to short intronic regions[8,21]. *Fifth*, in this study we only examine exons starting and ending on frame 0. Since skipping an exon that preserves frame might be less deleterious than for frame-shifting exons, our library selected here may suffer from selection bias, even though we find no appreciable differences in conservation profiles between the two (Supp. Fig. 14). We also found during this study that several of the plasmids developed for MFASS can be directly used to screen for frame-shifting exons. Finally, oligonucleotide libraries such as those used here are limited to ~200nt in length. This limits the size of exons we can explore, which can also lead to selection bias in that short exons of <100 bp may be more sequence constrained. This also limits the length of the surrounding intronic sequences, which could serve to buffer or alter the effects of sequence variation (Supp. Fig. 15). As oligonucleotide and gene library synthesis improves, we expect to include additional genetic context in the assays[49,50].

Despite the limitations, we see clear indications that many more rare variants than we expected can lead to large-effect splicing disruptions. More than >1000 SDVs discovered in this study are variants that directly eliminate exon recognition, and we reason that such large-effect SDVs seem to be the most likely to translate to other cell types and/or play a role in human traits and diseases. In addition, because almost all the candidate SDVs are extremely rare, genome-wide splicing quantitative trait loci (sQTL) studies may be underestimating much of how mutations affect traits through splicing[3,51]. More broadly, using multiplexed empirical models of important biological processes, such as ones derived from MFASS, can both help build and provide an alternative to improved computational models. Finally, given the propensity of large-effect regulatory variants that disrupt splicing discovered here, MFASS provides a scalable platform to functionally screen and aid precise clinical interpretation and prioritization of rare genetic variants[52].

Acknowledgments

This work was supported by the National Institutes of Health (5U01HG007912 & DP2GM114829 to S.K.), the NIH Biomedical Big Data Training Grant (T32CA201160 to C.B.), Searle Scholars Program [to S.K.], Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Ron Weiss for the original landing pad cell line, Felicia Codrea and Jessica Scholes (UCLA BSCRC flow cytometry core), and the BSCRC high throughput sequencing core for technical assistance. We thank Xinshu Grace Xiao, Douglas Black, and George Church for guidance while developing MFASS.

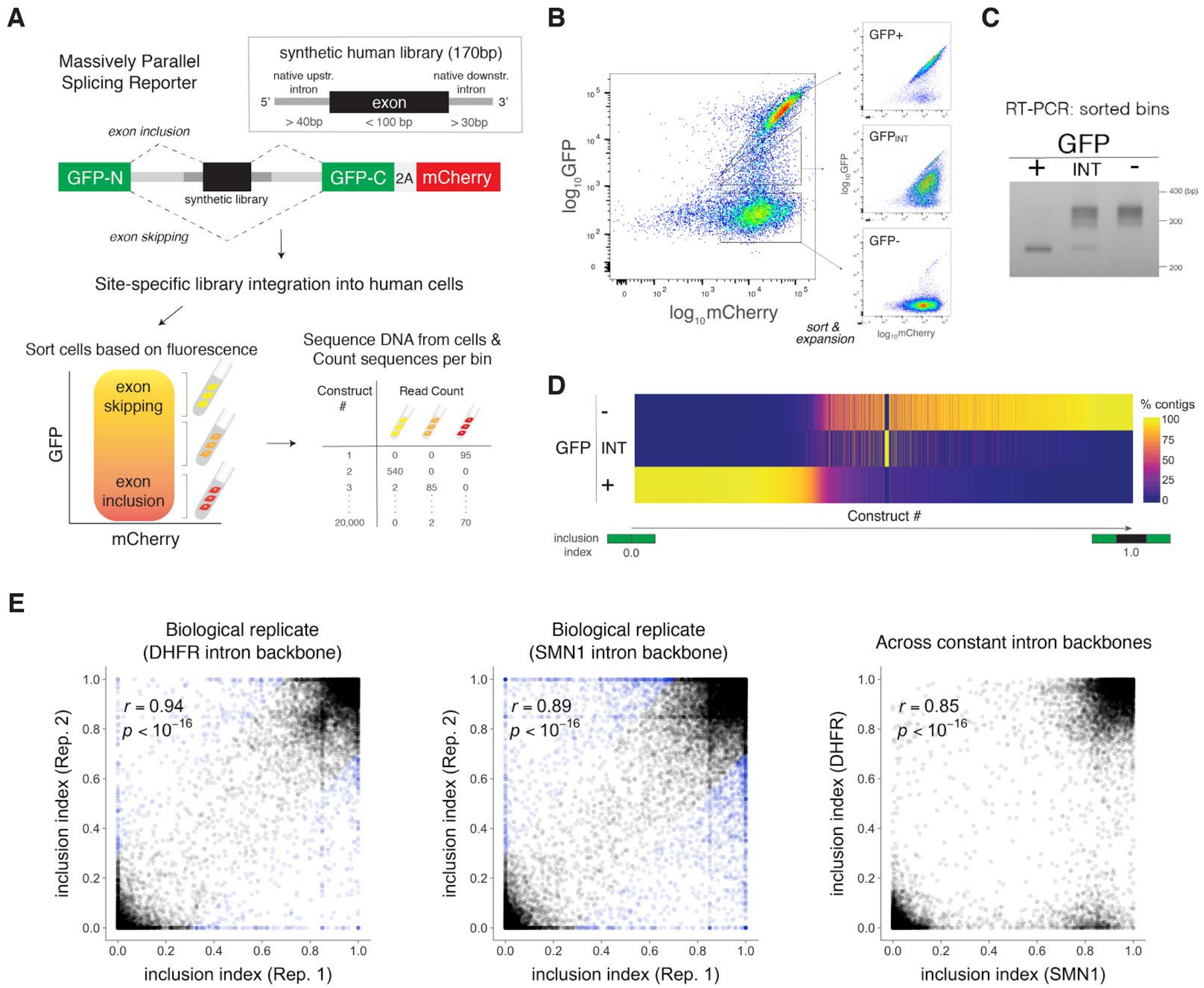

Figure 1. Multiplexed Functional Assay of Splicing by Sort-seq (MFASS). (A) We clone synthetic exons (black) and surrounding intronic sequences (dark grey) into our reporter plasmid containing a split-GFP reporter, followed by site-specific integration into HEK293T cells using Bxb1 integrase. Cells are sorted into bins based on GFP:mCherry fluorescence, followed by amplicon sequencing from cells in each sorted bin. The normalized, weighted average of sequence counts across bins reflects splicing efficiencies. **(B)** We used FACS to sort the genomically-integrated SRE library into three separate populations (left). After expansion, the sorted populations remained stable (right). **(C)** The observed RNA splicing efficiencies of the sorted bins as measured by RT-PCR correspond directly with observed fluorescence of the bins. **(D)** We plotted the percentage of reads for each construct in the SRE library (n = 10,683) and show that most fall predominantly into one bin, exhibiting either complete exon skipping or inclusion. **(E)** Exon inclusion indices show strong correlation between two independent biological replicates for *C. griseus* DHFR intron backbone ($r = 0.94$, $p < 10^{-16}$) and human SMN1 intron backbone ($r = 0.89$, $p < 10^{-16}$) (left and middle). (Points in blue indicate inclusion indices

for sequences that do not agree within 0.30.) Results are robust across different intron backbones (r = 0.85, $p < 10^{-16}$) (right). The data shown in B, C, and D are for the SMN1 backbone (for DHFR backbone see Supp. Fig. 7).

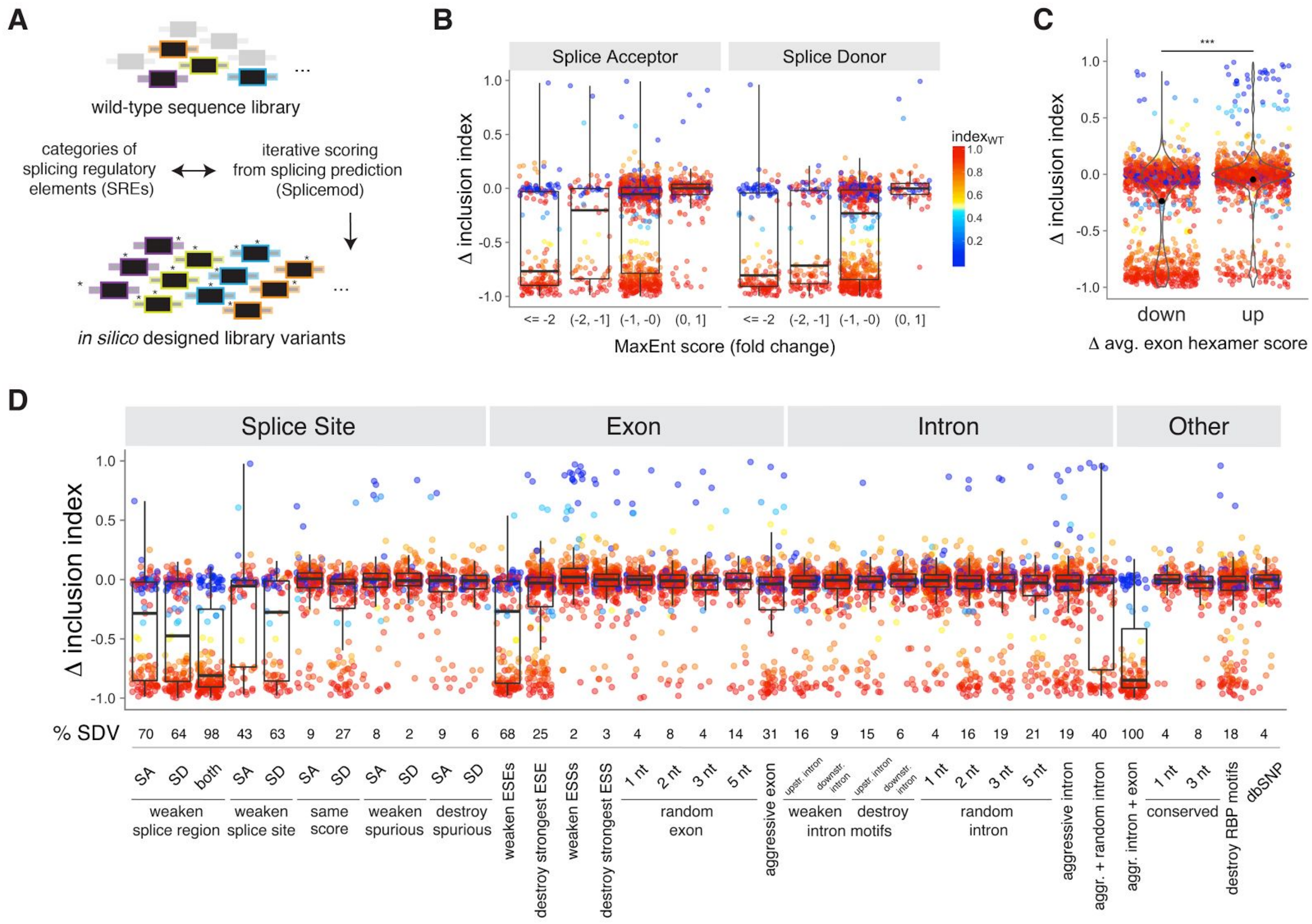

Figure 2. Analysis of SRE mutations across a broad range of human exon backgrounds. **(A)** We randomly chose 205 human exon backbones and used Splicemod to design variants intended to alter SREs. Splicemod is a software tool we developed to iteratively design mutations that are intended to alter SREs while accounting for other changes. We quantified 10,683 mutants (SMN1 intron backbone) and 8942 mutants (DHFR intron backbone) across these exons. **(B)** We quantify Δinclusion index for a mutant sequence as difference relative to wild-type (WT), and labeled the inclusion index of the corresponding WT sequence by color. Points colored blue indicate skipped WT exons, therefore mutations can only increase the inclusion index (+ Δinclusion index). Conversely, points colored red indicate fully included WT exons, and mutations can only decrease the inclusion index (- Δinclusion index). Data shown here is for the SMN1 dataset (see Supp. Fig. 8 for DHFR dataset; we only plot those points here that agree across both datasets). Median Δinclusion index for each class is indicated by overlaid boxplots. We find that weakening splice acceptor and donor sequences adversely affects exon inclusion based on MaxEnt prediction. **(C)** We find a significant difference in Δinclusion index between sequences that increase (up) or decrease (down) average exon hexamer score (Mann-Whitney *U* test, $p < 10^{-16}$). Decreasing hexamer strength leads to more exon skipping. Scores are based on the HAL model[8], and an alternative exon hexamer score metric is evaluated in Supp. Fig. 9[14]. **(D)** Quantitative measures of exon inclusion for mutations across

multiple classes of splicing regulatory elements. Splice-disrupting variants (SDVs) defined as Δinclusion index ≤ -0.50, and percentage of SDVs indicated below each class (only those natural exons that are >50% inclusion without mutation are used for this calculation). ESE, exonic splicing enhancer. ESS, exonic splicing suppressor. RBP, RNA-binding protein. SA, splice acceptor. SD, splice donor.

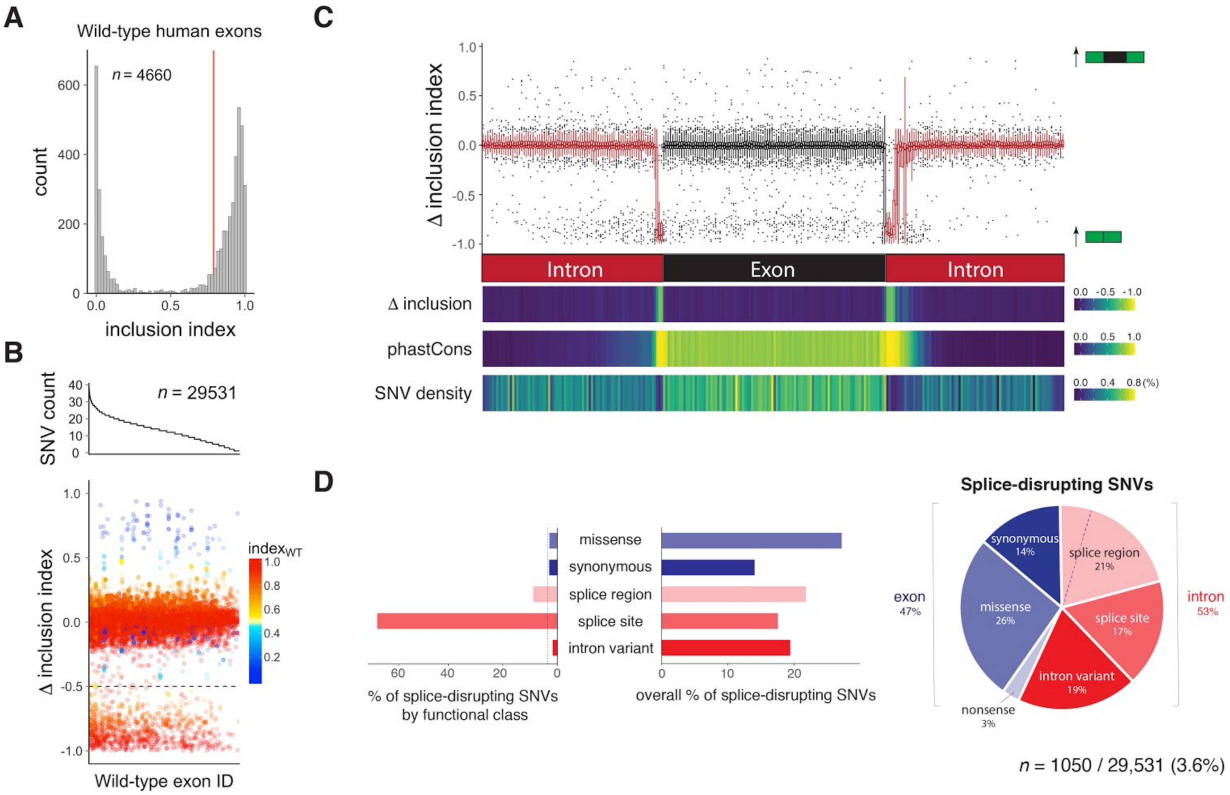

Figure 3. Testing 29,531 SNVs in ExAC finds SDVs broadly spread across human exons and surrounding intron regions. (A) We tested 4660 human exons for inclusion, and generated all SNVs found in the Exome Aggregation Consortium[2] (ExAC) for the 2902 exons with inclusion index ≥ 0.8. **(B)** The number of SNVs per variant (top panel) and the Δinclusion index (bottom panel) of the 29,531 ExAC SNVs plotted against wild-type exon ID (*n* = 2393). Top and bottom plots are ordered in decreasing number of variants from 44 to 1 per exon background, with an average of 13.9 human variants. Dashed line indicates threshold (Δinclusion index = -0.50) below which we call splice-disrupting variants (SDVs). **(C)** We plot Δinclusion index versus relative location of SNVs (top panel). The whiskers indicate 1.5-fold interquartile ranges. Δinclusion index, phastCons scores and SNV density averaged per scaled position for our SNV library (bottom panel). Each bin corresponds to 1-2 nucleotide per position. **(D)** Proportion of SDVs by functional class (left) and overall contribution (%) to SDVs by absolute number (middle), based on the Variant Effect Predictor[40]. Despite the higher sensitivity at the splice site, the number of SNVs in other regions dominates such that SDVs at splice sites only comprise 17% of total SDVs. Right panel shows contribution of functional classes to SDVs (*n* = 1050) segregated by exon and intron regions, which contribute roughly equally to SDVs. Splice region variants in exons (4%) and introns (17%) are separated by a dashed line.

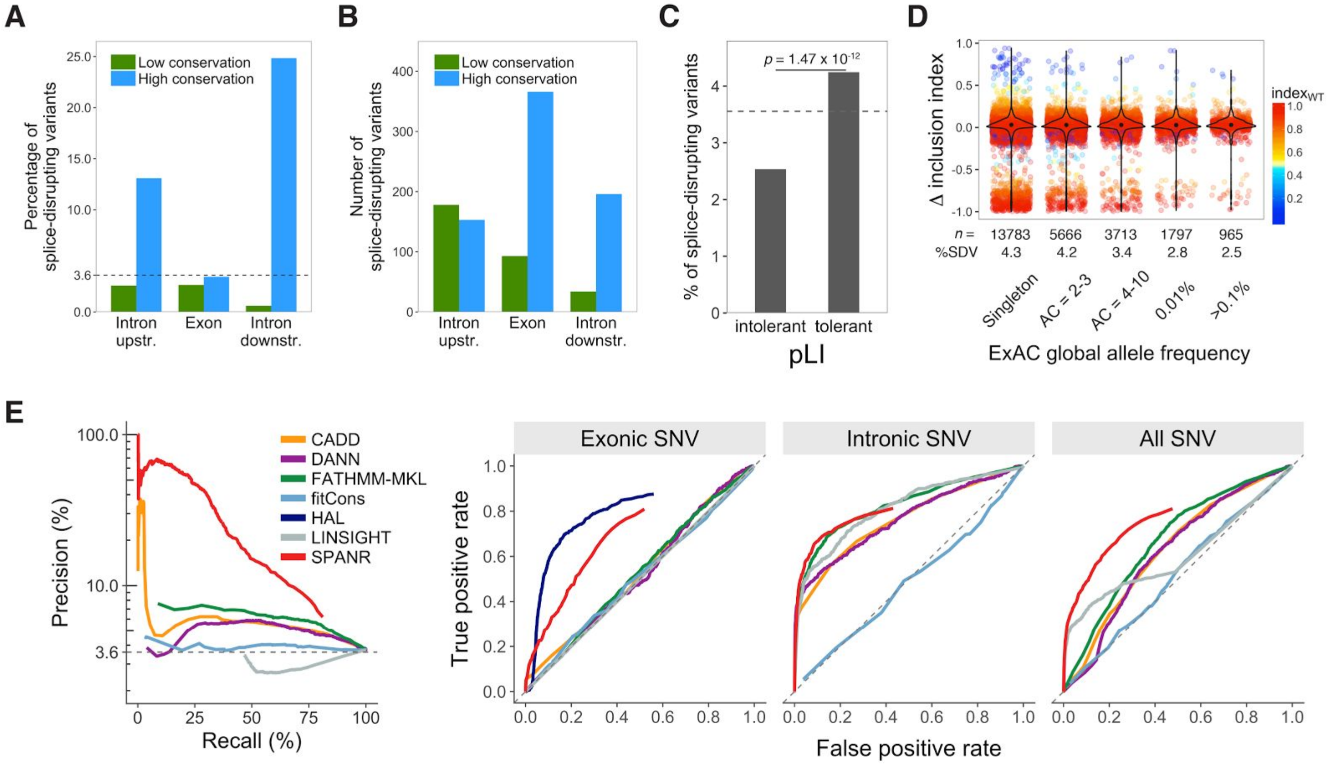

Figure 4. Characteristics of SDVs. (A) The proportion of SDVs with low or high phastCons conservation (<0.5 and ≥0.5). The overall percentage of SDVs is shown as a dashed line (3.6%). Introns, but not exons, are enriched for highly conserved SDVs (two-sided Fisher's exact test, $p < 10^{-16}$). **(B)** We plot the number of SDVs with low or high phastCons conservation, as opposed to percentage. **(C)** We observe significantly fewer SDVs for exons within those genes that are predicted to be intolerant to loss-of-function (pLI ≥ 0.9) (two-sided Fisher's exact test, $p = 2.67 \times 10^{-12}$). The overall percentage of SDVs is 3.6% (dashed line). **(D)** The change in exon inclusion index (Δinclusion index) plotted against the allele frequency spectrum. For each allele frequency bin, number of tested variants and % SDV in that bin is indicated at the bottom, and violin plot shows distribution of variants across Δinclusion index with median Δinclusion index (dot) indicated for each bin. The percentage of SDVs as a function of observed allele frequency shows a negative trend and is significantly different across allele frequencies (chi-squared test, $p = 1.12 \times 10^{-3}$). **(E)** Precision-recall curves (left) and receiver operating characteristic (ROC) curves (right) for algorithms that can predict splicing or non-coding genetic variants. For ROC curves (right), colors for each algorithm match those in the left panel, with the addition of the HAL predictor that evaluates exonic changes only. Dashed line (left) indicates overall percentage of SDVs determined by MFASS.

References

1. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

2. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.

3. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660.

4. Zhang X, Joehanes R, Chen BH, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet*. 2015;47(4):345-352.

5. Li YI, van de Geijn B, Raj A, et al. RNA splicing is a primary link between genetic variation and disease. *Science*. 2016;352(6285):600-604.

6. Telenti A, Pierce LCT, Biggs WH, et al. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A*. 2016;113(42):11901-11906.

7. Xiong HY, Alipanahi B, Lee LJ, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. 2015;347(6218):1254806.

8. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*. 2015;163(3):698-711.

9. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. *Nature*. 2010;465(7294):53-59.

10. Ast G. How did alternative splicing evolve? *Nat Rev Genet*. 2004;5(10):773-782.

11. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA*. 2013;4(1):49-60.

12. Sun H, Chasin LA. Multiple Splicing Defects in an Intronic False Exon. *Mol Cell Biol*. 2000;20(17):6414-6425.

13. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. *Cell*. 2004;119(6):831-845.

14. Ke S, Shang S, Kalachikov SM, et al. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res*. 2011;21(8):1360-1374.

15. Goren A, Ram O, Amit M, et al. Comparative Analysis Identifies Exonic Splicing Regulatory Sequences—The Complex Definition of Enhancers and Silencers. *Mol Cell*. 2006;22(6):769-781.

16. Chasin LA. Searching for splicing motifs. *Adv Exp Med Biol*. 2007;623:85-106.

17. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human

genetic diseases. *Trends Biochem Sci*. 2000;25(3):106-110.

18. Wainberg M, Alipanahi B, Frey B. Does conservation account for splicing patterns? *BMC Genomics*. 2016;17(1):787.

19. Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics*. 2014;30(12):i121-i129.

20. Soemedi R, Cygan KJ, Rhine CL, et al. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. 2017;49(6):848-855.

21. Adamson SI, Zhan L, Graveley BR. High-Throughput Identification of Genetic Variation Impact on pre-mRNA Splicing Efficiency. 2017. doi:10.1101/191122.

22. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. *Nat Rev Genet*. 2016;17(7):407-421.

23. Inoue F, Kircher M, Martin B, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res*. 2017;27(1):38-52.

24. Berg MG, Singh LN, Younis I, et al. U1 snRNP determines mRNA length and regulates isoform expression. *Cell*. 2012;150(1):53-64.

25. Munding EM, Shiue L, Katzman S, Donohue JP, Ares M Jr. Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. *Mol Cell*. 2013;51(3):338-348.

26. Duportet X, Wroblewska L, Guye P, et al. A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res*. 2014;42(21):13440-13451.

27. Matreyek KA, Stephany JJ, Fowler DM. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res*. 2017;45(11):e102.

28. Kinney JB, Murugan A, Callan CG Jr, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A*. 2010;107(20):9158-9163.

29. Sharon E, Kalma Y, Sharp A, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012;30(6):521-530.

30. Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. *Science*. 2013;342(6157):475-479.

31. Kosuri S, Goodman DB, Cambray G, et al. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A*. 2013;110(34):14024-14029.

32. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon

definition and function. *Nat Rev Genet*. 2010;11(5):345-355.

33. Arias MA, Lubkin A, Chasin LA. Splicing of designer exons informs a biophysical model for exon definition. *RNA*. 2015;21(2):213-229.

34. Cho S, Moon H, Loh TJ, et al. Splicing inhibition of U2AF65 leads to alternative exon skipping. *Proc Natl Acad Sci U S A*. 2015;112(32):9926-9931.

35. Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. 2013;498(7453):236-240.

36. Faigenbloom L, Rubinstein ND, Kloog Y, Mayrose I, Pupko T, Stein R. Regulation of alternative splicing at the single-cell level. *Mol Syst Biol*. 2015;11(12):845.

37. Song Y, Botvinnik OB, Lovci MT, et al. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Mol Cell*. 2017;67(1):148-161.e5.

38. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394.

39. Sherry ST. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311.

40. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.

41. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.

42. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2015;31(5):761-763.

43. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015;31(10):1536-1543.

44. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet*. 2017;49(4):618-624.

45. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet*. 2015;47(3):276-283.

46. Singh G, Cooper TA. Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques*. 2006;41(2):177-181.

47. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*. 2012;338(6114):1593-1599.

48. Elkon R, Agami R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol*. 2017;35(8):732-746.

49. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods*. 2014;11(5):499-507.

50. Plesa C, Sidore AM, Lubock N, Zhang D, Kosuri S. Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes. 2017. doi:10.1101/163550.

51. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. 2013;24(1):14-24.

52. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469-476.