8834–8851 *Nucleic Acids Research, 2022, Vol. 50, No. 15* *Published online 10 August 2022*
https://doi.org/10.1093/nar/gkac663

Modeling splicing outcome by combining 5′ss strength and splicing regulatory elements

Lisa Müller[1,†], Johannes Ptok[1,†], Azlan Nisar[1,2], Jennifer Antemann[1], Ramona Grothmann[1], Frank Hillebrand[1], Anna-Lena Brillen[1], Anastasia Ritchie[1], Stephan Theiss[1,*] and Heiner Schaal[1,*]

[1]Institute of Virology, Medical Faculty, Heinrich-Heine-University Düsseldorf, Düsseldorf 40225, Germany and [2]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, Recklinghausen 45665, Germany

Received November 30, 2021; Revised June 23, 2022; Editorial Decision July 12, 2022; Accepted July 27, 2022

ABSTRACT

Correct pre-mRNA processing in higher eukaryotes vastly depends on splice site recognition. Beyond conserved 5′ss and 3′ss motifs, splicing regulatory elements (SREs) play a pivotal role in this recognition process. Here, we present *in silico* designed sequences with arbitrary *a priori* prescribed splicing regulatory HEXplorer properties that can be concatenated to arbitrary length without changing their regulatory properties. We experimentally validated *in silico* predictions in a massively parallel splicing reporter assay on more than 3000 sequences and exemplarily identified some SRE binding proteins. Aiming at a unified 'functional splice site strength' encompassing both U1 snRNA complementarity and impact from neighboring SREs, we developed a novel RNA-seq based 5′ss usage landscape, mapping the competition of pairs of *high confidence* 5′ss and neighboring exonic GT sites along HBond and HEXplorer score coordinate axes on human fibroblast and endothelium transcriptome datasets. These RNA-seq data served as basis for a logistic 5′ss usage prediction model, which greatly improved discrimination between strong but unused exonic GT sites and annotated highly used 5′ss. Our 5′ss usage landscape offers a unified view on 5′ss and SRE neighborhood impact on splice site recognition, and may contribute to improved mutation assessment in human genetics.

INTRODUCTION

For almost all human primary protein coding transcripts recognition of splice sites, the borders between exons and introns, is key in deciphering their open reading frames. In order to accurately ligate exons after intron removal, splice sites at exon-intron-borders need to be recognized with single nucleotide precision during early assembly of the spliceosome. Splice site recognition depends upon conserved sequence motifs at both intron ends, and the first step in the splicing process is splice donor recognition by the U1 snRNP at a highly conserved GT dinucleotide (1).

Formation of an RNA duplex between up to 11 nucleotides (nt) of the splice donor (5′ss) with the 5′ end of U1 snRNA is a main determinant in 5′ss selection (2–4). The statistical likelihood of a 9 nt long potential 5′ss sequence being used as 5′ss is frequently quantified by its maximum entropy based MaxEnt (ME) score (5), while the HBond score (HBS) algorithm based on all 11nt quantifies the U1 snRNA complementarity of a potential 5′ss (https://www2.hhu.de/rna// (6,7)). However, exons and introns contain numerous GT sites with high MaxEnt and HBond scores indicating potential 5′ss, which under physiological conditions are not used as exon-intron-borders.

Thus, the proper 5′ss sequence cannot be the sole determinant of 5′ splice site use (8). The efficiency with which splice sites are recognized additionally depends on proximal *cis*-acting splicing regulatory elements (SREs) and their protein binding partners including SR (serine-arginine-rich) (9,10) and hnRNP (heterogeneous nuclear ribonucleoparticle) proteins (11,12). Generally, proteins bound by SREs act in a direction dependent way: SR proteins have enhancing properties on downstream located 5′ss and repress upstream located 5′ss, while hnRNP proteins act reversely (13,14). Mechanistically, splicing regulatory proteins (SRPs) may impact U1 snRNA duplex stability due to

*To whom correspondence should be addressed. Tel: +49 211 81 12393; Fax: +49 211 81 10856; Email: schaal@uni-duesseldorf.de
Correspondence may also be addressed to Stephan Theiss. Email: theiss@uni-duesseldorf.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present address: Anna-Lena Brillen, Institute of Virology, Faculty of Medicine, University of Bonn, Bonn 53127, Germany.

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

allosteric regulation of U1 snRNP structure (15). Through these combined SRP binding effects, the sequence neighborhood of a splice site can have a significant impact on splice site recognition and hence splicing efficiency (16–19). Especially with regard to an estimated at least 25% of human inherited diseases caused by mutations either directly altering splice sites or disrupting SREs in their vicinity (20,21), computational evaluation of a possibly pathogenic impact of individual SNVs is important for human genetics (22–27).

Various algorithms and corresponding computational tools have been developed and made publicly available to analyze splicing regulatory elements: some algorithms identify previously described hexamer or octamer motifs (e.g. ESEfinder, FAS-ESS, RESCUE-ESE, PESX, cf. e.g. (24,28)), others provide e.g. hexamer weights quantifying their splice enhancing or silencing properties, and enabling the calculation of SRE profiles in moving windows along genomic sequences (ESR-seq (29), HEXplorer (30)). Most recently, neural network or deep-learning based algorithms for splicing prediction have been developed that take splice sites and their neighborhoods or very wide sequence contexts into account (MMSplice (31), SpliceAI (32)) (4,33).

These algorithms have recently been complemented by an experimentally obtained database of RNA elements as part of the *Encyclopedia of DNA Elements* (ENCODE) project phase III. This dataset contains binding motifs for RNA-binding proteins, including splicing regulatory proteins (34).

Minigene splicing reporters are widely used model systems to experimentally examine splicing. In particular, massively parallel splicing assays (MPSA) permit screening the impact on splicing for a large number of randomly generated sequences in a single experiment. These random sequences can e.g. cover a 5′ss position, various specific exonic k-*mer* positions, or be spread out across an entire exon. For each individual 'input' sequence, an RNA-seq based enrichment index quantifies the sequence impact on splicing, the 'output' (3,29,35).

Here, we followed the inverse route of an *in-silico* design process for sequences with *a priori* prescribed splicing regulatory properties, represented by approximately constant HEXplorer profiles. We experimentally validated this HEXplorer guided design in an MPSA on more than 3000 sequences inserted between two competing 5′ss in a splicing reporter. Complementarily, we examined splice site competition in two large whole transcriptome RNA-seq datasets and derived a two-dimensional 5′ splice site usage landscape dependent on intrinsic 5′ss strength and SRE neighborhood. Introduction of a novel unified 5′ss score taking both factors into account improved discrimination accuracy between annotated 5′ss and exonic GT sites.

MATERIALS AND METHODS

Expression plasmids

pXGH5 (hGH) (36) was cotransfected to monitor transfection efficiency.

Oligonucleotides

All oligonucleotides used were obtained from Metabion GmbH (Planegg, Germany) (see Supplementary File S1).

Cloning

A reporter construct based on the HIV-1 glycoprotein/eGFP expression plasmid (6,13) as well as a 3-exon minigene based on the fibrinogen Bß subunit under the control of a cytomegalovirus immediate early (CMVie) promoter (37) were used in this study. All sequences were cloned using either PCR-products of the respective forward and reverse primer pairs or DNA fragments. Detailed cloning strategies and primer sequences can be found in Supplementary File S1.

Cell culture and RT-PCR analysis

HeLa cells (ATCC® CCL-2™, mycoplasma free) were cultivated in Dulbecco's high-glucose modified Eagle's medium (Gibco #41966) supplemented with 10% fetal calf serum (PAN Biotech #P30-3031) and 50 μg/ml penicillin and streptomycin each (Gibco #15140-122). Transient transfection experiments were performed with six-well plates at 2.5×10^5 cells per well by using TransIT®-LT1 transfection reagent (Mirus Bio LLC US #MIR2305) according to the manufacturer's instructions. Total RNA was isolated 24 h post-transfection by using acid guanidinium thiocyanate-phenol-chloroform as described previously (38). For (q)RT-PCR analyses, RNA was reversely transcribed by using Superscript III Reverse Transcriptase (Invitrogen #18080–085) and Oligo(dT) primer (Roche #10814270001). For the analyses of the splicing constructs either primer pair #3210/#3211(#640) or #2648/2649 was used and PCRs were separated on non-denaturing 10% polyacrylamide gels. Quantitative RT-PCR analysis was performed by using the qPCR MasterMix (PrimerDesign Ltd #PPLUS-CLSY-10ML) and Roche LightCycler 1.5. For normalization, primers #1224/#1225 were used to monitor the level of the transfection control hGH present in each sample.

Protein isolation by RNA affinity chromatography

Substrate RNAs were *in vitro* transcribed using theT7 RiboMaxTM Express Large Scale RNA Production System (Promega #P1320) according to the manufacturer's recommendations. Three thousand picomoles of the substrate RNA oligonucleotides for each octamer (+10.32 #5648, –0.15 #5647, –10.35 #5846) were covalently coupled to adipic acid dihydrazideagarose beads (Sigma #40802-10ML). 60% of HeLa nuclear extract (SKU: CC-01-20-50, Cilbiotech/now Ipracell #CC-01-20-50) was added to the immobilized RNAs. After stringent washing with buffer D containing different concentrations of KCl (20 mM HEPES–KOH [pH 7.9], 5% [vol/vol] glycerol, 0.1–0.5 M KCl, 0.2 M ethylenediaminetetraacetic acid, 0.5 mM dithiothreitol, 0.4M $MgCl_2$), precipitated proteins were eluted in protein sample buffer. Samples were heated up to 95°C for 10 min and either submitted to LC–MS/MS-analysis or loaded onto sodium dodecyl sulphate-polyacrylamide

gel electrophoresis (SDS PAGE) for western blot analysis. Samples were transferred to a nitrocellulose membrane probed with primary and secondary antibodies (SRSF3 (Abcam ab198291, 1:000), PTB (kind gift from Douglas Black, 1:1000), hnRNPD (Merck Millipore AUF-1 07-260, 1:1000), MS2 (Tetracore TC-7004-002, 1:1000), Goat anti-Rabbit IgG Superclonal™ Secondary Antibody (Invitrogen A27036, 1:2500) and developed with ECL chemiluminescence reagent (GE Healthcare #RPN2106).

HEXplorer score algorithm and splice site HEXplorer weight (SSHW)

Based on a RESCUE-type approach, the HEXplorer score HZ_{EI} is calculated from different hexamer occurrences in exonic and intronic sequences in the neighborhood of splice donors, and it has been successfully used for the identification of exonic splicing regulatory elements (30,37,39). Briefly, from 43 464 constitutively spliced human exons with canonical 5′ss collected from ENSEMBL (24), *Z*-scores for all 4096 hexamers were calculated from normalized hexamer frequency differences up- and downstream of weak and strong splice donors, ranging from −73 for TTTTTT to + 34 for GAAGAA.

The HEXplorer score HZ_{EI} of any index nucleotide in a genomic sequence is then calculated as average hexamer *Z*-score of all six hexamers overlapping with this index nucleotide. This algorithm permits plotting HEXplorer score profiles along genomic sequences, and these profiles reflect splice enhancing or silencing properties in the neighborhood of a splice donor: HEXplorer score positive regions support downstream splice donors and repress upstream ones, and HZ_{EI} negative regions *vice versa*. HEXplorer score profiles of genomic sequences were calculated using the web interface (https://www2.hhu.de/rna/html/hexplorer_score.php).

As measure of SRE impact on 5′ss recognition, we calculated the 5′ splice site HEXplorer weight SSHW as the total HZ_{EI} sum ($\sum_{up} HZ_{EI}$) in a 50 nt upstream minus the symmetrical 50 nt downstream neighborhood ($\sum_{dn} HZ_{EI}$) (37,40), excluding all 11 nt of the 5′ss from the HZ_{EI} calculation: the 50 nt wide neighborhoods ended at exonic position −4 and started at intronic position +9, respectively. This definition has been made analogous to the 'exonic splicing motif difference' ESMD introduced by Ke *et al.* and to the 'splice site enhancer weight' by Brillen *et al.* (37,40), and it captures both enhancing and silencing properties of 50 nt wide up- and downstream regions that have been used before and are plausibly considered to contain relevant SREs.

When comparing SSHW of pairs of exonic GT-sites and 5′ss, we carefully adapted the selection of appropriate neighborhoods depending on the GT-site-to-5′ss distance, excluding the 11 nt long proper 5′ss or exonic GT-site sequence: If GT-site and 5′ss were >60 nt apart, we used 50 nt wide neighborhoods A, B1, B2 and C as depicted in Supplementary Figure S4C. For pairs of GT-site and 5′ss that were between 61 nt and 111 nt apart, the neighborhoods B1 and B2 consequently overlapped. If GT-site and 5′ss were closer than 61 nt, we chose B1 = B2 as the entire region between but excluding the two sites. We then calculated the SSHW difference between GT site and 5′ss as $\Delta SSHW = (\sum_A - \sum_{B1} - \sum_{B2} + \sum_C) HZ_{EI}$ (Supplementary Figure S4C).

Mass spectrometric analysis

Protein samples were shortly separated over about 4 mm running distance in a 4–12% polyacrylamide gel. After silver staining, protein containing bands were excised and prepared for liquid chromatography–tandem mass spectrometry (LC–MS/MS) as described previously (37). *P*-values on the vertical axis of the volcano plot (Supplementary Figure S2A) give the probability that a given $\log_2$-fold change in protein binding detected by mass spectrometry may have occurred by chance. Smaller *P*-values correspond to more reliably detected protein binding differences. P-values do not only depend on the $\log_2$-fold change, but also on the absolute detection levels.

Preparation of octamer library

For the generation of the octamer library, a PCR fragment was generated with a primer containing a random 8-mer (#6576). PCR fragments were inserted into the respective backbone (see Supplementary File S1), and the plasmid library was amplified after transformation of *E. coli*. The library containing plasmids were then used for transfection, followed by RNA isolation and analysis via RT-PCR using primers #3210/#3211 and subsequent PAA-gel analysis. For amplicon sequencing, the desired band was excised and purified via the QIAquick Gel Extraction Kit (Qiagen #28704) and re-amplified using the same primers. For the sequencing of the plasmid library, plasmid DNA was amplified using primers #6654/#6655 and samples were purified via Monarch PCR & DNA Cleanup Kit (NEB #T1030L). NGS amplicon sequencing was carried out by Eurofins Genomics, Konstanz, Germany.

Sequencing of octamer libraries

The library was sequenced by the company Eurofins Genomics, using Illumina NovaSeq 6000 PE150, generating 9 917 080 and 8 418 350 reads for the plasmid sample (rev primer: #6654 and fwd primer: #6655) and the band sample (rev primer: #3211 and fwd primer: #6655), respectively.

Quantification of octamer frequencies

First, quality metrics of the reads, stored in FASTQ files, were assessed using the tools FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (https://academic.oup.com/bioinformatics/article/32/19/3047/2196507). Read pairs were corrected and merged using the bbmerge.sh script of the tool bbmerge (version 38.00) (41). Since human cells were transfected with the reporter construct, we aligned the reads against the reference genome and the reporter plasmid sequence simultaneously with STAR (version 2.5.4b) (42). From the reads of the gel-electrophoresis band sample, we selected only those reads, which showed usage of the downstream splice donor for further analysis, since those reads still hold the sequence within the octamer library.

The sequence of every read within the octamer library was determined, using regular expressions containing the flanking anchor sequences 5′: ATTGG upstream and 5′: CCTAT downstream of the octamer library (NNNNNNNN). Read pairs were discarded, when the determined octamer sequence was not identical in either the forward or reverse read, or when the anchor sequences could not be found, resulting in 1 002 322 reads from RNA fragments with downstream SD usage and 8 576 066 reads containing an octamer in plasmid sequencing data.

Single octamer detection frequencies in the octamer library were calculated from sequencing the transfected plasmids ('*input*') and the isolated band after gel-electrophoresis ('*output*'). The latter contained RNA fragments with usage of the downstream splice donor after transfection with the reporter plasmid. Octamer sequences more frequently found in the band indicate enhanced downstream splice donor usage. We calculated a normalized enrichment index (NEI) that quantifies octamer enrichment in the band relative to the plasmid input, corrected for different sample sequencing depth: $\mathrm{NEI} = (n_{\mathrm{band}}/n_{\mathrm{plasmid}})/(N_{\mathrm{band}}/N_{\mathrm{plasmid}})$, where n_{band} and n_{plasmid} denote the number of reads holding a given octamer in band or plasmid, whereas N_{band} and N_{plasmid} denote the total number of reads for the respective samples. To reduce the impact of technical fluctuations, we excluded octamers with $n_{\mathrm{band}} < 9$ reads and $n_{\mathrm{plasmid}} < 5$ reads.

RNA sequencing data generation and processing

We re-analyzed two RNA sequencing data sets: one originating from 46 samples of primary fibroblasts (previously described in (43)), and one from four samples of cardiovascular endothelial cells (18). Briefly, the cDNA libraries were created using TruSeq RNA SamplePrep kit (Illumina) after poly(A) enrichment according to the manufacturer's protocol. Afterwards, the samples were amplified on nine Illumina flow cells and sequenced on a Illumina HiSeq 2000 sequencer. Subsequently the resulting 101-nt sequence segments were converted to FASTQ by CASAVA (1.8.2). The samples were checked for base calling quality during sequencing, sub-sequences of a single read with low average base calling quality as well as left over adapters from library preparation were removed using Trimmomatic version 0.36 (44). Trimmed reads shorter than 75 bases were discarded since this length is an established threshold in the analysis concerning exon junctions (45). The tool sortMeRNA was used to validate complete rRNA removal during poly(A) RNA enrichment (46). Throughout the different steps of FASTQ file processing, the quality of the reads was assessed using the tools FASTQC and MultiQC. After processing the FASTQ files, the reads were mapped to the ENSEMBL human reference genome (version 91) using the STAR software package (2.5.4b). The reads were aligned to the reference following the two-pass mapping protocol recommended for splice site usage analysis (42,47). After alignment with STAR, the BAM files were summarized to a single gap file using CRAN package rbamtools (48) and Bioconductor package spliceSites (49). Additional packages were used during the analysis. FASTQ file preparation and alignment, as well as the first part of BAM file processing in R was accomplished using custom BASH shell scripts in the environment of the High Performance Computing Cluster of Heinrich-Heine University Düsseldorf. Computational support and infrastructure was provided by the 'Centre for Information and Media Technology' (ZIM) at Heinrich-Heine University Düsseldorf (Germany).

Gene-and-sample normalization of RNA-seq reads

Comparing RNA-seq reads across many genes from different samples requires careful normalization of reads and removal of potentially noisy read counts, which we address below.

In each human—46 fibroblast and 4 endothelium—sample, we separately collected (gapped) exon junction reads that had gap quality score gqs ≥400 and gap length <26 914 (95% of human introns are shorter) as described in (43,49). From here on, we denote such gapped reads detected at any given genomic site as '5′ss reads' on the corresponding 5′ splice sites, irrespective of *Ensembl* annotation.

The majority of genes were very reliably expressed in most samples. For the 46 fibroblast samples e.g. 12 850 genes (47.3%) containing 99.7% of all reads were detected in all 46 samples. The number of samples a gene was detected in followed a U-shaped distribution (Supplementary Figure S3A, black squares), and those genes detected in few samples each had very few reads. Genes detected in more samples also had more reads *per sample*, not just in total (Supplementary Figure S3A, gray bars).

Normalization of 5′ss reads then proceeded in three steps. In order to account for differential RNA-seq detection between samples, we normalized all 5′ss reads by the total number (in millions) of exon junction reads in each individual sample, obtaining sample normalized RPMG (*reads per million gapped reads*) values for the 5′ss usage in each sample.

In the second normalization step, we factored in differential gene expression in each sample. For each specific gene in a given sample, we determined the MRIGS (*maximum RPMG in gene and sample*) of the most used 5′ss in this gene as gene-expression measure. If genes with very few reads were detected in samples with an overall high level of technical RNA-seq read coverage (large sequencing library size), they may have been false-positive detections due to RNA-seq technique limitations, and could be identified by low MRIGS values. We subsequently kept *high-confidence genes* (with 99.1% of all exon junction reads) in our analysis only from those samples, where they were detected with MRIGS ≥1 (Supplementary Figure S3B, black arrow). Thus, a specific gene may be kept in one sample and discarded as *noise candidate* in another. By definition, in a gene with MRIGS <1, the most used 5′ss had less than one read for every million exon junction reads in the entire sample. To permit an appropriate 5′ss selection, we eventually extended the '*high-confidence*' criterion from genes to splice sites.

In order to allow 5′ss usage comparison across genes with different expression levels in a single sample, we normalized all 5′ss reads by the individual gene expression MRIGS in the specific sample. We thus obtained gene-and-sample normalized reads (GSNR) for each 5′ss in each sample, val-

ues between 0 and 100%, and in each sample each gene contained one 5′ss with GSNR = 100%: the 5′ss with this gene's maximum (MRIGS) number of reads in this sample. Finally, we averaged the different GSNRs of a 5′ss across all samples with sufficient (MRIGS ≥ 1) gene expression, obtaining gene normalized reads (GNR) as measure of the overall 5′ss usage in our RNA-seq dataset. Since the 'most-used' 5′ss of a given gene could differ from sample to sample, there was not necessarily a single 5′ss with GNR = 100% in every gene.

The above analysis steps were independently performed for both fibroblast and endothelium RNA-seq datasets. Here, we present summary data for the larger fibroblast dataset; the respective data for endothelium are shown in direct comparison to fibroblast data in Suppl. File S2. From the fibroblast dataset, we obtained 92,493 internal exons of high-confidence genes with canonical 5′ss that were *Ensembl* annotated in at least one TSL1 transcript and contained at least one exonic GT site. These exons had a median exon length of 166 nt (average 417 nt), and the 5′ss GNR distribution was composed of three parts (Supplementary Figure S4A: fibroblast dataset, B: endothelium dataset): (i) a narrow peak at low GNR indicating noisy reads, (ii) a Gaussian part between 20% and 97% with mean 72% and standard deviation 18% ($r^2 = 0.995$), and (iii) a peak at 98–100% reflecting the maximally used 5′ss in each gene. Similar to our approach in (1), we considered 3240 5′ss (3.5%) detected below 2% of gene expression level (GNR < 2%) as potential noise candidates. For further analysis, we retained 89 253 *high-confidence* 5′ss (96.5%) with GNR ≥ 2% from genes with MRIGS ≥ 1.

Expected relative enhancement of GT-site usage next to mutation-weakened 5′ss

Our original log-GNR ratio (LGNRr) landscape was built from human fibroblast RNA-seq data of 320 601 pairs of mostly inactive exonic GT-sites and their corresponding high-confidence TSL1 annotated 5′ss. However, this dataset contained many GT-sites with very low HBond scores unlikely to support any actual usage as splice site. Therefore, for 5′ss mutation assessment with respect to activation of cryptic GT-sites, we first determined an adapted LGNRr landscape using only 45 561 GT-sites with HBond score ≥10, applying the same procedure as detailed before. We then used this adapted landscape to determine LGNRr values for pairs of GT-site and wild type or mutated 5′ss from their respective coordinates ΔHBS(GT–5′ss) and ΔSSHW(GT–5′ss).

For each GT-site in the exonic or 150 nt intronic neighborhood of a documented 5′ss mutation, we determined their corresponding LGNRr values as measures of landscape-predicted GT-site usage relative to both wild type and mutant 5′ss. Numerically, we determined these LGNRr values from the lookup tables for the GT-site/5′ss pair coordinates ΔHBS and ΔSSHW: LGNRr(GT/wt) = LGNRr(ΔHBS(GT–wt), ΔSSHW(GT–wt)) and LGNRr(GT/mt) = LGNRr(ΔHBS(GT–mt), ΔSSHW(GT–mt)). From these two LGNRr values, we determined the *expected relative enhancement* (ERE) of GT-site usage next to the mutated 5′ss relative to its usage next to the wild type 5′ss as $\mathrm{ERE} = 10^{\mathrm{LGNRr(GT/mt)}-\mathrm{LGNRr(GT/wt)}}$.

For GT-site/5′ss pairs the ΔHBS-ΔSSHW lookup range covered by the LGNRr landscape, we exchanged GT-site and 5′ss, determining LGNRr(ΔHBS, ΔSSHW) = –LGNRr(–ΔHBS, –ΔSSHW) instead. This was particularly relevant for mutations that considerably weakened a 5′ss, so that ΔHBS(GT–mt) >2 was outside the original lookup range.

Receiver operating characteristic curves

For the classification task of separating TSL1 annotated 5′ss from exonic GT sites based on their HBond score or SSHW, we developed three different logistic regression models, either depending (i) only on SSHW, (ii) only on HBS or (iii) depending on both scores including an interaction term. For the most general logistic regression (3), we determined four parameters α, β, γ, δ from fitting a logistic function with values between *zero*, corresponding to a GT site, and *one*, referring to an annotated 5′ss:

$$f\,(HBS,\ SSHW) = 1/\left(1+\exp\left(-\alpha\cdot HBS-\beta\cdot SSHW-\gamma\cdot HBS\cdot SSHW-\delta\right)\right)$$

to the training dataset of 45 165 GT sites and 45 411 annotated 5′ss. The value of the (for $\gamma \ll 1$ approximately linear) fit function in the exponent can be considered as a generalized splice site score combining HBS and SSHW, and discriminating between annotated 5′ss and exonic GT sites.

Both the logistic regression models and the receiver operating characteristic curves (ROC) obtained for the three regression models were generated using the R-package ROCit (version 1.1.1).

RESULTS

Inserting SRSF3 binding motif CANC in 'splicing neutral' reference sequence

In the first part of this work, we aim at *in-silico* designing—and experimentally validating—sequence segments with controlled splicing regulatory properties, computationally represented by their HEXplorer profiles. In principle, such 'designer exons' can be created by inserting single or multiple known SRE motifs into reference sequences that are ideally splicing neutral with respect to a specific genomic or reporter context (19,50).

Following this approach, we first characterized a reference exon composed of repeats of the octamer CCTATTGG that presents a nearly constant average $\mathrm{HZ_{EI}}$ amplitude of –0.15 suggesting it is splicing neutral. In a three-exon splicing reporter (Figure 1A; previously described in (37)), we used five repeats of this 'octamer –0.15' as central exon with a strong splice acceptor (MaxEnt 11.07) and splice donors of varying strength (HBS 17.5 down to 10.7; Figure 1D) (http://www2.hhu.de/rna/html/hbond_score.php (7)). We found inclusion of the reference exon only for the strongest 5′ss (HBS 17.5) (Figure 1B, lane 1), while slightly weaker 5′ss with HBS of 16.3 or less led to full exon skipping (Figure 1B,

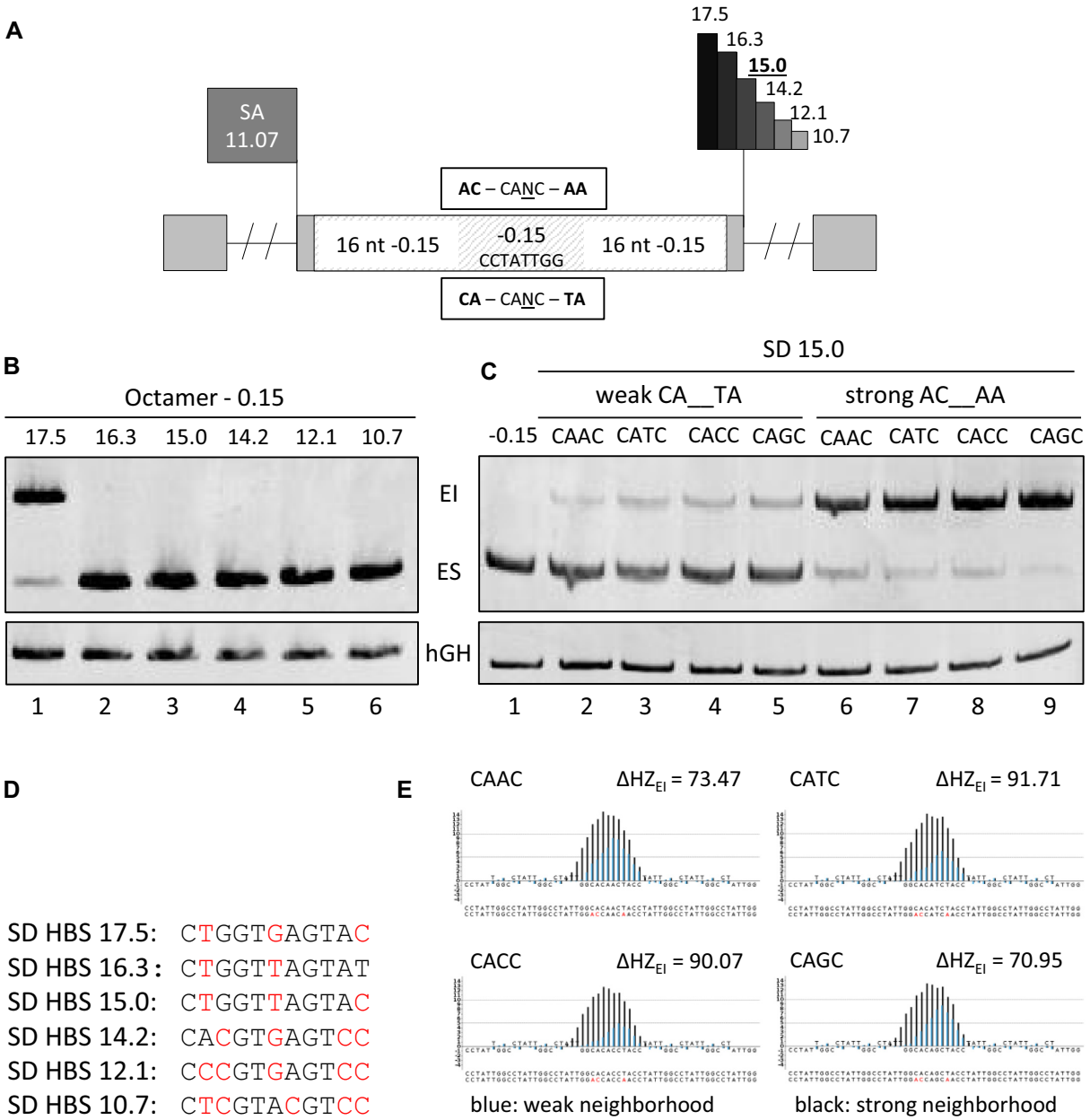

Figure 1. SRSF3 binding motifs CANC mediate exon inclusion in splicing reporter. (**A**) Sketch of the 3-exon minigene reporter plasmid. The middle exon contains an insertion site for different SREs which are flanked by an intrinsically strong splice acceptor (SA MaxEnt 11.07) and splice donor with varying intrinsic strength. (**B**) 2.5×10^5 HeLa cells were transfected with 1 μg of the reporter plasmids and 1 μg of pXGH5 (hGH) that was used for monitoring transfection efficiency. RNA was harvested and reverse transcribed into cDNA 24 h post transfection with primer pair #2648/#2649. PCR products were run on a 10% non-denaturing polyacrylamide gel to analyze exon inclusion in the presence of the neutral octamer –0.15 upstream of six different splice donors with HBond scores ranging from 17.5 down to 10.7. Without SRE support, lowering the HBond score from 17.5 to 16.3 resulted in full exon skipping. (**C**) A single repeat of an SRSF3 binding motif (CANC, N = all nucleotides) was inserted in the central octamer either flanked by AC–AA to maximize the total HEXplorer score or CA–TA in order to minimize the total HEXplorer score. In this construct, the intrinsic splice donor strength was set to 15.0. To analyze the splicing pattern, 2.5×10^5 HeLa cells were transiently transfected with 1 μg of each construct together with 1 μg of pXGH5 (hGH) to monitor transfection efficiency. Twenty-four hours after transfection, RNA was isolated and subjected to RT-PCR analysis using primer pairs #2648/#2649 and #1224/#1225 (hGH). PCR products were separated by 10% non-denaturing polyacrylamide gel electrophoresis and stained with ethidium bromide. The reduction of intrinsic splice donor strength resulted in full exon skipping upon insertion of the splicing neutral octamer –0.15. Depending on their neighboring dinucleotides, the SRSF3 binding motifs either induced a low level of exon inclusion with predominant exon skipping (CA–TA), or a high level of exon inclusion (AC–AA). (**D**) 5′ss sequences for (B). (**E**) HEXplorer plots show the comparison of CANC embedded in weak (blue) and strong (black) dinucleotide neighborhoods.

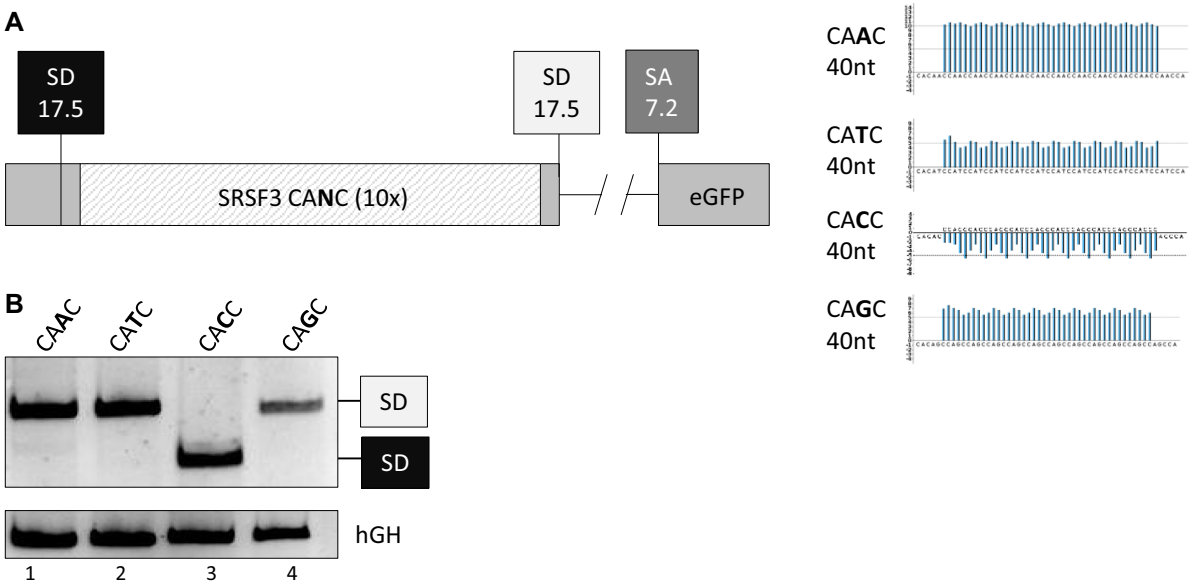

Figure 2. Splicing regulatory effects of concatenated SRSF3 binding sites. (**A**) Schematic drawing of the reporter construct that contains two equally strong splice donors SD with an HBond score of 17.5 (MaxEnt 10.10) and is used to detect up- or downstream enhancing or silencing properties of the concatenated SRSF3 binding motifs (CANC, N = all nucleotides). HEXplorer plots of the sequences show positive areas for the CAAC, CATC and CAGC repeats that indicate the likelihood of SR protein binding, while the CACC repeat displays a negative area that indicates putative hnRNP protein binding. (**B**) 2.5×10^5 HeLa cells were transiently transfected with 1μg of each construct together with 1 μg of pXGH5 (hGH) to monitor transfection efficiency. Twenty-four hours after transfection, RNA was isolated and subjected to RT-PCR analysis using primer pairs #3210/#3211 and #1224/#1225 (hGH). PCR products were separated by 10% non-denaturing polyacrylamide gel electrophoresis and stained with ethidium bromide. While the insertion of CAAC, CATC and CAGC repeats led to the use of the downstream located donor as expected upon the insertion of an SR protein binding site, concatenating the SRSF3 binding motif CACC led to the use of the upstream located splice donor.

lanes 2–6), marking the transition threshold between exon inclusion and skipping.

In order to test the insertion of a splicing enhancer motif in an instructive example, we therefore used a moderately strong 5′ss with HBS 15.0, and inserted the well-examined SRSF3 binding motif CANC (N = A, C, G, T) (51) into the center of the reference exon by replacing the middle octamer –0.15. In order to keep the length of the reference exon constant, we extended the CANC motif by two flanking nucleotides on either side. We chose two variants of flanking nucleotides that either maximized or minimized total HEXplorer score in this exon (Figure 1E). HZ_{EI} was maximized on average for ACCANCAA ('strong flanking nucleotides') and minimized for CACANCTA ('weak flanking nucleotides') as central octamers.

While the reference central octamer –0.15 led to complete exon skipping (as expected from the calibration experiment), in the weak neighborhood CA–TA each CANC SRSF3 binding site in the central octamer primarily resulted in exon skipping and only a low level of exon inclusion (Figure 1C, lanes 1, 2–5). Strengthening the neighborhood by substituting AC–AA as flanking dinucleotides around the same CANC sites increased total HZ_{EI} by between ∼70 and ∼92 (Figure 1E), and resulted in a high level of exon inclusion (Figure 1C, lanes 6–9). These experiments confirmed that all four CANC sites act as exonic splicing enhancers, in line with the solution structures of SRSF3 RNA-recognition motifs (RRM) in complex with the RNA sequence (51). Furthermore, the neighboring dinucleotides enclosing the central CANC motif additionally impacted exon inclusion level, in accordance with HEXplorer prediction for the *in silico* designed weak and strong neighborhoods.

Different splicing regulatory properties upon CANC concatenation for different 'N'

From the insertion of single SRSF3 binding sites (CANC), we now proceeded to using longer exonic splicing regulatory sequences by concatenating multiple copies of the CANC motifs. For a differential assessment of up- and downstream enhancing directions as well, we switched to a 5′ss competition reporter assay and inserted ten repeats of each CANC between two identical copies of a strong 5′ss with HBS 17.5. These competing 5′ss defined the 3′ end of the first exon of the HIV-based two-exon splicing reporter whose RNA level depends on U1 snRNP binding to either the upstream or downstream 5′ss (Figure 2A, (6)). The impact of the inserted 40 nt sequences on splice site selection was analyzed by RT-PCR following transient transfection assays.

We first determined HEXplorer score profiles for all four CANC repeats (ten repeats of every CANC). As expected for SRSF3 binding sites, HEXplorer score profiles were exclusively positive for CAAC (HZ_{EI} amplitude ∼10), CAGC (HZ_{EI} ∼7) and CATC (HZ_{EI} ∼5) repeats. Surprisingly however, ten repeats of CACC showed an entirely negative HEXplorer score profile with HZ_{EI} amplitude ∼–4, suggesting upstream splice enhancing and downstream splice suppressing properties (Figure 2A, right panels).

Consistent with the unexpected HEXplorer score prediction, insertion of CAAC, CATC and CAGC repeats led to the exclusive use of the downstream located donor (Fig-

ure 2B, lanes 1, 2 and 4), while insertion of CACC repeats led to a complete switch to the upstream 5′ss (Figure 2B, lane 3). For the CACC motif, in fact, concatenation creates a cytosine-rich CACCC motif which may be bound by the exonic splicing silencer hnRNP K (52), consistent with the negative HEXplorer profile.

This example strikingly demonstrates that concatenation of an enhancer sequence may even invert the original sequence's splicing regulatory properties. We therefore systematically searched for sequences with unaltered splicing regulatory properties when multiply concatenated.

HEXplorer profiles of periodic *k*-mer sequences

The previous experiments demonstrated that HEXplorer score profiles may accurately reflect unexpected experimental outcome of concatenating single splicing regulatory sequences. By systematically analyzing HEXplorer score profiles, we therefore computationally searched for *k*-mer sequences with specific *a priori* prescribed HZ_{EI} amplitude that retained splicing regulatory properties of the single *k*-mer upon concatenation. Since single RNA-recognition motifs (RRMs) of splicing regulatory proteins are thought to bind up to eight nucleotides (53), and in line with motif lengths applied by various computational tools, we searched for periodic *octamer* sequences with approximately constant HEXplorer score amplitude ($HZ_{EI} \approx$ const.). By definition, HEXplorer score profiles of periodic sequences (with period $\geq$6 nt) have the same periodicity as these sequences. Thus, for octamer repeats, up to eight different HZ_{EI} values can occur in the HEXplorer profile, and they repeat every eight nucleotides.

We therefore systematically searched for octamer sequences that upon concatenation show little HEXplorer score amplitude variation around their average. To this end, we calculated average and standard deviation of HZ_{EI} amplitudes for all 65 536 possible octamers from 5-fold concatenations. In order to avoid accidentally creating 5′ss or 3′ss in the designed sequences, we excluded octamers containing a GT or AG dinucleotide, or creating one by concatenation, with 23 120 octamers remaining. Limiting HZ_{EI} variation to standard deviation <2 still left 18 925 octamers in the average HZ_{EI} amplitude range from –20 to + 14. The octamer count histogram in Supplementary Figure S1A displays the number of different octamers for all HZ_{EI} intervals in this range (gray bars). Note that each bin contains sequences with low standard deviation <0.5 (open squares show the minimum HZ_{EI} standard deviation in each bin).

From this set of *in silico* designed, extremely low HZ_{EI} variability octamers, we selected a total of fifteen test octamers in addition to our reference octamer (CCTATTGG, average HZ_{EI} amplitude –0.15): eight downstream enhancing octamers with HZ_{EI} amplitude +10.32, and seven upstream enhancing octamers with HZ_{EI} amplitude –10.35.

Splicing reporter test of *in silico* designed octamers

In order to experimentally validate the HEXplorer predictions for all fifteen +10.32 and –10.35 octamers, as well as for the reference octamer –0.15, we tested five repeats of each in the above splicing competition reporter between two identical strong 5′ss (HBS 17.5). Figure 3 gives a representative example of one up- and one downstream enhancing octamer, while the remaining results are shown in Supplementary Figure 1B.

For each of the +10.32 octamers, insertion of repeats resulted in exclusive recognition of the downstream 5′ss, while the use of the upstream donor was completely repressed, confirming their predicted directional splicing regulatory activity (Figure 3C, lane 1; Supplementary Figure S1B, lanes A–G). While the splicing neutral octamer –0.15 mediated between the two splice donors on a basal level (Figure 3C, lane 2), for all but one –10.35 octamer, insertion of repeats resulted in exclusive selection of the upstream located 5′ss (Figure 3C, lane 3; Supplementary Figure S1B, lanes H–M), in agreement with their predicted splicing regulatory activities. One of the –10.35 octamers, however, exhibited neutral splicing of both competing 5′ss rather than only upstream enhancing behavior: GCATTTAT led to equal amounts of up- and downstream 5′ss use (Supplementary Figure S1B, lane J). This may be due to the joint effects of potential hnRNP D and SRSF6 binding sites (ENCODE (34), ESEFinder (54)) or to different protein RNA binding affinities that were not represented in the exonic and intronic datasets constituting the basis of the HEXplorer score algorithm.

In general, insertion of octamers +10.32 and octamers –10.35 drastically elevated overall (up- or downstream) splice donor recognition following the direction dependent action of splicing regulatory elements, whereas the splicing neutral octamer –0.15 did not show any splice donor preference in this reporter. The lower total amount of RNA found with the splicing neutral octamer –0.15 (Figure 3C, lane 2, 3E, lane 1) was in agreement with U1 snRNA dependent reduced transcription initiation, regardless of whether the U1 snRNA binding site was splicing active (55).

SR- and hnRNP proteins bind to HEXplorer-designed octamer sequences

To further analyze the mechanism of splicing regulation conducted by the non-evolutionary *in silico* designed artificial octamer sequences, we performed an RNA affinity purification assay to identify splicing regulatory proteins binding to the sequences. To this end, we incubated 40 nt long RNA oligonucleotides (five octamer repeats of the two ±10.3 octamers shown in Figure 3A, as well as the reference octamer –0.15) with HeLa nuclear extract (56). After several washing steps, the remaining specifically bound proteins were eluted and subjected to MS-analysis. Results were analyzed using Perseus software (57). When filtering for highest MS/MS counts and searching for splicing related proteins, a binding preference of SRSF3 to the downstream enhancing splicing regulatory octamer +10.32 was revealed (Supplementary Figure S2A). The negative octamer –10.35 was preferably bound by the PTB isoforms PTBP1 and PTBP2, as well as hnRNPDL and TIA-1, all known repressors of downstream splice donors (13). The neutral octamer –0.15 showed no preferred binding for any splicing related proteins (Suppl. File S3). Validation of these results was performed via western blot using antibodies specifically detecting the splicing related binding proteins SRSF3 for oc-

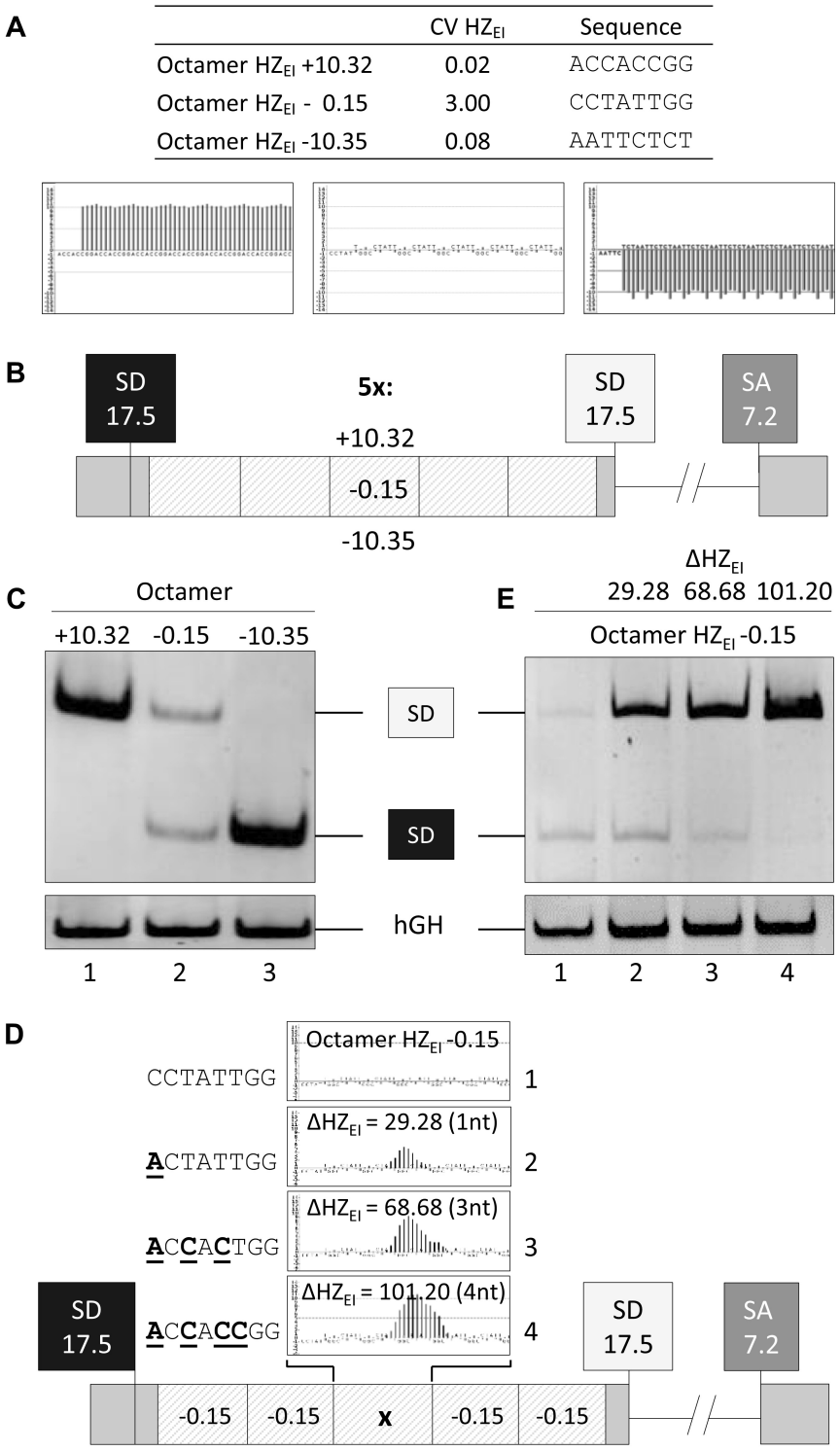

Figure 3. HEXplorer guided sequences shift splice donor use. (**A**) HEXplorer predicted positive, neutral and negative periodic octamer sequences. CV HZ_{EI} denotes standard deviation of HZ_{EI} values divided by their average. (**B**) Schematic drawing of the reporter construct that contains two equally strong splice donors with an HBond score of 17.5 (MaxEnt 10.10) and is used to detect up- or downstream enhancing or silencing properties of periodic, concatenated HEXplorer predicted octamers. (**C**) Five repeats of either octamer –10.35 or +10.32 completely shifted 5′ss usage to the up- or downstream SD. (**D**) Schematic drawing of the same reporter construct used to detect up- or downstream enhancing or silencing properties of single HEXplorer predicted octamers. Point mutations in the central splicing neutral octamer –0.15 sequence increase the positive HEXplorer plot area indicated by the positive ΔHZ_{EI} and morph the neutral reference octamer into octamer +10.32. (**C**, **E**) 2.5×10^5 HeLa cells were transiently transfected with 1 μg of each construct together with 1 μg of pXGH5 (hGH) to monitor transfection efficiency. Twenty-four hours after transfection, RNA was isolated and subjected to RT-PCR analysis using primer pairs #3210/#3211 and #1224/#1225 (hGH). PCR products were separated by a 10% non-denaturing polyacrylamide gel electrophoresis and stained with ethidium bromide (left).

tamer +10.32, PTB and hnRNP D for octamer –10.35 and the control MS2 coat (Supplementary Figure S2B, C).

Deep sequencing of octamer library inserted in splicing competition reporter

Having confirmed HEXplorer predicted impact on 5′ splice site usage for fourteen *in silico* designed 40 nt long octamer concatenates, we next sought to vary the single central octamer flanked by two reference octamers on either side. In order to systematically examine the impact on downstream 5′ splice site usage for a large octamer set, we eventually applied a massively parallel splicing assay (MPSA), using our established splicing competition assay with two identical strong 5′ss (HBS 17.5).

In a first step, we tested sensitivity to point mutations in the central octamer of our splicing reporter. Observing that octamer + 10.32 differed from the reference by only four nucleotide substitutions, we morphed the reference octamer into octamer +10.32 by successive point mutations (Figure 3D). The first single nt substitution increased the HEXplorer score by $\Delta HZ_{EI} = 29.28$, a three-nt substitution by $\Delta HZ_{EI} = 68.68$, and the final four-nt substitution by $\Delta HZ_{EI} = 101.2$, obtaining octamer +10.32.

Increasing the HEXplorer score HZ_{EI} of the reference octamer by ∼30 led to an increase of overall splicing efficiency and shifted 5′ss usage to the downstream 5′ss (Figure 3E, lane 2). Further increasing HZ_{EI} (total change ΔHZ_{EI} ∼70 from reference), reduced upstream and increased downstream 5′ss usage even more (Figure 3E, lane 3). Finally, the fourth point mutation morphed the central reference octamer into the + 10.32 octamer (total change ΔHZ_{EI} ∼100 from reference) and led to the exclusive usage of the downstream splice donor site, while upstream donor usage could not be detected (Figure 3E, lane 4). Thus, in this setting even a single octamer +10.32 within the otherwise HEXplorer neutral reference sequence led to a complete switch to the downstream 5′ss, similar to the previously tested five octamer copies (cf. Figure 3C, lane 1).

Having confirmed the splicing competition assay sensitivity to changes only in the central octamer, we prepared a minigene library incorporating a central random octamer in our reference exon between two identical copies of a strong 5′ss (HBS 17.5). Amplifying this library in *E. coli* yielded a total of 20 767 different octamers out of 65 536 possible octamers, as determined by amplicon sequencing. HeLa cells were subsequently transfected with this library, total RNA was isolated and amplified with primer pair #3210/#3211 enclosing both competing 5′ss. Bands corresponding to up- and downstream 5′ss usage were separated by PAGE. Octamer occurrence frequencies were again determined by amplicon sequencing. For each octamer, the number of reads both in the plasmid library and in the downstream 5′ss band were determined, and the normalized enrichment index (NEI) was calculated (cf. Materials and Methods). Excluding octamers with very low read counts in either library or band, we kept 3127 octamers with more than eight reads in the plasmid library and more than four reads in the band.

We then grouped these 3127 octamers in logarithmically equidistant intervals of 0.1 $\log_{10}$(NEI) units ('bins'). The

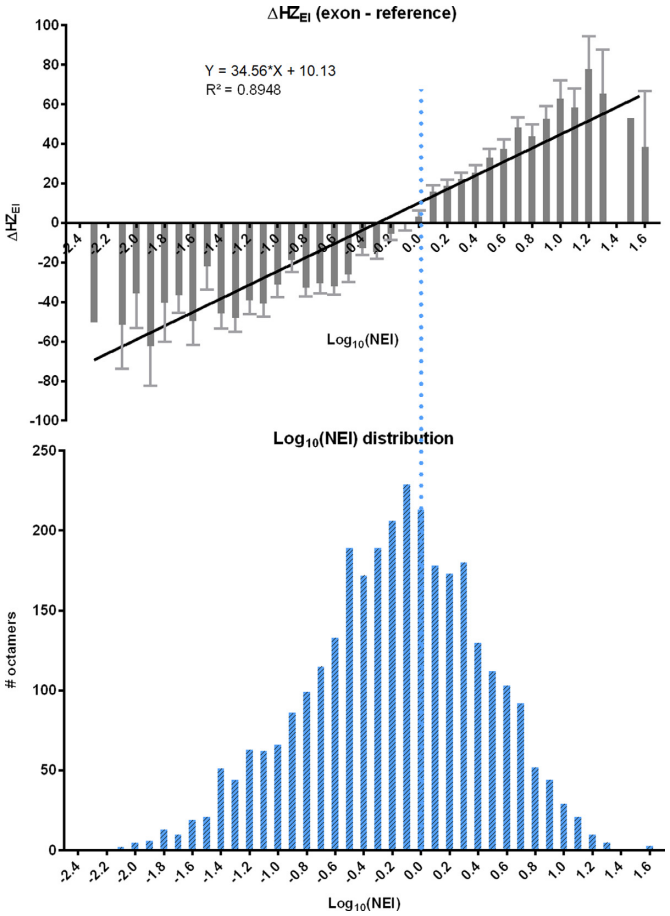

Figure 4. HEXplorer score increases with downstream 5′ss usage in random octamer library assay. Analysis of massively parallel splicing assay with random octamer library inserted in the center of the splicing neutral reference sequence in our splicing competition reporter. Normalized enrichment index (NEI) of octamers in gel band corresponding to downstream 5′ss usage. is log-normal distributed around NEI = 1 (lower panel). Average HEXplorer score (difference w.r.to the reference sequence) of all octamers in a bin with given $\log_{10}$(NEI) shows linear increase with $\log_{10}$(NEI). Whiskers depict standard error of mean.

NEI distribution was approximately log-normal and symmetrical around NEI = 0.7 in these octamers (Figure 4, lower panel). Searching for a relation between HEXplorer score and enrichment index for each octamer, we calculated the HEXplorer score difference ΔHZ_{EI} between the exons containing this central octamer and the reference exon. These individual ΔHZ_{EI} values still exhibited considerable scatter and were subsequently averaged for all octamers in a given $\log_{10}$(NEI) bin. For NEI > 1, average ΔHZ_{EI} were positive and showed a linear increase with $\log_{10}$(NEI) over two orders of magnitude for NEI (Figure 4, $r^2 = 0.89$ for the entire NEI range). Thus, more enriched octamers exhibited higher average ΔHZ_{EI}, and ΔHZ_{EI} was proportional to $\log_{10}$(NEI). For depleted octamers with NEI < 0.25 ($\log_{10}$(NEI) < −0.6), however, average ΔHZ_{EI} leveled off at about −40. Such depleted octamers originate from RNA with very little usage of the downstream 5′ss, and can thus be expected to have lower, negative ΔHZ_{EI} scores. However, in the MPSA approach used here, octamers sup-

porting upstream 5′ss usage are systematically underrepresented and average ΔHZ_{EI} values are thus less negative than expected.

The MPSA approach used here significantly extends our initial findings on ΔHZ_{EI} reflecting relative 5′ss usage in our splicing competition reporter from fourteen selected *in silico* designed octamers and five point mutations to more than three thousand random octamers.

While the presented experimental approaches reflect separate variation of either 5′ splice site or SRE neighborhood, we then sought to capture both factors simultaneously by analyzing 5′ss usage in two large human RNA-seq datasets.

320 601 pairs of high-confidence 5′ss and exonic GTs from exons of TSL1 transcripts

Complementary to our experimental analysis, we also examined 5′ss context impact on splice site competition using data from two large human RNA-seq transcriptome datasets: human fibroblasts (1,43) and endothelial cells (18). In order to mimic the 5′ss competition situation experimentally examined above (cf. Figure 3), we analyzed pairs of annotated 5′ss and nearby exonic GTs, using the ratio of RNA-seq reads detected on each as relative usage measure. Comparing RNA-seq reads across many genes from different samples, however, requires careful normalization of reads and removal of potentially noisy read counts, as detailed in the Methods section.

In particular, we applied a two-tier normalization process, taking both differential sequencing efficiency across samples (library size) and differential gene expression within a sample into account. To keep only reliably detected 5′ss RNA-seq reads above biological and sequencing noise, we discarded genes in those samples, where they were very weakly expressed, and additionally discarded 5′ss with gene-normalized reads (GNR, cf. Materials and Methods) below 2% of the gene expression level. In this way, we systematically improved the removal of noisy reads introduced in (1).

For these *high-confidence* 5′ss, we then extracted all GT dinucleotides between 12 nt downstream of the 3′ss and 17 nt upstream of the 5′ss. This GT search region was chosen to ensure that there was at least a one-hexamer wide potential SRP binding site not overlapping the 11 nt long 5′ss or GT-site, as well as the 23 nt long 3′ss.

We further excluded potential U12 splice donors, defined by the list of confirmed U12-dependent 5′ss reported in (58), and those 5′ss with a GTT trinucleotide at positions +1/+2/+3 which may bind U1 snRNP by bulging the T nucleotide in position +2. In order to better mimic our splice site competition experiments in splicing reporters with short exons, we only included GT sites less than 150 nt from the 5′ss. Collecting all GT dinucleotides in this search region (SA + 12 nt to SD-17 nt) while applying these strict filter conditions, we obtained a total of 320,601 GT-and-5′ss pairs in 89,008 exons of the fibroblast dataset. Note that actually 8833 exonic GT sites (2.8%) had RNA-seq reads. In each pair, we then compared U1 snRNA complementarity (HBS) and splice site HEXplorer weight (SSHW) between GT sites and annotated 5′ss.

Table 1. GT-site usage and SSHW for weaker vs. stronger GT-sites

GT-site/5′ss pairs	$\Delta HBS \leq 0$ *weaker* GT-site	$\Delta HBS > 0$ *stronger* GT-site	Total
Unused GT-sites (no reads)	309 678 (97.5%)	2090 (66.9%)	311 768
GT SSHW	– 59.32	– 165.6	– 60.04
Used GT-sites (with reads)	7801 (2.46%)	1032 (33.1%)	8833
GT SSHW	+ 72.17	–9.43	+ 62.64
Total	317 479 (100%)	3122 (100%)	320 601

Exonic GT sites have lower U1 snRNA complementarity than annotated 5′ss used in fibroblasts

As expected, exonic GT-sites had much lower U1 snRNA complementarity than 5′ss (GT HBS 6.2 ± 3.0, mean ± SD, versus 5′ss HBS 15.1 ± 2.5; $N = 320\,601$ pairs; cf. Figure 5i for individual GT- and 5′ss-HBS distributions). In Figure 5ii, light gray bars show the HBond score difference distribution $\Delta HBS = HBS_{GT} - HBS_{5'ss}$ in all individual pairs, and indeed, in 98.9% of pairs, the exonic GT-site was weaker than the 5′ss. For the subset of GT-sites with RNA-seq reads, e.g. from lower transcript levels, the ΔHBS distribution was significantly shifted to higher values (Figure 5ii, dark versus light gray bars).

Exonic GT sites have weaker SRE neighborhood than annotated 5′ss used in fibroblasts

In the 320 601 GT-and-5′ss pairs, exonic GT-sites also had lower splice site HEXplorer weights than 5′ss (GT SSHW –1.1 ± 4.8, mean ± SD, vs. 5′ss SSHW 5.8 ± 5.0; cf. Figure 5iii for individual SSHW distributions). However, the two SSHW distributions overlapped to a much higher degree than the respective HBS distributions, indicating higher importance of HBS for splice site recognition than SSHW (cf. Figure 5iii for individual GT- and 5′ss-SSHW distributions).

In Figure 5iv, light gray bars show the SSHW difference distribution $\Delta SSHW = SSHW_{GT} - SSHW_{5'ss}$, and in 82.0% of pairs, the exonic GT site had lower SSHW than the 5′ss. For the subset of GT sites with RNA-seq reads, the ΔSSHW distribution was only slightly shifted to higher values (Figure 5iv, dark vs. light gray bars).

Exonic GT sites with higher U1 snRNA complementarity than annotated 5′ss

While by far most GT-sites had lower U1 snRNA complementarity than the respective annotated 5′ss (‘*weaker*’ GT-site, $\Delta HBS = HBS_{GT} - HBS_{5'ss} \leq 0$), we now focused on the unexpected cases of ‘*stronger*’ GT-sites ($\Delta HBS > 0$). To this end, we split all 320 601 pairs of GT-sites and 5′ss into four groups: *weaker* vs. *stronger* as well as *unused* (with RNA-seq reads) vs. *used* GT-sites. This procedure created four groups as shown in the fourfold table below (Table 1). In the terminology of fourfold tables, ΔHBS is an ‘antecedent factor’, and GT-site usage corresponds to an ‘outcome’. The fourfold table has a highly significant odds ratio of 19.6.

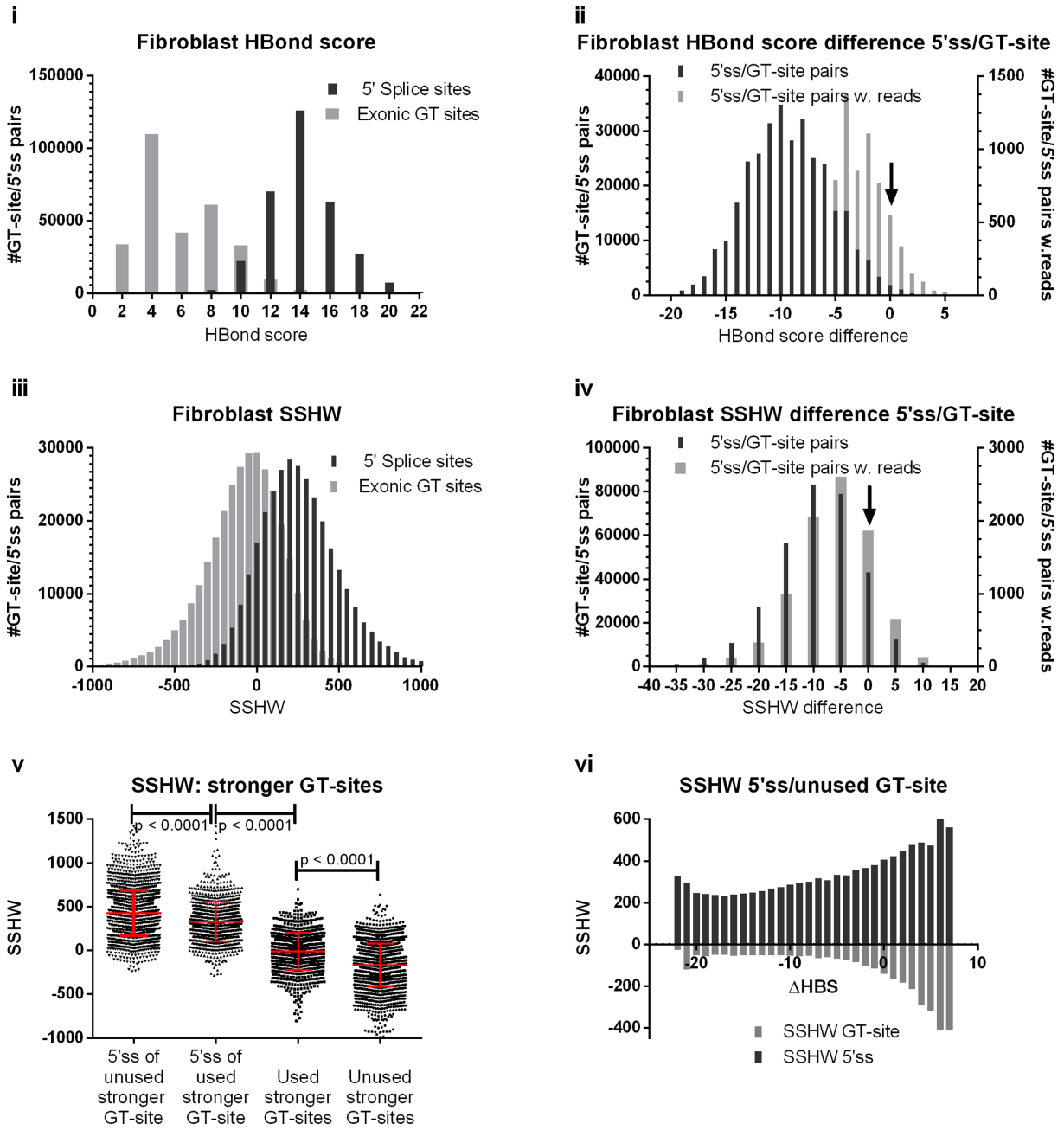

Figure 5. Exonic GT sites have lower U1 snRNA complementarity and weaker SRE support than nearby annotated 5′ss in fibroblast RNA-seq dataset. (**i**) HBond score distributions for 320 601 pairs of high-confidence annotated 5′ss and exonic GT sites closer than 150 nt. (**ii**) HBond score difference HBS_{GT}–$HBS_{5'ss}$ distribution. For 99% of all pairs, the 5′ss HBS was higher than the exonic GT HBS, indicating a stronger 5′ss compared to competing exonic GTs. Arrow indicates ΔHBS = 0. (**iii, iv**) SSHW distributions for the same datasets. (**v**) SSHW scatterplot for four groups of 5′ss and *stronger* exonic GT-sites (ΔHBS > 0). SSHW was higher in 2090 5′ss paired with *unused* GT-sites than in 1032 5′ss paired with *used* GT-sites. Conversely, 1032 *used* GT-sites had higher SSHW than 2090 *unused* GT-sites. (**vi**) SSHW of 311 768 5′ss paired with *unused* GT-sites, stratified by ΔHBS. 5′ss and GT-site SSHW strongly diverged for increasing ΔHBS, i.e. *stronger* GT-sites.

While only 2.5% of all weaker GT-sites were used (7801 GT-sites), 33% of stronger GT-sites (1032) were used—a 13.4-fold higher proportion. But not only was the proportion of used GT-sites higher among *stronger* versus *weaker* GT-sites, but on average *used stronger* GT-sites had 3.6-fold more reads than *used weaker* GT-sites.

We then examined the SRE support measure SSHW in the 3122 pairs of 5′ss and *stronger* GT-sites. Extending our previous results (37,40), we compared the SSHW distributions between four groups: 5′ss of *unused stronger* GT-sites (2090), 5′ss of *used stronger* GT-sites (1032), *used stronger* GT-sites (1032) and *unused stronger* GT-sites (2090).

While the SSHW distributions of the four groups overlapped, we found a clear trend: 5′ss SSHW was significantly higher than *used* GT-site SSHW in 1032 pairs, and *used* GT-site SSHW was in turn significantly higher than *unused*

GT-site SSHW (SSHW: –9.43 versus –165.6; cf. Figure 5v, Table 1). These findings confirm that *unused* GT-sites that are stronger than their respective 5′ss appear more repressed by their SRE neighborhood than *used* GT-sites, while 5′ss in both groups are enhanced by SREs (SSHW +328.5 and +427.8, respectively). *Weaker* GT-sites that are used are also enhanced (SSHW +72.17).

Finally, we systematically stratified all 311 768 *unused* GT-site-/5′ss-pairs by their HBond score differences ΔHBS(GT–5′ss), determining average SSHW for 5′ss and *unused* GT-sites in each ΔHBS bin. Overall, 5′ss SSHW was positive, i.e. enhancing splice site usage, and increased with increasing ΔHBS. In accordance with our expectation, *unused* GT-site SSHW was overall negative, indicative of GT-site repression by their SRE neighborhood. Plotted together, both SSHW graphs exhibited a trumpet shape with the trumpet bell in the region of *stronger* GT-sites (ΔHBS > 0). While for *weaker* GT-sites (ΔHBS ≤ 0), GT SSHW was only slightly negative and had little variation, for *stronger* GT-sites (ΔHBS > 0), GT SSHW was increasingly negative, suggesting that SSHW could compensate for ΔHBS > 0 and suppress GT-site usage in this region (Figure 5vi).

From these analyses, we conclude that in our RNA-seq fibroblast dataset, exonic GT-sites have significantly lower HBond scores than their associated 5′ss, and HBond scores of *used* GT-sites with RNA-seq reads are higher than those of GT sites without reads. Splicing regulatory properties of 50 nt wide neighborhoods, quantified by SSHW, exhibit the same tendencies, albeit to a much lower degree. In our endothelium RNA-seq dataset, we encounter the same findings as presented in Supplementary Figure S5.

5′ Splice site usage dependence on 5′ss strength and SRE support

After separately identifying HBond score and SSHW differences between GT-sites and 5′ss in 320 601 pairs, we set out to determine relative GT usage dependency both on U1 snRNA complementarity and SRE support simultaneously. This is a tentative approach to a comprehensive 'functional splice site strength' concept encompassing splice site U1 snRNA complementarity and SRE neighborhood.

In our RNA-seq dataset, gene-normalized reads (GNR) reflect GT-site or 5′ss usage likelihood, and we therefore quantified GT usage relative to 5′ss by their GNR log-odds ratio $\mathrm{LGNRr} = \log_{10}(\mathrm{GNR_{GT}}/\mathrm{GNR_{5'ss}})$. In order to tabulate LGNRr as a function of both ΔHBS and ΔSSHW, we first binned these variables to obtain GT-site-/5′ss-pair groups of approximately equal sizes. Rather than choosing equidistant ΔHBS- and ΔSSHW-bin intervals, we focused on adequate resolution in the important regime of GT-sites with RNA-seq reads. From the two ΔHBS and ΔSSHW distributions shown in Figure 5ii and Figure 5iv (gray bars), we obtained ten 10%-wide bins each for ΔHBS and ΔSSHW, splitting the sample of 8833 pairs with RNA-seq reads on the GT site into 10 × 10 two-dimensional bins containing about 8833/(10 × 10) GT-site-/5′ss-pairs each. On average, each 2D bin contained 3206 pairs overall and 88 pairs with RNA-seq reads. For every ΔHBS- and ΔSSHW-bin, we then calculated the average LGNRr of all pairs, and color-coded cells with low (high) relative GT-site usage in red (green). In this table, GT-site usage relative to 5′ss covered three orders of magnitude from 10^{-3} to 10^{-6} in statistically reliable values: the median coefficient of variation ($\mathrm{CV_{LGNRr}}$ = standard deviation/mean LGNRr) of the LGNRr values in each two-dimensional bin was 0.21 (average $\mathrm{CV_{LGNRr}} = 0.25$, standard deviation $\mathrm{CV_{LGNRr}} = 0.18$). We further averaged the 2D LGNRr table with an exponential smoothing algorithm using 0.7× average of all eight neighboring bins. Eventually, to obtain a LGNRr representation on an equidistant square grid, we applied cubic spline interpolation in ΔHBS steps of 0.2 and ΔSSHW steps of 25 (Figure 6A).

The two-dimensional surface plot (Figure 6A, and Supplementary Figure S6A for endothelial data set) showed a clear picture of relative GT-site-to-5′ss-usage dependence on both U1 snRNA complementarity and on SRE support. There is a region of low GT-site usage for both large negative ΔHBS and ΔSSHW (red), mirrored by a region of higher GT-site usage in the opposite corner with higher, positive ΔHBS and ΔSSHW (green), and a smooth, diagonal transition region (yellow). A sufficiently large negative ΔHBS cannot be compensated by even the strongest SRE-containing neighborhood (high SSHW), while for positive or only slightly negative ΔHBS, GT-sites can be used despite lack of SRE support (negative ΔSSHW). This result underscores that 5′ss complementarity to U1 snRNA is the dominant feature in splice site recognition, and SRE support plays a secondary, auxiliary part.

Activation of cryptic GT-sites versus exon skipping after 5′ss mutation

Finally, we tentatively assessed human 5′ splice site mutations using our LGNRr landscape. In particular, we examined 5′ss mutations leading to cryptic activation ('CA') of a GT-site in contrast to those leading to exon skipping ('ES').

We selected 5′ss mutations corresponding to these two types (CA, ES) from our own manually curated literature-based web database (https://www2.hhu.de/rna/html/viewmutationdatabase.php) containing 118 documented 5′ss mutations with RNA-level evidence. The control group ES comprised 78 5′ss mutations described to induce exon skipping, while there were only 19 mutations in the CA group, for which activation of a specific cryptic GT-site following 5′ss mutation was described in the literature, and where appropriate transcripts could be unambiguously identified. In the following, we denote these *mutation-activated* GT-sites as 'confirmed'.

For both groups of mutations, CA (19 mt) and ES (78 mt), we then determined all GT-sites within exons as well as 150 nt wide intronic regions. Eventually, for every GT-site we calculated its *expected relative enhancement* (ERE) describing how much more the GT-site is predicted by the landscape to be used, if it occurs next to the mutated (normally weakened) 5′ss instead of the wild type 5′ss (cf. Methods). In both CA and ES groups, GT-site *expected relative enhancement* values were distributed in two disjoint ranges of low (1—100) and high (5×10^8–10^{11}) ERE (Supplementary Figure S7). We surmised that if present, ERE values in

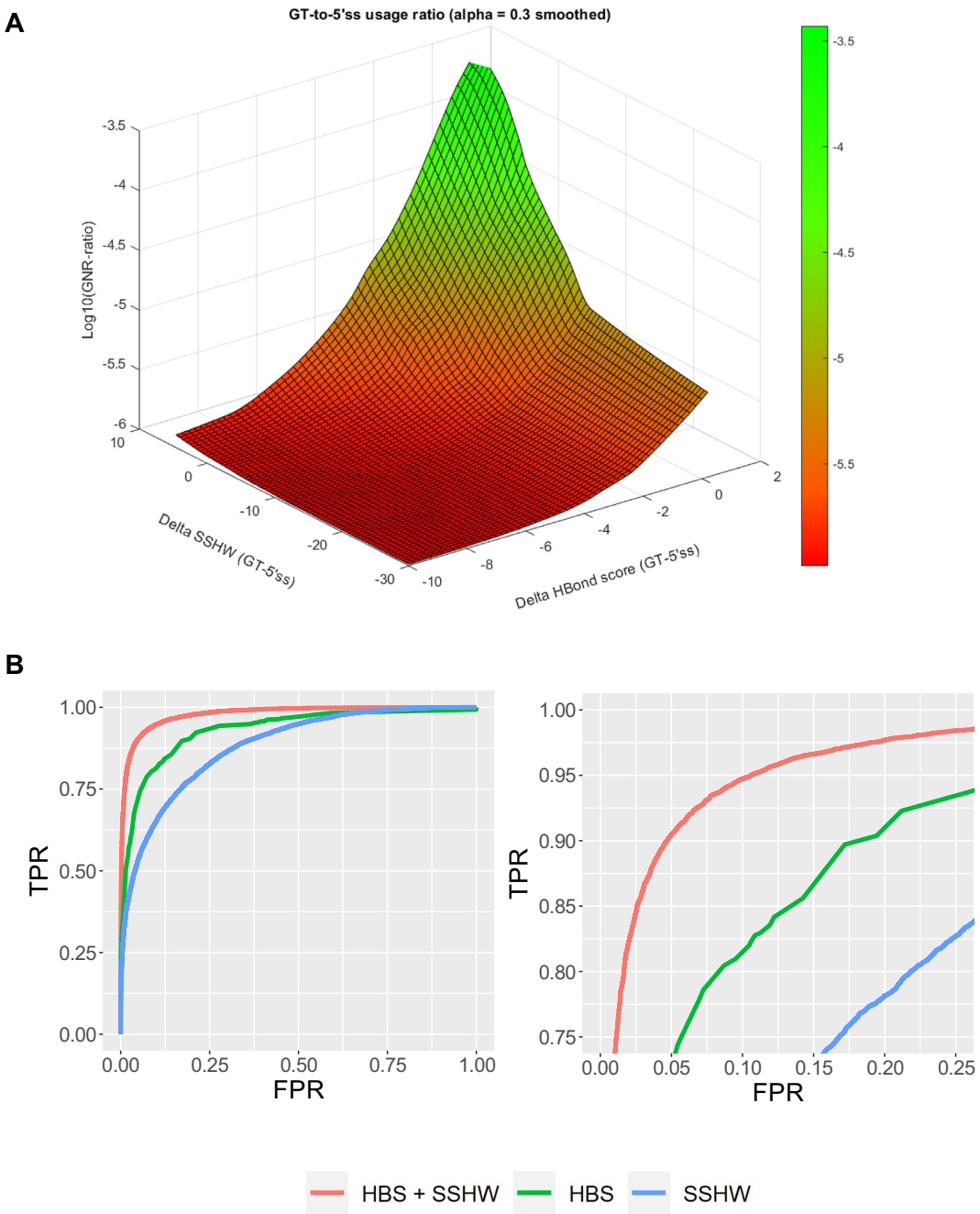

Figure 6. Combination of HBS and SSHW improves classification of GT sites and 5′ss in fibroblast RNA-seq dataset. (**A**) Average LGNRr = $\log_{10}(GNR_{GT}/GNR_{5'ss})$ as measure of GT-site usage relative to 5′ss (vertical z-axis), plotted as function of HBond score difference $\Delta HBS = HBS_{GT} - HBS_{5'ss}$ and splice site HEXplorer weight difference $\Delta SSHW = SSHW_{GT} - SSHW_{5'ss}$. Color-coding shows a monotonous transition from exclusive 5′ss usage (front corner, red) to higher GT-site usage (back corner, green). (**B**) Receiver operating characteristic curves of three logistic regression models for the classification of 14 401 annotated 5′ss and 14 405 exonic GT sites closer than 150 nt and with HBS ≥10, but <1% RNA-seq reads of the associated nearby 5′ss. ROC curves for logistic model based only on SSHW (blue, AUC 0.88), based only on HBS (green, AUC 0.93) and based on both HBS and SSHW (red, AUC 0.98) show stepwise improvement of classification accuracy.

the 'high' range were indicative of possible cryptic GT-site candidates.

Indeed, for 16 out of 19 mutations in the CA group, high ERE values were found for nearby GT-sites, and in 15 of these 16 mutations, the confirmed GT-site belonged to the set of high enhancement GT-sites (Supplementary Figure S7A; red symbols). In another two out of three CA mutations with ERE values only in the low range, the confirmed GT-site had the maximum ERE. Predicting a GT-site as candidate for cryptic GT-site activation by high or maximal enhancement would indeed retrieve 17 out of 19 confirmed GT-sites.

In the control group ES, only one third (26/78) of mutations had enhancement values in the high range, totaling 43 out of 880 GT-sites (Supplementary Figure S7B, showing only the first 26 mutations). Although there is a clear difference in the proportion of high-range *expected relative enhancement* values between CA and ES (16/19 versus 26/78), the LGNRr landscape does not permit specific discrimination of exon skipping from cryptic GT-site activation.

Combination of HBS and SSHW improves classification of GT-sites and 5′ss

In order to further examine the discriminatory power of HBond score and SSHW to distinguish annotated 5′ss from exonic GT-sites in a classification task, we selected 57,611 pairs with low usage GT-sites ($GNRr = GNR_{GT} / GNR_{5'ss} < 1\%$) that had medium-to-high U1 snRNA complementarity ($HBS \geq 10$). In competition with their respective 5′ss, these GT sites were barely used, although they had reasonable complementarity with an HBond score of at least 10. In this dataset, we expected SRE neighborhoods of both 5′ss and GT site to possibly play a stronger part in splice site selection.

We then split the pairs and pooled both GT-sites and 5′ss into a single set of 115 222 potential splice sites. Randomly splitting this entire dataset into a training set (75%) and a validation set (25%), we fit three different logistic models for the binary prediction of true 5′ss in a balanced sample of 43 206 GT sites and 43 201 5′ss. In the first model, we used only SSHW as single predictor variable, in the second model we used HBond score alone, and finally we entered both SSHW and HBS simultaneously into the regression model (cf. Materials and Methods). In all three regressions, the coefficients of SSHW and HBS were highly statistically significant ($P < 10^{-6}$), indicating that these variables significantly contributed to distinguishing true 5′ss from GT sites in the training dataset.

We then tested the three regression models on the remaining 25% of the entire dataset, containing 14,401 annotated 5′ss and 14 405 exonic GT-sites. Figure 6B – and Supplementary Figure S6B for endothelial data set—shows the receiver operating characteristic curves (ROC) obtained for the three regression models, plotting *sensitivity* (true positive rate, TPR) versus *1—specificity* (false positive rate, FPR) upon variation of the cutoff of the prediction scores obtained from the regressions. All three models achieved good classification results for discriminating true 5′ss from GT-sites in the validation dataset, indicated by all ROC curves extending far into the upper left corner of the diagram. Using the *area-under-the-curve* ($0 < AUC < 1$; $AUC = 0.5$ for random assignment) as overall measure to compare the regression models, we found a clear hierarchy for goodness of classification: the model using only the HBond score increased AUC to 0.93 from $AUC = 0.88$ for SSHW alone, and entering both variables into the model again improved the classification to $AUC = 0.98$. Thus, in terms of the ROC curves, there is a nearly even AUC spacing of 0.05 each from SSHW < HBS < SSHW + HBS. To complete the model, we also added an interaction term HBS × SSHW to the logistic regression, but this term did not acquire a significant coefficient and thus could not improve the classification. In the optimally discriminating regression model, we obtained a joint functional HBS-SSHW-score $X = 0.44 \cdot SSHW + 1.17 \cdot HBS - 16.3$ in the exponent.

This classification shows that for 5′ss and GT-sites, the HBond score is more informative than the 'SRE neighborhood parameter' SSHW alone, but SSHW adds as much classification value to HBS as HBS adds to SSHW.

DISCUSSION

In this manuscript, we present *in silico* designed sequences with arbitrary *a priori* prescribed splicing regulatory properties, quantitatively represented by a constant HEXplorer score profile. We comprehensively validated *in silico* predictions on splice site recognition in a massively parallel splicing assay on >3000 sequences. From an MS analysis of proteins binding to exemplary *in silico* designed SRE sequences, we confirmed splicing regulatory proteins binding specifically to enhancing, neutral or silencing sequences. We complementarily selected 320 601 pairs of high confidence 5′ss and neighboring exonic GT sites from our large human fibroblast RNA-seq dataset, as well as 285 441 pairs from our human endothelium RNA-seq dataset, and derived two-dimensional splice site usage landscapes from gene-and-sample normalized RNA-seq reads. These GNR landscapes served as basis for a logistic 5′ss usage prediction model, depending on both U1 snRNA complementarity and HEXplorer score. This model greatly improved 5′ss discrimination between strong but unused exonic GT sites and annotated highly used 5′ss by adding the splice site HEXplorer weight to the classification algorithm based exclusively on HBond score.

In principle, sequences with prescribed splicing regulatory properties could be obtained by inserting single known SRE motifs into assumed splicing neutral sequences, like the octamer 'CCAAACAA' that has been proposed and tested as a building block for splicing neutral sequences (19,50). However, even in this seemingly simple case, concatenation of the octamer 'CCAAACAA' accidentally creates a 'CANC' motif as potential SRSF3 binding site (51), altering the splicing regulatory properties of the single octamer (1). In this study, we used the HEXplorer algorithm (30) to design splice enhancing, silencing and neutral octamers, *ab initio* avoiding accidental HEXplorer profile fluctuations possibly introduced by concatenation. Reversing the above sketched process, we generated putative SRP binding sites by using the HEXplorer algorithm without restricting the sequences to single SR- or hnRNP binding sites, and we experimentally confirmed the splicing regula-

tory properties of *in silico* designed octamer sequences in a massively parallel splicing assay.

Assuming a proportional interplay between 5′ss strength (HBS) and SRE impact (ΔHZ_{EI}), a rough guesstimate of an equivalence between HBS and ΔHZ_{EI} can be gleaned from the experiments (Figure 1B): We observed that in the presence of just the splicing neutral octamer, an HBS of 17.5 was required for exon inclusion. For a weaker 5′ss with HBS = 15.0, SRE neighborhoods with $\Delta HZ_{EI} \leq 70$ did not suffice to support exon inclusion while $\Delta HZ_{EI} = 100$ did (data not shown), so that 2.5 HBS units seem to correspond to $\Delta HZ_{EI} \sim 100$. This conclusion is only valid in the context of our splicing reporter.

In a recent study, Wong *et al.* (3) systematically tested all possible 5′ss sequences in three genomic contexts, using an MPSA approach with a random 5′ss library. They conclude that 5′ss strength is the main determinant of 5′ss usage, while 5′ss context is less important. This is consistent with our findings. While Wong *et al.* systematically varied 5′ss sequences, we did so with exonic 5′ss neighborhoods in our random octamer library approach. Systematical co-variation of both 5′ss sequence and octamer context, however, would demand a considerably larger plasmid library with 65 536 × 32 768 possible different sequences, which exceeded our resources. Our RNA-seq analysis in samples from two different tissues, however, permitted systematic computation of splice site usage landscapes for a wide variety of naturally occurring 5′ splice sites and contexts, and it fully confirmed the dominance of 5′ss strength over neighborhood context. This is also reflected in the HBS and SSHW coefficients of the combined score derived from the 5′ss and exonic GT-site discrimination task.

Our novel RNA-seq based 5′ss usage landscape concept quantifies the usage of exonic GT-sites relative to their nearby 5′ss by their log-gene-normalized read ratio LGNRr, as function of both HBS and SSHW differences 'GT-site–5′ss'. We would expect a similar structure of the 5′ss usage landscape plotted vs. ΔMaxEnt score instead of ΔHBS (24). Necessarily, any choice of SRE neighborhood size is arbitrary. However, several studies indicate only weak dependence on neighborhood size: Putative exonic splicing enhancer and silencer octamer (PESX) frequencies have been shown to remain rather constant in 100 nt long composite exons (50 nt center and 25 nt ends) and introns (59). Similarly, the distributions of the top 400 ESEseqs and ESSseqs showed little variation in 100 nt long composite exons and introns (29). Eventually, individual hexamer weights used in the HEXplorer definition were highly correlated when derived from 100 nt or 30 nt wide 5′ss neighborhoods. Therefore, we expect to capture relevant SRP binding sites within the chosen 50 nt neighborhoods. Some RNA-binding proteins, however, may bind cooperatively to clusters of sites or interact with each other—effects that are not intrinsically reflected in any RESCUE-type algorithm based on n-mer frequencies. If such synergistic behavior had pronounced effects, it would be expected to be revealed in the extensive mapping and characterization of RNA elements recognized by the large collection of human RBPs, which consist of typically eight or less nucleotides (34).

In a tentative first evaluation of LGNRr landscape prediction of GT-site usage induced by 5′ss mutations, we found a significantly higher proportion of high enhancement values for mutations activating cryptic GT-sites than for those leading to exon skipping, although LGNRr landscape predictions did not permit specific discrimination between these groups. In the classification of 5′ss versus unused GT-sites, however, both sensitivity and specificity were significantly improved by using splice site HEXplorer weight in addition to HBond score. Thus, local sequence information on a potential splice site and its SRE neighborhood can be unified to a single 'functional 5′ss' description.

On the other hand, state-of-the-art machine learning algorithms for splice site prediction and mutation assessment have been developed and evaluated in recent years. Using a modular architecture, MMSplice encompasses six neural network modules covering donor and acceptor sites, as well as their respective exonic and intronic neighborhoods, and it outperformed previous splicing prediction models in the 'Critical Assessment of Genome Interpretation' (CAGI) challenge (31,60–62). Designed as a 32-layer deep neural network built from residual blocks, the deep learning tool SpliceAI achieved an impressive 95% top-*k* accuracy in identifying splice sites from DNA sequence alone, however using features from a very wide reference—and not the patient's own—genomic region of 10 000 nt around the index site (32). As all machine learning algorithms, these models appear as black boxes to the user, and their splice site usage predictions are not transparent in terms of biological mechanisms: they may well successfully apply features with no biological meaning. In contrast, our RNA-seq based GT-site-to-5′ss usage ratio landscape model clearly shows both effects of 5′ss strength and neighboring splicing regulatory elements.

DATA AVAILABILITY

Illumina sequencing data has been deposited on the NCBI Sequence Read Archive under accession number PRJNA782097. Computational analyses were performed using custom R scripts, which are available at https://github.com/caggtaagtat/SDusage. Liquid chromatography–tandem mass spectrometry (LC–MS/MS) data are included in Supplementary File S3 and have been deposited in PRIDE under project accession PXD030139.

ACCESSION NUMBERS

The fibroblast RNA-seq dataset (43) analyzed in this study is available through ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-4652.

The endothelium RNA-seq dataset (18) analyzed in this study is available through ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-7647.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Björn Wefers and Yvonne Dickschen for technical assistance and Philipp Peter for implementing the

HEXplorer algorithm on our RNA website. We also thank Gereon Poschmann for helping with MS-analysis evaluation. We thank all lab members for discussion and critical reading of the manuscript. We would like to thank Douglas Black for providing the PTB antibody. We also thank the anonymous reviewers for instructive criticism and suggestions.

FUNDING

Deutsche Forschungsgemeinschaft (DFG) [SCHA 909/4-1 to H.S.]; Jürgen Manchot Stiftung, Düsseldorf (to L.M., A.L.B., R.G., A.R., H.S.); Stiftung für AIDS-Forschung, Düsseldorf (to H.S.); Forschungskommission of the Medical Faculty, Heinrich Heine Universität Düsseldorf [2020-12 to H.S.]. Funding for open access charge: Heinrich Heine University, Düsseldorf.
Conflict of interest statement. None declared.

REFERENCES

1. Erkelenz,S., Theiss,S., Kaisers,W., Ptok,J., Walotka,L., Muller,L., Hillebrand,F., Brillen,A.L., Sladek,M. and Schaal,H. (2018) Ranking noncanonical 5′ splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. *Genome Res.*, **28**, 1826–1840.
2. Zhuang,Y. and Weiner,A.M. (1986) A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. *Cell*, **46**, 827–835.
3. Wong,M.S., Kinney,J.B. and Krainer,A.R. (2018) Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol. Cell*, **71**, 1012–1026.
4. Ptok,J., Muller,L., Theiss,S. and Schaal,H. (2019) Context matters: regulation of splice donor usage. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1862**, 194391.
5. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
6. Kammler,S., Leurs,C., Freund,M., Krummheuer,J., Seidel,K., Tange,T.O., Lund,M.K., Kjems,J., Scheid,A. and Schaal,H. (2001) 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*, **7**, 421–434.
7. Freund,M., Asang,C., Kammler,S., Konermann,C., Krummheuer,J., Hipp,M., Meyer,I., Gierling,W., Theiss,S., Preuss,T. *et al.* (2003) A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res.*, **31**, 6963–6975.
8. Sun,H. and Chasin,L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.*, **20**, 6414–6425.
9. Long,J.C. and Caceres,J.F. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.*, **417**, 15–27.
10. Anko,M.L. (2014) Regulation of gene expression programmes by serine-arginine rich splicing factors. *Semin. Cell Dev. Biol.*, **32**, 11–21.
11. Martinez-Contreras,R., Cloutier,P., Shkreta,L., Fisette,J.F., Revil,T. and Chabot,B. (2007) hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.*, **623**, 123–147.
12. Busch,A. and Hertel,K.J. (2012) Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip. Rev. RNA*, **3**, 1–12.
13. Erkelenz,S., Mueller,W.F., Evans,M.S., Busch,A., Schoneweis,K., Hertel,K.J. and Schaal,H. (2013) Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*, **19**, 96–102.
14. Reber,S., Stettler,J., Filosa,G., Colombo,M., Jutzi,D., Lenzken,S.C., Schweingruber,C., Bruggmann,R., Bachi,A., Barabino,S.M. *et al.* (2016) Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. *EMBO J.*, **35**, 1504–1521.
15. Shenasa,H., Movassat,M., Forouzmand,E. and Hertel,K.J. (2020) Allosteric regulation of U1 snRNP by splicing regulatory proteins controls spliceosomal assembly. *RNA*, **26**, 1389–1399.
16. Buratti,E., Baralle,M., De Conti,L., Baralle,D., Romano,M., Ayala,Y.M. and Baralle,F.E. (2004) hnRNP H binding at the 5′ splice site correlates with the pathological effect of two intronic mutations in the NF-1 and TSHbeta genes. *Nucleic Acids Res.*, **32**, 4224–4236.
17. Matera,A.G. and Wang,Z. (2014) A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.*, **15**, 108–121.
18. Merk,D., Ptok,J., Jakobs,P., von Ameln,F., Greulich,J., Kluge,P., Semperowitsch,K., Eckermann,O., Schaal,H., Ale-Agha,N. *et al.* (2021) Selenoprotein T protects endothelial cells against lipopolysaccharide-induced activation and apoptosis. *Antioxidants (Basel)*, **10**, 1427.
19. Zhang,X.H., Arias,M.A., Ke,S. and Chasin,L.A. (2009) Splicing of designer exons reveals unexpected complexity in pre-mRNA splicing. *RNA*, **15**, 367–376.
20. Lim,K.H., Ferraris,L., Filloux,M.E., Raphael,B.J. and Fairbrother,W.G. (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 11093–11098.
21. Sterne-Weiler,T., Howard,J., Mort,M., Cooper,D.N. and Sanford,J.R. (2011) Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.*, **21**, 1563–1571.
22. Caminsky,N., Mucaki,E.J. and Rogan,P.K. (2014) Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Res*, **3**, 282.
23. Soukarieh,O., Gaildrat,P., Hamieh,M., Drouet,A., Baert-Desurmont,S., Frebourg,T., Tosi,M. and Martins,A. (2016) Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS Genetics*, **12**, e1005756.
24. Hartmann,L., Theiss,S., Niederacher,D. and Schaal,H. (2008) Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases?*Front. Biosci.*, **13**, 3252–3272.
25. Wai,H.A., Lord,J., Lyon,M., Gunning,A., Kelly,H., Cibin,P., Seaby,E.G., Spiers-Fitzgerald,K., Lye,J., Ellard,S. *et al.* (2020) Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med.*, **22**, 1005–1014.
26. Grodecka,L., Buratti,E. and Freiberger,T. (2017) Mutations of Pre-mRNA splicing regulatory elements: are predictions moving forward to clinical diagnostics?*Int. J. Mol. Sci.*, **18**, 1668.
27. Canson,D., Glubb,D. and Spurdle,A.B. (2020) Variant effect on splicing regulatory elements, branchpoint usage, and pseudoexonization: strategies to enhance bioinformatic prediction using hereditary cancer genes as exemplars. *Hum. Mutat.*, **41**, 1705–1721.
28. 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.
29. Ke,S., Shang,S., Kalachikov,S.M., Morozova,I., Yu,L., Russo,J.J., Ju,J. and Chasin,L.A. (2011) Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.*, **21**, 1360–1374.
30. Erkelenz,S., Theiss,S., Otte,M., Widera,M., Peter,J.O. and Schaal,H. (2014) Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res.*, **42**, 10681–10697.
31. Cheng,J., Nguyen,T.Y.D., Cygan,K.J., Celik,M.H., Fairbrother,W.G., Avsec,Z. and Gagneur,J. (2019) MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.*, **20**, 48.
32. Jaganathan,K., Kyriazopoulou Panagiotopoulou,S., McRae,J.F., Darbandi,S.F., Knowles,D., Li,Y.I., Kosmicki,J.A., Arbelaez,J., Cui,W., Schwartz,G.B. *et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell*, **176**, 535–548.
33. Rowlands,C.F., Baralle,D. and Ellingford,J.M. (2019) Machine learning approaches for the prioritization of genomic variants impacting Pre-mRNA splicing. *Cells*, **8**, 1513.
34. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.Y., Cody,N.A.L., Dominguez,D. *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.
35. Braun,S., Enculescu,M., Setty,S.T., Cortes-Lopez,M., de Almeida,B.P., Sutandy,F.X.R., Schulz,L., Busch,A., Seiler,M., Ebersberger,S. *et al.* (2018) Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat. Commun.*, **9**, 3315.
36. Selden,R.F., Howie,K.B., Rowe,M.E., Goodman,H.M. and Moore,D.D. (1986) Human growth hormone as a reporter gene in

regulation studies employing transient gene expression. *Mol. Cell. Biol.*, **6**, 3173–3179.
37. Brillen,A.L., Schoneweis,K., Walotka,L., Hartmann,L., Muller,L., Ptok,J., Kaisers,W., Poschmann,G., Stuhler,K., Buratti,E. *et al.* (2017) Succession of splicing regulatory elements determines cryptic 5ss functionality. *Nucleic Acids Res.*, **45**, 4202–4216.
38. Chomczynski,P. and Sacchi,N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal.Biochem.*, **162**, 156–159.
39. Brillen,A.L., Walotka,L., Hillebrand,F., Muller,L., Widera,M., Theiss,S. and Schaal,H. (2017) Analysis of competing HIV-1 splice donor sites uncovers a tight cluster of splicing regulatory elements within exon 2/2b. *J. Virol.*, **91**, e00389-17.
40. Ke,S., Zhang,X.H. and Chasin,L.A. (2008) Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res.*, **18**, 533–543.
41. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.
42. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
43. Kaisers,W., Boukamp,P., Stark,H.J., Schwender,H., Tigges,J., Krutmann,J. and Schaal,H. (2017) Age, gender and UV-exposition related effects on gene expression in in vivo aged short term cultivated human dermal fibroblasts. *PLoS One*, **12**, e0175657.
44. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
45. Chhangawala,S., Rudy,G., Mason,C.E. and Rosenfeld,J.A. (2015) The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol*, **16**, 131.
46. Kopylova,E., Noe,L. and Touzet,H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.
47. Dobin,A. and Gingeras,T.R. (2015) Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics*, **51**, 11.14.1–11.14.19.
48. Kaisers,W., Schaal,H. and Schwender,H. (2015) rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples. *Bioinformatics*, **31**, 1663–1664.
49. Kaisers,W., Ptok,J., Schwender,H. and Schaal,H. (2017) Validation of splicing events in transcriptome sequencing data. *Int. J. Mol. Sci.*, **18**, 1110.
50. Arias,M.A., Lubkin,A. and Chasin,L.A. (2015) Splicing of designer exons informs a biophysical model for exon definition. *RNA*, **21**, 213–229.
51. Hargous,Y., Hautbergue,G.M., Tintaru,A.M., Skrisovska,L., Golovanov,A.P., Stevenin,J., Lian,L.Y., Wilson,S.A. and Allain,F.H. (2006) Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. *EMBO J.*, **25**, 5126–5137.
52. Cyphert,T.J., Suchanek,A.L., Griffith,B.N. and Salati,L.M. (2013) Starvation actively inhibits splicing of glucose-6-phosphate dehydrogenase mRNA via a bifunctional ESE/ESS element bound by hnRNP K. *Biochim. Biophys. Acta*, **1829**, 905–915.
53. Afroz,T., Cienikova,Z., Clery,A. and Allain,F.H.T. (2015) One, two, three, four! How multiple RRMs read the genome sequence. *Methods Enzymol.*, **558**, 235–278.
54. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
55. Damgaard,C.K., Kahns,S., Lykke-Andersen,S., Nielsen,A.L., Jensen,T.H. and Kjems,J. (2008) A 5′ splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol. Cell*, **29**, 271–278.
56. Erkelenz,S., Hillebrand,F., Widera,M., Theiss,S., Fayyaz,A., Degrandi,D., Pfeffer,K. and Schaal,H. (2015) Balanced splicing at the Tat-specific HIV-1 3′ss A3 is critical for HIV-1 replication. *Retrovirology*, **12**, 29.
57. Tyanova,S., Temu,T., Sinitcyn,P., Carlson,A., Hein,M.Y., Geiger,T., Mann,M. and Cox,J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, **13**, 731–740.
58. Alioto,T.S. (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.*, **35**, D110–D115.
59. Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.*, **18**, 1241–1250.
60. Rhine,C.L., Neil,C., Glidden,D.T., Cygan,K.J., Fredericks,A.M., Wang,J., Walton,N.A. and Fairbrother,W.G. (2019) Future directions for high-throughput splicing assays in precision medicine. *Hum. Mutat.*, **40**, 1225–1234.
61. Soemedi,R., Cygan,K.J., Rhine,C.L., Wang,J., Bulacan,C., Yang,J., Bayrak-Toydemir,P., McDonald,J. and Fairbrother,W.G. (2017) Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.*, **49**, 848–855.
62. Cheng,J., Celik,M.H., Nguyen,T.Y.D., Avsec,Z. and Gagneur,J. (2019) CAGI 5 splicing challenge: improved exon skipping and intron retention predictions with MMSplice. *Hum. Mutat.*, **40**, 1243–1251.