NAR Genomics and Bioinformatics, 2025, **7**, lqaf036
https://doi.org/10.1093/nargab/lqaf036
Standard Article

OXFORD

Fully haplotyped genome assemblies of healthy individuals reveal variability in 5′ss strength and support by splicing regulatory proteins

Johannes Ptok *, Stephan Theiss, Heiner Schaal

Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany

*To whom correspondence should be addressed. Email: Johannes.Ptok@hhu.de

Abstract

This work presents a comprehensive investigation of sequence variations at human 5′ splice sites (5′ss), exploring their impact on 5′ss strength and predicted splicing regulatory protein (SRP) binding. Leveraging 44 high-quality genomes, with fully haplotyped assemblies, we were able to fully assess homozygous and heterozygous sequence variations around and within 5′ss. Variations showed differing tolerance levels in protein-coding and non-coding transcripts. Around half of 5′ss variations did not alter 5′ss strength (measured by the HBond score, that estimates binding of the 11 nucleotides at the free U1 spliceosomal RNA 5′ end). Heterozygous variations resulted in stronger 5′ss strength reductions and less compensatory effects of multiple sequence variations at the same 5′ss, compared to homozygous variations. Additionally, we observed a slight balance between changes in 5′ss strength and predicted SRP binding. Strong 5′ss (HBond score >18.8) showed strength variations in both directions, as theoretically expected for random distributions, whereas weaker 5′ss consistently showed lower strength reductions than expected or achievable. Variations at 5′ss of essential genes were less frequent than in other genes and showed a higher amount of variations that did not alter 5′ss strength. Acceptable changes in predicted SRP binding sites were highly dependent on their respective 5′ss strength.

Introduction

Splicing is a crucial step in messenger RNA (mRNA) maturation of the majority of human genes, during which introns are removed from precursor RNA transcripts [1, 2]. The spliceosome is a complex of U small nuclear ribonucleoprotein particles and splicing proteins that assembles at exon–intron boundaries and recognizes key sequence elements, such as the 5′ splice site (splice donor, 5′ss) and the 3′ splice site (splice acceptor, 3′ss) [3].

The recognition of 5′ss is facilitated by RNA duplex formation between a 5′ss sequence and the 11 nucleotides long free 5′ end of the U1 snRNA (small nuclear RNA). Sequence variations within the 5′ss sequence itself can drastically affect 5′ss recognition, which can lead to alternative 5′ss usage and pathogenic changes in the amount of functional mRNA transcripts [4]. Several algorithms try to estimate 5′ss recognition and thus splice site strength by analyzing the 5′ss itself, like the HBond score (HBS) [5] or the most commonly used MaxEntScan score [6]. While the MaxEntScan for 5′ss only evaluates a 9 nt long sequence, the HBS considers all 11 nucleotides possibly complementary to the free 5′ end of the U1 snRNA. However, splice site usage is not only dependent on the 5′ss sequence itself but can be heavily influenced by splicing regulatory proteins (SRPs) that bind the RNA. Depending on their binding position relative to the splice site, they can significantly enhance or repress splice site usage [7]. The SplicePort tool, for instance, thus estimates splice site strength by additionally taking into account the whole immediate sequence surrounding of a splice site sequence (±80 nucleotides) [8]. SRP binding can also be estimated separately with the HEXplorer algorithm, which was developed to generally predict SRP binding sites in a genomic sequence. The HEXplorer-derived splice site HEXplorer weight (SSHW) summarizes the overall putative influence of SRPs bound within a ±60 nt window of a 5′ss [9, 10].

Correctly predicting splice site usage is very important during diagnosis and treatment development for various genetic diseases [11, 12]. Since around 97% of human genes contain intronic sequences that need to be removed during pre-mRNA maturation, aberrant splicing can potentially affect translation of almost every human gene [13]. Even single nucleotide variants (SNVs) can alter correct splicing of RNA transcripts when they affect 5′ss or SRP binding sites. Millions of SNVs directly changing the encoded amino acid sequence via nonsynonymous substitutions, frameshifts, or premature stop codons have already been described in the literature (ClinVar [14] or SNPdb [15]). However, within an RNA transcript, any SNV, whether silent or not, can alter RNA processing, which has been getting more and more attention in diagnostics in recent years. And indeed, ∼88% of human SNVs associated with disease do not directly affect the amino acid sequence, but result in non-functional transcript isoforms while located within intronic and intergenic sequence segments [16]. Depending on their genomic location, they can directly alter the sequences of annotated splice sites, alter the SRP-binding landscape or introduce strong cryptic

Received: September 27, 2024. **Revised:** February 18, 2025. **Editorial Decision:** March 18, 2025. **Accepted:** March 20, 2025

splice sites, potentially leading to splice site competition and aberrant splicing. To our knowledge, the global variance in splice site strength and the variance in estimated SPR binding landscape within the immediate 5′ss sequence surrounding was never analyzed.

Here, we comprehensively analyzed a specific dataset of 44 healthy individuals' haplotype genomes from the 1000 Genome Project to consider the entire individual sequences around all 5ss [17, 18]. The newly assembled chromosome haplotypes, generated using the latest technological advancements, offer a complete view of 5′ss sequence variations for the first time. Although the sample size of 44 individuals is relatively small, the dataset might be sufficiently diverse for the purpose of this work, since the 44 individuals were chosen from the 1000 Genomes Project to maximize ethnic diversity, as stated in the original paper describing the dataset. The apparent non-pathogenic variability of 5′ss strength or SSHW calculated in this dataset might help us to determine thresholds above which significant changes in splice site usage can be expected.

Materials and methods

A recently developed pipeline, combining long-read technology with single-cell template strand sequencing (Strand-seq), enables generation of fully phased diploid genome assemblies without use of parent-child trio information or reference genomes [19]. Variant calling from these high-quality haplotype assemblies increases sensitivity and correctly places variants into the right genetic context, which is essential when summarizing the effect of all relevant SNVs on a given splice site per haplotype. Some of the first genomes, newly assembled using this method, came from 44 individuals of the 1000 Genome Project [17]. To determine variants in proximity to splice sites, we analyzed the decomposed VCF file (GRCh38-VCF, version 1.1), from the github repository of the paper, which refers to coordinates of human reference genome GRCh38. A custom R-script then applied functions of our VarCon R package to reconstruct the sequence neighborhood of 5′ss holding the respective sequence variations [20] (see github.com/caggtaagtat/SNVimpact).

Results and discussion

High-resolution genomes reflect non-pathogenic sequence variations at 5′ss

To better understand the conditions under which a sequence variation near an exon–intron border—either by altering splice site strength or the binding landscape of SRPs—is strong enough to affect canonical splice site usage, we analyzed genomic variations in these regions. Our analysis focused on a subset of 88 high-quality haplotype assemblies from 44 individuals of the 1000 Genomes Project that were recently published [17]. The 44 genomes were originally selected to represent global genetic diversity, and for which consent had been given for unrestricted access [17].

Based on previously developed methods, we wrote a custom R-script that (i) identifies sequence variations occurring within a ±60 nt sequence-window around annotated 5′ss, (ii) integrates these into the reference sequence, and (iii) calculates changes in the 5′ss strength and SSHW from one or multiple sequence variations in the respective sequence region [20]. In total, we identified 2 574 001 single nucleotide variations (SNVs), 212 465 deletions, and 231 132 insertions that affected the ±60 nt sequence window of interest. Introducing insertion and deletion variations had to be carefully coordinated with the introduction of SNVs, since the former naturally changed the original genomic coordinates of the respective sequence window.

Reduction of functional protein concentrations due to specific sequence variations within the genetic sequence of only one copy of the gene (heterozygous) does not necessarily result in disease, since together with the second intact gene copy, enough protein is still being produced. Evolutionary pressure to avoid heterozygous mutations within regulatory sequence segments like splice sites might therefore be less stringent than on genetic variations that are present in both copies of a gene (homozygous), even if the encoded protein is essential. Following this principle, homozygous sequence variations within for instance the 5′ss itself could indicate, a general window of acceptable variation in 5′ss strength, that does not sufficiently affect 5′ss usage to be pathogenic. Defining this potential global threshold of variations in 5′ss strength or SSHW could be of clinical importance during identification and classification of individual genomic variations and their potential risk to induce disease. We therefore first analyzed annotated 5′ss that showed a homozygous difference in their strength or SSHW, in at least one individual and then compared the observations with differences originating from heterozygous variations.

Not all 5′ss annotated in transcript isoforms of a gene contribute equally to the production of functional proteins. For instance, some 5′ss may only be associated with transcript isoforms that are retained in the nucleus or targeted for degradation via nonsense-mediated decay. Hence, we generated groups of 5′ss that are either found (i) in exclusively non-coding transcript isoforms of a gene ("not protein coding") or (ii) in at least one protein-coding transcript ("protein coding"). Moreover, some transcripts annotated in the *ensembl* archive differ in splice site composition or are even completely missing from the independent pendant to the *ensembl* archive, called RefSeq, which is provided by the US-American National Center for Biotechnology Information (NCBI) [21]. This "confidence" in a transcript can be assessed by its transcript support level (TSL, levels ranging from 1 [highest confidence] to 5) provided by the *ensembl* archive. All annotated exon–intron borders of TSL1 transcript isoforms are found with the same coordinates in both the *ensembl* and RefSeq archives. We therefore additionally labeled 5′ss by whether they were either found in (i) exclusively non-TSL1 transcript isoforms of a gene ("Not TSL1") or (ii) in at least one TSL1 transcript ("TSL1").

Calculating the HBS of reference 5′ss of each category revealed that 5′ss generally show a mean HBS of around 14 to 15. 5′ss of TSL1 transcripts showed significantly higher HBond scores than 5′ss of "Not TSL1" transcripts (p-value of Mann–Whitney U test $< .001$). Additionally, TSL1 5′ss of protein-coding transcripts had slightly higher HBond scores than TSL1 5′ss of non-coding transcripts (Supplementary Fig. S1).

Homozygous variations within annotated 5′ss sequences

First, we selected annotated 5'ss that showed a homozygous sequence variation within the 5′ss sequence. From 311 433 5′ss annotated in *ensembl* (version 105), only 4044 (1.3%) showed a homozygous 5′ss sequence variation across the

Table 1. Groups of 5′ss belonging to TSL1- and/or protein-coding transcript isoforms or not, showing homozygous sequence variations within the 11 nucleotides of the 5′ss sequence

	Not TSL1	TSL1	Total
Not protein coding	1376 (1.6% of 83 910)	224 (1.9% of 11 404)	1600 (1.7% of 95 314)
Protein coding	648 (1.5% of 42 920)	1796 (1.0% of 173 199)	2444 (1.1% of 216 119)
Total	2024 (1.6% 126 830)	2020 (1.1% of 184 603)	4044 (1.3% of 311 433)

5′ss could either be (i) exclusively part of non-coding transcript isoforms ("not protein coding") or (ii) part of at least one protein-coding transcript ("protein coding"). Additionally, 5′ss were classified whether they were found in (i) transcripts of TSL level >1 ("not TSL1") or (ii) TSL1 transcripts ("TSL1"), with TSL1 transcripts representing high-confidence annotated transcripts that are found with the all same exon-junctions in the ensembl-pendant RefSeq.

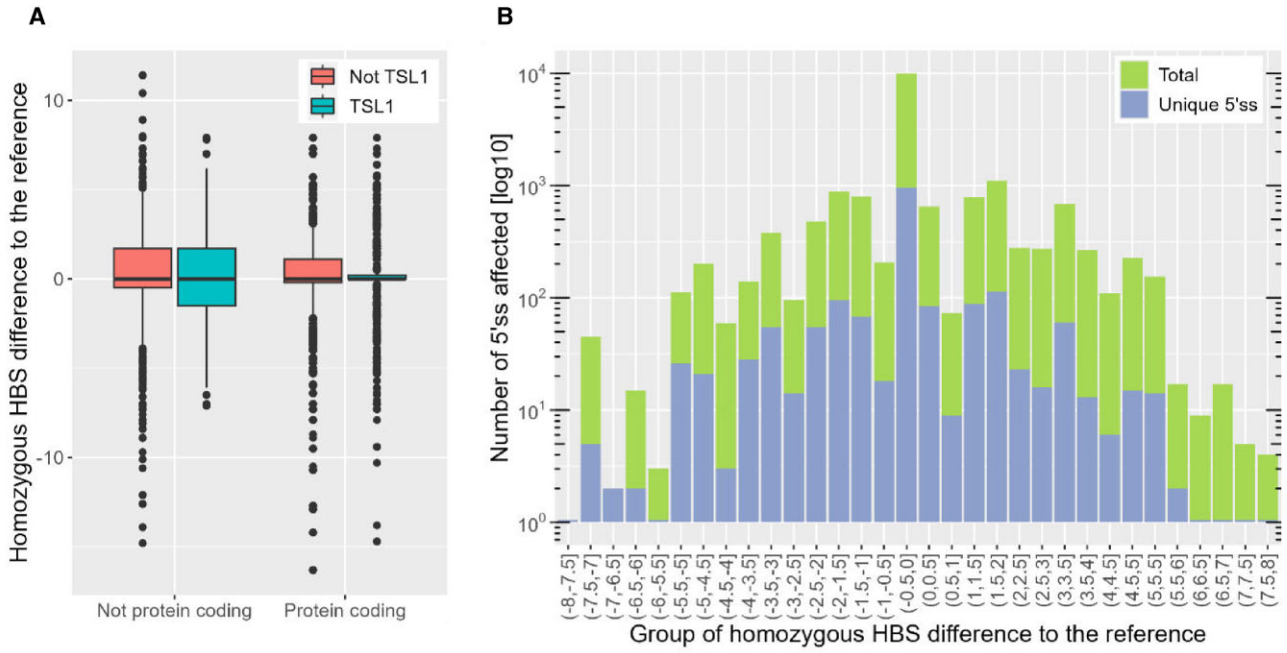

Figure 1. ΔHBS of 5′ss that are affected by homozygous sequence variations within the 11 nucleotides of the 5′ss sequence by TSL and "protein coding" category. (**A**) Boxplot depicting ΔHBS per category. 5′ss were grouped into those that were (i) exclusively part of non-coding transcript isoforms ("not protein coding") or (ii) part of protein-coding transcripts ("protein coding") and those, exclusively found in (i) transcripts of TSL level >1 ("not TSL1") or (ii) TSL1 transcripts ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts that are found with all the same exon-junctions in the ensembl-pendant RefSeq. (**B**) Histogram of ΔHBS upon homozygous sequence variation within the sequence of 5′ss, found in protein-coding TSL1 transcripts. Depicted is the number of affected 5′ss per ΔHBS group, either across all individuals (total = 18 156) or only considering unique 5′ss per ΔHBS group (total 1796). Since logarithmic transformation results in 5′ss groups containing only one 5′ss to show a value of 0, we added 0.02 to enable visualization in these cases.

44 individuals. 3011 (74.5%) of the 4044 5′ss showed a homozygous variation in multiple individuals, which additionally indicates that the occurrence of these variations might not be random, but predominantly within the same transcripts and 5′ss. 99% of those showed the same HBS difference across all affected individuals, potentially due to the presence of less frequent minor alleles in the reference genome sequence.

Grouping splice donor sites by the above stated "TSL1" and "protein coding" labels, we found that 5′ss belonging to protein-coding TSL1 transcript isoforms showed the lowest proportion of affected 5′ss, relative to the total number of annotated 5′ss in each category (Table 1).

To analyze changes in 5′ss strength induced by these variations, we calculated the HBond score difference (ΔHBS) by subtracting the HBS of the reference 5′ss from the HBS of an affected 5′ss and compared the ΔHBS distribution across the different 5′ss categories (Fig 1A).

Comparing homozygous ΔHBS between 5′ss categories, we saw the least deviation from an HBS difference of 0 for 5′ss that belong to protein-coding TSL1 transcripts (standard deviation of 1.8). Some of the 1796 5′ss of this category were affected in the genome of multiple individuals, resulting in a total of 18 156 instances. 9211 of 18 156 (51%) 5′ss of this category showed a ΔHBS of 0 across all individuals (Fig. 1B). The number of instances with no HBS difference to the reference despite a homozygous sequence variation is similar when accounting for 5′ss that show homozygous changes in multiple individuals (49%). 5′ss that do not belong to protein coding and not to TSL1 transcripts showed a slightly lesser proportion of ΔHBS of 0 (only 5079 of 13 013 instances at 1.37 unique 5′ss, 39%). 5′ss of non-TSL1 transcripts show a much higher ΔHBS compared to 5′ss of protein-coding TSL1 transcripts (standard deviation 2.8, *p*-value of Kolmogorov–Smirnov test < .001). However, 5′ss of non-coding TSL1 transcripts seem to show a greater ΔHBS variance, compared to 5′ss of coding TSL1 transcripts (standard deviation 2.6, *p*-value of Kolmogorov–Smirnov test < .001) (Fig. 1A).

Since most diagnostics based on the analysis of genomic sequences scan sequences of protein-coding genes, we next focused on protein-coding transcripts of highest TSL, TSL1.

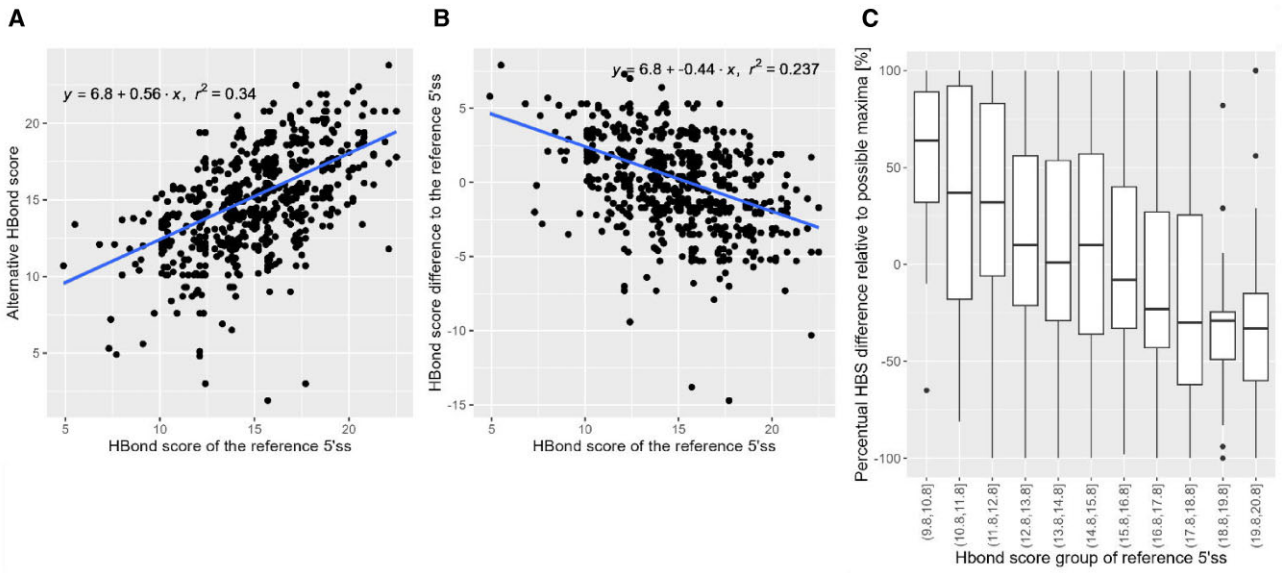

Figure 2. Observed HBS difference range depends on HBS of the reference 5′ss. Depicted is the HBS difference (**B**) or alternative HBS (**A**), caused by homozygous genetic variations from the reference, within the sequence of annotated 5′ss. 5′ss that were much weaker than the reference 5′ss were increasingly found for 5′ss of average strength or below (HBS). 5′ss of relatively high strength showed strong reductions in HBS compared to the reference. Each regression line is depicted with its formula stated above. The 95% confidence interval of the regression line is marked by a dark gray area. R^2 is 0.34 (**A**) and 0.24 (**B**), respectively. (**C**) HBS difference of 1743 5′ss with only one SNV within the 5′ss, relative to the strongest HBS increase or decrease that is theoretically achievable by one SNV, for each affected reference 5′ss.

Around half of affected 5′ss showed a ΔHBS of 0, despite the sequence variation within the 5′ss sequence. Left and right of the peak at ΔHBS of 0, one can detect two additional smaller peaks ranging from −2.5 to −1 and 1 to 2.5, respectively. This could indicate that 5′ss of "highest importance" (only in protein-coding TSL1 transcripts) might normally tolerate only a slight HBS variation of around ±1; however, some 5′ss seem to be less restricted in HBS variations without resulting in disease. Reasons for this could for instance be gene duplications, non-essential alternatively used 5′ss, a compensating SSHW, or the strength of the reference 5′ss. Since changes in strength measured by the HBS can be present in more than one subject, we additionally checked, how many unique 5′ss were found in each group and saw no significant aberrations from the previously described distributions (Fig. 1B). The ΔHBS of heterozygous 5′ss variants showed a slight shift to stronger HBS decreases (p-value of Kolmogorov–Smirnov test < .001) (Supplementary Fig. S2).

Previously, we saw that the reduction of 5′ss strength can induce usage of nearby competing 5′ss that are usually not used, independent of the strength of the original 5′ss. This effect was, however, dependent on the HBS and SSHW differences between the 5′ss and the competing 5′ss in proximity to it [22]. We therefore tested whether this observation would also fit to our data set of 5′ss from protein-coding TSL1 transcripts that showed homozygous variations within the donor sequence (Fig. 2).

We saw a small negative correlation between ΔHBS and the HBS of the reference 5′ss itself (R^2 of linear regression = 0.237). Stronger 5′ss had a higher tolerance with respect to SNVs, i.e. they permitted larger HBS differences. 5′ss with intermediate HBS (between ∼12 and 15) tolerated SNVs, that were only slightly weakening the 5′ss. Weak 5′ss (HBS < 12) mostly tolerated SNVs that actually increased their strength. The same holds true to a slightly lesser extent for homozygous and heterozygous sequence variations within 5′ss of other categories. However, no correlation could be detected with the SSHW of the affected 5′ss. One possible explanation for the correlation between reference HBS and ΔHBS is that nucleotide exchanges within high-complementarity 5′ss sequences may have a greater likelihood of randomly decreasing the HBS compared to exchanges within low-complementarity 5′ss sequences.

In order to test, whether the observed sequence variations induced a lower HBS reduction than expected by chance, we first grouped the reference 5′ss with homozygous ΔHBS by their HBS in steps of 1. Subsequently, we removed 5′ss with >1 nucleotide differences to the reference, resulting in 1743 5′ss. Per HBS group, we then collected every unique 5′ss sequence and generated every potential 5′ss that could result from a single nucleotide exchange (27 = 9*3 per unique 5′ss). This set of alternative sequences was then used as a data set that describes the chance-expected ΔHBS frequency, which was then compared to the observed real data.

For reference HBS groups of sufficient size (at least 2% of unique 5'ss in this HBS group), we compared the theoretically expected frequency of positive and negative ΔHBS with Fisher's exact test (two-sided) for a 2 by 2 table per HBS group (exemplary dataset for one HBS group in Supplementary Table S1). All HBS groups from 9.8–10.8 up to 17.8–18.8 showed significantly less negative ΔHBS (ΔHBS < −1) than expected (p-value < .05). Both remaining HBS groups from 18.8 to 19.8 and from 19.8 to 20.8, however, showed no significant shifts from the expected ΔHBS frequency (Supplementary Fig. S3 and Supplementary Table S2). Thus, stronger 5′ss (HBS ≥ 18.8) seem to allow HBS reduction to a higher extent because even after moderate HBS reduction, they can still ensure sufficient "correctly" spliced transcripts and thus functional protein. This observation is further emphasized, when putting the observed HBS difference in rela-

Table 2. Groups of 5′ss belonging to TSL1- and/or protein-coding transcript isoforms or not, showing heterozygous sequence variations within the 11 nucleotides of the 5′ss sequence

	Not TSL1	TSL1	Total
Not protein coding	6173 (7.4% of 83 910)	1001 (8.7% of 11 404)	7174 (7.5% of 95 314)
Protein coding	3082 (7.1% of 42 920)	7200 (4.1% of 173 199)	10 282 (4.8% of 216 119)
Total	9255 (7.3% of 126 830)	8201 (4.4% of 184 603)	17 456 (5.6% of 311 433)

5′ss could either be (i) exclusively part of non-coding transcript isoforms ("not protein coding") or (ii) part of at least one protein coding transcript ("protein coding"). Additionally, 5′ss were classified whether they were found (i) exclusively in transcripts of TSL level > 1 ("not TSL1") or (ii) in at least one TSL1 transcript ("TSL1"), with TSL1 transcripts representing high-confidence annotated transcripts that are found with the all same exon-junctions in the ensembl-pendant RefSeq.

tion to the maximal increase or decrease that is achievable by only one SNV within the sequence of the reference 5′ss. Across 5′ss categories, we saw a significant correlation between HBS of the reference 5′ss and a decrease in the relative HBS difference, that is theoretically possible by single nucleotide substitutions (Fig. 2C) (Spearman's rank correlation coefficient = −0.34, p-value < .001).

One set of 5′ss that was described to exhibit higher 5′ss strength compared to other 5′ss, are 5′ss of micro exons (exons of length ≤30 nucleotides) [23]. Following the observation from above, we would expect a higher degree of HBS variations for these 5′ss. Indeed, 5′ss exclusively found at micro exons showed significantly higher HBSs than 5′ss of longer exons, as well as stronger differences upon sequence variation (Supplementary Fig. S4, p-value of Kolmogorov–Smirnov test < .001). This further emphasizes the observation that stronger 5′ss seem to show HBS reductions to a higher degree than weaker 5′ss.

Analyzing which 5′ss sequence positions were the most affected by sequence variations revealed an increase in sequence variations with increasing distance to the GT dinucleotide into the respective intron (Supplementary Fig. S5A). Either 5′ss position +7 or +8 was affected in around 31% of all homozygous 5′ss variations. Since the widely used MaxEntScan score does not consider the last two 5'ss positions (+7 and + 8) in its 5′ss score calculation, we calculated the MaxEntScan score change of every homozygous 5′ss variation, excluding 5′ss that only differed to the reference 5′ss in these positions, reducing the number of unique affected 5′ss from 4044 to 2908 (Supplementary Fig. S5B). Interestingly, in only 0.06% of protein-coding 5′ss with homozygous variations, the MaxEntScan score did not change, whereas in the same dataset around 42% still showed no HBS change at all. However, for 40% of 5′ss, the difference in MaxEntScan score did not exceed a decrease of −1 or an increase of +1. The reason for the difference between the two scores might be that the HBS specifically assesses the number and density of H-Bonds between the potential 5′ss and the 5′ end of the U1 snRNA over 11 nucleotides. Mutations that do not affect the estimated formation of H-Bonds or H-Bond density thus not change the HBS but could potentially still lead to changes in MaxEntScan scores, which are based on the maximum entropy principle and consider both adjacent and non-adjacent dependencies between positions within a 9-nucleotide sequence. However, previous studies have shown, that complementary at 5′ss positions +7 and/or +8 is essential for some 5′ss [24], since positions +7 and +8 aid at exon recognition [5, 25–27]. Consequently, we chose the HBS for our analysis, as it accounts for these two positions.

Next, we analyzed the ΔHBS of canonical 5′ss of genes that have previously been described to be essential across human cell lines [28]. We expected no drastic changes in 5′ss strength for these genes, since the haplotyped assemblies came from healthy individuals. Indeed, most 5′ss were unaffected with only 19 (0.3%) of 5826 canonical 5′ss showing homozygous sequence variations within the 5′ss sequence, compared to 0.7% for canonical 5′ss of other genes. 5′ss of essential genes additionally showed significantly more instances with ΔHBS of 0, with 56% of 337 instances, compared to 51% for "non-essential" genes (p-value of Kolmogorov–Smirnov Test < .001). However, 8 unique 5′ss of canonical 5′ss of essential genes still showed stronger HBS differences to the reference ranging from −3.2 to even −5.3. This emphasizes that, although global analysis of physiologically occurring variation in 5′ss strength can give us an idea of the general acceptable variance, 5′ss strength variation introduced by a sequence variation has to be carefully analyzed in the respective genetic context of a given gene. For instance, one 5′ss belonging to the essential ribosomal protein RPS9, showed a ΔHBS of −5.3 in five donors. However, since it is the 5′ss of the first non-coding exon, reductions in recognition of this particular 5'ss might not be as of importance than recognition of coding exons further downstream.

Heterozygous 5′ss sequence variations

The number of 5′ss with heterozygous variations was understandably higher than the number of homozygous variations. From 311 433 5′ss annotated in *ensembl* (version 105), 17 456 (5.6%) showed a heterozygous 5′ss sequence variation across the 44 individuals. Of the 17 456 5′ss, 10 623 (60.1%) showed a heterozygous variation in multiple individuals (ranging from 2 to 44), further suggesting that these variations may not occur randomly but rather appear predominantly within the same transcripts and 5′ss. Moreover, 22.5% of these cases exhibited the same HBS difference across all affected individuals.

Grouping these splice donor sites by the above defined "TSL1" and "protein coding" categories, we found that 5′ss belonging to protein-coding TSL1 transcript isoforms again showed the smallest percentage of affected annotated 5′ss (Table 2). Generally, the percentage of affected unique 5′ss seemed to be on average four times larger than the number of 5′ss affected by homozygous variations.

As observed for homozygous HBS differences, 5′ss belonging to protein-coding TSL1 transcript isoforms showed the smallest proportion of affected 5′ss, relative to the total number of annotated 5′ss in each category (Table 2). We again checked the observed HBS differences per category, splitting 5′ss into a large set, where one allele is still the same as the reference (17 451 unique 5′ss), and one much smaller set, where both alleles were different from the reference, but not the same (92 unique 5′ss). We then compared the observed HBS differ-

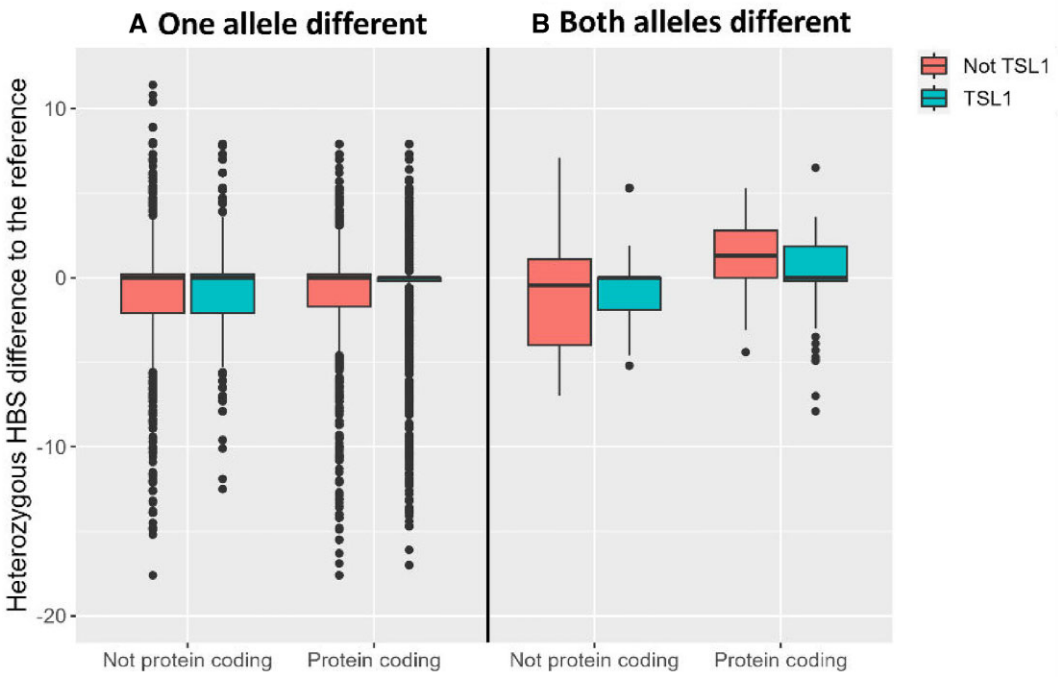

Figure 3. Differences in 5′ss strength to the reference 5′ss of 5′ss that are affected by heterozygous sequence variations within the 11 nucleotides of the 5′ss sequence by TSL and "protein coding" category. Boxplots depicting the HBS differences to the reference 5′ss per category, where either only one allele (**A**) or both alleles (**B**) are different from the reference 5′ss sequence. 5′ss were grouped into those that were (i) exclusively part of non-coding transcript isoforms ("not protein coding") or (ii) part of at least one protein coding transcript ("protein coding") and those, either found (i) exclusively in transcripts of TSL level > 1 ("not TSL1") or (ii) in at least one TSL1 transcript ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts that are found with the all same exon-junctions in the ensembl-pendant RefSeq.

ences of both datasets (Fig. 3) to the distribution of HBS differences found in homozygous sequence variations (Fig. 1A).

5′ss with only one allele different from the reference (Fig. 3A) showed a shift towards weaker HBSs compared to homozygous situations, although the median ΔHBS still lies at around zero in the heterozygous data. This trend to an HBS reduction was a little less expressed for protein-coding transcripts. The generally stronger reductions in HBSs observed in comparison to homozygous variations may be attributed to the presence of a remaining reference allele without HBS reduction (Kolmogorov–Smirnov test p-value < .001). This reference allele could compensate for the variation, suggesting that its unaffected transcript levels might, in some cases, be sufficient to prevent disease. 5′ss with heterozygous 5′ss differences in both alleles (Fig. 3B) showed a significant HBS increase for 5′ss of protein-coding transcripts compared to homozygous variations or heterozygous variations affecting only one allele (p-value of Kolmogorov–Smirnov test < .001). As expected, variations that introduce HBS reductions thus seem to be generally more prevalent in non-coding transcripts and for heterozygous variations, since they potentially do not disrupt protein-coding potential to an extent that is pathological.

Next, we focused in more detail on heterozygous 5′ss variations of protein-coding TSL1 transcripts, which again showed a similar HBS difference distribution compared to homozygous variations (Fig. 4). Single-allele changes again seemed to rather preserve the exact HBS as the reference, with 57% showing an H-bond score difference of 0, compared to 51% for homozygous variations. This could indicate that there might be similar evolutionary pressure to preserve the 5′ss strength for both heterozygous and homozygous 5′ss sequence variations.

However, when comparing the additive or compensatory effect on the HBS per SNV within 5′ss affected by multiple SNVs, heterozygous 5′ss variations showed a significantly lower number of instances where the single effect on the HBS per variation counteracted each other, compared to homozygous 5′ss variations. Here, we selected 648 unique heterozygous and 342 homozygous instances, where a 5′ss sequence showed multiple SNVs. For every 5′ss variation, we determined the maximal absolute HBond sore change, induced by only one of the respective multiple SNVs and then compared, how all remaining SNVs further effected the resulting total HBS difference (Supplementary Fig. S6A and B). We classified the instances into six groups, namely 5′ss, where the maximal absolute HBS difference of a single SNV led to (i) an accumulative HBS increase, (ii) an HBS increase that was partially compensated, and (iii) an HBS increase that was completely compensated to HBS 0 or even overcompensated to an HBS decrease, or iv–vi) the equivalent groups for HBS reductions (Supplementary Fig. S6C). Heterozygous 5′ss variations showed a higher amount of variations that resulted in a cumulative total decrease in HBSs (Supplementary Fig. S6, p-value of Chi-squared test < .001).

Since the SSHW previously showed to be different but less distinctive for silent or used 5′ss [22], we next measured SSHW differences to the reference 5′ss, expecting more variance in detected SSHW variations compared to ΔHBS distributions.

Homozygous variations in proximity of annotated 5′ss sequences

Homozygous sequence variations were also found in the neighboring sequence segment of ±60 nucleotides around an-

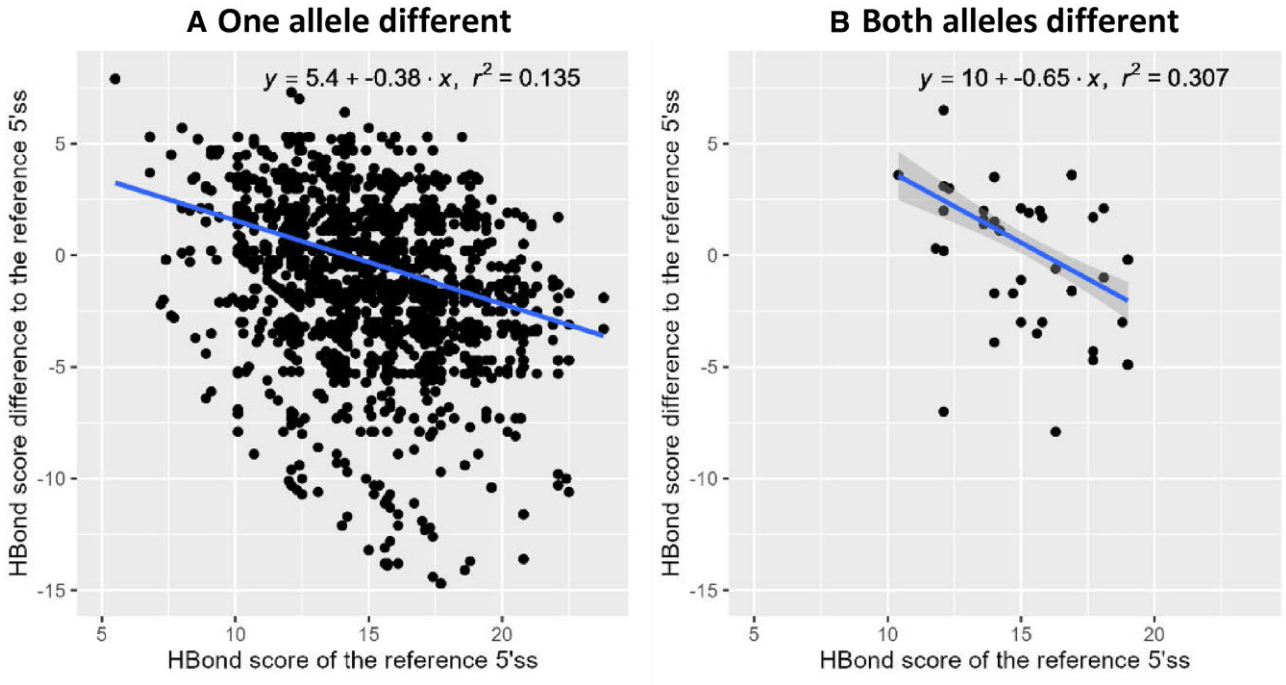

Figure 4. Observed HBS difference range depends on HBS of the reference 5′ss. Depicted is the HBS difference, caused by heterozygous genetic variations from the reference, within the sequence of annotated 5′ss of either one (**A**) or both alleles (**B**). 5′ss that were much more negative than the reference 5′ss were increasingly found for 5′ss of average strength (HBS). 5′ss of relatively high strength showed strong reductions in HBS compared to the reference. Each regression line is depicted with its formula stated above. The 95% confidence interval of the regression line is marked by a dark gray area. R^2 as a measure of how close the single data points are to the regression line is 0.13 (**A**) and 0.11 (**B**), respectively.

Table 3. TSL1 and protein-coding labels for every 5′ss that showed homozygous sequence variations within the ±60 nucleotides sequence surrounding of the 5′ss sequence

	Not TSL1	TSL1	Total
Not protein coding	16 750 (20.0% of 83 910)	2596 (22.7% of 11 404)	19 346 (20.3% of 95 314)
Protein coding	9128 (21.3% of 42 920)	30 842 (17.8% of 173 199)	39 970 (18.5% of 216 119)
Total	25 878 (20.4% of 126 830)	33 438 (18.1% of 184 603)	59 316 (19.0% of 311 433)

5′ss could either be (i) exclusively part of non-coding transcript isoforms ("not protein coding") or (ii) part of at least one protein-coding transcript ("protein coding") and those either found (i) exclusively in transcripts of TSL level > 1 ("not TSL1") or (ii) in at least one TSL1 transcript ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts that are found with the all same exon junctions in the ensembl-pendant RefSeq.

notated 5′ss. With 59 316 (19%) of the 311 433 *ensembl* (version 105) annotated 5′ss, a much higher number of 5′ss was affected than by variations within the 5′ss itself. This might be due to the much longer sequence segment, comparing a 120 nt long sequence (2 × 60 nt) with an 11 nt long sequence (Table 3).

With 17.8%, we again saw the smallest fraction of occurrences within 5′ss of the categories protein coding & TSL1, whereas 5′ss of the other categories showed a slightly higher fraction of variations. Overall, homozygous variations in the ±60 nt window around splice sites affect significantly more unique 5′ss than variations of the splice site sequence itself. This might be due to (i) the larger sequence of interest, which increases the probability to harbor sequence variations, and (ii) due to sequence variations resulting in a much lower impact on splice site usage than if positioned directly within the 5′ss sequence.

As with variations within the 5′ss sequence itself, we also analyzed the SSHW difference distribution across the 5′ss categories (Fig. 5A). To analyze variability in 5′ss SSHW, we calculated the SSHW difference (ΔSSHW) by subtracting the SSHW of the reference 5′ss from the SSHW of an affected 5′ss and compared the ΔSSHW distribution across the different 5′ss categories. Across all categories, ΔSSHW seemed to be similarly distributed around ΔSSHW of zero. ΔSSHW of 5′ss from protein-coding TSL1 transcripts ranged from −250 to +250 with a median of ∼0 (Fig. 5B). This broad variance around zero emphasizes again that 5′ss neighborhoods seem to be less subjected to evolutionary pressure than 5′ss. Applying the Anderson–Darling test for normality, however, showed that the SSHW difference values were still not normally distributed, testing either the total or unique 5′ss counts ($p_{\text{total}} < .05$, $p_{\text{unique}} < .05$). More extreme SSHW differences seem to be overrepresented.

Defining strong SSHW differences as lower −55 or greater 55, we selected the outer most 10% of the SSHW difference values, which are located at the two tails of the ΔSSHW distribution. Starting from 30 842 5′ss only found within coding TSL1 transcripts that had variations in the ±60 nt neighborhood, we generated a set of 6840 5′ss (21%) with stronger ΔSSHW. We subsequently analyzed whether stronger SSHW differences might correlate with reference SSHW value of the respective 5′ss, the allele-specific 5′ss strength, or changes of 5′ss strength. To compare reference SSHW with associ-

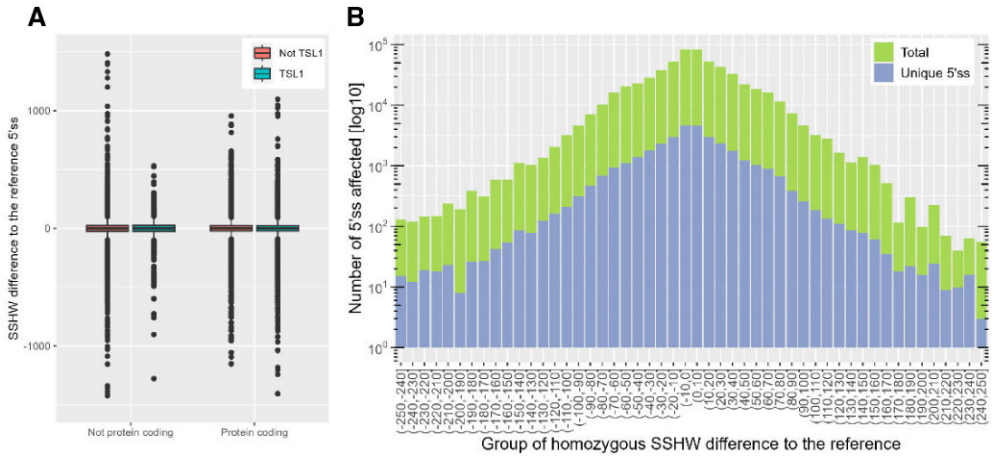

Figure 5. Homozygous SSHW differences to the reference 5′ss via sequence variations within the ±60 nucleotide window around the 5′ss sequence grouped by TSL and "protein coding" category. (**A**) Boxplot depicting the SSHW differences to the reference 5′ss per category. 5′ss were grouped into those that were part of protein-coding transcripts or not part of protein-coding transcripts and whether the 5′ss was part of TSL1 transcripts or not part of TSL1 transcripts, with TSL representing the confidence of splice site annotation by comparison with the RefSeq archive. (**B**) Histogram of SSHW differences to the reference 5′ss upon homozygous sequence variation within the ±60 nucleotide sequence window around 5′ss, exclusively found in protein-coding TSL1 transcripts. Depicted in red are the 5′ss counts per SSHW-difference group, counting affected 5′ss across all individuals. Depicted in blue are the counts of unique 5′ss per SSHW-difference group.

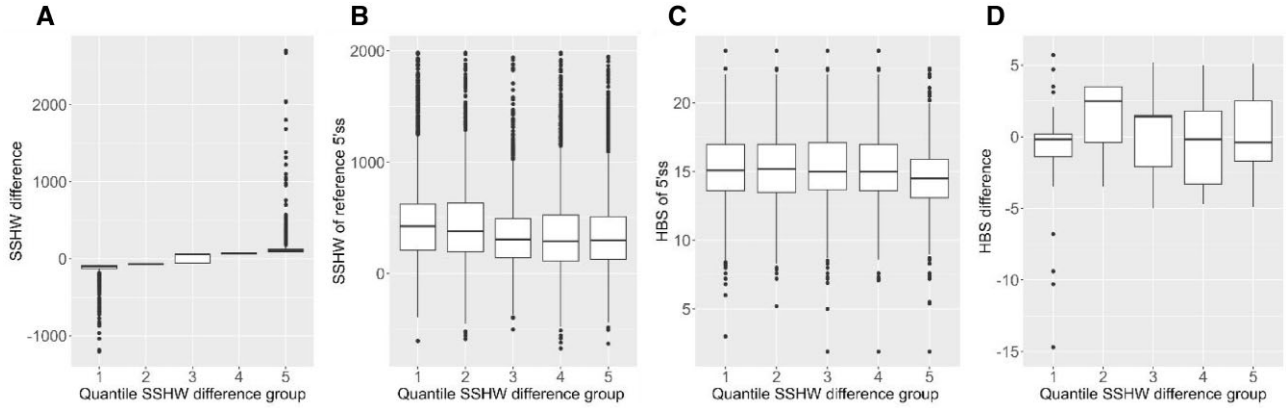

Figure 6. Strong homozygous SSHW differences to the 5′ss reference and corresponding factors. SSHW differences were grouped in five equally sized quintile groups starting from strong negative SSHW differences in the first quintile to strong positive SSHW differences in the fifth quintile. (**A**) SSHW difference groups of 5′ss with SSHW difference lower −55 or greater 55. (**B**) Boxplot depicting the reference SSHW of 5′ss affected by SSHW differences, showing a tendency for strong negative SSHW differences, to occur more often at 5′ss with a higher SSHW reference baseline as strong positive SSHW differences. (**C**) Boxplot depicting the reference HBond score of 5′ss affected by strong SSHW differences, showing prevalence of very strong SSHW increases to occur at slightly weaker 5′ss. (**D**) Boxplot depicting the HBond-score difference of 5′ss affected by strong SSHW differences of 5′ss, whose strength was altered by sequence variations. A slight tendency of strong negative SSHW differences can be detected to occur with negative HBond-score differences, which, however, is not supported by enough data points to be significant.

ated stronger SSHW differences, we binned ΔSSHW into five equally large groups (quintiles), with lowest ΔSSHW starting in group 1 and ΔSSHW greater 0 starting in group 3 with a mean ΔSSHW of 5 (Fig. 6A).

Similar to the negative linear correlation between HBS differences and reference HBS, we saw that negative ΔSSHW values from −1500 to −65 (belonging to quintile 1–2) seemed to occur more often in 5′ss with a slightly stronger reference SSHW baseline (reference SSHW average around 470), than ΔSSHW greater −60 (belonging to quintile 3–5 with a reference SSHW average of 352) (Fig. 6B). This observation was slightly less pronounced including heterozygous SSHW differences. Analyzing the allele-specific 5′ss strength per ΔSSHW quintile revealed no observable correlation (Fig. 6C).

Looking at a very small specific group of 5′ss that showed nucleotide exchanges within the ±60 nt neighborhood as well as within 5′ss, we saw a slight tendency of stronger negative ΔHBS to predominantly occur at 5′ss that showed higher ΔSSHW, whereas 5′ss groups with strong negative ΔSSHW seemed to almost exclusively harbor 5′ss, whose strength was not affected by the nucleotide exchange within the 5′ss sequence (Fig. 6D). It has to be emphasized that this particular dataset, due to its strict selection, however, only consisted of 1018 cases, affecting an even lower number of only 94 unique 5′ss (13–27 per quintile). The correlation between increasing ΔSSHW and decreasing ΔHBS is not significant.

Finally, we also analyzed canonical 5′ss of essential genes again to see which SSHW changes still seemed to be tolerated. Since the impact of the SSHW on 5′ss usage is determined by

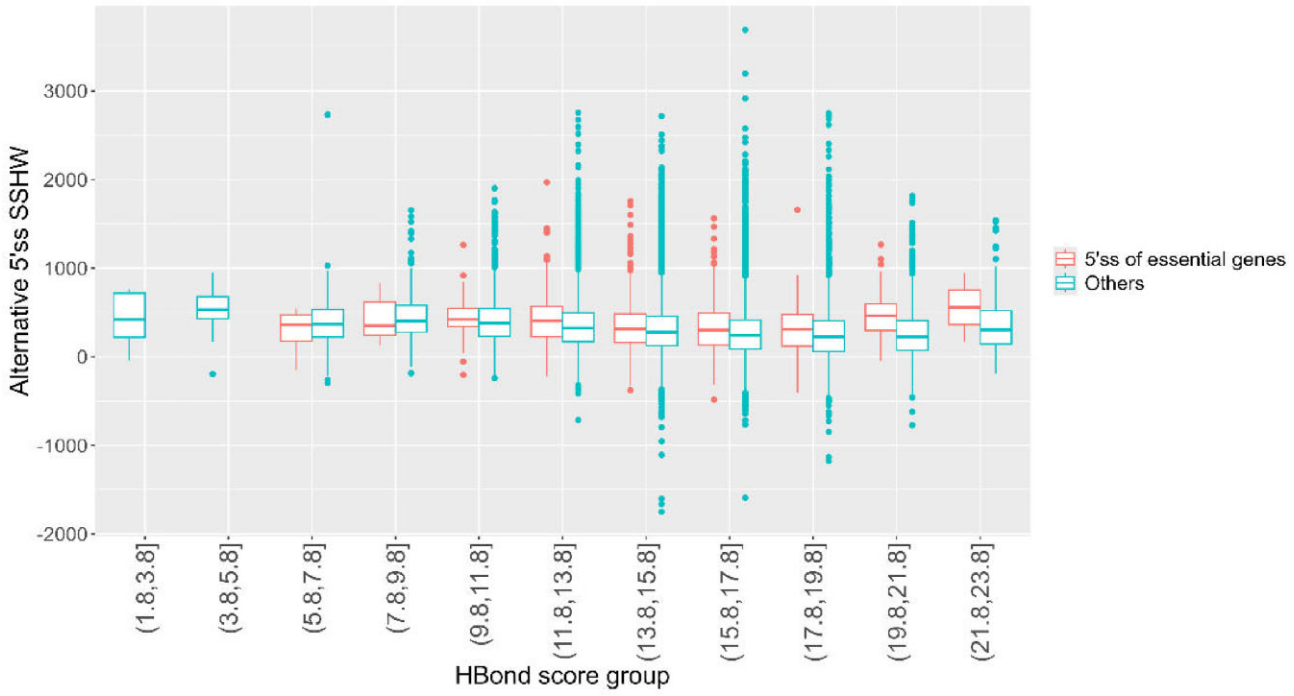

Figure 7. Alternative SSHW per HBS group of canonical 5′ss of essential genes. Depicted are the SSHW distributions per HBS group for a set of 5′ss from a set of 519 essential genes in human. In 15 572 instances, genomes showed a sequence variation within the genomic neighborhood of these 5′ss.

the HBS of the 5′ss sequence, we analyzed the SSHW upon sequence variation (Fig. 7).

As expected, the general need for a high SSHW to ensure 5'ss usage decreased with an increasing 5'ss strength (HBS), leveling out at around HBS of 16. The SSHW of constitutive 5′ss from other genes was significantly lower for 5′ss of HBS 11.8 to 17.8 (p-value of two-sided t-test $< .05$). This might indicate that a decrease in SSHW of essential 5′ss within this HBS range should be considered during genetic screening in the future to improve risk assessment for genetic diseases.

Conclusion

In this work, we analyzed sequence variations within the immediate vicinity of annotated human exon–intron borders and their impact on splice site strength and the predicted binding profile of SRPs, within a curated subset of high-fidelity genomes derived from 44 participants of the 1000 Genome Project [18]. First, a custom R-script mapped sequence variations within a 60-nucleotide window around annotated 5′ss to the reference sequence, allowing us to quantify changes in splice site strength and SSHW due to these variations within the respective genomic region.

A critical aspect of our study was the recognition that not all 5′ss annotated in transcript isoforms hold equal importance in terms of functional protein expression. In instances where 5′ss are exclusive to non-coding transcript isoforms, their impact on protein expression is minimal or negligible. In contrast, 5′ss featured in protein-coding transcripts significantly influence the expression of at least some splice isoforms. We additionally considered the confidence level associated with annotated transcripts, leveraging the TSL scale provided by the Ensembl archive, which ranges from 1 (highest confidence) to 5. This scale allowed us to distinguish between transcripts that were robustly supported and those with lower levels of confidence.

Homozygous sequence variations within the 5′ss sequence itself were relatively rare, occurring in just around 1.3% of annotated 5′ss sites, which reflects previous observations that splice site sequences seem to be conserved sometimes even across species [29]. Crucially, over half of these sites exhibited variations in multiple individuals, with most displaying consistent differences in H-Bond scores across all affected individuals. Heterozygous sequence variations within the 5′ss sequence itself were somewhat more frequent, occurring in around 5.6% of annotated 5′ss sites. The evolutionary pressure governing the tolerance of heterozygous mutations within the splice site sequence, therefore, might be somewhat less stringent than in the case of homozygous mutations, particularly in protein-coding transcripts. Interestingly, for both homozygous and heterozygous sequence variations within 5′ss sequences, around one half did not result in changes of predicted 5′ss strength. Since for some proteins, as for instance the β-globin gene in sickle cell anemia, a reduced expression is not *per se* pathogenic [30], we would have expected this percentage to be significantly lower for heterozygous variations, since here one allele would still remain intact [31].

Grouping 5′ss into categories based on coding potential (protein coding versus not protein coding) and TSL (TSL1 versus not TSL1) revealed that protein-coding TSL1 transcripts showed the lowest incidence of affected sites when normalized to the total number of annotated 5′ss in each category. We further focused on this particular subset of 5′ss since they are of special importance and annotated with high accuracy.

Analyzing the interplay between HBSs of reference 5′ss and detected HBS differences due to homozygous sequence variations, we saw that 5′ss with HBS lower 18.8 showed a clear prevalence for sequence variations that do not decrease 5′ss strength, compared with what would be randomly expected. 5′ss of HBS >18.8 showed randomly distributed increase or

decrease of 5′ss strength, indicating that variant-induced differences in strength have to be taken into consideration with the 5′ss baseline.

Widening the window around 5′ss, we also analyzed sequence variations within the ±60 nt window around splice sites, excluding the splice site sequence itself. This somewhat arbitrary sequence neighborhood was repeatedly described to hold an extensive amount of predicted splicing regulatory elements [32, 33]. Sequence variations within these regions were quite frequent, with 19% of annotated 5′ss sites affected. Like with HBS differences, SSHW reductions by sequence variations were slightly enriched in 5′ss with a higher SSHW baseline around 400.

Large-scale CRISPR and RNAi screens like The Cancer Dependency Map (DepMap) [34] and Project Score [35] investigate the impact of gene knockouts on cancer cell line fitness. These projects assign scores to indicate cell growth inhibition or death, but essentiality scores can vary across cell lines. Sharma *et al*. identified a cluster of genes highly likely to be essential across all analyzed cell lines from the mentioned projects, assigning 942 and 650 genes to this core essential gene set, with 519 genes shared between both projects. For 5′ss of the main transcript variant of these 519 genes, we expected significantly less variation in HBSs and SSHW. Indeed, homozygous sequence variations within the 5′ss sequence of essential genes were only half as common as homozygous variations within other 5′ss (0.3%) and additionally showed a higher amount of variations that did not alter the HBS at all. The SSHW resulting from sequence variations was still significantly higher for 5′ss of HBS 11.8–17.8, compared to 5′ss of other genes.

However, determining a general threshold for HBS or SSHW changes with a high probability to disrupt correct splicing remains challenging. Since around 61% of homozygous 5′ss variations did not decrease or increase the HBS by >1, one could assume that sequence variants of this category might show a quit low capacity to significantly affect splice site usage. Though in some cases, mild changes in splice site usage are able to induce disease, like the Hutchinson-Gilford progeria syndrome, which is caused by sequence variations that activate a cryptic exonic 5′ss, leading to aberrant splicing of the LMNA pre-mRNA and production of a protein isoform that misses 50 amino acids near the carboxy terminus. Although functional protein still makes up the majority of LMNA expression, the amount of pathogenic LMNA protein is sufficient to induce disease. We could show that physiologically occurring sequence variations that reduce the HBS are correlate with the HBS base level of the respective 5′ss and whether they occur on both or only one allele. However, even homozygous variations within 5′ss of essential genes showed strong HBS reductions by up to −5.3. Some of these strong HBS differences could be explained by a high HBS base level of the 5′ss or the genetic context. The change in for instance splice site strength thus has to be carefully taken into account with various additional factors as (i) base level splice site strength, (ii) the SSHW, (iii) the genetic context, (iv) exon length and potential competing splice sites in proximity, or even cell-type dependent expression of SRPs.

Acknowledgements

We thank Tobias Marschall, Jana Ebler, and Peter Ebert for helping us find a suitable dataset.

Author contributions: J.P. performed the analysis and the visualization of the results. H.S. supervised the project and advised the analysis together with S.T. The manuscript was written by J.P. with contributions and reviews from all the authors.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

We acknowledge funding by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; Netzwerk Universitätsmedizin, GenSurv/MolTraX), award number 01KX2021.

Data availability

Code and data to reproduce the analysis are available at zonodo.org under DOI 10.5281/zenodo.14730589 (https://zenodo.org/records/14730589).

References

1. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *ProcNatl Acad Sci U S A* 1977;**74**:3171–75. https://doi.org/10.1073/pnas.74.8.3171
2. Khodor YL, Rodriguez J, Abruzzi KC *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev* 2011;**25**:2502–12. https://doi.org/10.1101/gad.178962.111
3. Aebi M, Hornig H, Padgett RA *et al.* Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell* 1986;**47**:555–65. https://doi.org/10.1016/0092-8674(86)90620-3
4. Lord J, Baralle D. Splicing in the diagnosis of rare disease: advances and challenges. *Front Genet* 2021;**12**:689892. https://doi.org/10.3389/fgene.2021.689892
5. Freund M, Asang C, Kammler S *et al.* A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res* 2003;**31**:6963–75. https://doi.org/10.1093/nar/gkg901
6. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;**11**:377–94. https://doi.org/10.1089/1066527041410418
7. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 2008;**14**:802–13. https://doi.org/10.1261/rna.876308
8. Dogan RI, Getoor L, Wilbur WJ *et al.* SplicePort—an interactive splice-site analysis tool. *Nucleic Acids Res* 2007;**35**:W285–91. https://doi.org/10.1093/nar/gkm407
9. Erkelenz S, Theiss S, Otte M *et al.* Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res* 2014;**42**:10681–97. https://doi.org/10.1093/nar/gku736
10. Brillen AL, Schoneweis K, Walotka L *et al.* Succession of splicing regulatory elements determines cryptic 5′ss functionality. *Nucleic Acids Res* 2017;**45**:4202–16. https://doi.org/10.1093/nar/gkw1317
11. More D.A., Kumar A. SRSF3: newly discovered functions and roles in human health and diseases. *Eur J Cell Biol* 2020;**99**:151099. https://doi.org/10.1016/j.ejcb.2020.151099

12. Geuens T, Bouhy D, Timmerman V. The hnRNP family: insights into their role in health and disease. *Hum Genet* 2016;**135**:851–67. https://doi.org/10.1007/s00439-016-1683-5
13. Grzybowska EA. Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem Biophys Res Commun* 2012;**424**:1–6. https://doi.org/10.1016/j.bbrc.2012.06.092
14. Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;**42**:D980–5. https://doi.org/10.1093/nar/gkt1113
15. Sherry ST, Ward MH, Kholodov M *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**:308–11. https://doi.org/10.1093/nar/29.1.308
16. Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009;**106**:9362–7. https://doi.org/10.1073/pnas.0903103106
17. Liao WW, Asri M, Ebler J *et al.* A draft human pangenome reference. *Nature* 2023;**617**:312–24. https://doi.org/10.1038/s41586-023-05896-x
18. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53. https://doi.org/10.1126/science.abj6987
19. Porubsky D, Ebert P, Audano PA *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* 2021;**39**:302–8. https://doi.org/10.1038/s41587-020-0719-5
20. Ptok J, Theiss S, Schaal H. VarCon: an R package for retrieving neighboring nucleotides of an SNV. *Cancer Inform* 2020;**19**:1176935120976399. https://doi.org/10.1177/1176935120976399
21. O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45. https://doi.org/10.1093/nar/gkv1189
22. Muller L, Ptok J, Nisar A *et al.* Modeling splicing outcome by combining 5′ss strength and splicing regulatory elements. *Nucleic Acids Res* 2022;**50**:8834–51. https://doi.org/10.1093/nar/gkac663
23. Ustianenko D, Weyn-Vanhentenryck SM, Zhang C. Microexons: discovery, regulation, and function. *Wiley Interdiscip Rev RNA* 2017;**8**. https://doi.org/10.1002/wrna.1418
24. Hartmann L, Theiss S, Niederacher D *et al.* Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? *Front Biosci* 2008;**13**:3252–72. https://doi.org/10.2741/2924
25. Hibbert CS, Gontarek RR, Beemon KL. The role of overlapping U1 and U11 5′ splice site sequences in a negative regulator of splicing. *RNA* 1999;**5**:333–43. https://doi.org/10.1017/S1355838299981347
26. Kammler S, Leurs C, Freund M *et al.* The sequence complementarity between HIV-1 5′ splice site SD4 and U1 snRNA determines the steady-state level of an unstable env pre-mRNA. *RNA* 2001;**7**:421–34. https://doi.org/10.1017/S1355838201001212
27. Lund M, Kjems J. Defining a 5′ splice site by functional selection in the presence and absence of U1 snRNA 5′ end. *RNA* 2002;**8**:166–79. https://doi.org/10.1017/S1355838202010786
28. Sharma S, Dincer C, Weidemuller P *et al.* CEN-tools: an integrative platform to identify the contexts of essential genes. *Mol Syst Biol* 2020;**16**:e9698. https://doi.org/10.15252/msb.20209698
29. Abril JF, Castelo R, Guigo R. Comparison of splice sites in mammals and chicken. *Genome Res* 2005;**15**:111–9. https://doi.org/10.1101/gr.3108805
30. Williams TN, Thein SL. Sickle cell anemia and its phenotypes. *Annu Rev Genomics Hum Genet* 2018;**19**:113–47. https://doi.org/10.1146/annurev-genom-083117-021320
31. Bartha I, di Iulio J, Venter JC *et al.* Human gene essentiality. *Nat Rev Genet* 2018;**19**:51–62. https://doi.org/10.1038/nrg.2017.75
32. Ke S, Shang S, Kalachikov SM *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 2011;**21**:1360–74. https://doi.org/10.1101/gr.119628.110
33. Zhang XH, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* 2004;**18**:1241–50. https://doi.org/10.1101/gad.1195304
34. Tsherniak A, Vazquez F, Montgomery PG *et al.* Defining a Cancer Dependency Map. *Cell* 2017;**170**:564–76. https://doi.org/10.1016/j.cell.2017.06.010
35. Dwane L, Behan FM, Goncalves E *et al.* Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets. *Nucleic Acids Res* 2021;**49**:D1365–72. https://doi.org/10.1093/nar/gkaa882

Received: September 27, 2024. **Revised:** February 18, 2025. **Editorial Decision:** March 18, 2025. **Accepted:** March 20, 2025