HMG Advance Access published May 2, 2013

Functional implications of splicing polymorphisms in the human genome

Yerbol Z. Kurmangaliyev[1], Roman A. Sutormin[2], Sergey A. Naumenko[1,2], Georgii A. Bazykin[1,2], Mikhail S. Gelfand[1,2,*]

[1]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, 127994, Russia

[2]Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, 119992, Russia

[*]To whom correspondence should be addressed. Institute for Information Transmission Problems (Kharkevich Institute) RAS, 127994, GSP-4, Russia, Moscow, Bol'shoi Karetnyi per., 19. Tel: +7 495 650 42 25; Fax: +7 495 650 05 79; Email: gelfand@iitp.ru

Abstract

Proper splicing is often crucial for gene functioning and its disruption may be strongly deleterious. Nevertheless, even the essential for splicing canonical dinucleotides of the splice sites are often polymorphic. Here, we use data from The 1000 Genomes Project to study SNPs in the canonical dinucleotides. Splice sites carrying SNPs are enriched in weakly expressed genes and in rarely-used alternative splice sites. Genes with disrupted splice sites tend to have low selective constraint, and the splice sites disrupted by SNPs are less likely to be conserved in mouse. Furthermore, SNPs are enriched in splice sites whose effects on gene function are minor: splice sites located outside of protein-coding regions, in shorter exons, closer to the 3'-ends of proteins, and outside of functional protein domains. Most of these effects are more pronounced for high-frequency SNPs. Despite these trends, many of the polymorphic sites may still substantially affect the function of the corresponding genes. A number of the observed splice site-disrupting SNPs, including several high-frequency ones, were found among mutations described in OMIM.

Introduction

Alternative splicing is a major mechanism for expanding transcript diversity in higher eukaryotes [1]. It is a complex process which is regulated by the interaction of spliceosomal components with multiple trans-acting factors, and involves a variety of cis-regulatory splice sites. The latter include the main splicing signals (donor and acceptor sites, polypyrimidine tract, branch point) and auxiliary elements (enhancers and silencers of splicing) [2]. Mutations at these cis-regulatory elements could lead to serious alterations in splicing patterns. According to published estimates, up to 50% of human disease-causing mutations may affect the splicing of genes [3].

However, not all mutations that affect splicing patterns will necessarily be deleterious. Several studies have revealed the existence of widespread natural variation in splicing patterns within the human population [4-7]. They identified many single-nucleotide polymorphisms (SNP) that are correlated with changes in splicing patterns (both qualitatively and quantitatively). Such polymorphisms are usually located in close proximity to the affected exon-intron boundaries [5, 8, 9]. Most likely, these mutations either affect the existing splice sites or other cis-regulatory elements of splicing, or create novel ones [8, 9, 10, 11].

Usually, however, it is difficult to find a causative relationship between a particular mutation and splicing alterations. Existing methods for the prediction of splice sites and the analysis of effects of mutations are not sufficiently precise [11, 12], and many of the mutations computationally predicted to disrupt splicing turn out to be harmless. For instance, the analysis of SNPs overlapping donor splice sites demonstrated that only a small fraction of mutations occurring outside of the canonical dinucleotide GT affected the pattern of splicing [11]. Moreover, the observed correlations between SNPs and splicing patterns could be caused by linkage disequilibrium with causative alleles.

On the other hand, almost all splice sites in eukaryotes (>99%) contain the canonical conserved dinucleotides (GT and AG in the donor and acceptor sites, respectively). Mutations at these positions almost certainly abolish splicing at these sites [11]. The majority of disease-causing mutations of splice sites occur at these dinucleotides [13-15]. Traditionally, splice site disruptions are treated as loss-of-function variants, along with premature stops and frameshifts [16], although accurate analysis has shown that not all of them lead to a complete loss of function [17]. Previously, Shimada et al. conducted an analysis of SNPs from the dbSNP database that overlapped the canonical dinucleotides in human splice sites [18], and found 212 SNPs disrupting the latter. The only observed difference between the polymorphic and invariant splice sites was that SNPs usually disrupt splice sites of young exons not conserved in mouse [18].

Recently, large amounts of new data on human genetic variation became available. The 1000 Genomes Project Consortium identified the genotypes of more than 1000 individuals from a number of ethnic groups [19], and uncovered more than 30 million SNPs, most of them previously unknown. Here, we use these data to evaluate the functional impact of splice site-disrupting SNPs on gene function. We show that the splice sites carrying SNPs differ from the invariant splice sites in many parameters, and that these differences are more pronounced for variants with a high frequency of the noncanonical nucleotide.

Results

Sets of splice sites with SNPs at conserved dinucleotides

We compiled a set of polymorphisms disrupting the canonical dinucleotides at splice sites by mapping SNPs from the 1000 Genomes Project onto the Gencode annotation of human transcripts. We considered only cases where one of the alleles carried a proper canonical dinucleotide (GT and AG in donor and acceptor splice sites, respectively). In total, 2259 SNPs overlapping the canonical dinucleotides of 2249 splice sites were identified (in 10 splice sites, both positions were polymorphic).

Not all of these SNPs are necessarily mutations in pre-existing splice sites, as some of them could be mutations that create novel splice sites. To distinguish between the two possibilities, we checked for the existence of orthologous splice sites in genomes of phylogenetically close primates. 2090 of the 2249 (92.9%) polymorphic splice sites carried the canonical dinucleotide in at least one of the four primate genomes considered (*Pan troglodytes, Gorilla gorilla, Pongo abelii, Macaca mulatta*), and we assumed that these SNPs disrupted ancestral splice sites (Figure 1). The remaining 159 splice sites were probably newly acquired in the human lineage, and SNPs observed in them could represent mutations creating *de novo* splice sites (high values of non-functional allele frequencies (NAF), Figure 1) or reverse mutations of newly generated splice sites (low values of NAF, Figure 1). Since the number of such cases was too low for a meaningful statistical analysis, and we could not formally distinguish between the above two scenarios, only SNPs disrupting ancestral splice sites were used further.

We assigned splice site disruption events to two categories according to their NAFs, and separately considered low-frequency mutations (**lfSNP**, NAF<0.001) and high-frequency mutations (**hfSNP**, NAF>0.001). Correspondingly, splice sites were classified as lfSNP-sites (1109) or hfSNP-sites (981). For the 10 splice sites with SNPs at both canonical positions, frequencies of the non-functional alleles were summed. Allelic frequencies were calculated among 1094 diploid genotypes; therefore, the lfSNP-sites corresponded to cases when the non-functional allele was observed in only one or two chromosomes. The functional allele was more common (0.001<NAF<0.5) in 948 hfSNP-sites, while it was the minor allele (NAF>0.5) in 33 of them. In several hfSNP-sites, the functional allele was very rare (seven cases had NAF>0.95), probably corresponding to an almost complete loss of the ancestral splice site. Supplementary Table 1 provides the data on each polymorphic splice site, including the NAF values for each of the 14 human populations used in the 1000 Genome Project Interim Phase 1 [19]; Supplementary Table 2 provides the data on all individual genotypes.

All remaining splice sites inferred from the annotation of human transcripts, i.e. those in which no SNPs were observed in the canonical dinucleotide positions, comprised the control set (**noSNP**-sites). A small fraction of these sites carried non-canonical dinucleotides. These splice sites could still be functional: in some cases, splicing can be performed with the rare non-canonical dinucleotide variants. For instance, non-canonical GC-sites comprise about 1% of all human donor splice sites [20]. However, a fraction of the splice sites with non-canonical dinucleotides could also represent as-yet-unidentified SNPs if the individual whose transcriptomic data was used carried the canonical dinucleotide at this site. Since we could not distinguish between these two scenarios, we excluded the noSNP-sites with non-canonical dinucleotides from further analysis.

The final datasets are described in Table 1. Splice site disruptions more frequently occur in donor sites than in acceptor sites (the chi-squared test, p-value $<10^{-15}$, Table 2). A similar excess of splice site-disrupting mutations among donor sites, compared to the acceptor sites, is also observed among the disease-causing mutations [13], and is similar between the hfSNPs and lfSNPs, implying that this phenomenon has a mutational, rather than selectional, basis. Indeed, when GT and AG dinucleotides in intronic or intergenic regions were considered, an excess of mutations in GT dinucleotides was also observed (data not shown).

SNPs usually disrupt rarely-used splice sites

Splice site disruption can have little effect on gene function if it can be easily compensated for by the presence of a different functional isoform. As expected, alternative splice sites were more frequently disrupted by SNPs than the constitutively spliced ones (the chi-squared test, p-value $<10^{-15}$, Table 2). Overall, 0.26% and 0.19% of constitutive splice sites, and 0.35% and 0.41% of alternative splice sites, were disrupted by lfSNPs and hfSNPs, respectively. Hence, the distribution of derived allelic frequencies at alternative splice sites is shifted towards higher frequencies, compared to constitutive sites. Negative selection is expected to reduce the prevalence of hfSNPs to a larger extent than that of lfSNPs. Therefore, the larger difference between the fractions of constitutive and alternative splice sites disrupted by hfSNPs (0.19% vs. 0.41%), compared with that for lfSNPs (0.26% vs. 0.35%), can be due to stronger effect of selection on hfSNPs. However, it can also arise if a fraction of splicing events that are annotated as alternative actually represent allele-specific splicing. The probability of such pseudoalternatives is higher in the case of hfSNP-sites, because in such sites, different alleles are more likely to have contributed to the annotation. Since the properties of constitutive and alternative splice sites and gene segments can be very different, further analysis was performed for these two classes separately (Table 1).

The efficiency of the use of a splice site is correlated with its score, i.e., its similarity to the consensus sequence. We calculated splice site scores using the positional weight matrix as described in Data and Methods. In the case of polymorphic splice sites, scores were calculated for the functional alleles (i.e., those carrying the canonical dinucleotides). Polymorphic splice sites were, on average, weaker than the splice sites from the control set (Figure 2, Table 2).

A low score of a splice site suggests lower efficiency and frequency of its use. We measured frequencies of use of splice sites directly from transcript expression, using expression as a proxy for functional importance. To account for possible differences in expression between cell lines and for the possibility of allele-specific splicing, we used data from 14 different RNA-Seq experiments carried out on various types of cell lines (ENCODE). We considered two characteristics of the frequency of splice site usage: expression and inclusion level of a splice site. The expression level of a splice site is the total expression level of all transcripts spliced at this site, while the inclusion level is the expression level of this site divided by the total expression level of all transcripts spanning this genomic area (see Materials and Methods for details). Note that, even for constitutive sites (which by definition have inclusion level of 100%), the expression level of a site may be substantially lower than the expression level of the entire gene: for example, a constitutive site is not expressed if the terminal exon flanked by this site is not expressed due to alternative transcription initiation and/or termination.

Splice sites disrupted by SNPs had extremely low expression and inclusion levels, and this tendency was stronger in the case of hfSNPs (Figure 3). Both of these differences were statistically significant (the Kruskal-Wallis test, p-values $<10^{-15}$, Table 2). The number of fragments per kilobase per million mapped reads (FPKM) of alternative hfSNP splice sites (3.34) was an order of magnitude lower than the FPKM of the entire corresponding genes (22.37), indicating that the affected splice sites were often located in minor splicing isoforms. Furthermore, the FPKM of constitutive hfSNP splice sites (3.38) was similarly low, indicating that the affected splice sites were often located in alternatively used terminal exons. This is in line with their enrichment in UTRs (see below).

Many human exons were derived by exonization of transposable elements (particularly from Alu-elements [21, 22]). Ambiguous mapping of short reads on repetitive elements may complicate accurate quantification of expression level of such exons from RNA-Seq data. Moreover, Alu-derived exons are often alternatively spliced and have low inclusion levels [21]. To account for these effects, we searched for repetitive elements overlapping the splice sites in our datasets. In total, 57 (5%) hfSNP- , 40 (4%) lfSNP- and 4900 (1.3%) noSNP-sites overlapped some element annotated in RepeatMasker track of the UCSC genome browser [23, 24]. Among them, 14 hfSNP-, 8 lfSNP- and 1319 noSNP-sites overlapped Alu-elements. The repetitive elements that overlapped splice sites are listed in Supplementary Table 1. As expected, splice sites within repetitive elements had significantly lower expression levels (with median FPKM values: 3.15 for noSNP-, 2.39 for lfSNP- and 0.73 for hfSNP-sites). Nevertheless, excluding from analysis splice sites that overlapped repetitive elements had very little effect on the values in Table 2 (data not shown).

Weakly expressed gene segments with a low level of inclusion are known to be less conserved [25, 26]. A previous study on a small dataset demonstrated that polymorphic splice sites more frequently flank exons with low interspecies conservation [18]. This result is reproduced in our dataset: canonical dinucleotides of polymorphic splice sites were less conserved in the mouse genome than those of noSNP-sites, and this tendency was much stronger in hfSNP-sites (the chi-squared test, p-values $<10^{-15}$, Figure 4A, Table 2).

Similar patterns can be observed on the level of entire genes. Consensus dinucleotides tend to be polymorphic more often in genes with lower expression levels (Figure 3C). Genes carrying polymorphic splice sites also tend to be less constrained on the level of amino acid sequence, which can be seen from higher dN/dS values in these genes (Figure 4B).

Disruptions of polymorphic splice sites have weak effect on the protein structure

We analyzed the distribution of splice sites between the coding (CDS) and untranslated regions (UTR) of genes. As expected, disruptions of splice sites were more frequent in UTRs (the chi-squared test, p-values $<10^{-15}$, Figure 5A, Table 2). Still, a considerable number of mutated splice sites were located in CDS regions. Those SNPs have probably influenced the structures of the encoded proteins. However, not all changes in protein sequence necessarily result in serious structural alterations and/or loss of function.

To investigate the possible influence that polymorphisms which disrupt splice sites may have on the encoded proteins, we tried to estimate the extent to which the encoded amino acid sequence could be affected by these mutations (Table 1). Firstly, we compared the lengths of coding exons spliced at the considered splice sites. For partially coding exons, we used only the length of the CDS region. Exons with flanking polymorphic splice sites were on average shorter than exons flanked by noSNP-sites (Figure 5B). This difference was weak but statistically significant (Table 2) and was stronger for alternatively spliced sites (the Kruskal-Wallis test, p-values $<10^{-3}$ and $<10^{-4}$ for constitutive and alternative sites, respectively).

If a mutation at a splice site causes a frameshift, the downstream part of the protein is totally disrupted. Exon skipping is the most common type of alternative splicing in human [27]; therefore, frameshifts should be mainly caused by mutations of those splice sites that flank exons whose lengths are not multiples of three. 59% of noSNP-sites, 57% of lfSNP-sites, and 57% of hfSNP-sites were located at flanks of exons with lengths that are not multiples of three, but the differences between the three classes were not significant.

Assuming that mutations at splice sites that flank exons whose lengths are not multiples of three lead to frameshifts (which would be the case if the exon is skipped, but see the Discussion), and that the effect of a frameshift depends on the proximity of the frameshift site to the C-end of the encoded protein, we calculated the relative positions of splice sites flanking exons whose lengths are not multiples of three (Table 1). On average, lfSNP- and hfSNP-sites flanking such exons were located closer to the C-end of the proteins than the noSNP-sites. These differences were also weak but, in the case of alternative sites, statistically significant (the Kruskal-Wallis test, p-value = 0.003, Figure 5C, Table 2). No difference was observed for exons whose lengths are multiples of three.

To assess the effect of splice site disruptions on the protein structure, we asked how often polymorphic splice junctions lie within Pfam domains. Disruption of such splice sites would probably result in partial or complete disruption of functional domains, likely leading to the loss of function. Among the constitutive splice sites, hfSNP-sites overlapped Pfam domains much less frequently than either noSNP- or lfSNP-sites (the chi-squared test, p-value = 0.004; Figure 5D, Table 2), while no difference was observed between lfSNP- and noSNP-sites (Table 2). Among alternative splice sites, both lfSNP- and hfSNP-sites overlapped Pfam domains less frequently than noSNP-sites (the chi-squared test, p-value $<10^{-5}$, Figure 5D, Table 2). Pfam domains for all coding polymorphic splice sites are listed in Supplementary Table 1.

Some mutations disrupting splice sites are described in the OMIM database

Although most polymorphic splice sites seem to have a low functional load, many of such polymorphisms could still be deleterious. We conducted a semi-automatic search of splice site-disrupting SNPs among known mutations of genes from the OMIM Morbid Map (Online Mendelian Inheritance in Man) described in the (OMIM) repository [28], and found 16 mutations from OMIM that were also observed in our datasets (Table 3). Eight of them were hfSNPs, and the remaining eight were lfSNPs. Unexpectedly, some of these mutations were quite common in the human population: four of them had NAF > 1%. The most extreme example was the well-known mutation disrupting the acceptor site of the 7th exon of the OAS1 gene (antiviral enzyme 2,5-oligoadenylate synthetase, MIM:164350, rs10774671), which had NAF of 64%. This SNP affects the level of OAS1 activity and increases susceptibility of individuals to viral infections [29, 30]. Another striking example is the mutation of the acceptor site of the 4th exon of the CYP2D6 gene (hepatic cytochrome P450, MIM:124030, rs3892097) with NAF of 10% [31].

The characteristics of the 16 splice sites carrying the disrupting SNPs found in OMIM are given in Table 3. Most of these splice sites are constitutive (13 of 16), conserved in the mouse genome (14 of 16), and overlap the Pfam domains (11 of 16). In seven cases, the splicing disruption phenotype is known: splice site disruptions lead to cryptic site activation in five cases, and to exon skipping in the other two cases (Table 3).

Overall, 286 out of 1846 (15%) genes containing lfSNP- or hfSNP-sites are present in OMIM Morbid Map; this fraction was comparable to that observed for the control set of genes (2211 out of 15793, 14%). 45 out of 286 genes were marked as associated with susceptibility to complex diseases or with "non-disease” abnormal phenotypic variations. In total, these genes contained 323 polymorphic splice sites, 82 of which were conserved constitutive sites overlapping Pfam domains (Supplementary Table 1). Some of these splice site-disrupting SNPs may represent yet undescribed disease-causing alleles. Additionally we have performed search of splice site-disrupting mutations from our datasets among SNPs collected in the NGHRI GWAS catalog [32, 33]. The only case found in the NGHRI GWAS catalog was a SNP in the splice site of the GREB1 gene (rs13394619, [34]). Apparently, this mutation comprises not a disruption but a de novo creation of a splice site (this site is not present in primates), and was thus excluded from our main workflow.

Discussion

Splicing-disrupting mutations are often considered to result in the loss of function, under the assumption that improper splicing always gives rise to a defunct copy of the protein [16, 35]. However, the effects of splice site-disrupting mutations on the encoded protein may range from negligible to radical. On the genome level, the impairment inflicted by mutations within classes of sites can be assessed from the frequencies of SNPs affecting these sites, or from allelic frequencies of such polymorphisms. Previously, however, lack of whole-genome data on human polymorphism from a large number of individuals precluded such analyses for all but the largest classes of sites.

Here, we used the most extensive database of human polymorphism to date, originating from The 1000 Genomes Project, to characterize splice site-disrupting SNPs. Splice sites are constituted by ~9 functional positions for donor sites and ~15-20 positions for acceptor sites, and mutations in most of these positions have been shown to affect splicing [13-15, 36]. However, predicting the effect of mutations outside of the consensus dinucleotide is a challenge [11, 12]. Conversely, mutations in the consensus dinucleotide nearly always lead to disruption of splicing at this site. Here, by mapping the annotation of human transcripts to the polymorphism data, we identified more than 2000 SNPs overlapping the conserved dinucleotides of splice sites of protein-coding genes, which is ~10 times more than has been studied previously [18].

We find that SNP-carrying splice sites represent a biased subset of all splice sites in many respects. Most of the analyses show that the effects of such disruption on the proteome are weaker than would be expected by chance.

Firstly, the disruptions are overrepresented within alternative splice sites, compared to constitutive ones. A part of this difference may be due to misannotation of allele-specific splicing as bona fide alternative splicing. Indeed, since the reference gene models have been constructed from multiple genotypes, differences in splicing modes, either due to heterozygosity of a sampled individual or to pooling of genotypes from different individuals, may lead to annotation of multiple splicing isoforms if splicing is constitutive but gene models differ between genotypes. In the plant *Arabidopsis thaliana*, the fraction of disrupted genes dropped considerably after differences in gene models between genotypes were accounted for [37].

Still, it is unlikely that allele-specific splicing is the sole cause of the observed differences between constitutive and alternative sites. In particular, it can hardly explain the differences between the lfSNP- and noSNP-sets. As the lfSNPs have frequencies below 0.1%, it is highly unlikely that the minor variant has also contributed to the reference annotation. Moreover, even the hfSNPs have mean allelic frequency of ~1%. Since the average number of annotated isoforms per multi-exonic protein-coding gene in Gencode is only 3.3, most of the SNPs with such frequencies could not have contributed to it. Therefore, it appears safe to assume that the observed excess of SNPs in alternative splice sites is real.

Cases are known when the alternative splicing isoforms have radically different functions [38]. Still, our results show that disruption of an alternative splice site tends to affect the encoded protein less, probably because it can be compensated for by an alternative isoform.

Secondly, splice site-disrupting mutations tend to occur at weak and rarely used sites of weakly expressed genes that experience low selective constraint. The disrupted sites have low scores, suggesting that splicing at them is inefficient. Consistently, the inclusion levels of disrupted alternative sites were much lower than that of noSNP-sites. The disrupted sites also tended to occur within genes with low expression levels, and, among them, in rarely used isoforms.

Conceivably, the rarely used isoforms could be important, but rarely needed. Moreover, the expression data could have biases, as it comes from cultured cell lines, rather than from native tissues. Functions of some genes and splice sites could be restricted to narrow spectra of cell types and conditions, and minor isoforms may play a role in regulation of gene expression. Nevertheless, the high polymorphism levels at the splice sites corresponding to the rarely used isoforms undermine their credibility and suggest that some of such isoforms probably lack any function whatsoever, corresponding to splicing "noise" that is included in annotations [39]. Thus, the diversity of functional isoforms implied by the current gene annotations may be an overestimation. Still, the presence of additional isoforms expressed at low levels may facilitate subsequent adaptation: most new splicing variants originate as low-frequency [40], and minor isoforms experience a high degree of positive selection, suggestive of adaptation [41].

Thirdly, splice site-disrupting polymorphisms tend to affect splice sites that are not conserved between species [18] and to occur in genes that experience low selective constraint. Again, this argues against their functional importance.

Fourthly, the disruptions were enriched in those splice sites where they had little effect on the encoded protein. They were overrepresented in UTRs and correspondingly underrepresented in CDSs. If splice site disruptions lead to exon skipping, then mutations of splice sites located at the boundaries of longer exons, exons of length not a multiple of 3, and of exons located near the beginning of the protein will be more disrupting. Consistently, we observed that, within the CDSs, splice site disruptions tend to affect shorter exons, and, when they are expected to cause frameshifts, tend to occur closer to the protein C-termini, thus affecting only a short part of the protein. (Short exons are also skipped more often [36], probably contributing to the observed lower inclusion level of polymorphic splice sites.) Finally, splice site disruptions avoided structural domains, as revealed by the Pfam data.

The magnitude of the differences between polymorphic and non-polymorphic sites at the protein level is low. In particular, the bias of frameshifting variants towards the 3' protein ends is weaker than that previously observed for frameshifting indels and nonsense variants [17]. The main reason for this discrepancy is probably that not all of the variants that we consider to be frameshifting really are such. The above calculations were performed under the assumption that all splice site-disrupting mutations lead to exon skipping. In humans, however, disruption of splice sites usually leads either to exon skipping or to the use of an alternative/cryptic site leading to exon extension or shortening, and the frequencies of these two types or events are comparable (Table 3; [36]). In case of cryptic site activation, our strategy of only considering as frameshifting the disruptions of splice sites in exons with lengths not in multiples of 3 may fail: what is relevant instead is not the exon length, but the distance between the authentic and the cryptic splice sites. If we knew the exact exonic segments affected by these variants, the effects would probably have been stronger. Moreover, as we have shown, the polymorphic splice sites often belong to the rarely used isoforms, so only a small fraction of transcripts/proteins are affected by these mutations; therefore, the observed protein-level effects are weak.

Unfortunately, predicting the exact outcome of a splice site disruption is notoriously difficult. We have searched for potential cryptic splice sites in proximity to polymorphic splice sites. However, the closest GT/AG dinucleotide was no more likely to be in-frame for polymorphic splice sites than for noSNP-sites. Looking for cryptic splice sites with positional weight matrices also revealed no biases (data not shown). The question of exact molecular effect of splice site disruptions can be readdressed with the arrival of sufficient amounts of coupled genotype/individual-transcriptomics data.

The majority of the observed differences were the strongest between the hfSNP- and noSNP-sites, with the lfSNP-sites having intermediate values. Allelic frequencies are shaped by selection, and more deleterious alleles segregate at lower frequencies [42]. Therefore, these differences suggest that the observed features of polymorphic sites (detailed in Supplementary Table 1) are directly related to the functional effects of the corresponding SNPs.

The observed splice site disruptions could be slightly contaminated by functional donor splice sites with non-consensus GC dinucleotides. However, we believe that this contamination has little or no effect on our analysis, for the following reasons. Firstly, such sites are very rare (~1%) in humans [20]. Secondly, functional GC-donor sites are known to be closer to consensus at other positions [20]. Thus, GT-splice sites mutated to GC-splice sites would likely maintain their functionality only when the remainder of the site is strong. However, polymorphic donor sites from our sets were on average weaker than other sites (Figure 2, Table 2).

In summary, the splice site-disrupting mutations tend to have little effect on gene function. Still, many of the mutations, especially those observed at low frequencies, may in fact lead to the loss of function of a gene. Search in the OMIM database showed that many of the splice site disruptions, which occur at a range of frequencies in healthy individuals, can still be disease-causing (Table 3).

Materials and Methods

Initial sets of splice sites and transcripts

The list of splice sites belonging to transcripts of protein-coding genes was obtained from the GENCODE reference annotation of human transcripts (release 7, [43]) from the UCSC genome browser [24]. Genes from the X, Y and mitochondrial chromosomes were not considered in this study. We also removed common splice sites of overlapping genes.

A splice site was considered constitutive if all transcripts spanning this genomic area were spliced on it, and alternative otherwise. To calculate the attributes of splice sites related to protein structure (location with respect to the CDS, length of exons and its multiplicity by three, relative position in protein, and overlap with Pfam domains), we assigned, for each splice site, one representative transcript spliced at this site. To this end, the transcript with the highest expression level was used, calculated as described below.

SNP data

SNP data of Interim Phase 1 of the 1000 Genomes Project were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/ (June 2011 Data Release, [16, 19]). In total, the genotypes of 1094 individuals contained 37,852,169 autosomal SNPs.

Conservation of splice sites and genes

Multiz whole-genome multiple alignments [44] were downloaded from the UCSC genome browser [24]. Conservation of splice sites was determined between human and 5 other mamalian genomes: chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla gorilla), orangutan (Pongo pygmaeus abelii), rhesus macaque (Macaca mulatta) and mouse (Mus musculus). A splice site was considered conserved in a given species if the appropriate functional dinucleotide was present in its genome.

Selective constraint acting on amino acid sequences was determined by thecalculation of dN/dS between human and mouse genomes. For each gene, we chose the isoform with the highest expression level as representative. Genes not aligned between human and mouse were not considered in this analysis. dN/dS was calculated using a Bioperl [45] wrapper for the codeml program of the PAML ([46]) software package.

Splice site scores

Splice site scores were calculated using positional weight matrices covering positions from −3 to +6 (for donor sites) and from −15 to +2 (for acceptor sites) as in [36]. The positional weight matrices corresponding to splice sites of internal constitutive exons were obtained from [36]. In case of polymorphic splice sites, scores were calculated for the functional alleles.

Expression and inclusion level of splice sites and genes

Expression and inclusion levels of splice sites were measured using data from the Caltech ENCODE RNA-seq track [47]. Fourteen sets of alignments of paired-end reads (2Ч75, insertion length 200) obtained from fourteen different cell lines were downloaded via the UCSC Genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/) [24]. Detailed descriptions of the datasets are provided in Supplementary Table 3. Abundance of transcripts was measured in Fragments Per Kilobase per Million mapped reads (FPKM, [48]) using Cufflinks (version 1.3.0).

For each splice site, we measured two parameters: expression level and inclusion level. Expression level of a splice site was calculated as the total expression level of transcripts spliced on this site, summed across all cell lines. Inclusion level of an (alternative) splice site is the ratio of the total expression level of transcripts spliced on this site, to the total expression level of all transcripts spanning this genomic area. Inclusion level of constitutive splice sites is defined as 100%. Still, the expression level of constitutive splice sites may be lower than the overall gene expression due to alternative transcription starts and ends. Splice sites not expressed in any of the analysed cell types (with total FPKM=0) were excluded from the analyses that used the expression levels.

The expression level of a gene was defined as the total expression level of all its transcripts summed across all cell lines. Genes with null expression were excluded from analysis.

Search for Pfam domains

For each splice site, we used the PfamScan tool (version 1.3) [49] to search for functional domains in its representative transcripts. Search was restricted to Pfam-A entries including all types of functional regions (families, domains and repeats).

Search for mutations with known functional consequences

We performed a semi-automatic search of splice site-disrupting SNPs among the known mutations collected in the Online Mendelian Inheritance in Man (OMIM) database (12.04.2012, [28]). To this end, we extracted all intronic mutations described in OMIM entries corresponding to the genes listed in the OMIM Morbid Map with at least one polymorphic splice site, and searched among them for mutations corresponding to splice site-disrupting SNPs from our datasets. The original papers describing these mutations were used for exact verification of matched mutations.

Additionally we have performed a search of splice site-disrupting mutations from our datasets among SNPs collected in NGHRI GWAS catalog (15.01.2013, [32, 33]).

Statistical analysis

The statistical significance of observed differences between our sets of splice sites was calculated using the chi-squared test for categorical parameters, and the Kruskal-Wallis and Mann-Whitney tests for quantitative parameters. The main text uses results from the Kruskal-Wallis analysis; the results of the pairwise comparison of datasets by the Mann-Whitney test are provided in Supplementary Table 4. All of these tests were implemented in R [50]. Plots were created by the ggplot2 R package [51].

Acknowledgments

We are grateful to Andrei A. Mironov, Konstantin Y. Popadin, Alexander V. Favorov and Ekaterina O. Ermakova for useful discussions, and to Elena Nabieva for language editing. This work was partially supported by the Ministry of Education and Science of the Russian Federation (contracts 11.G34.31.0008, 14.512.11.0042, 8049 and 8814), the Russian Foundation for Basic Research (grant 12-04-33202), and the Russian Academy of Sciences (program “Molecular and Cellular Biology”).

Conflict of interest statement

We declare no conflicts of interest.

References

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

2 Wang, Z. and Burge, C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802-813.

3 Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749-761.

4 Nembaware,V., Wolfe,K.H., Bettoni,F., Kelso,J. and Seoighe,C. (2004) Allele-specific transcript isoforms in human. *FEBS Lett.*, **577**, 233-238.

5 Hull,J., Campino,S., Rowlands,K., Chan,M.S., Copley,R.R., Taylor,M.S., Rockett,K., Elvidge,G., Keating,B., Knight,J. and Kwiatkowski, D. (2007) Identification of common genetic variation that modulates alternative splicing. *PLoS Genet.*, **3**, e99.

6 Kwan,T., Benovoy,D., Dias,C., Gurd,S., Provencher,C., Beaulieu,P., Hudson,T.J., Sladek,R. and Majewski,J. (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.*, **40**, 225-231.

7 Graveley,B.R. (2008) The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. *Trends Genet.*, **24**, 5-7.

8 Coulombe-Huntington,J., Lam,K.C., Dias,C. and Majewski,J. (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet.*, **5**, e1000766.

9 Woolfe,A., Mullikin,J.C. and Elnitski,L. (2010) Genomic features defining exonic variants that modulate splicing. *Genome Biol.*, **11**, R20.

10 de Souza,J.E., Ramalho,R.F., Galante,P.A., Meyer,D. and de Souza,S.J. (2011) Alternative splicing and genetic diversity: silencers are more frequently modified by SNVs associated with alternative exon/intron borders. *Nucleic Acids Res.*, **39**, 4942-4948.

11 Lu,Z.X., Jiang,P. and Xing, Y. (2012) Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdiscip. Rev. RNA*, **3**, 581-592.

12 ElSharawy,A., Hundrieser,B., Brosch,M., Wittig,M., Huse,K., Platzer,M., Becker,A., Simon,M., Rosenstiel,P., Schreiber,S. et al. (2009) Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum. Mutat.*, **30**, 625-632.

13 Krawczak,M., Thomas,N.S., Hundrieser,B., Mort,M., Wittig,M., Hampe,J. and Cooper,D.N. (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.*, **28**, 150-158.

14 Vorechovskэ,I. (2006) Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **34**, 4630-4641.

15 Buratti,E., Chivers,M., Krбlovicovб,J., Romano,M., Baralle,M., Krainer,A.R. and Vorechovsky,I. (2007) Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Res.*, **35**, 4250-4263.

16 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.

17 MacArthur,D.G., Balasubramanian,S., Frankish,A., Huang,N., Morris,J., Walter,K., Jostins,L., Habegger,L., Pickrell,J.K., Montgomery,S.B., et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823-828.

18 Shimada,M.K., Hayakawa,Y., Takeda,J., Gojobori,T. and Imanishi,T. (2010) A comprehensive survey of human polymorphisms at conserved splice dinucleotides and its evolutionary relationship with alternative splicing. *BMC Evol. Biol.*, **10**, 122.

19 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.

20 Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364-4375.

21 Sorek,R., Ast,G. and Graur,D. (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060-1067.

22 Lev-Maor,G, Sorek,R, Shomron,N and Ast,G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, **300,** 1288-1291.

23 Smit,AFA, Hubley,R and Green,P. (1996-2010) RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

24 Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996-1006.

25 Drummond,D.A., Bloom, J.D., Adami,C., Wilke,C.O. and Arnold,F.H. (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U S A.*, **102**, 14338-14343.

26 Nurtdinov,R.N., Mironov,A.A. and Gelfand,M.S. (2009) Rodent-specific alternative exons are more frequent in rapidly evolving genes and in paralogs. *BMC Evol. Biol.*, **9**,142.

27 Sammeth,M., Foissac,S. and Guigy,R. (2008) A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, **4**, e1000147.

28 Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {12.04.2012}. URL: http://omim.org/.

29 Bonnevie-Nielsen,V., Field,L.L., Lu,S., Zheng,D.J., Li,M., Martensen,P.M., Nielsen,T.B., Beck-Nielsen,H., Lau,Y.L. and Pociot, F. (2005) Variation in antiviral 2',5'-oligoadenylate synthetase (2'5'AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene *Am. J. Hum. Genet.,* **76**, 623-633.

30 Lalonde,E., Ha,K.C., Wang,Z., Bemmo,A., Kleinman,C.L., Kwan,T., Pastinen,T. and Majewski,J. (2011) RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res*., **21**, 545-554.

31 Hanioka,N., Kimura,S., Meyer,U.A. and Gonzalez,F.J. (1990) The human CYP2D locus associated with a common genetic defect in drug oxidation: a G1934----A base change in intron 3 of a mutant CYP2D6 allele results in an aberrant 3' splice recognition site. *Am. J. Hum. Genet.,* **47**, 994-1001.

32 Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U S A.,* **106**, 9362-9367.

33 Hindorff,L.A., MacArthur,J., Morales,J., Junkins,H.A., Hall,P.N., Klemm,A.K., and Manolio,T.A. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 15.01.2013.

34 Nyholt,D.R., Low,S.K., Anderson,C.A., Painter,J.N., Uno,S., Morris,A.P., MacGregor,S., Gordon,S.D., Henders,A.K., Martin,N.G. et al. (2012) Genome-wide association meta-analysis identifies new endometriosis risk loci. Nat Genet., 44, 1355-1359.

35 Sanders,S.J., Murtha,M.T., Gupta,A.R., Murdoch,J.D., Raubeson,M.J., Willsey,A.J., Ercan-Sencicek,A.G., DiLullo,N.M., Parikshak,N.N., Stein,J.L. et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237-241.

36 Kurmangaliyev,Y.Z. and Gelfand,M.S. (2008) Computational analysis of splicing errors and mutations in human transcripts. *BMC Genomics*, **9**, 13.

37 Gan,X., Stegle,O., Behr,J., Steffen,J.G., Drewe,P., Hildebrand,K.L., Lyngsoe,R., Schultheiss,S.J., Osborne,E.J., Sreedharan,V.T. et al. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, ***477***, 419-423.

38 Xing,Y., Xu,Q. and Lee,C. (2003) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett.*, **555**, 572-578.

39 Pickrell,J.K., Pai,A.A., Gilad,Y., Pritchard,J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, ***6***, e1001236.

40 Nurtdinov,R.N., Artamonova,I.I., Mironov,A.A. and Gelfand,M.S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. Hum. Mol. Genet., 12, 1313-1320.

41 Ramensky,V.E., Nurtdinov,R.N., Neverov,A.D., Mironov,A.A. and Gelfand,M.S. (2008) Positive selection in alternatively spliced exons of human genes. *Am. J. Hum. Genet.* **83**, 94-98.

42 Charlesworth,B., Charlesworth,D. (2010) Elements of Evolutionary Genetics. Roberts and Company, Greenwoord Village, Colorado.

43 Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D., et al. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7**, Suppl 1, S4.1-9.

44 Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D., Haussler,D. and Miller,W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708-715.

45 Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. et al. (2002) The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611-1618.

46 Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586-1591.

47 Rosenbloom,K.R., Dreszer,T.R., Long,J.C., Malladi,V.S., Sloan,C.A., Raney,B.J., Cline,M.S., Karolchik,D., Barber,G.P., Clawson,H. et al. (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.*, **40**, D912-917.

48 Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.*, **28**, 511-515.

49 Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. et al. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290-301.

50 R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

51 Wickham,H. (2009) ggplot2: elegant graphics for data analysis. Springer, New York.

52 Wassif,C.A., Maslen,C., Kachilele-Linjewile,S., Lin,D., Linck,L.M., Connor,W.E., Steiner,R.D. and Porter,F.D. (1998) Mutations in the human sterol delta7-reductase gene at 11q12-13 cause Smith-Lemli-Opitz syndrome. *Am. J. Hum. Genet.*, **63**, 55-62.

53 Yu,H., Lee,M.H., Starck,L., Elias,E.R., Irons,M., Salen,G., Patel,S.B. and Tint,G.S. (2000) Spectrum of Delta(7)-dehydrocholesterol reductase mutations in patients with the Smith-Lemli-Opitz (RSH) syndrome. *Hum. Mol. Genet.*, **9,** 1385-1389.

54 Kobayashi,K., Sinasac,D.S., Iijima,M., Boright,A.P., Begum,L., Lee,J.R., Yasuda,T., Ikeda,S., Hirano,R., Terazono,H. et al. (1999) The gene mutated in adult-onset type II citrullinaemia encodes a putative mitochondrial carrier protein. *Nat. Genet.*, **22**, 159-163.

55 Torrents,D., Mykkдnen,J., Pineda,M., Feliubadaly,L., Estйvez,R., de Cid,R., Sanjurjo,P., Zorzano,A., Nunes,V., Huoponen,K. et al. (1999) Identification of SLC7A7, encoding y+LAT-1, as the lysinuric protein intolerance gene. *Nat. Genet.*, **21**, 293-296.

56 Meinsma,R., Fernandez-Salguero,P., Van Kuilenburg,A.B., Van Gennip,A.H. and Gonzalez,F.J. (1995) Human polymorphism in drug metabolism: mutation in the dihydropyrimidine dehydrogenase gene results in exon skipping and thymine uracilurea. *DNA Cell Biol.*, **14**, 1-6.

57 Ogorelkova,M., Gruber,A. and Utermann,G. (1999) Molecular basis of congenital lp(a) deficiency: a frequent apo(a) 'null' mutation in caucasians. *Hum. Mol. Genet.*, **8**, 2087-2096.

58 Akerman,B.R., Zielenski,J., Triggs-Raine,B.L., Prence,E.M., Natowicz,M.R., Lim-Steele,J.S., Kaback,M.M., Mules,E.H., Thomas,G.H., Clarke,J.T., et al. (1992) A mutation common in non-Jewish Tay-Sachs disease: frequency and RNA studies. *Hum. Mutat.*, **1**, 303-309.

Legends to Figures

Figure 1. Non-functional allele frequencies (NAF) of SNPs overlapping the conserved dinucleotides of splice sites. Grey, 2090 ancestral polymorphic splice sites; white, 159 human-specific polymorphic splice sites.

Figure 2. Splice site scores. Boxplot corresponds to the first and third quartiles. The dash in boxplot is the median, and the length of whiskers is 1.5 times the interquartile range.

Figure 3. SNPs more frequently disrupt rarely-used splice sites of lowly expressed genes. Total expression level of splice sites (**A**), inclusion levels of alternative splice sites (**B**), and expression level of genes (**C**), summed across 14 cell lines. Expression level was measured in Fragments Per Kilobase per Million mapped reads (FPKM). For the boxplot description, see the legend of Figure 2. Outliers are represented as dots.

Figure 4. Conservation of splice sites and genes. A: fraction of splice sites conserved in the mouse genome. B: dN/dS ratios for genes. For the boxplot description, see the legend of Figure 2. Outliers are represented as dots.

Figure 5. Disruptions of polymorphic splice sites rarely lead to serious changes in the structure of proteins. (**A**) Fraction of splice sites localized in CDS-regions of the genes. **(B)** Coding length of exons corresponding to splice sites. (**C**) Relative position along the CDS of splice sites flanking exons with lengths not in multiple of three. 0 and 1 corresponds to the N- and C-termini of proteins (**D**) Fraction of splice sites overlapping the Pfam domains, among those located in CDS-regions. For the boxplot description, see the legend of Figure 2.

Tables

Table 1. Sets of splice sites used in the analysis.

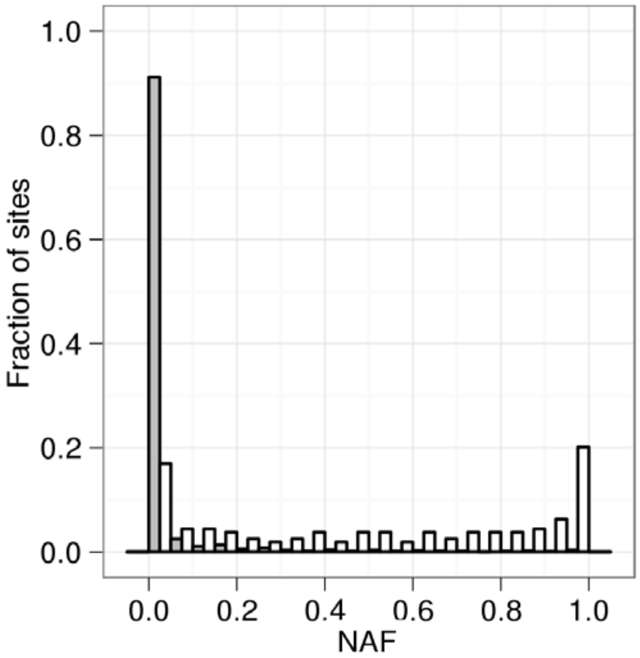
	noSNP	**lfSNP**	**hfSNP**
total number of genes	15793	941	905
total number of sites	380409	1109	981
by types			
acceptor sites	190694	459	390
donor sites	189715	650	591
by alternativety			
constitutive sites	275732	734	545
alternative sites	106677	375	436
within CDS regions			
constitutive sites	246713	606	363
alternative sites	88202	267	268
within CDS regions, flanking exons having length a multiple of three			
constitutive sites	147413	345	214
alternative sites	49916	156	147

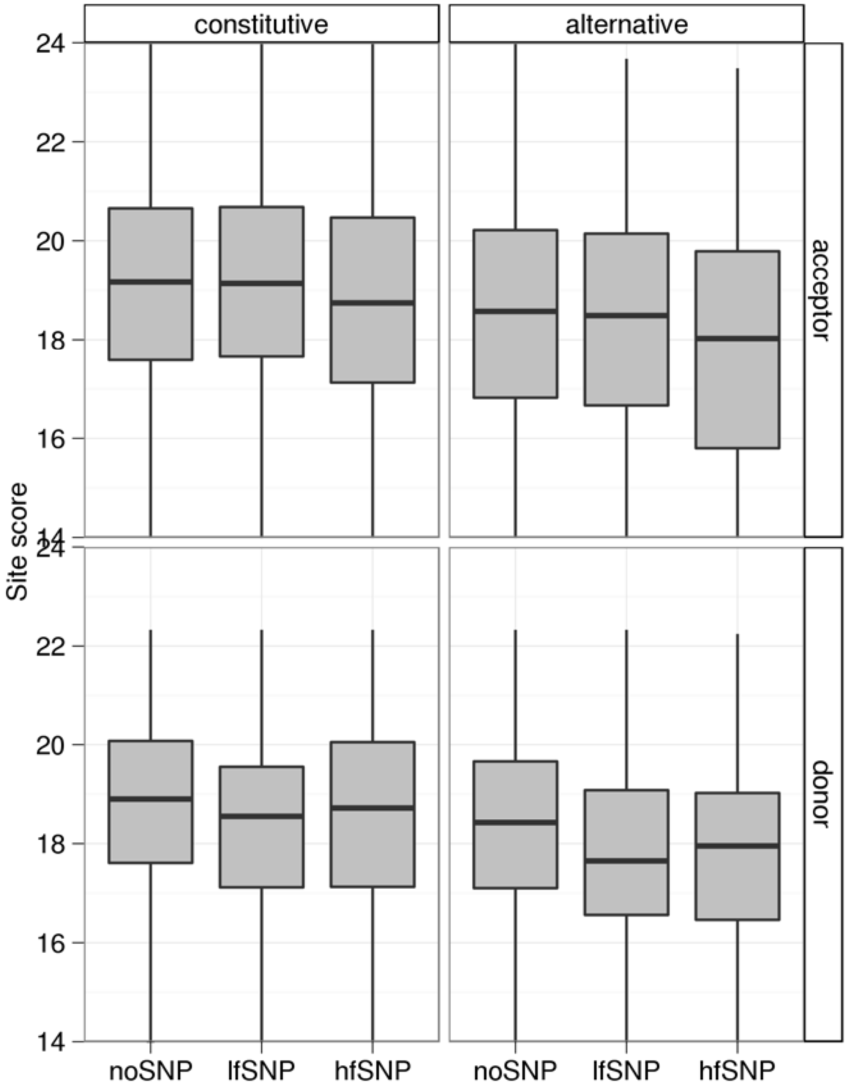
Table 2. Properties of splice sites. For quantitative parameters, the medians are reported. The last two columns report the statistical significance of the differences of the distributions, using the chi-squared test (CS) and the Kruskal-Wallis test (KW); n/s – non significant.

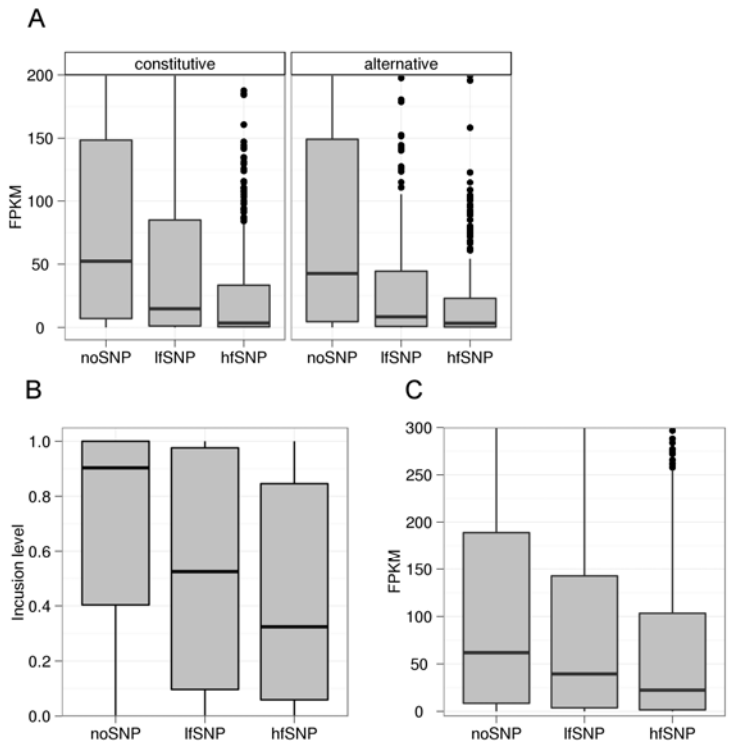
	noSNP	**lfSNP**	**hfSNP**	**CS**	**KW**
Fraction of donor sites	0.5	0.59	0.6	$<10^{-15}$	-
Fraction of constitutive sites	0.72	0.66	0.56	$<10^{-15}$	-
Site scores					
for acceptor sites					
constitutive sites	19.17	19.14	18.74	-	n/s
alternative sites	18.57	18.49	18.03	-	0.01
for donor sites					
constitutive sites	18.9	18.55	18.72	-	$<10^{-6}$
alternative sites	18.43	17.65	17.95	-	$<10^{-7}$
Expression level of genes (FPKM)	61.84	39.46	22.37	-	$<10^{-15}$
Expression level of sites (FPKM)					
constitutive sites	52.36	14.77	3.48	-	$<10^{-15}$
alternative sites	42.62	8.41	3.34	-	$<10^{-15}$
Inclusion level of alternative sites	0.9	0.52	0.32	-	$<10^{-15}$
dN/dS	0.10	0.14	0.18	-	$<10^{-15}$
Fraction of sites conserved in mouse					
constitutive sites	0.92	0.81	0.66	$<10^{-15}$	-
alternative sites	0.81	0.62	0.43	$<10^{-15}$	-
Fraction of sites localized in CDS regions					
constitutive sites	0.9	0.83	0.67	$<10^{-15}$	-
alternative sites	0.83	0.71	0.62	$<10^{-15}$	-
Coding length of exons (bp)					
constitutive sites	124	118	117	-	$<10^{-3}$
alternative sites	113	106	99	-	$<10^{-4}$
Relative position					
constitutive sites	0.49	0.53	0.53	-	n/s
alternative sites	0.48	0.61	0.58	-	0.003
Fraction of sites overlapping with Pfam domains					
constitutive sites	0.44	0.44	0.36	0.004	-
alternative sites	0.41	0.28	0.34	$<10^{-5}$	-

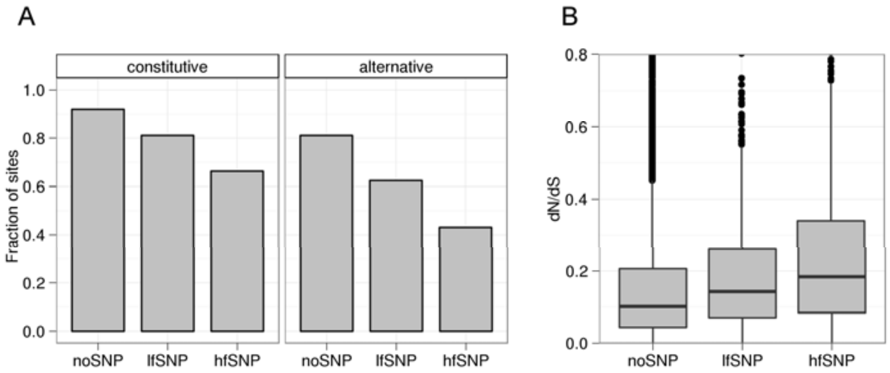
Table 3. Polymorphic splice sites carrying the disrupting SNPs found in OMIM. Braces "{}" and brackets "[]" identify susceptibility alleles to complex diseases and non-disease variants as in OMIM Morbid Map. For alleles, 2 bp of the splicing dinucleotide and 6 bp flanking it from each side are shown.

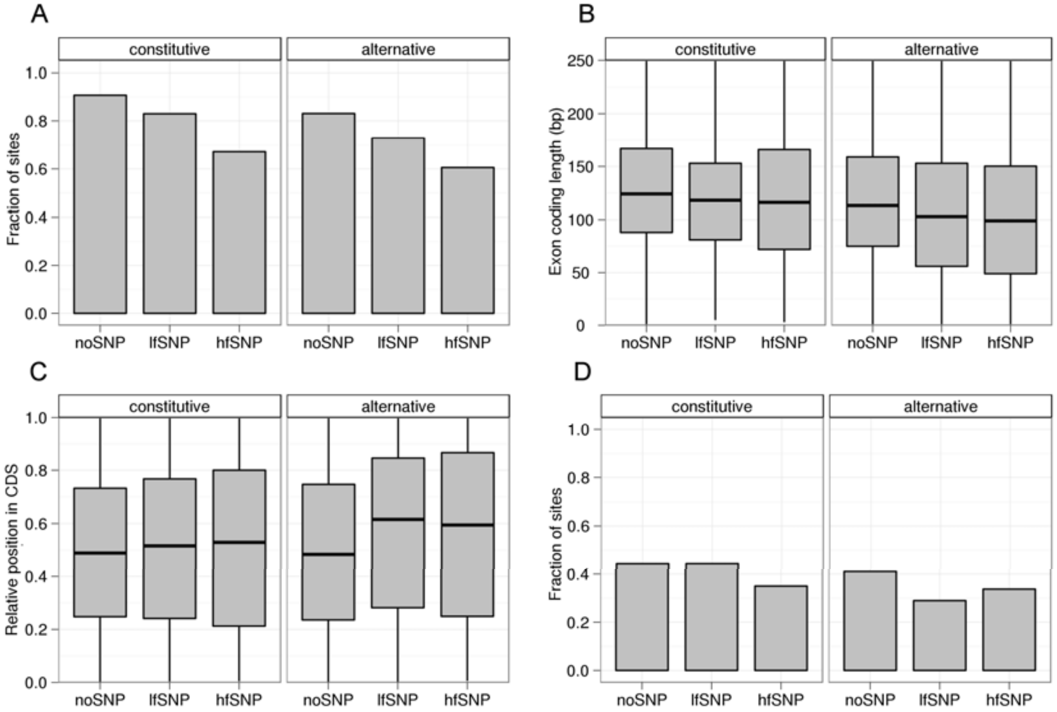
MIM	Gene	Exon	Site	Alleles (common/alternative)	NAF	Conservation in mouse	Gene segment	Overlap with Pfam domains	Phenotype (MIM)	Splicing disruption phenotype
238310	AMT	8	acceptor	ttctcc(AG/AA)ggaagc	0.0005	conserved	alternative	-	Glycine encephalopathy (605899)	not described
602858	DHCR7	9	acceptor	cccccc(AG/AC)ggtctg	0.0046	conserved	alternative	PfamA	Smith-Lemli-Opitz syndrome (270400)	activation of a cryptic splice site 134 bp upstream [52, 53]
139250	GH1	2	donor	aaatcc(GT/AT)gagtgg	0.0027	conserved	constitutive	PfamA	Growth hormone deficiency, type IA (262400), type IB (612781), type II (173100); Kowarski syndrome (262650)	not described
603859	SLC25A13	11	donor	atagag(GT/AT)tagtgc	0.0014	conserved	constitutive	PfamA	Citrullinemia, adult-onset type II (603471), neonatal-onset type II (605814)	in-frame exon skipping [54]
217050	C6	16	donor	ctgtag(GT/GC)aagaga	0.0009	conserved	constitutive	PfamA	C6 deficiency (612446); Combined C6/C7 deficiency	not described
606673	UPB1	9	acceptor	ctttaa(AG/AA)ctcacc	0.0009	conserved	constitutive	-	Beta-ureidopropionase deficiency (613161)	not described
603593	SLC7A7	6	acceptor	tccctt(AG/TG)actttt	0.0005	conserved	constitutive	PfamA	Lysinuric protein intolerance (222700)	activation of a cryptic splice site 10 bp downstream (next AG) [55]
124030	CYP2D6	4	acceptor	accccc(AG/AA)gacgcc	0.1024	conserved	constitutive	PfamA	{Codeine sensitivity} (608902); {Debrisoquine sensitivity} (608902)	not described
602044	UCP3	6	donor	caaggg(GT/AT)gagcct	0.0338	conserved	constitutive	PfamA	{Obesity, severe, and type II diabetes} (601665)	not described
603100	AGPAT2	5	acceptor	gcctgc(AG/GG)gtgccc	0.0009	conserved	constitutive	PfamA	Lipodystrophy, congenital generalized, type 1 (608594)	not descibed
164350	OAS1	6	acceptor	cctttc(AA/AG)gctgaa	0.6408	non-conserved	alternative	-	{Susceptibility to diabetes mellitus, type 1} (222100); {Susceptibility to viral infection}	activation of cryptic splice sites 1 bp and 98 bp downstream [29, 30]
612779	DPYD	14	donor	gacaac(GT/AT)aagtgt	0.0023	conserved	constitutive	PfamA	5-fluorouracil toxicity (274270); DPYD deficiency (274270)	in-frame exon skipping [56]
152200	LPA	26	donor	aaatgc(GT/AT)atgtct	0.0174	conserved	constitutive	PfamA	[LPA deficiency, congenital] ; {Suspectibility to coronary artery disease}	activation of a cryptic splice site 64 bp upstream [57]
606869	HEXA	9	donor	ccagac(GT/AT)gaggaa	0.0009	conserved	constitutive	PfamA	GM2-gangliosidosis (272800); Tay-Sachs disease (272800); [Hex A pseudodeficiency] (272800)	activation of cryptic splice sites 40 bp upstream and 17 bp downstream [58]
604629	MMP20	7	acceptor	tcccat(AG/TG)gatttt	0.0005	conserved	constitutive	-	Amelogenesis imperfecta (612529)	not described
118470	CETP	14	donor	tgtctc(GT/AT)aagtgt	0.0005	non-conserved	constitutive	PfamA	Hyperalphalipoproteinemia (143470); [High density lipoprotein cholesterol level QTL 10] (143470)	not described


1.0
0.8
0.6
0.4
0.2
0.0
Fraction of sites
0.0
0.2
0.4
0.6
0.8
1.0
NAF


constitutive
alternative
acceptor
donor
Site score
24
22
20
18
16
14
noSNP
lfSNP
hfSNP
noSNP
lfSNP
hfSNP


A
constitutive
alternative
FPKM
200
150
100
50
0
noSNP
lfSNP
hfSNP
noSNP
lfSNP
hfSNP
B
Incusion level
1.0
0.8
0.6
0.4
0.2
0.0
noSNP
lfSNP
hfSNP
C
FPKM
300
250
200
150
100
50
0
noSNP
lfSNP
hfSNP


A
constitutive
alternative
Fraction of sites
1.0
0.8
0.6
0.4
0.2
0.0
noSNP
lfSNP
hfSNP
noSNP
lfSNP
hfSNP
B
dN/dS
0.8
0.6
0.4
0.2
0.0
noSNP
lfSNP
hfSNP


A
constitutive
alternative
Fraction of sites
1.0
0.8
0.6
0.4
0.2
0.0
noSNP
lfSNP
hfSNP
B
Exon coding length (bp)
250
200
150
100
50
0
C
Relative position in CDS
D
Fraction of sites