## Title

A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions

Journal
Molecular Cell, 73(1)

ISSN
1097-2765

Authors
Cheung, Rocky
Insigne, Kimberly D
Yao, David
et al.

Publication Date
2019

DOI
10.1016/j.molcel.2018.10.037

Peer reviewed

Published in final edited form as:
*Mol Cell*. 2019 January 03; 73(1): 183–194.e8. doi:10.1016/j.molcel.2018.10.037.

# A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions

**Rocky Cheung**[1,9], **Kimberly D. Insigne**[2,9], **David Yao**[3], **Christina P. Burghard**[2], **Jeffrey Wang**[1], **Yun-Hua E. Hsiao**[4], **Eric M. Jones**[1], **Daniel B. Goodman**[5], **Xinshu Xiao**[2,6,7], and **Sriram Kosuri**[1,7,8,10,*]

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

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

[3]Department of Genetics, Stanford University, Stanford, CA 94035, USA

[4]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA 90095, USA

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

[6]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA 90095, USA

[7]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

[8]UCLA-DOE Institute for Genomics and Proteomics, 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, Los Angeles CA 90095, USA

[9]These authors contributed equally

[10]Lead Contact

## SUMMARY

[*]Correspondence: sri@ucla.edu.

AUTHOR CONTRIBUTIONS

R.C. and S.K. designed the study. R.C., D.Y., and J.W. developed and validated MFASS. K.D.I. and R.C. developed analysis methods and analyzed and interpreted the data. C.P.B., Y.-H.E.H., and X.X. provided additional bioinformatics analysis. D.B.G. and E.M.J. implemented early prototypes of MFASS. R.C., K.D.I., and S.K. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

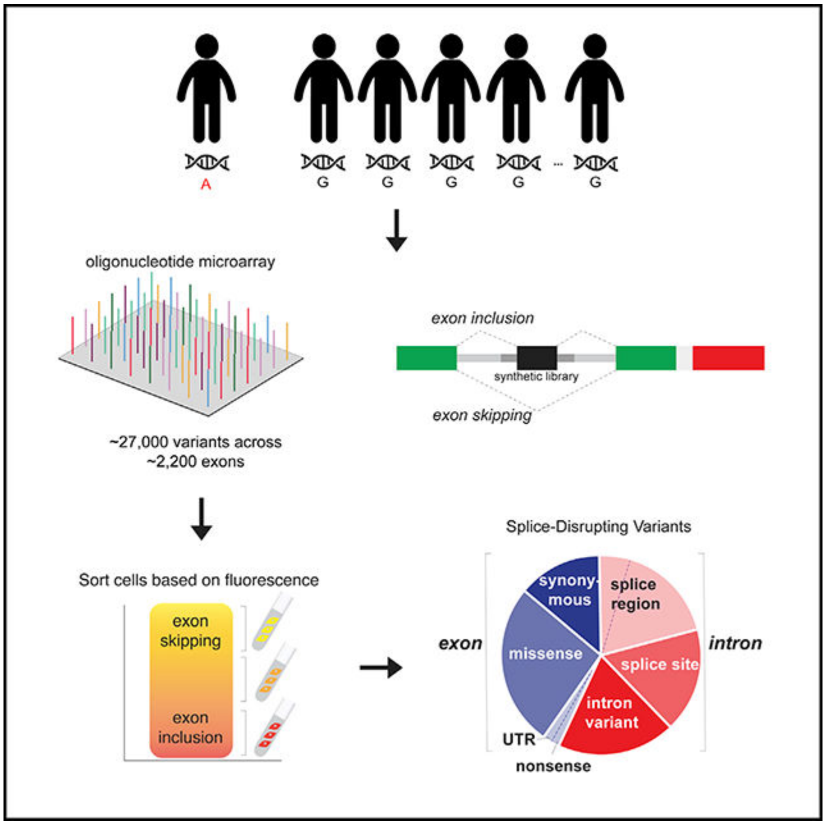
Supplemental Information includes six figures and four tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.10.037.

Mutations that lead to splicing defects can have severe consequences on gene function and cause disease. Here, we explore how human genetic variation affects exon recognition by developing a multiplexed functional assay of splicing using Sort-seq (MFASS). We assayed 27,733 variants in the Exome Aggregation Consortium (ExAC) within or adjacent to 2,198 human exons in the MFASS minigene reporter and found that 3.8% (1,050) of variants, most of which are extremely rare, led to large-effect splice-disrupting variants (SDVs). Importantly, we find that 83% of SDVs are located outside of canonical splice sites, are distributed evenly across distinct exonic and intronic regions, and are difficult to predict a *priori*. Our results indicate extant, rare genetic variants can have large functional effects on splicing at appreciable rates, even outside the context of disease, and MFASS enables their empirical assessment at scale.

## In Brief

Mutations that lead to splicing defects can have severe consequences on gene function and cause disease. Cheung et al. developed MFASS, which enables largescale screening for splicing defects. They tested tens of thousands of natural human genetic variants across a broad range of exons and revealed a surprising fraction of rare, large-effect variants that disrupt exon recognition.

## Graphical Abstract



## INTRODUCTION

Any individual's genome contains about 4 to 5 million genetic variants that differ from reference, and understanding how these variants give rise to trait diversity and disease susceptibility is a central goal of human genetics (Auton et al., 2015). A vast majority (96%–99%) of an individual's variants are common, though at the population level, the overwhelming majority of variants are rare (Montgomery et al., 2011; Nelson et al., 2012; Tennessen et al., 2012; UK10K Consortium et al., 2015). Common variants in the human population usually contribute small, additive effects toward complex traits, as negative selection has removed large-effect deleterious alleles (Altshuler et al., 2008). However, population expansion ~10,000 years ago left humans with an abundance of rare variation, and most Mendelian disease traits are caused by rare alleles with large effect sizes (Keinan and Clark, 2012). Because of their scarcity in an individual's genome, rare variants that play important roles in complex traits are likely to have large functional effects (Bomba et al., 2017), and traditional population or computational genomic methods cannot reliably estimate their contribution (Uricchio et al., 2016).

Recent whole-genome and transcriptome sequencing studies of large cohorts indicate that rare variation is playing an important role in shaping global gene expression (Battle et al., 2017; Hernandez et al., 2017; Li et al., 2017). However, new comprehensive reverse-genetic studies indicate that individual mutations in promoter and enhancer regions rarely have large effects (Canver et al., 2015; Diao et al., 2016; Gasperini et al., 2017; Rajagopal et al., 2016; Sanjana et al., 2016), which could be the result of functional redundancy between transcription control elements (Frankel et al., 2010; Hong et al., 2008; Osterwalder et al., 2018). How can individual rare variants be broadly shaping gene expression but at the same time rarely having large effects on transcriptional control? We can expect the mutational profiles of large-effect rare variants to mirror those that cause Mendelian traits, which are dominated by non-synonymous exonic mutations, structural and copy number variants, or mutations that affect splicing (Bamshad et al., 2011; Chong et al., 2015). Although copy number changes and non-synonymous mutations are easy to detect, splicing changes are more difficult to diagnose, as only mutations at canonical splice sites are easy to predict and interpret (Jian et al., 2014).

Recent evidence indicates that splicing is a major mechanism by which genetic variation influences traits. For common variants, large-cohort RNA sequencing (RNA-seq) studies that examine splicing are finding many splicing quantitative trait loci (sQTL), especially when considering exon-level expression differences (Battle et al., 2017; Ongen and Dermitzakis, 2015; Zhang et al., 2015). Moreover, a majority of eQTLs (expression quantitative trait loci) tend to act on an individual exon level rather than the gene level, indicating that *cis*-eQTLs might be broadly affecting exon recognition (Ramasamy et al., 2014). In addition, functional genomic measurements of GEUVADIS individuals indicate that common genetic variation influencing splicing is a primary mechanism that confers susceptibility to common diseases (Li et al., 2016). For rare variation, analysis of bottlenecked populations finds that many rare variants that segregate with large-effect expression changes are enriched at splice sites (Pala et al., 2017). In addition, prospective transcriptional profiling studies for Mendelian diseases are increasingly finding many rare variants that affect splicing are difficult to predict a *priori* (Cummings et al., 2017; Kremer et al., 2017). More broadly, computational splicing predictors trained on RNA-seq data and sequence features seem to indicate that many rare and disease variants are predicted to influence splicing levels (Xiong et al., 2015). Finally, mutations that cause an exon to be skipped can have severe functional consequences on gene function, and many known disease-causing mutations reduce or eliminate exon recognition (Soemedi et al., 2017; Baralle and Buratti, 2017).

We developed multiplexed functional assay of splicing using Sort-seq (MFASS) as a multiplexed, scalable platform to test the extent to which mutations, both within exons and introns, can lead to large-effect defects in exon recognition. MFASS uses a set of three-exon, two-intron minigene reporters, in which skipping of the middle exon leads to reconstitution of fluorescence (Figures 1A and S1A–S1D). We cloned libraries of microarray-derived oligonucleotides that encoded human exons and surrounding intronic sequences into these reporters *en* masse to construct reporter libraries. These libraries are then integrated into HEK293T human cell lines using high-efficiency, serine-integrase-based, site-specific integration (Figure 1A), ensuring one copy of library sequence per cell (Duportet et al., 2014). The pooled sequence library is then separated into bins using fluorescence-activated cell sorting (FACS), and we use DNA-seq of the constructs to quantify the variants. We used MFASS to functionally classify 27,733 exonic and intronic natural genetic variants from Exome Aggregation Consortium (ExAC) for exon recognition across 1,626 genes in 2,198 exon backgrounds, most of which are extremely rare variation in the human population. Here, we show that more than a thousand (3.8%) of these rare genetic variants leads to near complete loss of exon recognition, on par with the prevalence of protein-truncating variants within genomes. Most of the effects of rare variants on splicing are challenging to predict.

# RESULTS

## Optimization of MFASS

We tested human exons in several reporter designs. Our initial designs relied on the reconstitution of fluorescence using a pair of constant short dihydrofolate reductase (DHFR) introns (~100 bp) flanking the exon library (Figure S1D). However, we found that much of the library had little to no fluorescence above background (Figures S1E and S1F). These results were suggestive of intron retention, which is a process that dominates in lower eukaryotic organisms. In humans, due to long intron lengths, exons are first recognized by the splicing machinery in a process called exon definition (De Conti et al., 2013), and thus mutations that affect exon recognition often result in exon skipping rather than intron retention (Baralle and Buratti, 2017). Due to these concerns, we optimized our reporter designs with longer constant intron backbones and observed ~20- to 100-fold higher level of fluorescence overall.

In order for MFASS to work in a multiplexed, scalable format, the assay relies on a single copy of the reporter construct per cell before FACS, thereby ensuring that our splicing fluorescence readout corresponds to a single sequence. Each library sequence is integrated once per cell using high-efficiency site-specific genome integration (Figures 1A and S1G–S1I) and expressed at the AAVS1 locus to minimize any pleiotropic effects. However, we noticed upon transient transfection of the splicing reporter libraries that each cell contains hundreds of reporter copies on average (Figure S1J, top left). We characterized the copy number of the reporter library in human cells across culture passages by flow cytometry and RT-PCR and found 100,000-fold cell dilution to be sufficient without contaminating plasmids in single cells (Figures S1K and S1M).

Although episomal splicing reporter assays are commonly used, it has been reported that splicing outcomes can be more reproducible when sequences are genomically integrated (Smith and Lynch, 2014). We constructed reporters corresponding to individual library sequences and evaluated both fluorescence and RNA splicing under episomal and genomic expression (Figures S1N–S1Q). We selected nine sequence variants for further analysis by flow cytometry (Figures S1N and S1O; STAR Methods) and RT-PCR (Figures S1P and S1Q). Individual controls sorted from the library showed consistent behavior between inclusion rates estimated by RT-PCR and fluorescence output, and reporter fluorescence in stably integrated constructs is more consistent with RT-PCR results (Figures S1N–S1Q).

## Evaluating MFASS based on Known Splicing Regulatory Elements

To test and validate MFASS, we first designed, built, and assayed a test library of 6,713 mutations aimed at perturbing regulatory elements across a randomly chosen library of 205 natural in-frame human exons and surrounding intronic sequences (Splicing Regulatory Element library). We first developed this test library in order to evaluate the MFASS assay and test the effects of designed mutations in a large set of natural sequence contexts. To mutate sequences iteratively while accounting for the creation of unintentional motifs, we developed a custom software toolkit for the design of *in silico* splicing mutations (STAR Methods). In particular, this toolkit incorporates information about splicing regulatory elements from the literature to calculate a composite score for each sequence across different functional classes. We chose natural human exons that are less than 100 bp and begin and end on frame 0 and designed a 170-bp exon library with its surrounding intronic contexts, which includes at least 40 bp of upstream intron and at least 30 bp of downstream intron. Overall, we randomly chose a subset of ~200 human exons and iteratively designed 60–80 perturbations per sequence that weaken, strengthen, or destroy splicing motifs focused on three major motif types (Tables S1 and S2).

We used MFASS to assay the SRE library with biological replicates across two different intronic backbones (Figure 1A). We expanded these sorted bins over several passages and observed that the sorted populations remained stable (Figure 1B). We also performed bulk RT-PCR for each bin and found that the observed RNA splicing efficiencies corresponded with observed fluorescence of the bins (Figure 1C). To obtain an exon inclusion index for each sequence, we first considered reads that perfectly matched the SRE library and normalized based on read depth and weighted by the corresponding bin population percentage from FACS. Finally, we computed a weighted average of normalized read counts across all bins using the average exon inclusion level in each bin as measured by the GFP:RFP ratio and confirmed by bulk RT-PCR (STAR Methods). Overall, the inclusion indices for our library are bimodal, with most library sequences represented predominantly in one bin, showing either complete exon inclusion or skipping (Figure 1D).

We measured the replicability of inclusion indices across biological replicates using the tetrachoric correlation ($r_t$) due to the bimodality in our results (Pearson correlation provided as a comparison). We tested these libraries across two constant intron backbones (SMN1 and DHFR) and found that exon inclusion metrics are highly reproducible within the backbone across biological replicates (Figures 1E and 1F; $r_t = 1.00$, $p < 10^{-16}$, tetrachoric; $r = 0.94$, $p < 10^{-16}$, Pearson, DHFR intron backbone; $r_t = 0.97$, $p < 10^{-16}$, tetrachoric, $r = 0.89$, $p < 10^{-16}$, Pearson, SMN1 intron backbone) and between backbones (Figure 1G; $r_t = 0.96$, $p < 10^{-16}$, tetrachoric; $r = 0.85$, $p < 10^{-16}$, Pearson). We consider 6,713 designed mutations present across both backbones in subsequent analysis and highlight data for the SMN1 intron backbone (Figure 2).

Overall, we showed that, although the loss of exon recognition is consistent with known splicing motifs, the effects of these perturbations are not easily predicted for 6,713 designed mutations across 205 human exons (Figure 2). To focus on the mechanisms by which large-effect splicing changes can occur, we defined large-effect variants as Δinclusion index ≤ −0.5 (i.e., mutations to a wild-type exon with an inclusion index of ≥0.5, which is reduced by an absolute value of at least 0.5), which we term “splice-disrupting variants” (SDVs). We quantified the percentage of SDVs for designed mutations in each category (Figure 2A). As expected, we found that splice-site mutations to the nearly invariant dinucleotides cause SDVs at the highest rates (Figure 2A). Mutations to the splice site (splice acceptor, positions −20 to +3; splice donor, positions −3 to +6) individually result in SDVs 48%–73% of the time (Figure 2A; “acceptor site” and “donor site”) and 96% of the time when mutating simultaneously both splice donor and acceptor (Figure 2A; “acceptor + donor site”). This is likely an underestimate as mutations eliminating splice site recognition may be utilizing alternative splice acceptors or donors, which cannot be distinguished from exon inclusion by MFASS. Within exons, mutations can still have strong effects. Encoded synonymous mutations to all putative exonic splicing enhancers (ESEs) lead to SDVs ~72% of the time (Figure 2A; “all exonic splicing enhancers”). Although removing clusters of putative exonic splicing silencers (ESSs) results in increased exon inclusion (Figure S2A; all exonic splicing silencers), removing the strongest identified ESE alone results in 30% SDVs (Figure 2C; “strongest exonic splicing enhancer”). More generally, splicing metrics, such as MaxEnt for splice site strength (Figure 2B) or exon hexamer metrics (Figures 2C and S2B), are consistent with predicted effects on splicing behavior.

## Effects of Rare Human Variation on Exon Recognition

Although these results indicate that mutations intended to alter previously recognized motifs can commonly lead to loss of exon recognition, we wanted to explore the extent to which natural genetic variation in the human population results in SDVs. We generated the Single Nucleotide Variant library (SNV library), for which we designed and synthesized all cataloged exonic and intronic single-nucleotide variants (SNVs) from the ExAC, for wild-type human exons that demonstrated exon inclusion (inclusion index ≥ 0.8) in the SRE library (STAR Methods). From this SNV library, we first tested two reporter constructs that split at distinct positions of GFP. To evaluate the splicing reporter output across two versions of the SNV dataset from the MFASS assay, we monitored GFP and mCherry fluorescence from the initial library and sorted cells using flow cytometry (Figures S3A and S3B).

Overall, the two different contexts displayed high correlations for detecting splice-disrupting variants (Figure S3C, n = 5,740, $r_t = 1.00$, $p < 10^{-16}$, tetrachoric; $r = 0.94$, $p < 10^{-16}$, Pearson). Because the SNV library was examined in independent reporter constructs testing different frames, this indicates we will be able to use MFASS to screen for exons across in-frame and frameshifting exons for future studies.

Overall, we quantified the effects of more than half (52.4%; 27,733 of 52,965) of the ExAC SNVs found across 2,198 human exons and found that 1,050 of 27,733 (3.8%) ExAC variants assayed led to almost complete loss of exon recognition, are broadly spread across 543 exon backgrounds from 473 genes for 1,038 distinct genomic positions (Figure 3A), and show increased sensitivity at the splice region (Figure 3B). Correlations between biological replicates were high (n = 31,583, $r_t = 0.94$, $p < 10^{-16}$, tetrachoric; $r = 0.80$, $p < 10^{-16}$, Pearson; Figure S3D). To minimize false positives, we require replicate agreement within 0.20 instead of 0.30 used for the SRE library (STAR Methods). To ensure that MFASS-identified SDVs are robust to experimental artifacts, we additionally analyzed a number of controls. First, we tested the SNV library using three control sets (Figure 3C): 24 of 24 (100.0%) scrambled nucleotides; 70 of 71 (98.6%) skipped exons; and 945 of 977 (97.3%) broken splice-signal sequences result in loss of exon recognition (inclusion index < 0.5; Figure 3C), noting that alternative $5'$ and $3'$ splice site usage result in false negatives for MFASS. In addition, we also analyzed sequences containing synthetic errors resulting in single-nucleotide deletions (n = 9,801) from our designed sequence library (STAR Methods), and SDVs for these deletions are enriched across the exon-intron junction (Figure 3D).

Finally, we further validated MFASS results individually for 34 SDVs across multiple functional classes of splicing variation across the original tested context as well as longer intronic contexts (Figure 3E; STAR Methods). Our results suggest that MFASS is robust across a majority of rare genetic variants tested for splicing defects. We individually verified SDVs using transient expression assays and found that nine of 11 (81.8%) showed large-effect splicing defects, with all 11 (100.0%) showing reduced exon inclusion relative to their respective wild-type sequences (Figure 3E). Furthermore, we tested the effect of longer intronic context on individual SDVs and found that 17 of 23 (73.9%) showed large defects in splicing, with only one of 23 (4.3%) mutations showing no appreciable exon recognition defect (Figure 3E). Finally, to examine the cell-type specificity of SDVs, we further picked a subset of 15 SDVs with the strongest changes in exon inclusion and tested wild-type or matched SDV reporter constructs across three additional cell types in the ENCODE consortium. We found that large-effect splicing disruptions are consistent across four cell types in all 15 of the splice-disrupting variants assayed (15/15; 100.0%; Figures 3F and S3E).

Of the 1,050 SDVs detected, we observe almost equal contributions from introns (561; 54%) and exons (489; 46%) among the variants we tested (Figure 4A). We found that 76% of splice site variants are SDVs (Figure 4B, left). Variants in the broader splice region, synonymous exonic variants, non-synonymous exonic variants, and deeper intronic variants disrupt splicing more rarely at 8.5%, 3.0%, 3.1%, and 1.5%, respectively (Figures 4B, left, S4A, and S4B). The splice donor and acceptor regions show different patterns of sensitivity to splicing disruptions (Figure 4C), with splice donor regions being more sensitive than splice acceptor regions. Because SNVs are not equally distributed among these categories, splice site SDVs only constitute 17% of all SDVs, whereas intron variants, which are the least sensitive to splicing disruption, comprised 16% of SDVs (Figure 4B, right). SNVs at the splice sites are rare in our library (Figure 4C, bottom; SNV density) and also for all ~7.4 million ExAC variants (Figure S4C). The larger number of variants in regions away from the splice sites outweighs their reduced sensitivity (Figure 4C, bottom; SNV density) and contributes to 83% of the SDVs reported here.

## Population Genetic, Evolutionary, and Functional Analyses of Splice-Disrupting Variants

A number of population genetic, evolutionary, and functional characterizations indicate that our measured SDVs are relevant. First, the proportion of SNVs that are SDVs shows significant reductions as a function of allele frequency (chi-square test; $p = 1.03\ 3\ 10^{-4}$). Consistent with population genetic theory, a vast majority (98.8%) of our SDVs are extremely rare (allele frequency from the Genome Aggregation Database [gnomAD] < 0.5%; Figure 5A). Second, we find a significantly lower SDV rate (~2×) within genes that rarely have protein-truncating variants (PTVs) within ExAC, indicating strong functional constraint (probability of loss-of-function intolerant [pLI] ≥ 0.9; Lek et al., 2016; Figure 5B; two-tailed Fisher’s exact test; $p = 3.0 \times 10^{-11}$). Considering the rates of SDV and PTV overall, we conclude our SDV rate is at least on par to that of protein-truncating variants from ExAC. Third, SNVs that are SDVs show significantly stronger evolutionary conservation, suggesting purifying selection at these sites (Mann-Whitney U test; $p < 10^{-16}$; Figure 5C). Missense variants alone do not seem to drive the conservation signature, as the difference in mean phyloP conservation score is greater without missense variants ($\text{phyloP}_{\text{non-SDV}} = 0.04$ versus $\text{phyloP}_{\text{SDV}} = 2.7$) than with missense variants ($\text{phyloP}_{\text{non-SDV}} = 1.4$ versus $\text{phyloP}_{\text{non-SDV}} = 3.1$; Student’s two-sample t test; $p < 10^{-16}$; two-sided), suggesting that SDVs are under stronger evolutionary conservation independent of missense variation. Fourth, nucleotide positions under strong evolutionary conservation have higher rates of SDVs, and this is especially apparent within introns (two-tailed Fisher’s exact test; $p < 10^{-16}$; Figure 5D). However, this conservation has limited predictive power, because within introns there are many more SNVs at neutral sites than sites under strong conservation and within exons most sites are highly conserved (Figure 5E). Fifth, for exonic SNVs, we observed that SDVs significantly reduce exon hexamer scores when compared with nonSDVs, suggesting that SDVs are disrupting important functional sites for exon recognition (Student’s t test; $p < 10^{-16}$; Figure 5F). Sixth, motif enrichments at the splice acceptor suggests that SDVs enriched for T to A mutations disrupt the area near the mechanistically important polypyrimidine tract, and for splice donors, we find that guanine-rich motifs are less tolerated (Figure S5A). Seventh, we found several enriched gene ontology (GO) terms for SDVs comprising of four enriched categories (Table S3; STAR Methods). Two of the GO categories contain mostly collagen genes, many of which have large repeated protein domains. In addition, the “post-Golgi vesicle-mediated transport” GO category also contained a number of SDVs in genes with other repeat domains, such as ankyrin and spectrin repeats. Such repeat-expansion genes can often be variable between populations, and in-frame exon skipping events are likely to have fewer severe consequences (Chan et al., 2008).

## Cross-Validation of Individual SDVs

It is likely that some fraction of SDVs detected by MFASS do not reflect actual changes in humans because minigene reporters are widely used but imperfect models of endogenous exon recognition (Gaildrat et al., 2010). For example, we detect 11 SDVs with a minor allele frequency of greater than 0.5% that correspond to a set of common variants. Because common variants will likely overlap with other datasets, we first cross-referenced our ExAC library with the ClinVar database (Landrum et al., 2014). Only 0.5% (141/27,733) of the ExAC library is present in ClinVar, with eight SDVs and two annotated pathogenic variants in the *MTMR2* and *PARN* genes. To look more broadly in the datasets of healthy cohorts other than ExAC, we cross-referenced our assayed SNVs with the Genotype Tissue-Expression (GTEx) project (Battle et al., 2017). Overall, 9 of these 11 common variants have exon inclusion levels from GTEx (ΔPSI; Δpercent spliced in, Figure S5B), and three had globally significant differences (Figure S5B, i, ii, and v). If we extend this analysis to rare variants as well, we were able to determine PSI values for 1,471 assayed exons (STAR Methods), but only 28 are SDVs (including the common SDVs described above). Of these 28, seven (25%) show globally significant difference in exon inclusion levels from RNA-seq. In addition, two additional SNVs have large-effect splicing disruptions in the single tissue they were expressed in (Figure S5B, viii and vi). Overall, we consider magnitude instead of sign concordance, which allows more stringent comparison of splicing changes for specific variants and that there are some important caveats with this analysis. First, we only use pre-computed PSI values (STAR Methods), which cannot account for complex splicing defects like alternative splice donors or acceptors. Second, the limited intersection of the two sets of variants are enriched for the most common variants that we call as SDVs and are likely to be false positives because of the propensity of smaller effect changes in common variants.

To better understand how rare SDVs in ExAC replicate in their full gene context, we assembled 19 SDVs and associated wildtype controls for 12 full-length genes using isothermal gene assembly and examined splicing disruptions using RT-PCR upon episomal expression of the full gene (STAR Methods). We validated that 13 variants in nine genes cause splicing disruptions (Figures S5C and S5D; 68.4%, 13/19 variants or 75.0%, 9/12 genes), with nine of 19 variants (42.1%) having appreciable effects on exon recognition. Interestingly, five of the detected changes involved alternative 5′ and 3′ splice site usages in the broader full gene context, indicating that many of the identified exon skipping events in MFASS might have different consequences *in vivo*.

## Large-Effect Rare Variants on Splicing Are Challenging to Predict

Our results indicate that traditional metrics for assessing how mutations affect splicing are likely to fail, because although it is known that splice site variants are likely deleterious, it has been unclear to what extent rare genetic variation affects splicing outside of these sites. For example, the existing variant effect predictors for missense mutations, such as PolyPhen and SIFT, either largely provide no annotation for SDVs or call them benign (Figures 6A and S6A). Meanwhile, the SDV rate in synonymous mutations, which are usually assumed to be benign, is nearly equivalent to missense variants (3.0% versus 3.1%; Figure 3A).

We used a number of contemporary variant effect predictors that are capable of predicting the effects of non-coding variation based on both functional genomic and/or evolutionary information, CADD (Kircher et al., 2014), DANN (Quang et al., 2015), FATHMM-MKL (Shihab et al., 2015), fitCons (Huang et al., 2017), LINSIGHT (Gulko et al., 2015), phastCons (Siepel et al., 2005), and phyloP (Pollard et al., 2010), as well as two specifically designed for splicing, SPANR (Xiong et al., 2015) and HAL (Rosenberg et al., 2015; Figure 6B). Most predictors have low precision, with several providing no better prediction than random guessing. FATHMM-MKL, CADD, and DANN perform best among those not trained specifically for splicing but only achieve ~7% to 8% precision at any appreciable recall. Much of their power is the result of the ability to call intronic SDVs (Figures 6C and S6B), likely due to increased conservation or molecular function near or at those nucleotides. Not surprisingly, those predictors trained specifically for calling splice defects perform best. At equivalent effect size compared to our assay (>50% splicing disruption), SPANR achieves 44.5% precision, though only a minority of the SDVs are called (11.8%, Figure 6B). 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, though the predicted effect size is ~2% loss of inclusion. More generally, the SPANR effect sizes poorly predict our observed inclusion rates ($R^2 = 0.11$; Figure S6C). The increased power of SPANR over other predictions is largely due to its ability to predict exonic SDVs. HAL provides even better precision in these exonic regions (Figure 6C) but only calls SNVs within exons.

## DISCUSSION

In this work, we tested over half of the variants found in 2,198 human exons across 60,000 individuals and observed that 3.8% of these variants (1,050 of 27,733) cause loss of exon recognition. The rate of SDVs we find here is surprisingly high. Our SDV rate (3.8%) is ~73% of the rate of probably damaging variants predicted by PolyPhen for the same set of SNVs (5.2%; 1,437 of 27,733) and ~3-fold higher than the observed rate of protein-truncating variants found in ExAC as a whole (1.3%; 121,309 of 7,404,909; Lek et al., 2016). We would expect such exon skipping events to be detrimental not only to protein function but, if our results generalize to exons that do not preserve frame, also cause large changes to mRNA stability through nonsense-mediated decay (Lewis et al., 2003). This may help explain why extremely rare variation seems to have large predicted effects on gene expression, even though we rarely observe mutations with large effects on transcription control elements (Hernandez et al., 2017; Li et al., 2017).

In MFASS, most of the assayed SNVs result in either no effect or near complete exon skipping, in contrast with the magnitude of variant effects from sQTLs (Battle et al., 2017; Pala et al., 2017; Takata et al., 2017). We speculate this apparent discrepancy is for several reasons. First, MFASS is not well suited to detect small-effect variations due to the limitations of flow cytometry. Second, the variant context in MFASS makes it unlikely that small-effect changes we observe reflect *in vivo* changes. Third, most sQTL studies analyze transcripts from heterogeneous cell types, as compared with more bimodal splicing events from single-cell sequencing studies (Shalek et al., 2013, Faigenbloom et al., 2015). Fourth, current sQTL studies are limited by small sample sizes and thus only powered to study common variation. Meanwhile, studies of rare variants find large-effect mutations that affect splicing, most notably in GTEx (Li et al., 2017) and in Mendelian diseases (Kremer et al., 2017).

We may be over- or underestimating the rate of SDVs using MFASS. First, although minigene reporters represent an important standard for the evaluation of clinically relevant splicing mutations, they do not always capture the necessary context for splicing. Second, we only chose in-frame exons that are less than 100 bp in length. Although we see no appreciable difference in average conservation in ExAC SNVs for in-frame and out-offrame exons (Figure S4C), these constraints do enrich for genes with repeat expansions, where an individual skipped exon may have fewer functional consequences. Third, we may not be including enough intronic context to correctly diagnose mutations, even though most of the intronic conservation signal was contained within the tested intron sizes (Figure S6D). Because the intronic variation in our genome is ~3-fold greater than exonic variation, we might be missing a substantial number of SDVs contained within untested intron regions. In addition, because ExAC is an aggregation of exome data, surrounding introns have lower coverage and thus fewer covered SNVs. Fourth, any alternative 5′ and 3′ splice site usages are false negatives from MFASS, and SDVs might also manifest as alternative splice sites *in vivo*. Finally, although the large-effect SDVs appear to transfer across cell types (Figure 3F), some SDVs would likely be cell-type specific.

Our results suggest loss of exon recognition by rare human variants may be a major source of functional and expression variation, and their effects are particularly difficult to predict a *priori* using computational prediction. We show most of the large-effect rare variation on splicing would not be easily recognized, as only ~17% of such functional rare variation we found are in canonical splice sites. Compared to other multiplexed splicing reporters, MFASS is unique in that it screens both exonic and intronic variants, is applicable to a broad spectrum of human exons, uses long constant intron backbones, site-specifically integrates reporters at single copy, and provides increased power for detecting large-effect loss-of-function variants (Julien et al., 2016; Ke et al., 2011; Rosenberg et al., 2015; Soemedi et al., 2017; Adamson et al., 2018). MFASS is best suited for screening large numbers of large-effect rare variants, which is especially useful for the analysis of mutations in Mendelian diseases, cancer, and population genetics. MFASS is the largest study of splicing defects in SNVs of natural human exons to date by ~10-fold and can likely be scaled substantially. More broadly, MFASS can help interpret variants found in large-exome datasets to obtain a broader understanding for how rare, *de novo*, and somatic variants shape complex traits and diseases (MacArthur et al., 2014).

# STAR★METHODS

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sriram Kosuri (sri@ucla.edu).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

**Human Cell Lines—**All cell culture reagents were obtained from Thermo Fisher Scientific. HEK293T chromosomal landing pad cells and derivatives, HepG2 cells, and HeLa S3 cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 0.1 mg/mL streptomycin. K562 cells were cultured and maintained in RPMI supplemented with 10% FBS, 100 U/mL penicillin, and 0.1 mg/mL streptomycin. All cells except K562 cells were passaged using 1× TrypLE Express. All restriction enzymes were obtained from New England Biolabs. Plasmid modifications were performed either by restriction cloning or Gibson assembly (SGI-DNA). Synthesized genes were obtained as sequence fragments from either Gen9 or Twist Biosciences. All oligonucleotides indicated below were obtained from IDT Technologies or Eurofins.

## METHOD DETAILS

**Splicing Reporter Design—**The organization and key features of our MFASS splicing reporter constructs are as follows: emerald GFP (emGFP) coding sequence is split into two exons that flank a constant intron backbone sequence (Figures S1A–S1D). emGFP is split at two different locations for various reporter designs without disrupting the downstream reading frame. For the SRE library and version 1 of the SNV library, the reporter library contains exons that start and end on phase 1. For version 2 of the SNV library, the reporter library contains exons that start and end on phase 0. The synthetic sequence library is cloned into a pair of restriction sites, AgeI and NheI, or AscI and PacI, in the middle of the backbone. The expression of the splicing reporter module is driven by the CAG-GS promoter. For selection of genomic integrants, we included a B×b1 attB site and promoterless puromycin such that drug resistance is conferred in the HEK293T cell library following site-specific recombination, due to a CAGGS promoter adjacent to the B×b1 attP site in the landing pad cell line. We tested two sets of longer constant intron backbones with >250bp of sequence for each intron, which have both been previously characterized as more faithful intron backbones in the context of such three-exon, two-intron reporters (Figures S1A–S1D). These two backbones were the C. griseus long DHFR intron backbone (Arias et al., 2015) and human SMN1 intron backbone (Cho et al., 2015) (Figures S1A–S1D). In particular, the long DHFR introns were the same introns used in previous characterizations of exon definition (Arias et al., 2015).

**Microarray-Derived Oligonucleotide Library Design—**We obtained microarray-derived oligonucleotides of 200 to 212 bp from Agilent Technologies to generate synthetic DNA libraries. We selected human exons that are less than 100 bp and begin and end on frame 0 from the Ensembl mySQL server. We designed a 170-bp intron-exon-intron sequence library *in silico* containing all 9,634 human exons fulfilling above criteria (Ensembl release 73, hg19 assembly), which includes at least 40 bp of upstream intron and at least 30 bp of downstream intron, with the exon in the middle. We added extra native intronic sequences as length limitations allowed (i.e., if exons were shorter), split between the upstream and downstream equally with an extra base added to the donor side for odd number of bases added. Finally, a pair of 15-mer amplification primer sequences, containing either AscI and PacI or AgeI and NheI restriction sites, were added to yield 200-mer or 212-mer sequences for DNA synthesis respectively for the SRE or SNV libraries.

**Design of SRE Library—**For the SRE library, we obtained 9,634 human exons that are less than 100 bp and begin and end on frame 0 and designed a 170-bp exon library with its surrounding intronic contexts, that includes at least 40 bp of upstream intron and at least 30 bp of downstream intron. Overall, we randomly chose 230 exons from this set and designed 60–80 synonymous mutations per sequence that correspond to specific functional classes of regulatory elements governing splicing using a toolkit of custom Python scripts we developed for scoring these mutations using defined scoring criteria as detailed below. We focused on three major motif types related to splicing in our custom scoring algorithm (Table S1). The first major motif type is the splice acceptors and donors. These sequences are scored with MaxEntScan (Yeo and Burge, 2004), an algorithm based on the maximum entropy principle that learns splice site motif strength. The second major motif type is the exonic splicing enhancers/silencers (ESEs/ESSs) (Ke et al., 2011). The third major motif type is the conserved intronic sequences that affect splicing in either the acceptor or donor side of the intron (Voelker and Berglund, 2007). Next, we iteratively designed synonymous mutations in exons and/or introns that affect splicing (Table S2). Mutations made to sequences were scored in the same fashion as wild-type sequences, with a higher score as a proxy for increased exon inclusion. Mutations were scored and generated to weaken, strengthen or destroy splicing motifs. We define functional classes of mutants that differ in score requirements, minimum base separation between mutants and the number of mutants per class. Mutations were made iteratively until we generate the desired number of mutants or reach the maximum number of iterations. For splice sites or splice regions, the invariant positions of the splice donor or acceptor are not mutated, with the exception for the “weaken splice site” category. In addition, we tested 53 RNA-binding protein motifs obtained from the RNA-binding protein database (RBPDB) (Cook et al., 2011) as position frequency matrices and thresholded at 1% false positive rate, and 109 human single-nucleotide polymorphisms (SNPs) obtained from dbSNP (build 133) (Smigielski et al., 2000).

**Design of SNV Library—**For the SNV library, we started with a library of 2,920 natural exons that exhibited exon inclusion using MFASS (inclusion index ≥ 0.8; SRE library, DHFR intron backbone). We designed single nucleotide variants (SNVs) from the Exome Aggregation Consortium (Lek et al., 2016) (ExAC, version 0.3.1). We stored hg19 genomic coordinates of each sequence in BED file format and used bcftools to intersect the ExAC variants with our library of wild-type human exons to subset all relevant SNVs. We only synthesized variants with a filter status of “PASS,” and generated all alternate alleles (up to 3) if more than one alternate allele was indicated. These sequences were filtered to (i) exclude sequences containing unique NheI or AgeI restriction sites used for library cloning and (ii) include SNVs only within nucleotides 11 through 160 of each 170 bp library sequence to avoid possible spurious interactions with restriction sites, resulting in 2,902 exons as template with their associated variants that fit above criteria.

We designed two library subpools with redundancy for wild-type that enables separate retrieval of sublibraries from the microarray. We transfected these pools at the stage of plasmid reporters at the ratio of 1:3 that enables increased representation of natural sequences. From the initial design carried through to the completion of MFASS, 80.5% of the designed natural sequences (2,339 of 2,902) were represented in the final cell reporter library. 2,198 out of 2,339 natural sequences have at least one corresponding SNV, while an additional 30 sequences represented in the control library. Ultimately, we only report and include SNV data for which data for natural sequences are available, have replicable data across two biological replicates, and have an inclusion index of greater than or equal to 0.5 for wild-type. For these 2,198 exon backgrounds, we obtained the corresponding paired variant data for 27,733 SNVs, from which 1,050 SDVs are observed (Figures 3A and 4A).

**Library Amplification and Cloning—**The splicing regulatory element (SRE) library was amplified with KAPA HiFi HotStart (KK2701) in eight 50 mL reactions, each with 500 pg of oligonucleotide library, and 0.4 mM of ORC405 and ORC406 primers. The reaction and cycling conditions are: 95°C for 3 min, 5 cycles of 95°C for 3 s, 50°C for 20 s, 60°C for 10 s, 15 cycles of 95°C for 3 s, 60°C for 30 s, followed by an extension of 60°C for 5 min. The SRE library was amplified similarly as above with ORC403 and ORC404 primers, as well as the following cycling conditions: 95°C for 3 min, 5 cycles of 95°C for 3 s, 50°C for 20 s, 60°C for 10 s, 11 cycles of 95°C for 3 s, 60°C for 30 s, followed by an extension of 60°C for 5 min. Splicing reporter plasmids and SRE library were digested with AscI and PacI. Reporter plasmid and library were ligated with T4 DNA ligase (New England Biolabs).

For the SNV library, we performed similar procedures as above with the following alterations: we performed emulsion PCR for the two subpools (35 cycles) containing both natural exons and SNVs with biotinylated primers. The second subpool was amplified similarly (40 cycles), with biotinylated ORC513 and ORC514 primers, and both pools were processed with AgeI and NheI at 37°C before ligation-based cloning in *E. coli*.

**Generation of Landing Pad Cell Lines and Integration—**For site-specific integration of exon libraries in HEK293T cells, we engineered a chromosomal landing pad cell line which allows stable expression of splicing reporter library at the AAVS1 locus, which is modified from Duportet et al. by CRISPR-Cas9 in order to remove expression of the endogenous YFP gene (Duportet et al., 2014). We characterized 25 clones expanded from single cells by flow cytometry, microscopy, and genomic PCR, and selected a clone (which we termed RCA7) that does not express any YFP or mCherry fluorescence for our current study.

We site-specifically integrated the splicing reporter using Bxb1 integrase into cells containing the chromosomal landing pad (Figures 1 and S1), first without any exon library sequences between the intron backbones, and later with individual exons and/or synthetic sequence libraries cloned in between. For the SRE library, we transfected HEK293T chromosomal landing pad cells, grown in six T-225 flasks (BD) per biological replicate that were processed in tandem. Each T-225 flask was transfected at 80% confluency with 50 μg of plasmids containing exon library and Bxb1 integrase, and 150 μL Polyethylenimine (Polysciences) or 75 μL Lipofectamine 3000 (Thermo Fisher Scientific). Cells were transfected for 72 hr, and then selected with 5 μg/mL puromycin (Thermo Fisher Scientific).

Cells were subsequently passaged serially for at least 18 days before cell sorting. For the SNV library, we transfected HEK293T chromosomal landing pad cells, grown in sixteen 150 $cm^2$ plates (45 μg plasmids per plate) for 3 days, pooled and transferred to two 4500 $cm^2$ roller bottles (BD Biosciences) or equivalent volume for 150 $cm^2$ plates per biological replicate, selected for integrants as above, and maintained in eight 150 $cm^2$ plates per biological replicate for 20 days before cell sorting.

**Fluorescence-Activated Cell Sorting—**We measured cell samples for GFP and mCherry fluorescence intensities by flow cytometry (BD LSRFortessa or LSRII) across passages. Cells harboring variant libraries were sorted using a FACSAria III (BD Biosciences) into bins based on GFP fluorescence, given a minimal amount of mCherry fluorescence (as thresholded using a genome-integrated mCherry driven by the pCAGGS promoter as a positive expression control, Figure 1A). For the SRE library (DHFR intron backbone), we sorted ~7.5 million cells for $GFP_+$ and $GFP_{neg}$ bins, and 7.5 × $10^5$ cells for $GFP_{int}$ bin. For the SRE library (SMN1 intron backbone), we obtained ~4 million cells for $GFP_+$ and $GFP_{neg}$ bins, and 4.2 × $10^5$ cells for $GFP_{int}$ bin. Sorted sub-libraries for each replicate were grown separately and passaged. We eliminated dead cells, debris, and doublets based on forward and side scatter, and single-color and double-negative controls were used for gating and calibration. For the SNV library (v1), we performed two sorts to ensure purity of the final populations of $GFP_+$, $GFP_{int}$ and $GFP_{neg}$ cells (Figure S3A). For the first sort, we obtained 16 million cells for $GFP_{neg}$ library, 2.6 million cells for $GFP_+$ library and 2.7 million cells for $GFP_{int}$ library (biological replicate 1), 15 million cells for $GFP_{neg}$ library, 2 million cells for $GFP_+$ library and 2.8 million cells for $GFP_{int}$ library (biological replicate 2). For the purifying sort, we further sub-sorted the libraries from the first sort, and obtained ~2 million cells for $GFP_{neg}$ library, 1 million cells for $GFP_+$ library and 2.5 million cells for $GFP_{int}$ library (biological replicate 1), and 1 million cells for $GFP_{neg}$ library, 1 million cells for $GFP_+$ library and 2.5 million cells for $GFP_{int}$ library (biological replicate 2).

For the SNV library (v2), we sorted cells based on GFP fluorescence into four bins: $GFP_+$, $GFP_{int\text{-}hi}$, $GFP_{int\text{-}lo}$, and $GFP_{neg}$ bins (Figure S3B). For both biological replicates, we obtained 16 million cells for $GFP_{neg}$ library, 2 million cells for $GFP_+$ library, 2 million cells for GFPint-hi and GFPint-lo library.

**DNA-seq of FACS-Sorted Libraries—**To obtain cells containing a single individual reporter construct, we first sorted single cells by FACS from individual bins, with GFP fluorescence gates defined from library sort, and expanded homogeneous clones from single cell sort. For the SRE library, we extracted genomic DNA from 10 million cells for the sorted populations using blood and cell culture DNA midi kit (QIAGEN). We amplified each sublibrary for ~300-fold amplicon coverage, and reactions were performed in 96-well format in three to nine 50 μL reactions for each sublibrary proportional to bin size. Per biological replicate, we amplified library variants from genomic DNA with KAPA HiFi HotStart, using 5 μg of template for $GFP_+$ and $GFP_{neg}$ sub-libraries, and 2 mg of template for the $GFP_{int}$ sublibrary, with 500 nM of the primers ODY093 and ODY028 for the DHFR intron backbone, or the primers ODY088 and ODY089 for the SMN1 intron backbone. The following cycling conditions were used: for the DHFR intron backbone, 98°C for 45 s, 23 cycles for $GFP_{int}$, or 22 cycles for $GFP_+$ and $GFP_{neg}$ using: 98°C for 15 s, 68°C for 30 s, 72°C for 30 s, followed by an extension of 72°C for 1 min; for the SMN1 intron backbone: 98°C for 45 s, 24 cycles for $GFP_{int}$, or 29 cycles for $GFP_+$ and $GFP_{neg}$ of: 98°C for 15 s, 68°C for 30 s, 72°C for 30 s, followed by an extension of 72°C for 1 min. The reactions for each population were pooled separately, purified and gel-extracted on 1% agarose gel and quantified using Tapestation 2200 (Agilent).

For the SNV library, procedures were performed similarly to the SRE library in the DHFR intron backbone, with the following optimizations. Library variants was amplified from genomic DNA (ORC515 and ODY028), and genomic DNA was extracted similar to procedures for the SRE library. Sorted libraries were indexed by PCR amplification, in twenty-four 50 μL reactions for $GFP_{neg}$ and eight 50 μL reactions for all other sublibraries, using the forward primer ORC522, and the reverse primers ODY32 through ODY41, and ORC531 through ORC534.

**Validation of MFASS Using Individual Exon Controls—**We performed individual controls to assess the correspondence to sequences in our library and to observe consistent splicing behavior across RNA and fluorescence output. For the data from Figures S1N–S1Q, we characterized more than 20 cell clones expanded from single cells, and only 9 individual sequences that perfectly match the reference SRE library were used for RT-PCR and flow cytometry analysis.

RNA from sorted sub-libraries as well as individual control exons were extracted using RNEasy MiniKit (QIAGEN). Reverse transcription-PCR was performed using Superscript III or Superscript IV (Thermo Fisher Scientific) according to manufacturer's protocol using reverse transcription primer (Table S4), which binds to a region in exon 2 of emGFP, and PCR was performed with extracted cDNA. The reaction and cycling conditions are optimized as follows: 95°C for 2 min, 18 cycles of 98°C for 3 s, 62°C for 15 s, 72°C for 10 s, followed by an extension of 72°C for 2 min.

34 SDVs were tested for exon inclusion by transient transfection using Lipofectamine 3000 (Life Technologies) in HEK293T cells for 24 hr. A ratio of GFP:mCherry fluorescence was obtained in linear mode (BD LSRII or BD LSRFortessa) for the comparison of exon inclusion rates across samples. We subtracted background fluorescence based on a transfected empty vector control, and only consider GFP:mCherry fluorescence above the threshold. We tested sequences either exactly in the original sequence context in the reporter construct examined in MFASS, or with an additional 130 bp of endogenous intronic contexts (65 bp upstream and 65 bp downstream). Percent inclusion is calculated for both the individual SDV and its respective wild-type sequence, with the change in percent inclusion calculated as the absolute difference between the mutant and the wild-type sequence. All mutants were normalized to a no-insert control as a baseline for complete exon skipping for assessment of change in exon inclusion.

**Cell-type Specificity of SDVs across Four Human Cell Types—**We tested 29 human exons with its surrounding intronic contexts (15 SDVs with the 14 corresponding wild-type sequences) across 4 human cell types. The four human cell lines tested are HEK293T (RCA7 cell line established in this study), HeLa S3 (ATCC CCL-2.2), HepG2 (ATCC HB-8065) and K562 (ATCC CCL-243). We validated these constructs across cell types in the same manner that we validated individual exon controls in above section.

**Validation of Rare SDVs in Full Genes—**We considered rare 61 SNVs in 34 genes that have a change in inclusion index of ≤ −0.50 across both replicates from MFASS (i.e., SDVs) under 15kb. From these, we were able to assemble complete 12 wild-type full genes (up to ~13kb) with at least one corresponding SDV (19 SDVs total, Figures S5C and S5D). Using isothermal gene assembly, mutations were introduced in the middle of the oligonucleotide with ~40bp overlap on each overlapping fragment, and assembled without the mutations for the wild-type gene sequences. Genomic sequences with wild-type and matched SNVs were amplified from the same human genomic DNA template (NIST, SRM 2372, or Promega, G1521) using PrimeSTAR GXL polymerase (R050, Takara). Each partial gene fragment was amplified using 25ng of genomic DNA in a single 50 μL PCR reaction, and purified with either the DNA Clean and Concentrator Kit (Zymo Research) or Agencourt AmPURE XP beads (Beckman Coulter). The reaction and cycling conditions are optimized as follows: 94°C for 1 min, 28 to 30 cycles of 98°C for 10 s, 68°C for 5 min, followed by an extension of 72°C for 5 min. A linear plasmid backbone fragment (~5.2kb) was prepared for isothermal assembly using BamHI and SacI, purified and concentrated using DNA Clean and Concentrator Kit (Zymo Research), and further gel purified using Zymoclean Gel Recovery Kit (Zymo Research). We expressed a subset of these fully assembled genes between the BamHI and SacI sites of the splicing reporter plasmid backbone in this study, in place of the MFASS splicing reporter (see Splicing Reporter Design section). We performed isothermal assembly of 3 to 4 gene fragments of interest and the plasmid backbone using the Gibson Assembly Ultra Kit (SGI-DNA), and transformed into electrocompetent DH10B *E. coli* cells (New England Biolabs, or Life Technologies) to select for correct gene assembly. We confirmed the sequence for each gene with or without splice-disrupting variants using Sanger sequencing, before transfection into HEK293T cells for testing of mutation effects. We extracted and performed reverse transcription from RNA using the Cells to cDNA II kit (Thermo Fisher Scientific) and corresponding gene-specific primer for each exon (Table S4) according to manufacturer's protocol. For each tested exon, qPCR was performed with SYBR FAST qPCR Mastermix (Kapa Biosystems), using 1 μL of reverse-transcribed cDNA in a 20 μL PCR reaction, as well as primers flanking the upstream and downstream exons, and compared RT-PCR gene products of wild-type and mutant sequences for each gene of interest. Fragments of interest were further PCR purified and verified using Sanger sequencing.

## QUANTIFICATION AND STATISTICAL ANALYSIS

**DNA-seq Read Processing and Filtering—**SRE library datasets were generated from two Illumina MiSeq 300-bp paired-end sequencing runs and a Illumina HiSeq 2500 150-bp paired-end sequencing run. SNV library version 1 dataset was generated from Illumina MiSeq 300-bp paired-end sequencing. SNV library version 2 dataset was generated from Illumina NextSeq 2500 150-bp paired-end sequencing. We removed read pairs with any ambiguous "N" base calls, followed by read pair merging with *bbmerge* from the BBMap suite (BBtools package version 37). We developed custom Python and bash scripts to filter for perfect reads aligned to our reference, from which we can aggregate read counts for sequences from each sorted bin. We then further process these read counts to calculate inclusion index (see below section on the quantification of inclusion index).

To allow for stringent analysis of replicable data for SNVs, we require a coverage of at least 5 reads for the SRE library and at least 10 reads across all bins for the SNV library for the two biological replicates. Our SRE library size was 16,717 (5,975 wildtype sequences, 10,683 mutants, 59 controls) for the SMN1 intron backbone, and 13,922 (4,920 wild-type sequences, 8,942 mutants, 60 controls) for the DHFR intron backbone. We additionally require that inclusion indices agree between biological replicates within 0.30 (SRE library) and 0.20 (SNV library). For the SNV library, we only analyzed a mutant sequence if its corresponding wild-type sequence has an inclusion index of ≥ 0.5. The final library size after all filtering steps for the SRE library is 10,482 (3,714 wild-type sequences, 6,713 mutants, 55 controls). The final library size after all filtering steps for the SNV library size (version 1) is 6,768 (1,981 wild-type sequences, 3,853 mutants, 934 controls). The SNV library size (version 2) is 31,144 (2,339 wild-type sequences, 27,733 mutants that correspond to 2,198 wild-type sequences, 1,072 controls).

**Exon Inclusion Quantification—**We normalized bin counts based on read depth (reads per million, RPM) and corresponding bin population percentage after FACS using the following formula:

$$Normalized\ read\ count\ GFP_{bin,\ i} = \frac{percentage\ sorted \times raw\ read\ count\ GFP_{bin,\ i}}{reads\ per\ million}$$

We calculated exon inclusion index for each sequence based on a weighted average of normalized counts across all bins. Bin weights are assigned based on GFP fluorescence measurements of individual bins that correspond to the extent of exon inclusion or skipping. For the splicing regulatory element (SRE) library and single nucleotide variant (SNV) library, version 1:

$$\frac{(0 \times GFP_{+}) + (0.85 \times GFP_{int}) + (1 \times GFP_{neg})}{GFP_{+} + GFP_{int} + GFP_{neg}}$$

For the SNV library, version 2:

$$\frac{(0 \times GFP_{+}) + (0.80 \times GFP_{int-hi}) + (0.95 \times GFP_{int-lo}) + (1 \times GFP_{neg})}{GFP_{+} + GFP_{int-hi} + GFP_{int-lo} + GFP_{neg}}$$

The change in inclusion index for an individual library sequence between wild-type (WT) and mutant is computed as follows:

$$\Delta\ inclusion\ index = inclusion\ index_{mutant} - inclusion\ index_{WT}$$

A positive Δinclusion index denotes increased exon inclusion for the mutant relative to WT, while a negative Δinclusion index denotes increased exon skipping for the mutant relative to WT.

**ExAC and gnomAD Data Analysis—**Annotation of variants for individual human samples in VCF format were obtained from the Exome Aggregation Consortium (Lek et al., 2016) (ExAC, version 0.3.1), including global allele frequencies. We further obtained global allele frequencies of individual variants from the Genome Aggregation Database (gnomAD). We binned gnomAD global allele frequency similar to the ExAC study (Lek et al., 2016), and tested for significant difference between allele frequency bins using chi-square test of independence. We obtained the rate of protein-truncating variants from ExAC. We also obtained gene level evolutionary constraint estimates from ExAC based on probability of loss-of-function intolerance (pLI), and defined genes that are extremely intolerant of loss-of-function as those with a pLI score ≥ 0.9. We then tested for genes with enrichment in splice-disrupting variants (SDVs) using Fisher's exact test.

**Functional Genomic Analysis of SNVs—**We functionally classified our variants using the Ensembl variant effect predictor (McLaren et al., 2016) (VEP v80), and filtered the most severe sequence ontology (SO) term for a given variant. We obtained phyloP 100-way (v1.4) nucleotide conservation for the hg38 genome for the SNV library, and classified quickly evolving regions of the genome (accelerating, phyloP < −2.0), neutral selection (−1.2 ≤ phyloP ≤ 1.2) and highly conserved region of the genome (deleterious, phyloP > 2.0). To compute genomewide locations of ExAC SNVs by gene regions, we used GENCODE (release 27, GRCh38 reference assembly) for exon annotation, and bedtools (Quinlan, 2014) to annotate intronic regions by subtracting exon coordinates from gene coordinates. To determine the density of SNVs for each genomic position, we determined the number of SNVs averaged at each relative position for the SNV library across exons and upstream/downstream introns, and relative position is set such that the boundary of upstream intron/5′ exon = 0, and the boundary of 3′ exon/downstream intron boundary = 1. In addition, we incorporated scaled positions to normalize for variable intron and exon lengths. We performed similar positional SNV density analysis for genome-wide SNVs from the ExAC consortium across gene regions.

**Motif Analysis—**To define potential disruption of *k*-mer motifs by ExAC SNVs, we performed *k*-mer based motif enrichment analysis using *k*pLogo (git/e2fac18) for both splice acceptor (positions −20 to +3, upstream *intron-exon* junction) and splice donor (positions −3 to +6, downstream *exon-intron* junction). Based on our SNV dataset, SDVs are background-corrected against non-SDVs to obtain motif logos that are enriched or depleted at each nucleotide. We used a p value cutoff of p < 0.01, gapped *k*-mer length of *k* = 1,2,3,4 and fixation frequency of 0.75 (Wu and Bartel, 2017). We scored splice sites, exonic splicing enhancers/silencers, and conserved acceptor and donor intronic sequences based on metrics in Table S1.

In addition, we implemented the hexamer additive linear (HAL) model, which estimates a splicing strength score for every possible exon hexamer (Rosenberg et al., 2015). For each variant, we calculated the change in score at each position relative to the wild-type sequence.

We compared the distribution of maximum score change between SDVs and non-SDVs using the Mann-Whitney *U* test.

**Assessment of Variant Prediction Algorithms—**To computationally predict the effects of rare genetic variants on splicing, we used various prediction algorithms that are able to assess coding and/or non-coding SNVs in our assay. We selected Δinclusion index ≤ −0.5 as the threshold for splice-disrupting variant (SDV) and designate our calls as true positives. We assessed performance by varying the score threshold at which a variant is called splice-disrupting (considering whether the score is positively or negatively correlated to Δinclusion index). We assessed various genomic predictors that use a variety of machine learning methods, annotations, and training sets to predict the functional impact of coding and non-coding variants. These methods incorporate a variety of functional data, including conservation, histone modifications, DNase hypersensitivity, transcription factor binding, transcript abundance, and protein-level scores.

We obtained functional scores of single nucleotide variants from four genomic predictors based on the hg19 assembly: raw CADD scores from CADD v1.3 (r0.3 Exome Aggregation Consortium dataset), DANN whole-genome SNV scores (Nov. 2014), FATHMMMKL (git/908d865), fitCons multi-cell (i6 dataset, git/20f336d) highly significant scores ($p < \sim 0.003$), and LINSIGHT (git/58fe558). For SPANR (splicing-based analysis of variants) (Xiong et al., 2015), we obtained the predicted change in percent spliced in ($\Delta\psi$, or ΔPSI) for single nucleotide variants in our SNV library across the genome. The hexamer additive linear model (HAL) (Rosenberg et al., 2015) can only assess exonic variants.

To consider the predictive power of conservation alone, we obtained phyloP 100-way (v1.4) nucleotide conservation for the hg19 genome for the SNV library. In addition, we obtained phastCons (v1.4) scores for 100-way eutherian mammalian nucleotide conservation for our SNV library and genome-wide SNVs from the ExAC consortium (Siepel et al., 2005). To assess the functional effects of missense, exonic single nucleotide variants from the SNV library, we used variant annotations from PolyPhen (v2.2.2) and SIFT (v5.2.2).

We assessed above predictors using receiver operating characteristic and precision-recall analysis. We used the pROC package version 1.10.0 to compute and plot the ROC curves, calculate the 95% confidence interval, and calculate the area under the curve. The precision recall curves were plotted with a custom function which evaluates each method by varying the score threshold at which a sequence is classified as an SDV, and calculating the corresponding precision and recall. The area under the precision recall curve is calculated with the trapz function in R.

**Analysis of SDVs from GTEx RNA-seq—**Genotype data (from Illumina SNP arrays, whole exome sequencing, or whole genome sequencing) and RNA-seq data were obtained from the GTEx database (v6p release). To get a list of high-quality SNVs for further analyses, we used a quality filter of GQ ≥ 20 for whole-genome sequencing and whole-exome sequencing and a quality filter of IGC ≥ 0.2 for Illumina SNP arrays, all of which were provided by GTEx. These cutoffs are similar as recommended by the GATK package (Van der Auwera et al., 2013). In addition to the genotyped SNPs, we also identified dbSNPs (version 146) that are expressed in the RNA-seq data by requiring a minimum total read coverage of 10 and a minimum read coverage of 3 for the alternative allele.

The RNA-seq data in FASTQ format were first adaptor-trimmed. Subsequently, the reads were aligned to the hg19 genome and transcriptome (Ensembl Release 75) using HISAT2 (Kim et al., 2015) with parameters–mp 6,4–no-softclip–no-mixed–no-discordant. Only uniquely mapped read pairs were retained for further analyses. Samples with fewer than 25 million uniquely aligned read pairs were excluded due to low depth for splicing analysis. In total, 7822 RNA-seq datasets from 47 tissues and 515 donors were retained.

Percent-spliced-in (PSI) values were calculated using the method described in Schafer et al. (Schafer et al., 2015). This analysis was carried out for all internal exons from the GENCODE comprehensive annotation (v24lift37). To ensure the accuracy of PSI estimation, we required the exons to be covered by ≥ 15 total reads (inclusion reads + exclusion reads) or ≥ 2 exclusion reads per sample.

We compared PSI values from tissues expressing the gene containing an SDV with a cutoff of transcript per million (TPM) ≥ 1 based on median gene TPM values. After filtering on expression, exon PSI for 28 SDVs (out of 1,050 ExAC SDVs in this study) were available in at least one tissue sample. The distribution of PSI values across tissues was compared for individuals with the alternative SDV alleles versus those with the corresponding reference alleles. Comparisons were made with the Mann-Whitney *U* test, and adjusted p values were calculated using the Benjamini-Hochberg procedure at an FDR of 5%.

**Gene Ontology Enrichment Analysis—**We performed Gene Ontology (GO) enrichment analysis between SDV-containing genes (n = 473, for 1,050 SDVs) and all genes in the ExAC SNV library (n = 1,616, for 27,733 SNVs) using topGO. We determined over-representation of GO terms for SDV genes based on gene counts using Fisher's exact test. Each GO term is tested independently and only terms with p < 0.01 are shown (see Table S3).

**Software—***bbmerge* from the BBMap suite (v37) was used to merge raw paired-end sequencing files. Custom python and bash scripts used for read processing, and mapping reference and synthetic error read counts. Further analysis was performed with Python 2.7, using Pandas v0.21.0 and Numpy v1.13.3, and R v3.4.2, using tidyverse including dplyr v0.7.4 and ggplot2 v2.2.1. Variant analyses were performed using Ensembl variant effect predictor (v80), CADD (v1.3), MaxEntScan (Yeo and Burge, 2004), DANN (https://cbcl.ics.uci.edu/public_data/DANN/), FATHMM-MKL (git/908d865), fitCons (i6 dataset, git/20f336d), HAL (git/ca54d11), kpLogo (git/e2fac18), LINSIGHT (git/58fe558), phastCons (v1.4), phyloP (v1.4), PolyPhen (v2.2.2), SIFT (v5.2.2), and SPANR/SPIDEX (v1.0) (http://annovar.openbioinformatics.org/en/latest/).

## DATA AND SOFTWARE AVAILABILITY

The accession number for the sequencing data reported in this paper is GEO: GSE120695. Pre-processed datasets are available upon request. All code needed to reproduce the analyses is included in the following repository: https://github.com/KosuriLab/MFASS

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

This work was supported by the NIH (5U01HG007912 and DP2GM114829 to S.K. and U01HG009417 and R01AG056476 to X.X.), the NIH Biomedical Big Data Training Grant (T32LM012424 to C.P.B.), Searle Scholars Program (to S.K.), Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Felicia Codrea, Jeffrey Calimlim, and Jessica Scholes (UCLA BSCRC flow cytometry core) for technical assistance; Ron Weiss for the HEK293T landing pad cell line; Christopher D. Sundberg for the HepG2 cell line; Jason Ernst and Nathan B. Lubock for advice on bioinformatics analysis; and Douglas L. Black and George M. Church for guidance while developing MFASS. We would also like to thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource.

## REFERENCES

Adamson SI, Zhan L, and Graveley BR (2018). Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol. 19, 71. [PubMed: 29859120]

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. [PubMed: 20354512]

Altshuler D, Daly MJ, and Lander ES (2008). Genetic mapping in human disease. Science 322, 881–888. [PubMed: 18988837]

Arias MA, Lubkin A, and Chasin LA (2015). Splicing of designer exons informs a biophysical model for exon definition. RNA 21, 213–229. [PubMed: 25492963]

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526, 68–74. [PubMed: 26432245]

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, and Shendure J (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet 12, 745–755. [PubMed: 21946919]

Baralle D, and Buratti E (2017). RNA splicing in human disease and in the clinic. Clin. Sci. (Lond.) 131, 355–368. [PubMed: 28202748]

Battle A, Brown CD, Engelhardt BE, and Montgomery SB; GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/ NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration & Visualization—EBI; Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature 550, 204–213. [PubMed: 29022597]

Bomba L, Walter K, and Soranzo N (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 18, 77. [PubMed: 28449691]

Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, Shalem O, Chen DD, Schupp PG, Vinjamur DS, Garcia SP, et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197. [PubMed: 26375006]

Chan T-F, Poon A, Basu A, Addleman NR, Chen J, Phong A, Byers PH, Klein TE, and Kwok P-Y (2008). Natural variation in four human collagen genes across an ethnically diverse population. Genomics 91, 307–314. [PubMed: 18272325]

Cho S, Moon H, Loh TJ, Jang HN, Liu Y, Zhou J, Ohn T, Zheng X, and Shen H (2015). Splicing inhibition of U2AF65 leads to alternative exon skipping. Proc. Natl. Acad. Sci. USA 112, 9926–9931. [PubMed: 26216990]

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, et al.; Centers for Mendelian Genomics (2015). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am. J. Hum. Genet 97, 199–215. [PubMed: 26166479]

Cook KB, Kazan H, Zuberi K, Morris Q, and Hughes TR (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 39, D301–D308. [PubMed: 21036867]

Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL, et al.; Genotype-Tissue Expression Consortium (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med 9, eaal5209. [PubMed: 28424332]

De Conti L, Baralle M, and Buratti E (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA 4, 49–60. [PubMed: 23044818]

Diao Y, Li B, Meng Z, Jung I, Lee AY, Dixon J, Maliskova L, Guan K-L, Shen Y, and Ren B (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. 26, 397–405. [PubMed: 26813977]

Duportet X, Wroblewska L, Guye P, Li Y, Eyquem J, Rieders J, Rimchala T, Batt G, and Weiss R (2014). A platform for rapid prototyping of synthetic gene networks in mammalian cells. Nucleic Acids Res. 42, 13440–13451. [PubMed: 25378321]

Faigenbloom L, Rubinstein ND, Kloog Y, Mayrose I, Pupko T, and Stein R (2015). Regulation of alternative splicing at the single-cell level. Mol. Syst. Biol 11, 845. [PubMed: 26712315]

Frankel N, Davis GK, Vargas D, Wang S, Payre F, and Stern DL (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466, 490–493. [PubMed: 20512118]

Gaildrat P, Killian A, Martins A, Tournier I, Frébourg T, and Tosi M (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol 653, 249–257. [PubMed: 20721748]

Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, Cusanovich DA, and Shendure J (2017). CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. Am. J. Hum. Genet 101, 192–205. [PubMed: 28712454]

Gulko B, Hubisz MJ, Gronau I, and Siepel A (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet 47, 276–283. [PubMed: 25599402]

Hernandez RD, Uricchio LH, Hartman K, Ye J, Dahl A, and Zaitlen N (2017). Singleton variants dominate the genetic architecture of human gene expression. bioRxiv. 10.1101/219238.

Hong J-W, Hendrix DA, and Levine MS (2008). Shadow enhancers as a source of evolutionary novelty. Science 321, 1314. [PubMed: 18772429]

Huang Y-F, Gulko B, and Siepel A (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet 49, 618–624. [PubMed: 28288115]

Jian X, Boerwinkle E, and Liu X (2014). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet. Med 16, 497–503. [PubMed: 24263461]

Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, and Lehner B (2016). The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat. Commun 7, 11558. [PubMed: 27161764]

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, and Chasin LA (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 21, 1360–1374. [PubMed: 21659425]

Keinan A, and Clark AG (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740–743. [PubMed: 22582263]

Kim D, Langmead B, and Salzberg SL (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. [PubMed: 25751142]

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet 46, 310–315. [PubMed: 24487276]

Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, Haack TB, Graf E, Schwarzmayr T, Terrile C, et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun 8, 15824. [PubMed: 28604674]

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, and Maglott DR (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42 (Database issue, D1), D980–D985. [PubMed: 24234437]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. [PubMed: 27535533]

Lewis BP, Green RE, and Brenner SE (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. USA 100, 189–192. [PubMed: 12502788]

Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, and Pritchard JK (2016). RNA splicing is a primary link between genetic variation and disease. Science 352, 600–604. [PubMed: 27126046]

Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT, Zappala Z, Strober BJ, Scott AJ, et al.; GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration & Visualization—EBI; Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz (2017). The impact of rare variation on gene expression across tissues. Nature 550, 239–243. [PubMed: 29022581]

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469–476. [PubMed: 24759409]

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, and Cunningham F (2016). The Ensembl variant effect predictor. Genome Biol. 17, 122. [PubMed: 27268795]

Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, and Dermitzakis ET (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. 7, e1002144. [PubMed: 21811411]

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100–104. [PubMed: 22604722]

Ongen H, and Dermitzakis ET (2015). Alternative splicing QTLs in European and African populations using Altrans, a novel method for splice junction quantification. bioRxiv. 10.1101/014126.

Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239–243. [PubMed: 29420474]

Pala M, Zappala Z, Marongiu M, Li X, Davis JR, Cusano R, Crobu F, Kukurba KR, Gloudemans MJ, Reinier F, et al. (2017). Population- and individual-specific regulatory variation in Sardinia. Nat. Genet 49, 700–707. [PubMed: 28394350]

Pollard KS, Hubisz MJ, Rosenbloom KR, and Siepel A (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121. [PubMed: 19858363]

Quang D, Chen Y, and Xie X (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31, 761–763. [PubMed: 25338716]

Quinlan AR (2014). BEDTools: the Swiss-Army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.12.1–11.12.34.

Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJM, Gifford DK, and Sherwood RI (2016). High throughput mapping of regulatory DNA. Nat. Biotechnol 34, 167–174. [PubMed: 26807528]

Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, De T, Coin L, de Silva R, Cookson MR, et al.; UK Brain Expression Consortium; North American Brain Expression Consortium (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci 17, 1418–1428. [PubMed: 25174004]

Rosenberg AB, Patwardhan RP, Shendure J, and Seelig G (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell 163, 698–711. [PubMed: 26496609]

Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, Cheng C, Regev A, and Zhang F (2016). High-resolution interrogation of functional elements in the noncoding genome. Science 353, 1545–1549. [PubMed: 27708104]

Schafer S, Miao K, Benson CC, Heinig M, Cook SA, and Hubner N (2015). Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). Curr. Protoc. Hum. Genet 87, 11.16.1–11.16.14. [PubMed: 26439713]

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240. [PubMed: 23685454]

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, and Campbell C (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 31, 1536–1543. [PubMed: 25583119]

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050. [PubMed: 16024819]

Smigielski EM, Sirotkin K, Ward M, and Sherry ST (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 28, 352–355. [PubMed: 10592272]

Smith SA, and Lynch KW (2014). Cell-based splicing of minigenes. Methods Mol. Biol 1126, 243–255. [PubMed: 24549669]

Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, and Fairbrother WG (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet 49, 848–855. [PubMed: 28416821]

Takata A, Matsumoto N, and Kato T (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat. Commun 8, 14519. [PubMed: 28240266]

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69. [PubMed: 22604720]

UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82–90. [PubMed: 26367797]

Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, and Hernandez RD (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res. 26, 863–873. [PubMed: 27197206]

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, LevyMoonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33. [PubMed: 25431634]

Voelker RB, and Berglund JA (2007). A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. Genome Res. 17, 1023–1033. [PubMed: 17525134]

Wu X, and Bartel DP (2017). kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res. 45 (W1), W534–W538. [PubMed: 28460012]

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806. [PubMed: 25525159]

Yeo G, and Burge CB (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol 11, 377–394. [PubMed: 15285897]

Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, Johnson AD, Levy D, and O'Donnell CJ (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat. Genet 47, 345–352. [PubMed: 25685889]

## Highlights

- MFASS: massively parallel splicing minigene reporter for exonic and intronic variants
- Tested 27,733 natural human variants in 2,198 exons for defects in exon recognition
- Most splice-disrupting variants are rare, not at splice sites, and hard to predict
- MFASS enables variant assessment of large-effect splicing defects at scale

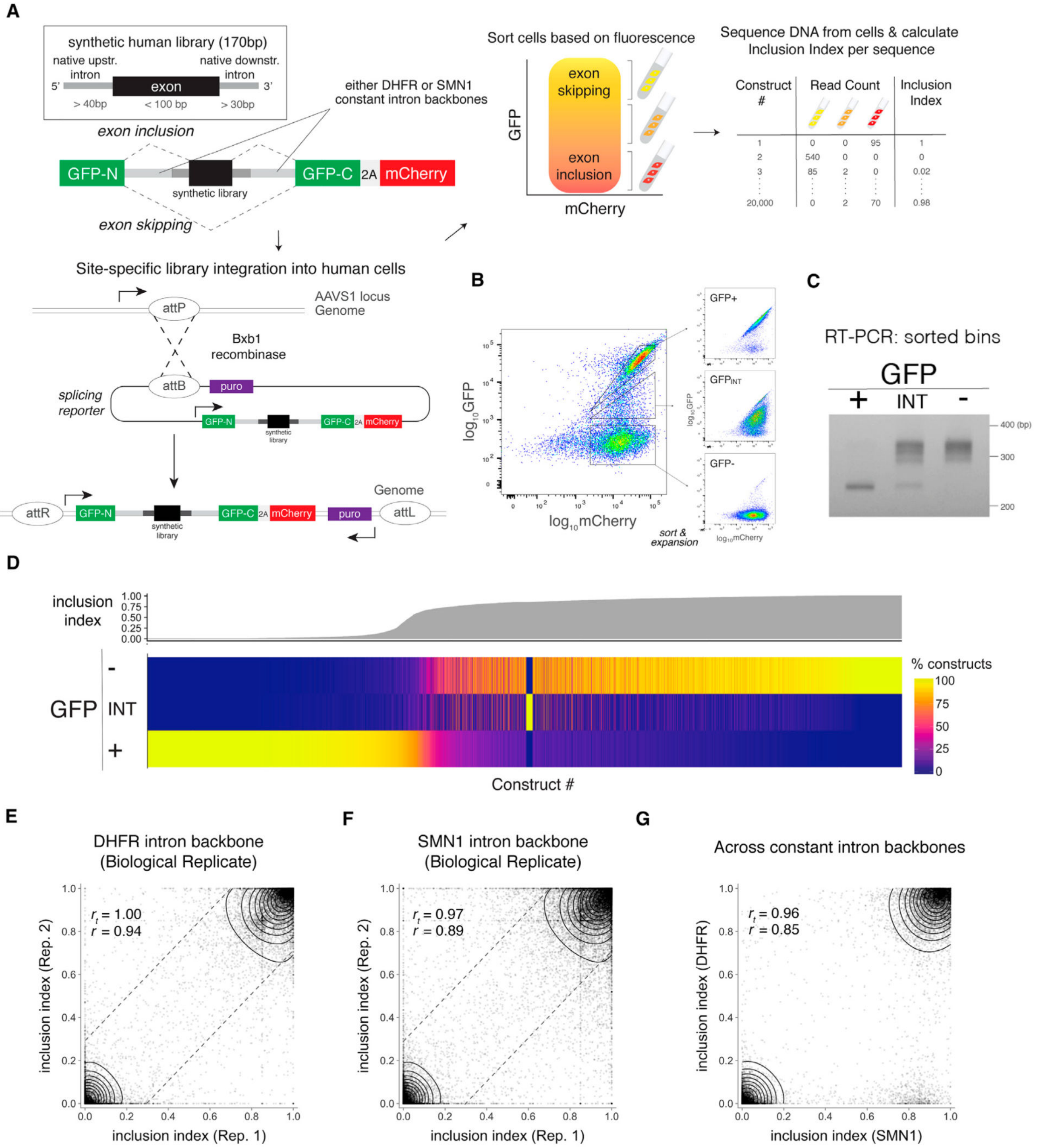


**Figure 1. Multiplexed Functional Assay of Splicing using Sort-Seq**

(A) We cloned synthetic human exons (black) and surrounding intronic sequences (dark gray) into our reporter plasmid containing a split-GFP reporter with flanking constant intron backbones (light gray), followed by site-specific integration into HEK293T cells using Bxb1 integrase. Cells are sorted into bins based on fluorescence, followed by amplicon sequencing of DNA from cells in each sorted bin. We calculated exon inclusion index for each sequence using a weighted average of normalized read counts based on exon inclusion level from bins (STAR Methods).

(B) We used FACS to sort the genomically integrated SRE library into three separate populations (left). After expansion, the sorted populations remained stable(right). GFP-int, GFP-intermediate. For this library (SMN1 intron backbone), we obtained ~4 million cells for $GFP_{+}$ and $GFP_{neg}$ bins and $4.2 \times 10^{5}$ cells for $GFP_{int}$ bin. The percentage of cells sorted is as follows: $GFP_{+}$ (33.3%); $GFP_{neg}$ (44.5%); and $GFP_{int}$ (5.6%).

(C) The observed RNA splicing efficiencies of the sorted bins as measured by RT-PCR correspond almost directly with observed fluorescence of the bins.

(D) We plotted the percentage of reads for each construct in the SRE library containing both natural and mutant exons (n = 10,477). We showed that most sequences fall predominantly into one bin, exhibiting either complete exon skipping or inclusion, allowing for facile classification of exon skipping variants of large effects (Δinclusion index %0.5). Corresponding exon inclusion indices for each bin are indicated at top panel. The data shown in (D) correspond to the SMN1 backbone.

(E–G) SRE library splicing behavior replicates between individual biological replicates and across two constant intron backbones. Tetrachoric correlation indicates whether two distinct measurements are concordant in one of the four quadrants and is more suited to assess large-effect variants.

(E) Exon inclusion indices show strong correlation between two independent biological replicates for C. griseus DHFR intron backbone ($r_t = 1.00$, $p < 10^{16}$, tetrachoric; $r = 0.94$, $p < 10^{16}$, Pearson).

(F) Exon inclusion indices show strong correlation between two independent biological replicates for human SMN1 intron backbone ($r_t = 0.97$, $p < 10^{16}$, tetrachoric; $r = 0.89$, $p < 10^{16}$, Pearson). For (E) and (F), after calculation of correlation coefficients, sequences for which inclusion indices do not agree within 0.30 (outside the dashed lines) are excluded from subsequent analysis.

(G) Results are robust across different intron backbones ($r_t = 0.96$, $p < 10^{16}$, tetrachoric; $r = 0.85$, $p < 10^{16}$, Pearson).

See also Figure S1.

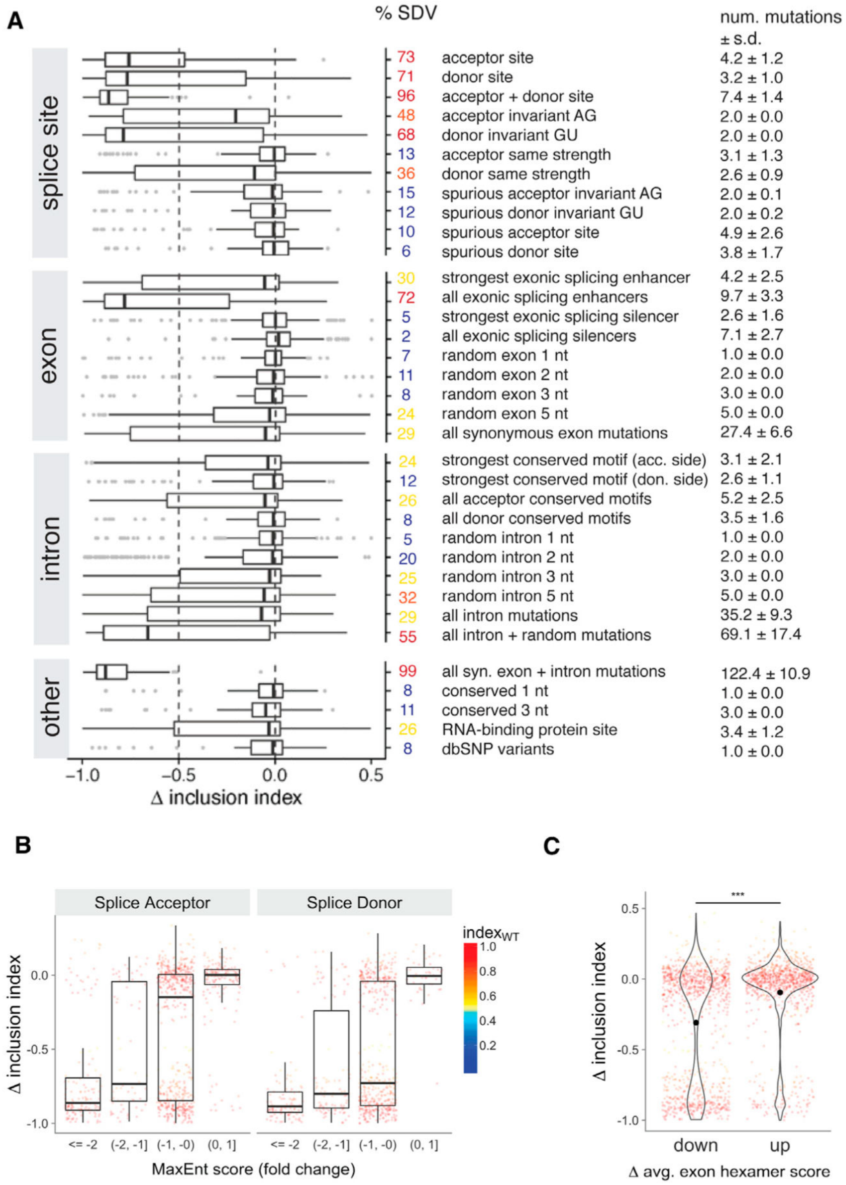


**Figure 2. Effects on Exon Recognition Are Not Easily Predicted across 6,713 Designed Mutations in Splicing Regulatory Elements**

(A) We quantitatively measured exon inclusion for iteratively designed mutations (n = 6,713) across categories of splicing regulatory elements from 205 human exons (see Tables S1 and S2 for categorical explanations and definitions). We defined SDVs as variants that result in a Δinclusion index ≤ −0.5, relative to the wild-type sequence (STAR Methods). We only consider SNVs when the corresponding wild-type sequence is also detected, requiring that the wild-type exons demonstrate inclusion in our assay (inclusion index of ≥0.5) for variants to be considered an SDV. Here, we highlight the data for the SMN1 intron backbone and detected 21.3% (1,428/6,713) of variants as SDVs across all categories. See also Figure S2A for mutations to exons that are skipped in MFASS (inclusion index of <0.5) across designed categories. Splice acceptor, positions −20 to +3; splice donor, positions −3 to +6.

(B) Mutating the splice acceptor and splice donor sites adversely affects exon inclusion based on MaxEnt prediction for included exons (inclusion index of ≥0.5; Yeo and Burge, 2004).

(C) Decreasing overall exon hexamer score leads to more exon skipping. Hexamer scores are based on the HAL model (Rosenberg et al., 2015). An alternative score metric is evaluated in Figure S2B (Ke et al., 2011). ***p < 0.001.

See also Figure S2 and Tables S1 and S2.

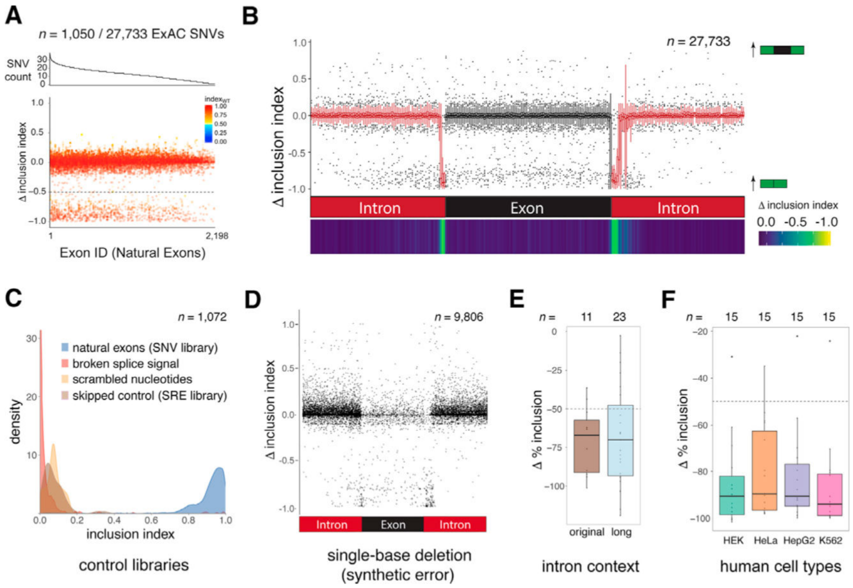


**Figure 3. MFASS Enables Functional Characterization of Variant Effect on Splicing at Scale across Libraries of Human Exons and Variants**

(A) The number of SNVs per exon sequence (top) and the Δinclusion index (bottom) of the 27,733 ExAC SNVs are plotted against the wild-type exon backgrounds (n = 2,198) and colored by the inclusion index of the corresponding wild-type (WT) sequence. Both the top and bottom panels are ordered in decreasing number of variants tested from 44 to 1 per human exon background, with an average of 12.6 human variants and 3.8 SDVs per assayed wild-type exon sequence background. We found 1,050 of 27,733 SNVs tested (3.8%) are SDVs (Δinclusion index ≤ −0.5) and are broadly spread across the 543 human exon backgrounds in 473 genes. Dashed line indicates the threshold (Δinclusion index = −0.5), below which we call SDVs.

(B) The change in inclusion index as a function of relative position for our SNV library across 2,198 human exon sequences shows that the splice donor and acceptor sites are most sensitive to mutations. Intron-exon boundary on the left corresponds to the splice acceptor, and the intron-exon boundary on the right corresponds to the splice donor. The splice donor is more sensitive to mutation because its consensus site is longer and more conserved. The bottom panel displays the relative sensitivity of each position. Each bin corresponds to 1 or 2 nucleotides per position, and locations are relative as we test a range of exon lengths.

(C) Three control sets for validating the SNV library (n = 1,072). Most control sequences that were designed to cause exon skipping led to almost complete loss of exon recognition. The three control sets were (1) scrambled sequences (n = 24), (2) a previously tested subset of exons that were skipped in the SRE library (n = 71), and (3) breakage of the splice sites (n = 977). The broken splice-signal control library mutates 5′ splice sites (SD) at the downstream intron from GT to CC and 3′ splice sites (SA) at the upstream intron from AG to TT. SA, splice acceptor; SD, splice donor. We included the distribution of wild-type sequences (i.e., natural exons; n = 2,339, of which 2,198 sequences have relevant SNV data; STAR Methods). These exons initially demonstrated exon inclusion in the SRE library (inclusion index ≥ 0.8), and we subsequently retested them and their associated SNVs in the SNV library.

(D) We analyzed the effects of single-base deletions derived from synthetic errors on exon inclusion. We showed the effect of exon inclusion for synthetic deletions (n = 9,801) across replicates, with an SDV rate of 3.59%. We observed an enrichment of SDVs at or near the splice sites.

(E) We validated large-effect rare variants detected by MFASS (n = 34) and their corresponding wild-type sequences. We measured exon inclusion in either the original sequence context examined in MFASS (n = 11) or as a more stringent test with an additional 130 bp of longer intronic contexts (n = 23) in HEK293T cells. For the longer set, we tested SDVs that represent variant classes in Figure 4B: missense variants (n = 3); synonymous variants (n = 3); intron variants (n = 4); splice donor (n = 4); splice acceptor (n = 5); and splice region variants (n = 4). The levels of exon inclusion were calculated for both the individual SDV and its respective wild-type sequence. All mutants were normalized to a no-insert control as a baseline of complete exon skipping for the assessment of change in exon inclusion. Dashed line indicates the threshold (D% inclusion = 50%), below which we call splicing-disrupting variants (SDVs).

(F) To examine the cell-type specificity of SDVs, we further picked a subset of 15 SDVs from the long intronic context with the strongest change in inclusion levels for testing their effects across 4 cell types and validated reporter constructs for WT or the corresponding SDVs. n = WT, SDV: 14,15 (HEK293T), 14,15 (HeLa S3), 14,15 (HepG2), 14,15 (K562). We found that large-effect splicing disruptions are consistent across 4 cell types in all 15 of the splice-disrupting variants assayed (15 of 15; 100%). The generalizability of per variant exon inclusion measurements across cell types is included in Figure S3E.

See also Figure S3.

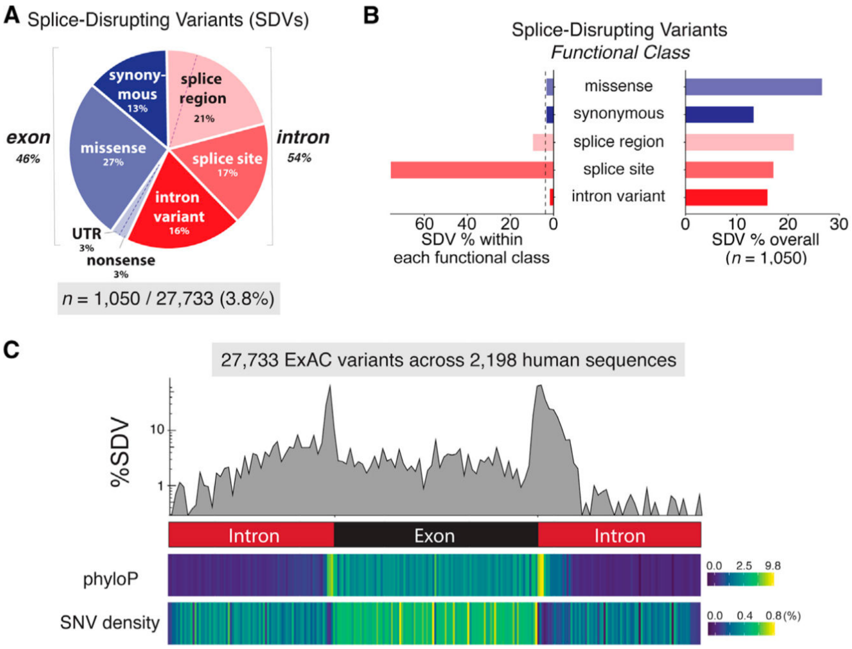


**Figure 4. Global Analysis of Splice-Disrupting Variants across 27,733 ExAC SNVs in or near 2,198 Human Exons**

(A) We functionally classified our variants by variant class from the Ensembl variant effect predictor (STAR Methods). SDVs (n = 1,050) from natural genetic variation are split almost equally between exonic and intronic regions (blue and red, respectively). Dashed line separates the exonic regions (4%) and intronic regions (17%) of the splice region variants. Splice site variants are defined as those within 2 bp of intron adjacent to exon, whereas splice region variants are located 3 bp into the exon and 8 bp into the intron, excluding splice sites.

(B) Splice site mutations are by far the most likely region to result in an SDV (left). However, because SNVs at splice sites are relatively rare, SDVs in regions other than the splice site constitute 83% of all SDVs (right). The distributions for non-SDVs across variant classes and the distribution of SDV effect sizes are shown in Figure S4A.

(C) The percentage of SDVs as a function of position along the exon and surrounding intron sequence shows that splice donor regions are more sensitive Than splice acceptor regions (top panel).Plotted below is the average change in mammalian evolutionary conservation (phyloP score averages) and ExAC SNV density as a function of location. Each bin corresponds to 1 to 2 nucleotides per position, and locations are relative to account for variable exon length.

See also Figure S4.

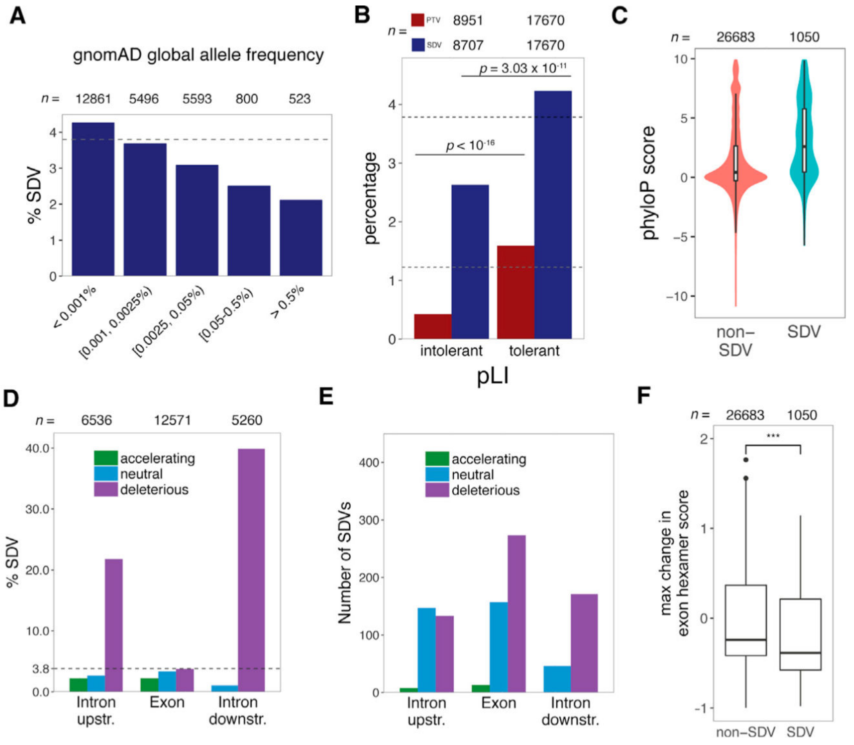


**Figure 5. Population Genetics, Evolutionary, and Functional Analyses of SDVs across 27,733 ExAC SNVs**

(A) The percentage of SDVs as a function of allele frequency shows significant reductions across allele frequencies from the Genome Aggregation Database (gnomAD) (chi-square test; $p = 1.03 \times 10^{-4}$). A vast majority (97.9%) of the ExAC variants assayed were rare (gnomAD global minor allele frequencies [MAF] ≤ 0.5%). Allele frequencies are not available for 2,460 variants because of insufficient coverage in gnomAD.

(B) We analyzed the proportion of SDVs and PTVs in genes predicted to be intolerant to loss-of-function alleles (pLI ≥ 0.9) and tolerant genes. We observe both significantly fewer SDVs (two-tailed Fisher's exact test; $p = 3.03 \times 10^{-11}$) and significant fewer PTVs (two-tailed Fisher's exact test; $p < 10^{-16}$) for exons within intolerant genes. Dashed lines mark the overall percentage of SDVs (3.8%) and PTVs (1.2%) in our dataset without considering the pLI metric.

(C) SDVs are under stronger evolutionary conservation as evidenced by higher overall phyloP scores (Mann-Whitney U test; $p < 10^{-16}$).

(D) Within introns, we found that positions that are evolutionarily conserved (deleterious; phyloP > 2.0; purple) have a higher SDV rate than those under neutral (−1.2 ≤ phyloP ≤ 1.2; blue) or accelerating selection (phyloP < −2.0; green; two-tailed Fisher's exact test; $p < 10^{-16}$).

(E) There are more SNVs outside of regions of high intron conservation, which leads to many SDVs located within nucleotides that display neutral selection.

(F) We observed a significantly higher negative maximum change in predicted exonic hexamer scores within exonic SDVs than non-SDVs (Student's t test; $p < 10^{-16}$). ***$p < 0.001$.

See also Figure S5.

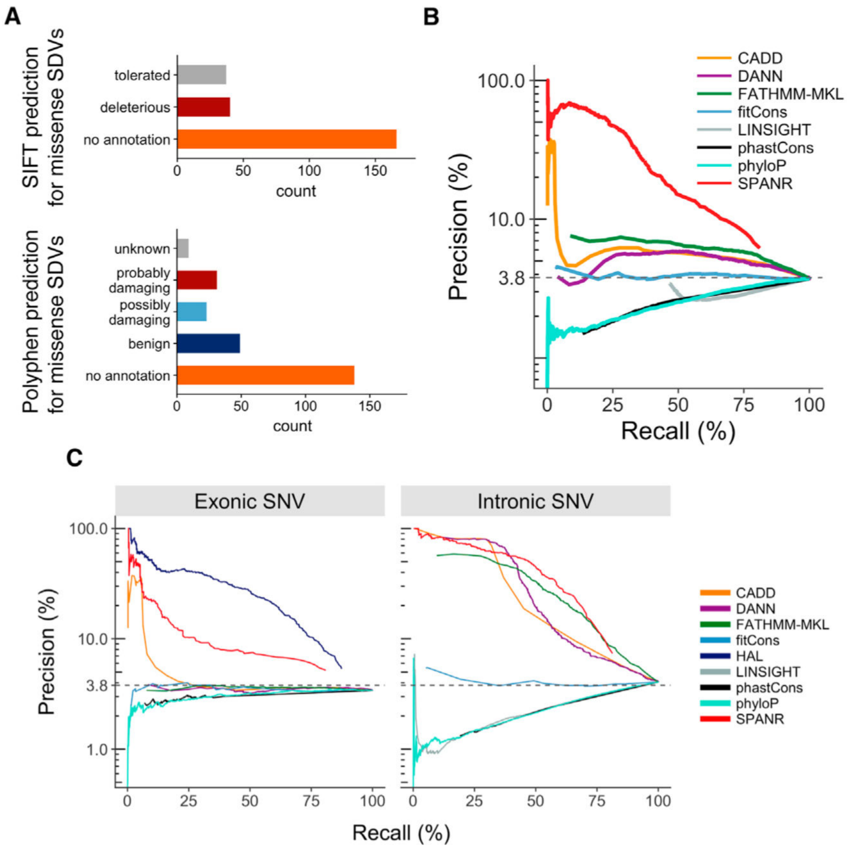


**Figure 6. Evaluation of Genomic and DeepLearning Predictors for Rare Variation on Splicing**

(A) Functional prediction from SIFT and PolyPhen for missense SDVs (n = 250) show few are predicted to be loss-of-function variants. The distributions for missense non-SDVs for SIFT and PolyPhen are shown in Figure S6A.

(B) Precision-recall curves for algorithms that can predict splicing or non-coding genetic variants. Dashed line represents the overall percentage of SDVs (3.8%) from MFASS. Corresponding receiver operating characteristic (ROC) curves are shown in Figure S6B.

(C) Precision-recall curves for algorithms that can predict splicing or non-coding genetic variants, focusing on either intronic or exonic variants only.

See also Figure S6.

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Bacterial and Virus Strains | | |
| Electrocompetent DH10B | Life Technologies | Cat#: 18290015 |
| Electrocompetent 10-beta | New England Biolabs | Cat#: C3020K |
| Chemicals, Peptides, and Recombinant Proteins | | |
| Polyethylenimine, Linear, MW 25000 | Polysciences | Cat#: 23966–1 |
| Lipofectamine 3000 | Thermo Fisher Scientific | Cat#: L3000–015 |
| Critical Commercial Assays | | |
| Gibson Assembly Kit | SGI-DNA | Cat#: GA1100–50MM |
| KAPA HiFi HotStart PCR kit | Kapa Biosystems | Cat#: KK2602 |
| QIAGEN blood and cell culture DNA midi Kit | QIAGEN | Cat#: 13343 |
| QIAGEN RNEasy MiniKit | QIAGEN | Cat#: 74104 |
| Illumina MiSeq Reagent Kit v3 600 cycles | Illumina | Cat#: MS-102–3003 |
| Cells to cDNA II kit | Thermo Fisher Scientific | Cat#: AM1722 |
| SYBR FAST qPCR Mastermix | Kapa Biosystems | Cat#: KK4601 |
| Deposited Data | | |
| Exome Aggregation Consortium (ExAC, version 0.3.1) | Lek et al., 2016 | http://exac.broadinstitute.org/ |
| Genome Aggregation Database (gnomAD) | Lek et al., 2016 | http://gnomad.broadinstitute.org/ |
| Raw sequencing data and processed files | This paper | GEO: GSE120695 |
| Experimental Models: Cell Lines | | |
| Human: RCA7 | This paper | N/A |
| Human: HepG2 | ATCC | Cat#: HB-8065 |
| Human: HeLa S3 | ATCC | Cat#: CCL-2.2 |
| Human: K562 | ATCC | Cat#: CCL-243 |
| Oligonucleotides | | |
| Splicing Regulatory Element (SRE) Library | Agilent Technologies | G4893A |
| Single Nucleotide Variant (SNV) Library | Agilent Technologies | G7223A |
| SRE Library Primers, see Table S4 | This paper | N/A |
| SNV Library Primers, see Table S4 | This paper | N/A |
| RT-PCR Primers, see Table S4 | This paper | N/A |
| Gene assembly and reverse transcription primers, see Table S4 | This paper | N/A |
| Recombinant DNA | | |
| Plasmid: Bxb1 Integrase | Duportet et al., 2014 | N/A |
| Plasmid: DYP1 (MFASS reporter with DHFR intron backbone & AscI-PacI multiple cloning site) | This paper | N/A |

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Plasmid: DYP4 (MFASS reporter with DHFR intron backbone & AscI-PacI multiple cloning site) | This paper | N/A |
| Plasmid: RCP124 (MFASS reporter with DHFR intron backbone & AgeI-NheI multiple cloning site) | This paper | N/A |
| Plasmid: RCP125 (MFASS reporter with SMN1 intron backbone & AgeI-NheI multiple cloning site) | This paper | N/A |
| Software and Algorithms | | |
| Python 2.7 | https://www.python.org | RRID: SCR_008394 |
| R (v3.4.2) | https://www.r-project.org | RRID: SCR_001905 |
| Ensembl Variant Effect Predictor (v80) | McLaren et al., 2016 | http://uswest.ensembl.org//useast.ensembl.org/info/docs/tools/vep/index.html?redirectsrc=//uswest.ensembl.org%2Finfo%2Fdocs%2Ftools%2Fvep%2Findex.html |
| FlowJo (v10.5.2) | https://www.flowjo.com/solutions/flowjo | RRID: SCR_008520 |
| SIFT (v5.2.2) | http://sift.bii.a-star.edu.sg/ | RRID: SCR_012813 |
| PolyPhen (v2.2.2) | Adzhubei et al., 2010 | http://genetics.bwh.harvard.edu/pph2/ |
| CADD (v1.3) | Kircher et al., 2014 | https://cadd.gs.washington.edu/ |
| DANN | Quang et al., 2015 | https://cbcl.ics.uci.edu/public_data/DANN/ |
| SPANR/SPIDEX | Xiong et al., 2015 | http://annovar.openbioinformatics.org/en/latest/ |
| FATHMM-MKL (git/908d865) | Shihab et al., 2015 | http://fathmm.biocompute.org.uk/downloads.html |
| LINSIGHT (git/58fe558) | Gulko et al., 2015 | https://github.com/CshlSiepelLab/LINSIGHT |
| kpLogo (git/e2fac18) | Wu and Bartel, 2017 | http://kplogo.wi.mit.edu/ |
| phastCons (v1.4) | Siepel et al., 2005 | http://compgen.cshl.edu/phast/ |
| HAL (git/ca54d11) | Rosenberg et al., 2015 | http://splicing.cs.washington.edu/ |
| phyloP (v1.4) | Pollard et al., 2010 | http://compgen.cshl.edu/phast/helppages/phyloP.txt |
| BBMap (v37) | Bbtools package | https://jgi.doe.gov/data-and-tools/bbtools/ |
| MaxEntScan | Yeo and Burge, 2004 | http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html |
| fitCons (i6 dataset, git/20f336d) | Huang et al., 2017 | http://compgen.cshl.edu/fitCons/ |
| Other | | |
| Agilent Tapestation 2200 | Agilent Technologies | Cat#: G2964AA |
| FACSAria III flow sorter | BD Biosciences | Cat#: 644832 |
| LSRII flow cytometer | BD Biosciences | Cat#: 744821 |
| LSRFortessa flow cytometer | BD Biosciences | Cat#: 751752 |