nature genetics

Pathogenic variants that alter protein code often disrupt splicing

Rachel Soemedi[1,2,7], Kamil J Cygan[1,2,7], Christy L Rhine[2], Jing Wang[2], Charlston Bulacan[3], John Yang[4], Pinar Bayrak-Toydemir[5], Jamie McDonald[5] & William G Fairbrother[1,2,6]

The lack of tools to identify causative variants from sequencing data greatly limits the promise of precision medicine. Previous studies suggest that one-third of disease-associated alleles alter splicing. We discovered that the alleles causing splicing defects cluster in disease-associated genes (for example, haploinsufficient genes). We analyzed 4,964 published disease-causing exonic mutations using a massively parallel splicing assay (MaPSy), which showed an 81% concordance rate with splicing in patient tissue. Approximately 10% of exonic mutations altered splicing, mostly by disrupting multiple stages of spliceosome assembly. We present a large-scale characterization of exonic splicing mutations using a new technology that facilitates variant classification and keeps pace with variant discovery.

Human genetic disorders occur in ~8% of the population[1]. Major technological advancements in the past decade have made it possible to detect all sequence variations in individual genomes in a cost-effective manner. In combination with capture technologies, targeted sequencing of all protein-coding regions of the human genome (the exome) has been increasingly used for routine diagnostics in Mendelian disorders[2,3]. Unfortunately, the tremendous progress that has been made in variant detection has outpaced the capacity to characterize sequence variations. Recent deep sequencing of human exomes detected ~14,000 single-nucleotide variants (SNVs) per individual, 47% of which were predicted to be deleterious by one or more *in silico* prediction tools, but there was very little agreement (<1%) between the commonly used methods[4].

Large-scale sequencing has identified many loss-of-function variants in asymptomatic individuals that are thought to cause severe genetic disorders[5,6]. These variants could represent annotation or sequencing errors, partial penetrance or recessive alleles carried by asymptomatic individuals. This uncertainty illustrates the urgency for better characterization of sequence variation. Although it is difficult to predict the effect of an SNV on protein function, the characterization of splicing mutations is a tractable problem. Splicing mutations are easily detected and quantified. They are deleterious, and one-third of the alleles that cause hereditary disease are predicted to confer some degree of missplicing[7]. Some of these mutations disrupt canonical splice sites, whereas others disrupt the multitude of enhancers and silencers that can modulate splice-site usage. Any change in an exonic sequence may therefore disrupt or create *cis*-acting elements that facilitate exon recognition, resulting in aberrant splicing. Here we present a new parallel splicing reporter system to characterize 4,964 published disease-causing exonic mutations for effects on splicing. The present study identified an allelic splicing imbalance caused by these exonic mutations and provided insights into the determinants and mechanisms of splicing aberrations.

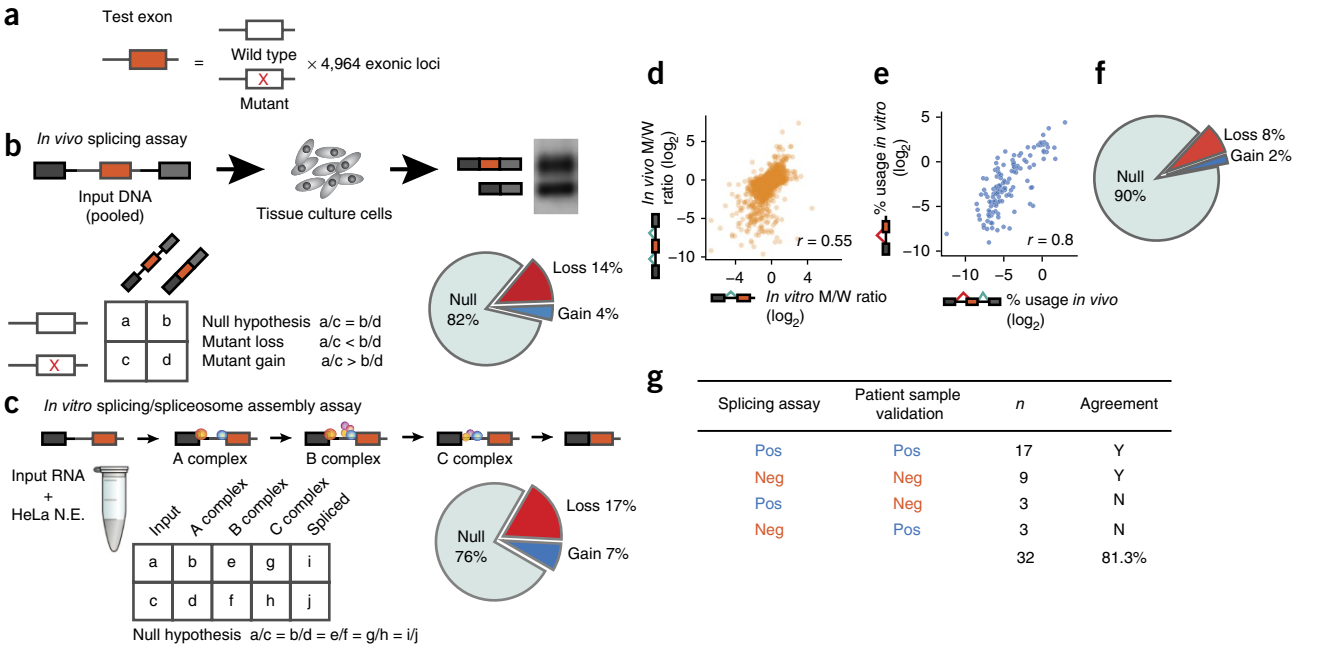
RESULTS

Massively parallel splicing assays

We developed a massively parallel splicing assay (MaPSy) to screen a panel of 4,964 exonic disease mutations (5K panel) reported in the Human Gene Mutation Database[8] (HGMD) for mutations causing splicing defects. One library was designed to evaluate the effects of the mutations on splicing *in vivo* via transfection in cells grown in tissue culture. The second library comprised RNA substrates designed to evaluate the mutations' effects on splicing *in vitro* via incubation in cell nuclear extract. Solid-phase oligonucleotide synthesis technology and PCR were used to manufacture the *in vivo* library and the template for the *in vitro* library (**Fig. 1**). Each reporter in the library contains a 170-mer genomic fragment of either the mutant or wild-type (reference) sequence, each of which consists of an exon, at least 55 nt of the upstream intron and 15 nt of the downstream intron (**Fig. 1a**)[9]. The allelic ratio for each mutant/wild-type (M/W) pair was determined from the allelic counts obtained from deep sequencing of the input libraries, the output spliced fractions and the RNA pools isolated from different *in vitro* spliceosomal intermediates (**Fig. 1b**,**c**). The most common outcome of disrupted splicing *in vivo* is exon skipping, whereas most pre-mRNAs with splicing mutations *in vitro* remain unspliced. While changes in transcription or stability may account for

[1]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA. [2]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island, USA. [3]Department of Computer Engineering, Brown University, Providence, Rhode Island, USA. [4]Department of Computer Science, Brown University, Providence, Rhode Island, USA. [5]Department of Pathology, University of Utah, School of Medicine, Salt Lake City, Utah, USA. [6]Hassenfeld Child Health Innovation Institute of Brown University, Brown University, Providence, Rhode Island, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to W.G.F. (william_fairbrother@brown.edu).

Received 4 September 2016; accepted 16 March 2017; published online 17 April 2017; doi:10.1038/ng.3837



g

Splicing assay	Patient sample validation	*n*	Agreement
Pos	Pos	17	Y
Neg	Neg	9	Y
Pos	Neg	3	N
Neg	Pos	3	N
		32	81.3%

Figure 1 MaPSy on the 5K panel. (**a**) The panel consists of 4,964 mutant–wild type pairs. (**b**) The panel was incorporated into a three-exon *in vivo* library. The allelic ratios of both the input and output library were determined by deep sequencing. The result from RT–PCR of output RNA (spliced species) is shown (see also **Supplementary Fig. 2f**). Splicing aberrations were found in 18% of mutants. (**c**) Allelic ratios were determined in spliceosomal intermediates; ~24% of species disrupt splicing *in vitro*. N.E., nuclear extract. (**d**) Allelic splicing M/W ratios *in vivo* versus *in vitro*. (**e**) Cryptic splice-site usage *in vivo* versus *in vitro*. (**f**) Exonic splicing mutations were identified in ~10% of the 5K panel. (**g**) Summary of MaPSy validations in tissue samples from patients with mutations tested with MaPSy.

an altered allelic ratio in the spliced fraction *in vivo*, the *in vitro* assay is a direct measure of splicing. Despite substantial differences in processing and substrate design, general agreements were observed between the allelic splicing ratios obtained from the two assays (**Fig. 1d**; Pearson's $r = 0.55$). Approximately 10% of the exonic mutations in the 5K panel altered splicing in both systems (**Fig. 1f**; >1.5-fold change, two-sided Fisher's exact test, adjusted with 5% false discovery rate (FDR)) and thus were regarded as unambiguous splicing changes and were classified as exonic splicing mutations (ESMs). We also performed MaPSy on a control panel of common SNPs, which disrupted splicing at a significantly lower level (8/228 or 3% of common SNPs, $P = 9.94 \times 10^{-5}$, two-sided Fisher's exact test; **Supplementary Table 1**). Additionally, cryptic 3′-splice-site usage was identified in both assays (**Fig. 1e**; Pearson's $r = 0.8$). Although most bona fide cryptic splicing events (74%) were caused by the creation of an AG (i.e., a splice acceptor site), a substantial number of disease-associated alleles caused dramatic shifts in the usage of an existing AG (**Supplementary Fig. 1**).

MaPSy was found to be robust (Pearson's $r = 0.85$–0.89 between allelic splicing ratios from experimental replicates; **Supplementary Fig. 2a–d**). In order to assess the validity and relevance of the splicing aberrations detected by MaPSy, we performed RT–PCR validations in RNA extracted from patient samples consisting of lymphoblastoid cell lines, fibroblasts, whole blood and postmortem brain tissues (**Supplementary Fig. 3a–f** and **Supplementary Table 2**). The validation samples were chosen solely on the basis of availability. In addition, we searched the literature for follow-up studies involving the mutations in the 5K panel that included RNA splicing analyses in patient tissue samples. A summary of the validations can be found in **Supplementary Table 2**. Overall, ~81% (26/32) of MaPSy-detected ESMs were validated in patient tissue samples (**Fig. 1g**). Furthermore, we compared the splice-site usage in 19 different cell lines that are part of the Encyclopedia of DNA Elements (ENCODE) data set with wild-type (reference) splicing in our 5K panel. Exons that spliced most efficiently in the 5K panel also had the highest average splice-site usage in the ENCODE cell lines, whereas exons that spliced least efficiently in the 5K panel also had the lowest average splice-site usage in the ENCODE data (**Supplementary Fig. 3g**).

Nonuniform distribution of splicing mutations

Some exons appeared to have a higher fraction of splicing mutations than others (for example, exon 8 of *MLH1* and exon 18 of *BRCA1*, adjusted $P = 2.26 \times 10^{-3}$ and 4.18×10^{-6}, respectively, two-sided binomial test). Interestingly, the set of (mostly) intronic splice-site mutations (SSMs) were also not distributed uniformly in disease-associated genes. Analyses of 2,314 disease-causing gene loci identified 64 genes that are predisposed to SSMs (**Fig. 2a**, left and **Supplementary Table 3**)[8]. SSMs often result in exon skipping. Not surprisingly, SSMs and nonsense mutations in human disease-associated transcripts were positively correlated, as they both result in loss of function of the proteins that they encode. This correlation was not observed between missense mutations and SSMs (**Fig. 2a**, middle and right). We found that ESMs were more abundant in genes that were also enriched for SSMs ($P = 3 \times 10^{-6}$, Kruskal–Wallis; **Fig. 2b** and Online Methods). This effect was more pronounced at the level of the individual exons ($P = 2.1 \times 10^{-34}$, Kruskal–Wallis; **Fig. 2c** and Online Methods). Moreover, disease-causing mutations with autosomal dominant inheritance showed a twofold ESM enrichment in haploinsufficient genes as compared to haplosufficient genes ($P = 0.002$, Kruskal–Wallis; **Fig. 2d**). This finding is in agreement with splicing mutations acting mainly via a loss-of-function mechanism and further confirms the utility of MaPSy in identifying deleterious ESMs (**Supplementary Fig. 4**). The same enrichment was also observed in SSMs reported in the HGMD ($P = 0.02$, Kruskal–Wallis; **Fig. 2e**)[10]. Recently, the Exome Aggregation Consortium (ExAC) identified 3,230 genes that are depleted of

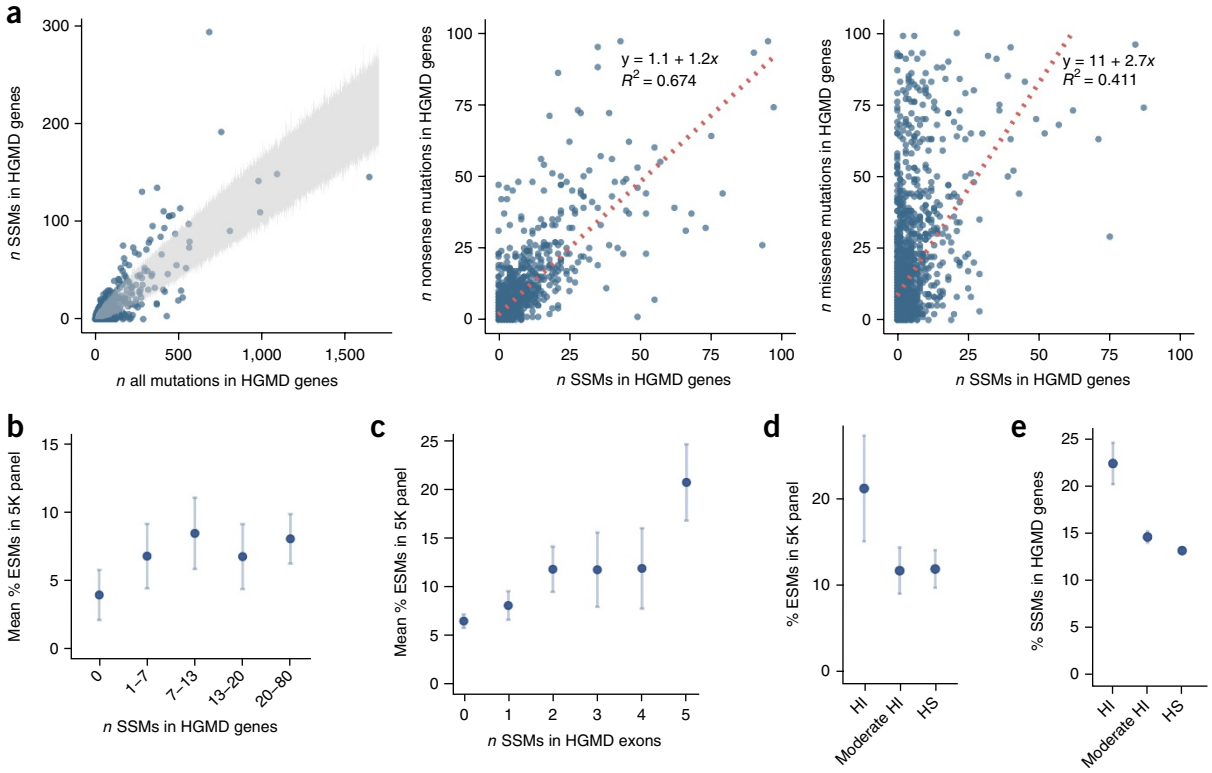


Figure 2 Prevalence of splicing mutations in disease-associated genes. (**a**) Left, SSMs versus all exonic mutations in the HGMD with the 99.9% confidence interval shown in gray. Middle and right, number of SSMs versus nonsense variants (middle) and missense variants (right) in all disease-associated genes. (**b**) Mean ESM percentage for each gene plotted against roughly equal bins of the percentage of SSMs in HGMD genes ($n = 708$). (**c**) Mean ESM percentage for each exon versus the number of SSMs per exon ($n = 2{,}048$). (**d**) Percentage of ESMs in haploinsufficient (HI; $n = 174$), moderately haploinsufficient ($n = 567$) and haplosufficient (HS; $n = 874$) genes in autosomal dominant diseases in the 5K panel[10]. (**e**) Percentage of SSMs in HGMD with autosomal dominant inheritance in haploinsufficient ($n = 1{,}383$), moderately haploinsufficient ($n = 14{,}059$) and haplosufficient ($n = 59{,}901$) genes[10]. Error bars, s.e.m. (**b**,**c**) and 95% confidence interval (**d**,**e**).

protein-truncating variants (PTVs) in 60,706 humans[6], thus providing evidence for extreme selective constraint. Because PTVs and splicing mutations often share the same loss-of-function mechanism, we examined disease-associated ESM occurrence in PTV-intolerant genes (probability that a gene is intolerant to a loss-of-function mutation (pLI) ≥ 0.9)[6] in comparison to other genes. In the 5K panel, we found a threefold excess of ESMs in PTV-intolerant genes ($n = 92$) as compared to PTV-tolerant genes ($n = 66$) that cause dominant disease traits (adjusted $P = 0.005$, Kruskal–Wallis; **Supplementary Fig. 5a**)[6]. These findings suggest that ESMs and SSMs are enriched in haploinsufficient genes, in which the loss of one functional copy likely leads to a disease phenotype.

Random forest classification of exonic splicing mutations

Various genomic and sequence features have been reported to affect splicing[10–14]. Although most of these studies were only done using a few substrates, MaPSy enables direct comparisons of the splicing performance of thousands of exons *in vivo* and *in vitro* (**Supplementary Fig. 2e**). Many of these features (for example, differential GC content between exons and introns and density of exonic splicing silencers (ESSs)) were confirmed with MaPSy (**Supplementary Fig. 6a**)[11,13]. We used random forest classification (Online Methods) on the ESM data set generated with MaPSy to further understand the different contributions of the various genomic and sequence features that may lead to ESM[15]. Performance of the random forest model was measured by mean area under the curve (AUC = 0.81, 0.755 and 0.816 for the *in vivo*, *in vitro* and combined approaches, respectively) (**Fig. 3a**). The *in vivo* assay performed better than the *in vitro* assay, but combining the two assays resulted in further increase in sensitivity to ESMs. Measures of feature importance were calculated as the mean decrease in accuracy (MDA). Each feature was categorized as a property of the mutation, the exon or the gene (**Fig. 3b**). It was surprising that the majority of the top predictors of ESMs that are not within the splice-site regions (~76%) were exon-level features, rather than some properties of the nucleotide substitutions (for example, exon splicing enhancer (ESE) disruption and ESS creation). In other words, some exon properties (for example, low ESE density and high ESS density) sensitize an exon to ESMs—variants in these exons are more likely to disrupt splicing (adjusted $P = 1.8 \times 10^{-12}$ and 7.8×10^{-18}, Kruskal–Wallis, for ESE and ESS density, respectively; **Supplementary Fig. 6b**). In addition, the random forest model suggests that ESMs are more likely to occur in genes with many introns. We found that PTV-intolerant genes[6] also contained more introns than the average for disease-associated genes ($P < 2.2 \times 10^{-16}$, Mann–Whitney), similar to ESM- and SSM-enriched genes (**Supplementary Fig. 5b**).

RNA-binding protein motifs in the 5K panel

Presumably, most mutations that alter splicing act by disrupting the binding site of an activator or by creating a binding site for a repressor. The loss or gain of previously characterized elements (i.e., the mutation being predicted to either promote or inhibit splicing) was compared

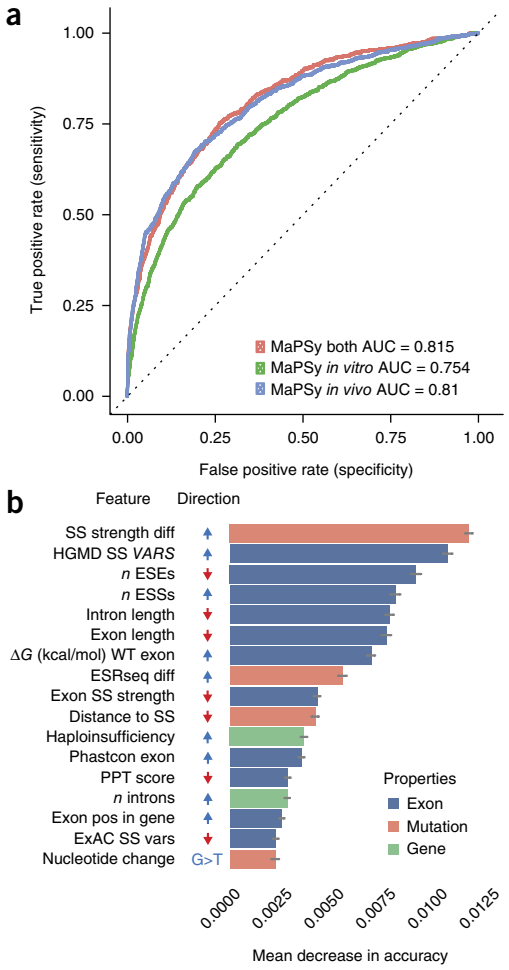


Figure 3 Random forest classification of exonic mutations that disrupt splicing. (**a**) The classification performance of the random forest model was calculated as the AUC in receiver operating characteristic (ROC) analysis. (**b**) The order of variable importance by mean decrease in accuracy. Error bars, s.d. The directions of changes that promote ESMs are shown; positive directions are colored blue, and negative directions are colored red. Variables include differences in splice-site strength[35] and hexamer splicing scores[11] (SS strength diff, ESRseq diff), the sum of the effects of splice-site variants in the HGMD and ExAC data sets (HGMD SS vars, ExAC SS vars)[6,8], numbers of ESEs and ESSs in the exon (*n* ESEs, *n* ESSs), the free-energy estimate in wild-type exon (Δ*G*(kcal/mol) for the wild-type (WT) exon)[31], exon conservation (Phastcon exon), number of introns (*n* introns) and relative exon position in the gene (Exon pos in gene). PPT, polypyrimidine tract.

to loss or gain of splicing in MaPSy[12,16–19] (**Fig. 4a**). A positive correlation was observed between gains of known exonic enhancing elements and relative splicing performance (i.e., mutant/wild type ratio, adjusted $P = 7.75 \times 10^{-25}$, linear regression; **Fig. 4b** and Online Methods). In contrast, a negative correlation was observed between gains of known exonic silencing elements and the relative performance of splicing (adjusted $P = 0.0001$, linear regression; **Fig. 4b**).

To predict which binding events of *trans*-acting factors were affected by exonic mutations, we compared the splicing effect of thousands of point mutations (using the relative splicing performance of the mutant versus wild-type sequence in MaPSy) with the predicted change of the binding affinity of 155 human RNA-binding proteins (RBPs) (determined bioinformatically using published data)[20]. Briefly, mutant–wild type pairs were ranked from the lowest to highest degree of exon inclusion for the mutant allele relative to the wild-type allele. The predicted changes in binding affinity were compared to the observed gain or loss of splicing activity (i.e., the mutant/wild type ratio)[21]. Levels of SRSF1, a well-characterized exonic splicing activator[22,23], showed a positive correlation with splicing (adjusted $P = 3.34 \times 10^{-27}$, linear regression; **Fig. 4b**), whereas levels of polypyrimidine tract–binding protein 1 (PTBP1), a known exonic splicing repressor, correlated negatively with splicing performance (adjusted $P = 3.26 \times 10^{-21}$, linear regression; **Fig. 4b**)[24,25]. As the presence of an RBP motif does not necessarily result in a binding event[20,26], it is necessary to validate the relationship between the increase or decrease of protein binding with the increase or decrease of splicing. An ESM in exon 20 of *COL1A2* (NM_000089.3:c.1045G>T) was predicted to create a PTBP1 motif. If PTBP1 binding were responsible for splicing repression, depletion of PTBP1 would be predicted to relieve the splicing defect. We found that, in the absence of PTBP1, rescue of splicing (i.e., ~0.5-fold less skipping) was observed in the mutant exon, but not in the wild-type exon ($P = 4.19 \times 10^{-5}$, two-sided Cochran–Mantel–Haenszel χ^2 test; **Fig. 4d**, right and **Supplementary Fig. 7a**). An ESM that was predicted to function by disrupting SRSF1 binding in exon 8 of *MLH1* (NM_000249.3:c.595G>C) was also selected for similar analysis. In the absence of SRSF1, the wild-type exon had a significant increase in skipping events ($P = 0.0002$, two-sided Cochran–Mantel–Haenszel χ^2; **Fig. 4d**, left and **Supplementary Fig. 7b**), but the mutant exon did not ($P = 0.07$, two-sided Cochran–Mantel–Haenszel χ^2). This result demonstrates how motif prediction can identify mutations where the gain of PTBP1 binding or the loss of SRSF1 binding can lead to the ESM phenotype.

Clustering the functional profiles of human RBP motifs in the 5K panel (Online Methods) resulted in 19 clusters, of which the 2 largest matched the profile of exonic splicing enhancers and repressors (**Fig. 4c**). The method was robust; >90% of all motifs that functioned as silencers or enhancers *in vivo* segregated into the same category *in vitro* ($P = 1 \times 10^{-16}$ and 1.5×10^{-10}, one-sided Fisher's exact test for Venn diagram overlap of exonic splicing repressors and activators, respectively; **Fig. 4c** and **Supplementary Fig. 8e**). Overall, 38 motifs corresponding to 35 RBPs consistently behaved as exonic repressors and 24 motifs corresponding to 25 RBPs behaved as exonic activators in both assays. Comparing the degree of predicted intronic binding with splicing performance suggests that most exonic repressors enhance splicing when bound in introns (57%; **Supplementary Fig. 8c**) and most exonic activators repress splicing when bound in introns (77%; **Supplementary Fig. 8d**). These findings reinforce the notion that splicing factors behave in highly position-dependent manners[7,27].

Mechanistic signatures of splicing mutants

During the development of the *in vitro* splicing assay in the 1980s, techniques were developed to isolate the biochemical intermediates in the stepwise assembly of the spliceosome[28]. A spliceosome is assembled from the A through the B to the C complex on the model adenovirus substrate, as previously described[29,30]. In accordance with catalysis occurring in the C complex, chemical intermediates of splicing co-migrated with the C complex during glycerol-gradient centrifugation (**Fig. 5a**). This same procedure was implemented on the 5K panel of mixed library substrates. Although each library member is the same length, greater heterogeneity in complex mobility was observed (**Fig. 5b**). Despite this increased heterogeneity, distinct splicing complexes were effectively partitioned, as the splicing intermediates and final products were found to segregate into the same fractions as seen in the control (**Fig. 5c**). Furthermore, each stage of spliceosome assembly had a distinct composition of library

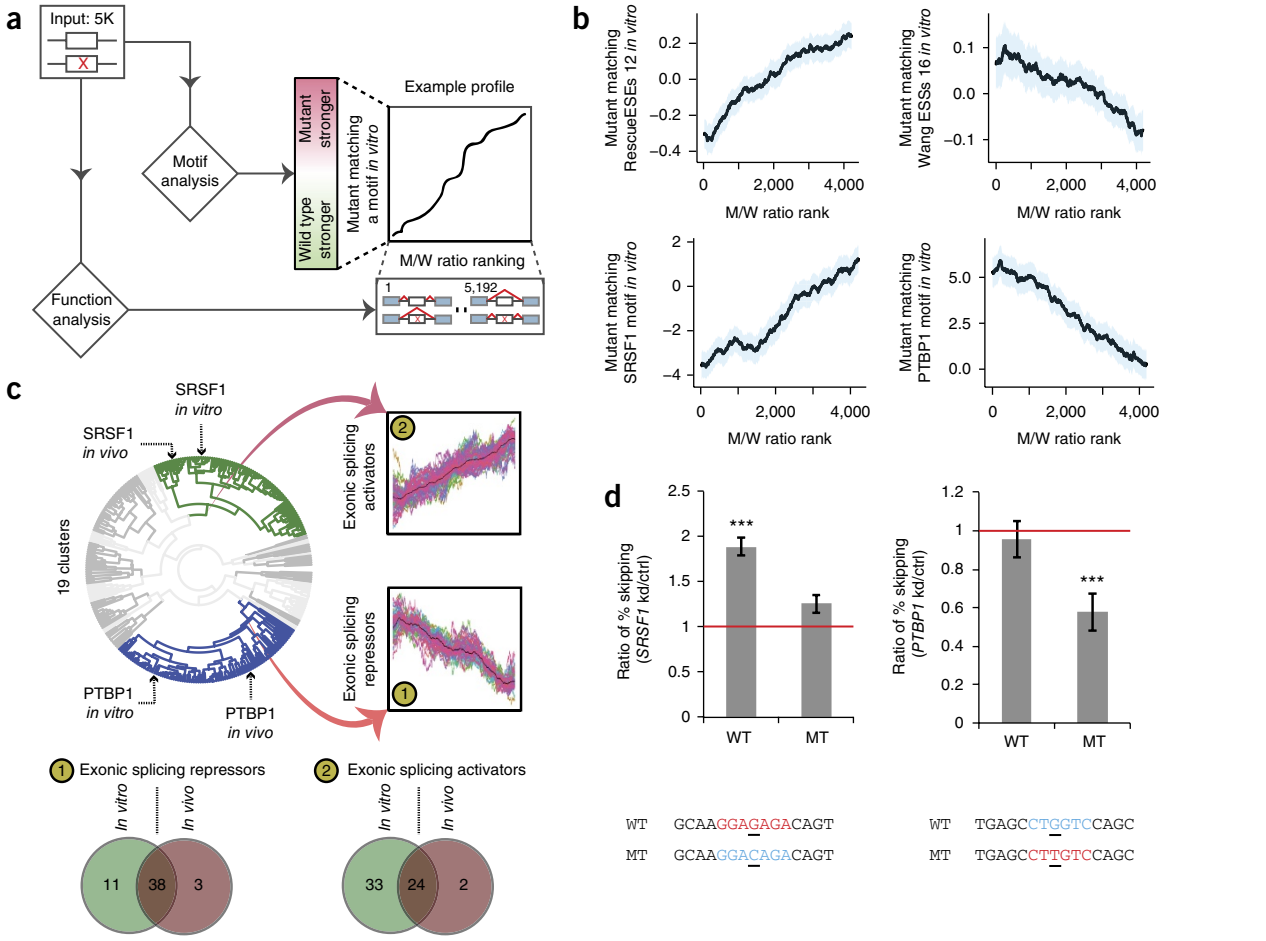


Figure 4 Detection of RBP motifs that affect splicing. (**a**) All mutant–wild type pairs were examined for difference in position weight matrices corresponding to 155 RBP motifs and known exonic *cis* elements. (**b**) Motif profiles show clear trends of agreement with previously defined functions. Shaded blue regions represent 95% confidence intervals. (**c**) Clustering of data shows similar functions for RBP motifs *in vivo* and *in vitro*. The plot was generated by sliding window (Online Methods). The mean values from each bin (sliding window) are colored black. (**d**) Left, in the absence of *SRSF1*, the mutant (MT) exon in which the SRSF1-binding motif is disrupted shows a modest but nonsignificant increase in exon skipping, whereas the wild-type (WT) exon with the SRSF1 motif has a twofold increase in exon skipping, compared to splicing in the presence of SRSF1. Right, the splicing phenotype of a mutation that creates a PTBP1-binding motif was rescued (~0.5-fold fewer skipping events) when *PTBP1* was knocked down, whereas the wild-type exon was not affected. The sequences represent the exonic sequence around the mutation site: red, sequence that highly matches the motif; blue, sequence that does not match the motif. ***$P < 0.001$, two-sided Cochran–Mantel–Haenszel test. Error bars, s.d. kd, knockdown; ctrl, control. Experiments were performed in two cell culture replicates.

species that could be further enriched by a systematic evolution of ligands by exponential enrichment (SELEX) approach (**Fig. 5d** and **Supplementary Fig. 9a**). For example, extracting RNA from the B/C fraction and repeating the spliceosome assembly assay returned a clear bias toward the B/C complex (**Fig. 5d**, middle), whereas reassembly of the A fraction resulted in a bias toward the A complex (**Fig. 5d**, bottom). By using glycerol-gradient centrifugation coupled with next-generation sequencing, the allelic ratio of each locus was determined at the different stages of spliceosome assembly: pre-assembly (t_0), A, B/C and spliced. In general, RNA species that were enriched in the early A complex were under-represented in the spliced fraction, suggesting that the species blocked from transitioning to the catalytic B/C complex were accumulating in the A complex. Conversely, RNA species that were enriched in the B/C complex were also enriched in the spliced fraction, suggesting that spliceosomes at the B/C stage were mostly committed to splicing (**Supplementary Fig. 9b**). Clustering the 5K panel by allelic ratios in the different spliceosomal fractions showed distinct patterns of disruptions. Most mutations affected multiple transitions of the spliceosome (**Fig. 6** and **Supplementary Fig. 9c**). We found that mutations in the same exon were more likely to cluster together ($P = 0.008$, permutation test). This result suggests that an exon disrupted by splicing mutations will tend to fail at the same stage of spliceosome assembly, a behavior that is consistent with the finding that exon properties are strong predictors of ESMs (**Fig. 3b**). The allelic ratio profiles in the different assemblies seem to represent mechanistically distinct scenarios of splicing disruption. For example, mutants in cluster 20 are strongly inhibited in each step of spliceosome assembly (**Fig. 6**). Interestingly, cluster 20 comprises mutations that are likely to trigger structural rearrangements (average $\Delta\Delta G$ = 1.95 kcal/mol, adjusted $P = 0.014$, permutation test)[31]. They are single substitutions that, on average, were predicted to trigger the formation of four new base pairs that contribute to a more closed RNA secondary structure. Cluster 15 contained mutations in weakly defined exons (low differential GC content and high numbers of ESSs, adjusted $P = 0.008$

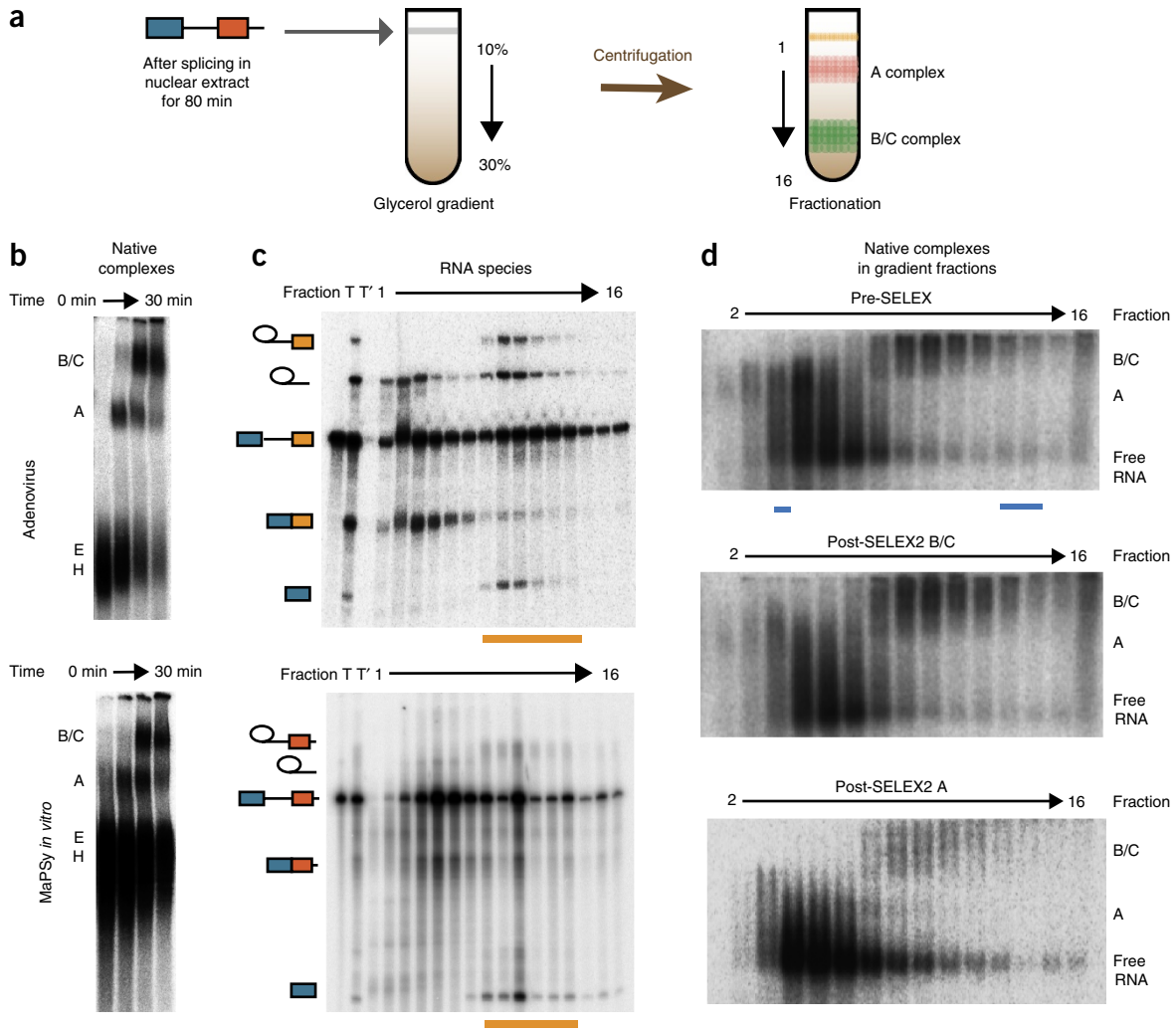


Figure 5 Isolation of spliceosomal intermediates. (**a**) After the *in vitro* MaPSy assay, the splicing reaction was loaded onto a 10–30% glycerol gradient followed by fractionation. Spliceosomal intermediates from different stages of assembly were retrieved from the different fractions. (**b**) Spliceosomal complexes (B/C, A, E, H) visualized in native gels for control (top) and heterogeneous library (bottom) substrates. (**c**) RNA splicing intermediates migrate to the same fractions in control and heterogeneous library substrates (underlined in orange). Total RNA before (T) and after (T′) splicing is indicated. (**d**) Reassembly of purified B/C and A fractions (middle and bottom) in comparison to assembly of original input (top). Fractions used for SELEX are underlined (cyan).

and 0.014, respectively, permutation test) and flanked by highly conserved introns (adjusted $P = 0.006$, permutation test). The splicing progression of these mutants was stalled in A and B/C, all of which significantly altered splicing *in vitro* and ~80% of which also significantly altered splicing *in vivo*. Exons in clusters 15 and 20 are also frequent targets of disease-causing SSMs[8], which is consistent with the finding that disease-causing ESMs and SSMs are often co-enriched in the same exons. In contrast, mutations in cluster 14 were associated with strongly defined exons (high differential GC content and low numbers of ESSs, adjusted $P = 0.001$ and 0.002, respectively, permutation test) and rarely disrupted splicing (**Fig. 6**). Mutants in cluster 7 were found in exons with strong splice sites (adjusted $P = 0.01$, permutation test), and their respective wild-type exons were strong splicers both *in vivo* and *in vitro*, having a mean splicing efficiency that was significantly higher than the mean splicing efficiency of wild-type exons from a random sampling of 10,000 (adjusted $P = 0.0008$ for both assays, permutation test). The splicing progressions of these mutants were mainly inhibited in the A complex. Whereas mutations in clusters 15, 16 and 20 represented ESMs with the most dramatic change in the splicing phenotype of the mutant substrate in comparison to the wild-type substrate, ESMs in clusters 7 and 14 had modest effects on splicing (**Supplementary Fig. 10**). It remains to be determined whether these distinct modes of splicing disruption are associated with the degree of severity or other aspects of disease phenotypes. We predict that a mechanism operating via structural changes (for example, cluster 20) is likely to function independently of tissues and cell types, as they seem more independent of *trans*-acting factors that may vary across tissues and cell types, whereas mutational mechanisms that involve *trans*-acting factors recognizing exonic-binding motifs (for example, cluster 15) are more likely to be tissue and cell type dependent.

Each mutation in the 5K panel represents a variant reported in a patient and/or family in the last four decades. We have established a large-scale collection of the effect of exonic mutations on splicing and created a public webserver that enables visualization of the MaPSy results (see URLs and **Supplementary Fig. 11**).

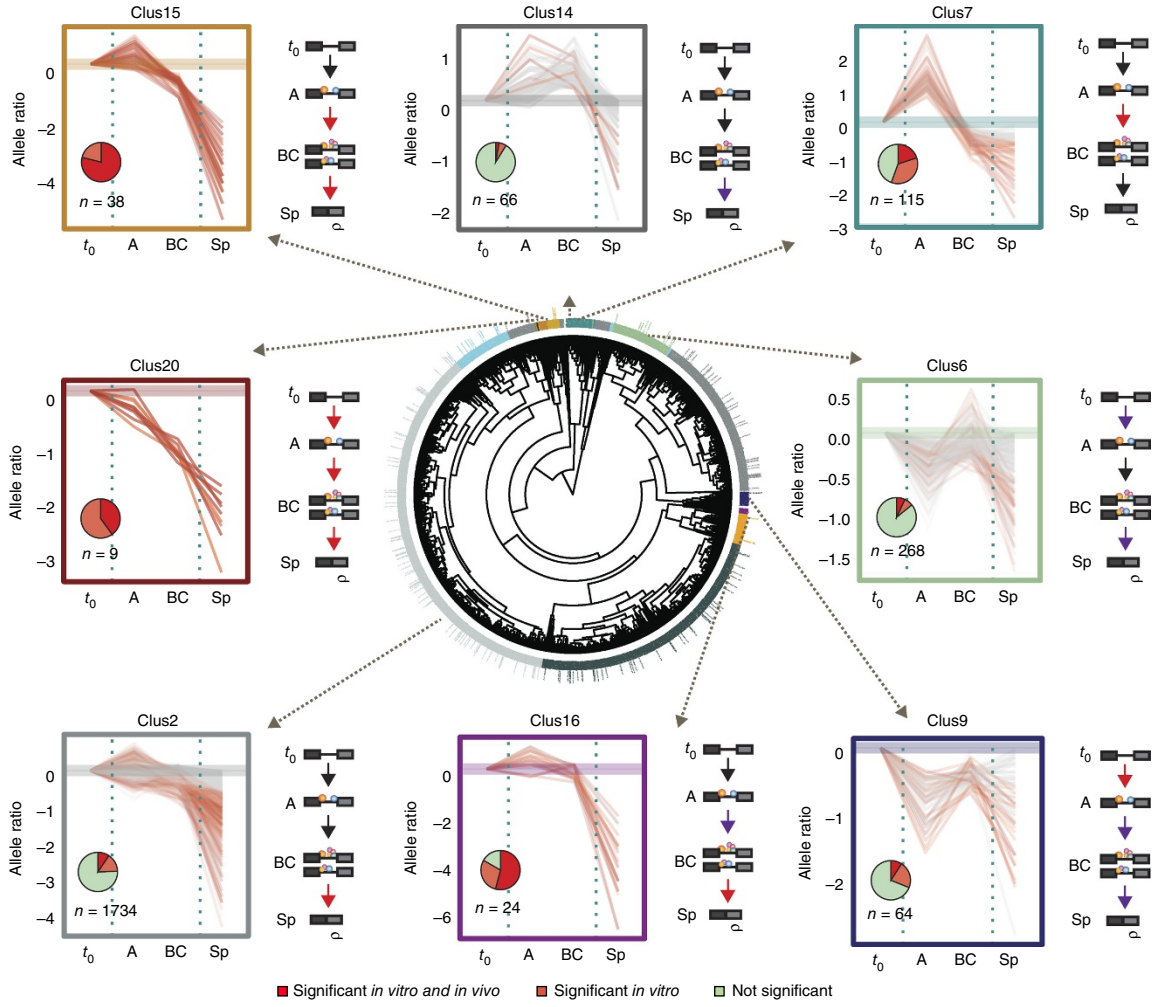


Figure 6 Clustering of allelic ratios provides ESM mechanistic insights. The result of hierarchical clustering of allelic ratios in spliceosomal fractions is shown (center plot) with representative clusters shown in different colors. The individual panels surrounding the center plot show the allelic ratios of each mutant–wild type pair in the different fractions (t_0, A, B/C and spliced (sp)) for the corresponding clusters. Each pair is colored according to its ESM classification (dark red for significance in both assays, orange for significance *in vitro* and gray for no significance). The complete profile of all clusters can be found in **Supplementary Figure 9c**. Pie charts in individual panels show the proportion of ESM classifications. Spliceosome stages are depicted at the right of the individual panels. Major disruptions in assembly transitions are indicated with red arrows, and minor disruptions are indicated with purple arrows.

DISCUSSION

The need for better characterization of sequence variation is ever more urgent with the increasing number of rare variants being discovered from many large-scale sequencing efforts[6,32]. Previous studies tested the effect of random *k*-mers in enhancing or silencing splicing[11,33,34]. We present the results of a survey of the effects of 4,964 point mutations on splicing using MaPSy, a new parallel splicing system. We further characterized the splicing aberrations by their stage of disruption in spliceosome assembly. We found that ~10% (513/4,964) of exonic disease-associated alleles disrupt splicing *in vivo* and *in vitro*. In contrast, only 3% (7/228) of common SNPs altered splicing in both assays. It is interesting that in diseases that are more frequently caused by splicing mutations, more exonic mutations were also found to disrupt splicing. This likely reflects disease processes that occur through loss-of-function mechanisms. We found that exonic features have a large role in forming ESMs. We also identified 24 exonic RBP motifs that are associated with increased splicing and 38 RBP motifs that are associated with decreased splicing.

MaPSy has certain limitations; particularly, only mutations in exons of fewer than 100 nt in length can be evaluated owing to the current limitation in oligonucleotide synthesis technology. Given that the average length of internal exons is around 130 nt, half of all human exons are not eligible for splicing characterization using MaPSy. We also cannot rule out the presence of other influences—for example, flanking splice sites, different transcription efficiencies and tissue-specific effects, all of which are not preserved in MaPSy. It is intriguing that some features previously shown to be predictors for SSMs but not present in MaPSy (for example, flanking intron length and number of introns) were also identified as predictors for ESMs (**Fig. 3b**)[14]. These findings, together with the high concordance rate with splicing phenotypes in corresponding patient tissue samples, suggest that, despite these limitations, MaPSy contains most of the critical elements required for splicing in native conditions and thus is a powerful tool for characterization of the sequence variation underlying splicing aberrations.

In conclusion, MaPSy facilitates large-scale identification and characterization of ESMs. The system effectively translates to 5K implementations of basic mutational approaches and can be further adapted to other mutation panels, thus accelerating efforts to characterize all sequence variation.

URLs. Visualization of MaSPy results, http://fairbrother.biomed.brown.edu/ESM_browser/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank K. Villanueva for generating the list of SNPs used in this study and A. Leblang for compiling the variants to make the oligonucleotide library. We thank M. Jurica and M. Moore for suggestions and protocols for the *in vitro* spliceosome assembly assay and nuclear extract preparation. We thank A. Janssens for contacting investigators for patient samples. We thank A. Toland (Ohio State University), J. Marini (NIH/NICHD) and A. Goate (Washington University Alzheimer's Disease Research Center) for contributing patient samples for validation. R.S. was supported by a Postdoctoral Fellowship from the Center for Computational Molecular Biology (CCMB), Brown University. C.R. was supported by a Graduate Research Fellowship from the National Science Foundation (NSF). This work was supported by US National Institutes of Health (NIH) grants R01GM095612 (to W.G.F.), R01GM105681 (to W.G.F.) and R21HG007905 (to W.G.F.) and by SFARI award 342705 (to W.G.F.). Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University and the Genomics Core Facility, Brown University.

AUTHOR CONTRIBUTIONS

W.G.F. and R.S. designed the experiments. R.S. performed MaPSy experiments. R.S., J.W., P.B.-T. and J.M. performed validation experiments. K.J.C. performed alignment, counting and RBP motif analyses. R.S. performed ESM analyses, machine learning and MaPSy SELEX analyses. C.L.R. performed HGMD gene analyses. C.B. and J.Y. developed the visualization web browser. W.G.F. and R.S. wrote the paper with contributions from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Baird, P.A., Anderson, T.W., Newcombe, H.B. & Lowry, R.B. Genetic disorders in children and young adults: a population study. *Am. J. Hum. Genet.* **42**, 677–693 (1988).
2. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *J. Am. Med. Assoc.* **312**, 1870–1879 (2014).
3. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
4. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
5. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
6. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
7. Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J. & Fairbrother, W.G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. USA* **108**, 11093–11098 (2011).
8. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
9. Taggart, A.J., DeSimone, A.M., Shih, J.S., Filloux, M.E. & Fairbrother, W.G. Large-scale mapping of branchpoints in human pre-mRNA transcripts *in vivo*. *Nat. Struct. Mol. Biol.* **19**, 719–721 (2012).
10. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
11. Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **21**, 1360–1374 (2011).
12. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
13. Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1**, 543–556 (2012).
14. Mort, M. *et al.* MutPred Splice: machine learning–based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15**, R19 (2014).
15. Breiman, L.Random forests. *Mach. Learn.* **45**, 5–32 (2001).
16. Wang, Z. *et al.* Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831–845 (2004).
17. Ke, S., Zhang, X.H. & Chasin, L.A. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res.* **18**, 533–543 (2008).
18. Smith, P.J. *et al.* An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum. Mol. Genet.* **15**, 2490–2508 (2006).
19. Zhang, X.H. & Chasin, L.A. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.* **18**, 1241–1250 (2004).
20. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
21. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
22. Long, J.C. & Caceres, J.F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* **417**, 15–27 (2009).
23. Rahman, M.A. *et al.* SRSF1 and hnRNP H antagonistically regulate splicing of *COLQ* exon 16 in a congenital myasthenic syndrome. *Sci. Rep.* **5**, 13208 (2015).
24. Shen, H., Kan, J.L., Ghigna, C., Biamonti, G. & Green, M.R. A single polypyrimidine tract binding protein (PTB) binding site mediates splicing inhibition at mouse IgM exons M1 and M2. *RNA* **10**, 787–794 (2004).
25. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N. & Sanford, J.R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **21**, 1563–1571 (2011).
26. Wang, J., Xiao, S.H. & Manley, J.L. Genetic analysis of the SR protein ASF/SF2: interchangeability of RS domains and negative control of splicing. *Genes Dev.* **12**, 2222–2233 (1998).
27. Lim, K.H. & Fairbrother, W.G.Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics* **28**, 1031–1032 (2012).
28. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S. & Sharp, P.A. Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55**, 1119–1150 (1986).
29. Konarska, M.M. & Sharp, P.A. Electrophoretic separation of complexes involved in the splicing of precursors to mRNAs. *Cell* **46**, 845–855 (1986).
30. Das, R. & Reed, R. Resolution of the mammalian E complex and the ATP-dependent spliceosomal complexes on native agarose mini-gels. *RNA* **5**, 1504–1508 (1999).
31. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
32. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
33. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–1052 (2012).
34. Rosenberg, A.B., Patwardhan, R.P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
35. Yeo, G. & Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).

ONLINE METHODS

Library design and synthesis. Nonsynonymous mutations classified as disease causing (DM) were downloaded from the Human Gene Mutation Database[8] (HGMD; accessed in May 2012) and mapped to the GRCh37/hg19 human reference sequence. Mutations were mapped to internal exons that were ≤100 nt in length, and the exons that fit into 170-nt genomic windows that included 15 nt of the downstream intron and ≥55 nt of the upstream intron (n = 4,964) were selected. The mutant and wild-type versions of the 170-mer genomic fragments were flanked by 15-mer common primer sequences and designed into a 200-mer oligonucleotide library. Solid-phase oligonucleotide synthesis was performed by Agilent Technologies and used to generate *in vivo* and *in vitro* reporters.

MaPSy *in vivo* assays. The *in vivo* splicing reporter includes a cytomegalovirus (CMV) promoter, an adenovirus (pHMS81)[36] exon with part of its downstream intron, the 200-mer oligonucleotide library, exon 16 of *ACTN1* with part of intron 15 and the bGH poly(A) signal sequence (**Supplementary Fig. 12**). Common sequences (everything except the 200-mer library) were concatenated by overlapping PCR and cloned with TOPO TA (Invitrogen) to generate a 5 common fragment and a 3 common fragment. Each cloned fragment was PCR amplified, and equimolar amounts of the common fragments and the oligonucleotide library were concatenated in a single PCR reaction and purified and size selected twice with a 0.4:1 ratio of Agencourt AMPure beads (Beckman Coulter) to PCR reaction. The resulting *in vivo* reporters were transfected into human embryonic kidney HEK293T cells (ATCC) in three cell culture replicates using Lipofectamine 2000 (Invitrogen) in a six-well plate. RNA was extracted 24 h after transfection using TRIzol (Thermo Fisher) and DNase treated. Random 9-mers were used to generate cDNA with SuperScript III Reverse Transcriptase (Invitrogen) followed by PCR (GoTaq, Promega). All PCR reactions were kept to the lowest possible number of cycles (15–20 cycles). Input reporters and spliced species were sequenced on an Illumina HiSeq 2500 (100-bp paired-end reads). Cultured cells were periodically tested for mycoplasma contamination.

MaPSy *in vitro* assays. The *in vitro* splicing reporter has a design similar to that of the *in vivo* reporters, but it lacks the *ACTN1* exon and a T7 promoter was used (**Supplementary Fig. 12**). *In vitro* reporters were obtained via transcription *in vitro* using T7 RNA polymerase (Stratagene) and internally labeled with [α-^{32}P]UTP (PerkinElmer) and were capped with G(5)ppp(5)G (New England BioLabs). The resulting RNA was gel purified and used for splicing reactions in 40% HeLa-S3 (NCCC) nuclear extracts containing 40% HeLa-S3 nuclear extract for 80 min at 30 °C (the salt conditions for splicing reactions have been previously described)[37]. Pools of input and spliced RNA were converted to cDNA (SuperScript III, Invitrogen) and used to generate an Illumina library (NEBNext kit, New England BioLabs) for deep sequencing. For glycerol-gradient fractionation, 120 µl of the splicing reaction was treated with 0.5 mg/ml heparin for 5 min at 30 °C and then loaded onto 3.75 ml of a 10–30% glycerol gradient and centrifuged at 175,000*g* using a SW55 rotor (Beckman Coulter) at 4 °C for 2.5 h. After centrifugation, the gradient was fractionated from top to bottom in 16 equal volumes, and the fractions were analyzed on 2.1% native agarose (UltraPure Low-Melting-Point Agarose, Invitrogen) or 8% denaturing polyacrylamide (29:1 cross-linking) gels. The *in vitro* MaPSy assays were done in two experimental replicates. Gels were visualized with a Typhoon PhosphorImager (GE Healthcare). Unspliced RNAs that were bound to different complexes were extracted from relevant fractions, converted to cDNA (SuperScript III, Invitrogen), reattached to the T7 promoter sequence by PCR, gel purified and used as template for subsequent *in vitro* transcription to make pre-mRNA substrates for the next round of SELEX (**Supplementary Fig. 9a**). RNA pools recovered from each purification step were converted to cDNA, PCR amplified and analyzed by deep sequencing (Illumina HiSeq 2500, 100-bp paired-end reads).

Library species alignment and counting. We generated 'reference genomes' for both the *in vivo* and *in vitro* libraries with each pair of wild-type (reference) and mutant species treated as its own 'chromosome'. Paired-end reads were mapped using the STAR aligner[38]. For input alignment, we did not allow for split reads and only uniquely mapped reads with a maximum of ten mismatches were allowed. We used the same settings for output alignment as we did for input alignment, with the exception that we allowed for split reads. Because there may be more than one mutation per exon in the 5K panel, the requirement for calling a species as wild type can be more stringent than the requirement for calling each of the mutants, given that calling the wild-type species would require the read pair to span all mutation positions in the same exon, whereas calling the mutant species would only require the read pair to span the respective mutant position. Thus, we also required all mapped reads to span all mutation positions in order to ensure balanced detection of wild-type and mutant species.

Allelic imbalance analyses. The allelic ratios for MaPSy analyses were calculated as

$$\log_2\left(\frac{m_o / m_i}{w_o / w_i}\right)$$

where m_o is the count of mutant spliced species, m_i is the count of mutant input, w_o is the count of wild-type spliced species and w_i is the count of wild-type input. To assess statistical significance, a two-sided Fisher's exact test was used and the resulting P values were adjusted to account for multiple comparisons using the p.adjust function in R (method = 'fdr'). A significance level of <0.05 and an allelic ratio of ≥1.5-fold were used to call ESMs.

Splicing efficiency analyses. To compare splicing performance between individual species, the following splicing index was calculated for each species

$$\log_2\left(\frac{\mathrm{spl}_i / \sum_{i=1}^{n} \mathrm{spl}}{\mathrm{inp}_i / \sum_{i=1}^{n} \mathrm{inp}}\right)$$

where spl_i is the count for spliced output for species i, inp_i is the count for the input for species i and n is the number of species in the library pool.

MaPSy validation in patient samples. Tissue samples (n = 13) were obtained from the University of Utah School of Medicine (Salt Lake City, UT), the Washington University School of Medicine Alzheimer's Disease Research Center (St. Louis, MO), Ohio State University (Columbus, OH), the National Institute of Child Health and Human Development (Bethesda, MD) and the Coriell Repository. Ethical approvals were granted by local institutional review boards, and informed consent was obtained from all participants. RNA was extracted using the PAXgene kit (Qiagen) for whole-blood samples, the RNAeasy kit (Qiagen) for postmortem brain samples and TRIzol (Life Technologies) for all other samples, using the respective manufacturer's protocols. SuperScript III Reverse Transcriptase (Invitrogen) was used to generate cDNA with random 9-mers, followed by PCR (GoTaq, Promega). PCR primers were designed to map to exons flanking the mutant exon. In the case of individuals with nonsense mutations for whom we had lymphoblastoid cell lines or fibroblasts available, the cells were also treated with 10 µg/ml cycloheximide for 3 h before RNA extraction.

MaPSy validation in ENCODE data. We downloaded 46 whole-cell RNA–seq long poly(A)$^+$ ENCODE data sets for 19 different cell lines (for accession numbers, see **Supplementary Table 4**). Reads were mapped to hg19 using the STAR[38] aligner with default parameters. Each STAR output generates a splice-junction file, which was used to calculate percentage usage at each splice junction as follows.

$$\%\,\text{usage at } 3'\text{ss} = \left(\frac{\#\,3'\text{ss reads}}{\#\,\text{upstream } 5'\text{ss reads}}\right) * 100\%$$

$$\%\,\text{usage at } 5'\text{ss} = \left(\frac{\#\,5'\text{ss reads}}{\#\,\text{downstream } 3'\text{ss reads}}\right) * 100\%$$

Results from multiple runs of the same cell lines were collapsed. The hg19 positions of the 3′ splice sites (ss), 5′ splice sites, upstream 5′ splice sites and downstream 3′ splice sites for all wild-type exons in the 5K panel were

doi:10.1038/ng.3837

retrieved and were binned into four groups of increasing splicing performance in MaPSy. Average percentage usage at both splice sites was plotted in each bin and compared.

HGMD mutation analyses. Disease-causing splicing and coding-sequence mutations were selected from HGMD (n = 77,943). The mutations were classified as splicing, missense or nonsense mutations, and the numbers of all classes of mutation were determined for each gene. The total number of mutations was plotted against the total number of SSMs in a gene (**Fig. 2a**). Weighted random sampling was then used to construct a 99.9% confidence interval that capitulates the expected number of SSMs given the total number of mutations within a gene. Using the proportion of total SSMs to total mutations in the HGMD as a weight for random sampling, the proportion of SSMs given the total number of mutations in each gene was simulated 1,000 times. Genes falling outside the simulated values represent genes that have more (above the confidence interval) or fewer (below the confidence interval) SSMs than expected ($P < 0.01$) based on the distribution of mutation types within the data set. Haploinsufficiency scores were obtained from published data[10]. HGMD genes were binned as haploinsufficient genes (haploinsufficiency (HI) score = 1), moderately haploinsufficient genes (HI score = 0.7–1) and haplosufficient genes (HI score ≤0.7).

Random forest classification. We used R implementation of random forest[15], a nonparametric ensemble learning method, to model the contribution of various genomic, sequence and functional features to the likelihood that an exonic mutation will have an impact on splicing. Each tree in the forest is constructed with a different bootstrap sample from the original data set, with approximately two-thirds of the bootstrap samples being used for construction of the kth tree and the remaining one-third (out-of-bag data) used for cross-validation. The results from all trees are then averaged to provide unbiased estimates of predicted values, error rates and measures of variable importance. Default parameters were used to build the random forest model, with the exception that the number of trees was specified as 1,000. As variable importance measures may vary depending on the parameters of the algorithm, and both the degree of correlation and the scale of the variables can influence them, we opted to use two different methods for feature selection and measures of importance. The first method created shuffled copies of all the features (shadow features) and trained a random forest classifier using the supplementary set while iteratively removing irrelevant features (those with z scores less than the maximum z score of the respective shadow features). This was done until all features were either confirmed or rejected, using the Boruta[39] package in R. For the second method, we generated the null distribution of the variable importance measures by permuting the response variable so that the relationship between the response and predictor variables was destroyed. This was done with 1,000 runs of random forest, and the empirical P values for importance measures were calculated by counting the number of occurrences in which each importance measure in the original data was either lower or equal to the respective importance measure in the permuted data. Features that are selected in both methods with significance level <0.05 were used for the final random forest model.

Random forest predictor variables. Splice-site strength was computed using Perl scripts downloaded from the MaxEntScan[35] package, which uses a maximum-entropy approach on large data sets of splice sites in humans while taking into account both adjacent and nonadjacent dependencies. The splice-site models assign log-odds ratios to 9-bp sequences (−3 to +6 positions) for the 5 splice-site scores and 23-bp sequences (−20 to +3 positions) for the 3 splice-site scores. 'SS vars' is the sum of the differences in wild type–mutant splice-site scores for all SSMs in the HGMD[8] and ExAC[6] datasets at each exon. ESEs and ESSs were downloaded from published data[11,16,40]. 'ESRseq diff' was computed as the wild type–mutant difference in hexamer splicing scores[11]. Haploinsufficiency scores were obtained from a previous study that developed a haploinsufficiency prediction model using a large deletion data set (Wellcome Trust Consortium Controls)[10]. PPT scores were computed as previously described[41]. 'Exon POS in gene' was calculated as exon number divided by the total number of exons in the gene (values between 0 and 1). The free energy estimate (ΔG) was computed using ViennaRNA package[31] version 1.8.5, using default settings with the --d2 and --noLP options.

Motif analyses. RBP, ESE and ESS motifs were obtained from published sources[11,21]. ESE and ESS hexamers were mapped and counted in each of the mutant and wild-type exons from the 5K panel. The contribution of known splicing elements to MaPSy splicing was evaluated by plotting the mutant–wild type difference in ESE and ESS counts against the mutant/wild type splicing ratio in sliding windows (size = 1,000, step = 1). RBP motifs were mapped to the exons and upstream introns of the 5K panel using the matchPWM function from the Bioconductor package[42] with default settings (minimum score = 0.8). We computed the maximum matchPWM score percentiles of all spanning n-mers at the mutation positions that overlapped the exonic motif maps and calculated the mutant–wild type difference for each mutation position (n = length of motif). The *in vitro* and *in vivo* splicing profiles of exonic motifs were generated by plotting the mean of the maximum score differences in rolling windows of increasing mutant allele inclusion of spliced species (i.e., mutant/wild type ratio, window size = 1,000, step = 1). Intronic motif maps of wild-type species (n = 2,086) were used to calculate intronic motif density for each RBP (**Supplementary Fig. 8a**). Wild-type splicing profiles of intronic motifs were generated by plotting the mean motif density in rolling windows of increasing splicing efficiency (window size = 200, step = 1). *In vitro* and *in vivo* profiles were combined and fitted using the smooth.spline function in R[43]. The Bayesian information criterion was used to determine the optimal number of clusters with the mclust function from the mclust R package[44]. Profiles were clustered on the basis of the coefficient values from spline fitting using the hclust function in R (**Fig. 4c** and **Supplementary Fig. 8b**).

RBP-binding motif validation. We ordered small interfering RNA (siRNA) for human *PTBP1* from Thermo Scientific (s11436) and siRNA for human *SRSF1* from Dharmacon as previously described[23]. For control siRNA, AllStar negative-control siRNA (Qiagen) was used. Minigenes were synthesized by Synbio Technologies. HeLa cells (ATCC) were plated 24 h before transfection. For *PTBP1* knockdown, 7.5 µl of Lipofectamine RNAiMAX (Invitrogen) was used to transfect siRNA for *PTBP1* (20 nM, final concentration) in a six-well plate for 48 h according to the manufacturer's protocol (Invitrogen). This was followed by a second transfection with 3.75 µl of Lipofectamine 3000 (Invitrogen) and the same siRNA in Opti-MEM (Life Technologies) and 500 ng of DNA in 100 µl of pure DMEM (Invitrogen). RNA was extracted 24 h later with TRIzol according to the manufacturer's protocol (Ambion), followed by DNase treatment and RT–PCR as described above. For *SRSF1* knockdown, 1.5 µl of Lipofectamine 3000 (Invitrogen) was used to transfect siRNA for *SRSF1* (20 nM final concentration) in Opti-MEM (Life Technologies) and 500 ng of DNA in 100 µl of pure DMEM (Invitrogen). After 72 h, RNA was isolated, followed by DNase treatment and RT–PCR. Knockdown efficiencies were evaluated with immunoblotting using anti-SRSF1 (sc-33652, Santa Cruz), anti-PTBP1 (32-4800, Thermo Fisher) and anti-GAPDH (sc-47724 and FL-335, Santa Cruz). All experiments were done in two cell culture replicates that had been periodically tested for mycoplasma contamination.

Functional SELEX analysis. The allele ratios were calculated as follows

$$\log_2\left(\frac{mi_e/mi_i}{mj_e/mj_i}\right)$$

where mi_e is the minor allele count in the enriched pool, mi_i is the minor allele count in input, mj_e is the major allele count in the enriched pool and mj_i is the major allele count in input. The minor allele was the allele that spliced less efficiently than the respective major allele; these alleles differed by one nucleotide. All analyses were performed in R. Hierarchical clustering was performed on all mutant–wild type pairs that were recovered in all purified fractions (n = 4,873) using the hclust function with the complete linkage method and Euclidean distances. Bayesian information criterion plots were generated for $k = 1$ to $k = 50$ using the mclust package to estimate the optimal number of clusters. The resulting clusters were visualized, and the tree was cut using the cutree function ($k = 32$). To determine the significance of the observation that mutations in the same exons were more often clustered together, we permuted the exon assignment in the 32 clusters 10,000 times and obtained the χ^2 distribution of the permuted data. The P value was obtained by counting the number of times

the computations of the permuted data exceeded or equaled that of the original data divided by the number of permutations. To examine whether certain genomic features may act as 'signatures' of the identified clusters, we plotted the distribution of each feature in the different clusters, and significance was determined by the mean difference in two-sided *t*-statistics on the actual data and permuted data 10,000 times using the flip function followed by flip.adjust (method = 'fdr') to account for multiple testing[45].

Data availability. The data generated from this study (raw allelic counts and allelic ratios from each mutant–wild type pair from MaPSy experiments with the corresponding genomic positions, variant allele and HGMD accession numbers) are available at http://fairbrother.biomed.brown.edu/ESM_browser/.

36. Gozani, O., Patton, J.G. & Reed, R. A novel set of spliceosome-associated proteins and the essential splicing factor PSF bind stably to pre-mRNA prior to catalytic step II of the splicing reaction. *EMBO J.* **13**, 3356–3367 (1994).
37. Reichert, V. & Moore, M.J. Better conditions for mammalian *in vitro* splicing provided by acetate and glutamate as potassium counterions. *Nucleic Acids Res.* **28**, 416–423 (2000).
38. Dobin, A. *et al.* STAR: ultrafast universal RNA–seq aligner. *Bioinformatics* **29**, 15–21 (2013).
39. Kursa, M.B., Jankowski, A. & Rudnicki, W.R. Boruta—a system for feature selection. *Fundam. Inform.* **101**, 271–285 (2010).
40. Fairbrother, W.G. *et al.* RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* **32**, W187–W190 (2004).
41. Lin, C.L. *et al.* RNA structure replaces the need for U2AF2 in splicing. *Genome Res.* **26**, 12–23 (2016).
42. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
43. Chambers, J.M. & Hastie, T. *Statistical Models in S* (Wadsworth & Brooks/Cole Advanced Books & Software, 1992).
44. Fraley, C. & Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97**, 611–631 (2002).
45. Pesarin, F. *Multivariate Permutation Tests: With Applications in Biostatistics* (J. Wiley, 2001).