Modelling splicing outcome by combining 5’ss strength and splicing regulatory elements

Modellierung des Spleißergebnisses durch Kombination von 5'ss Stärke und spleißregulierenden Elementen

Inaugural-Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Johannes Ptok

aus Ingolstadt

Düsseldorf, November 2023

aus dem Institut für Virologie

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. H. Schaal
2. Prof. Dr. M. Feldbrügge

Tag der mündlichen Prüfung:

Table of contents

Summary

Pre-mRNA splicing is an mRNA processing step in which intronic sequences are excised and exonic sequences are ligated. Variations in splice site selection, so-called alternative splicing, enables expression of different proteins, originating from the same pre-mRNA transcript. Initially, an RNA duplex is formed between the 5' splice site (5'ss) and the free 5' end of the U1 snRNA. Splicing regulatory elements (SREs) recruit splicing regulatory proteins (SRPs), like hnRNP or SR proteins, that can greatly influence 5'ss usage, depending on their position. Mutations affecting these regulatory elements can lead to aberrant splicing, which can induce several diseases. Thus, any algorithm that estimates only intrinsic splice site strength is insufficient to correctly capture the impact of mutations on splice site selection, and thus whether they potentially lead to a reduction in functional protein. Publication I reviewed the influence of SREs on 5'ss selection and predictive algorithms, like the HEXplorer. The Splice Site HEXplorer Weight summarizes the overall enhancing or repressing properties of the immediate sequence context of splice sites. Studying 5'ss usage competition between neighbouring 5'ss of a large RNA-seq data set of fibroblasts in Publication II showed, that the differences in HBS and SSHW had to be considered together, to best model predicted 5'ss usage. Intrinsic strength, however, had a greater impact on 5'ss recognition than the SSHW, which was also shown for most likely non-pathogenic mutations of healthy individuals of the 1000 Genome project (unpublished paper I). For fast analysis of sequence variations in context of splicing, the VarCon algorithm was developed in Publication III, which converts ambiguous positional information to genomic positions. To analyze why cardiovascular endothelial cells treated with high concentrations of low-density lipoprotein show low levels of functional NO-synthase 3 (NOS3) protein and high oxidative stress, alternative NOS3 splicing was studied in Publication IV. However, not miss-splicing, but an internal promotor most likely resulted in the truncated NOS3 protein, finally inducing apoptosis. The first 20 amino acids of the Apurinic/Apyrimidinic Endodeoxyribonuclease 1 (APEX1), were able to reduce stress-induced apoptosis in endothelial cells, via upregulation of SELENOT as described in Publication V. To manipulate the SSHW of 5'ss in reporter systems or expression vectors, the ModCon algorithm was developed in Publication VI, which applies a genetic algorithm to manipulate SRP binding via synonymous substitutions. Changes in SRP composition of RNAs, however, might also affect process like RNA export, as observed in Publication VII, analyzing export repressing viral sequence elements, that preferably recruited hnRNP proteins.

Deutsche Zusammenfassung

Prä-mRNA-Spleißen ist ein Schritt der mRNA Prozessierung, bei dem intronische Sequenzen entfernt und exonische Sequenzen ligiert werden. Variable Spleißstellen-Wahl, sogenanntes alternatives Spleißen, ermöglicht die Expression unterschiedlichster Proteine, ausgehen vom gleichen prä-mRNA-Transkript. Zunächst wird ein RNA-Duplex zwischen der 5'-Spleißstelle (5'ss) und dem freien Ende der U1 snRNA gebildet. Spleißregulatorische Elemente (SREs) rekrutieren spleißregulatorische Proteine (SRPs), wie hnRNP- oder SR-Proteine, die je nach ihrer Position die Erkennung von Spleißstellen stark beeinflussen können. Mutationen, die diese regulatorischen Elemente stören, können zu fehlerhaftem Spleißen führen, was verschiedensten Krankheiten auslösen kann. Algorithmen wie der HBond-Score (HBS), die die intrinsische Stärke von 5‘ss abschätzten, wie z.B. der HBond-Score (HBS), reichen daher alleine nicht immer aus, um die Auswirkungen von Mutationen auf die Nutzung von Spleißstellen korrekt zu erfassen und dadurch festzustellen, ob sie möglicherweise zu einer Verringerung an funktionalem Protein führen. In der Veröffentlichung I wurde der Einfluss von SREs auf die 5'ss-Nutzung und Algorithmen, die diese vorhersagen, wie z.B. der HEXplorer, untersucht. Das Splice Site HEXplorer Weight (SSHW) fasst die insgesamt fördernden oder hemmenden Eigenschaften des unmittelbaren Sequenzkontexts von Spleißstellen zusammen. Die Untersuchung von 5'ss Konkurrenz Situationen in einem großen RNA-seq-Datensatz von Fibroblasten in Publikation II zeigte, dass Unterschiede in HBS und SSHW zusammen berücksichtigt werden müssen, um die vorhergesagte 5'ss-Nutzung am besten zu modellieren. Die intrinsische Stärke hatte jedoch einen größeren Einfluss auf 5'ss-Erkennung als das SSHW, was auch für wahrscheinlich nicht-pathogene Mutationen gesunder Individuen des 1000-Genom-Projekts gezeigt werden konnte (unveröffentlichtes Manuskript I). Um schnell Sequenzvariationen auf ihren Einfluss auf Spleißstellen-Nutzung zu analysieren, wurde in Publikation III der VarCon-Algorithmus entwickelt, der mehrdeutige Positionsinformationen in genomische Positionen umwandelt. Um zu verstehen, warum kardiovaskuläre Endothelzellen, die mit hohen Konzentrationen von Lipoprotein niedriger Dichte (LDL) behandelt wurden, niedrige Konzentrationen von funktionellem NO-Synthase-3-Protein (NOS3) und hohen oxidativen Stress aufweisen, wurde in Publikation IV alternatives NOS3-Spleißen untersucht. Doch nicht Fehl-Spleißen, sondern Nutzung eines internen Promotors führte höchstwahrscheinlich zu einem verkürzten NOS3-Protein, das schließlich Apoptose auslöste. Die ersten 20 Aminosäuren der Apurin-/Apyrimidin-Endodeoxyribonuklease 1 (APEX1) waren

in der Lage, die stressinduzierte Apoptose in Endothelzellen über eine Hochregulierung von SELENOT zu reduzieren, wie in Publikation V beschrieben. Um das SSHW von 5'ss in Reportersystemen oder Expressions-vektoren zu manipulieren, wurde in Publikation VI der ModCon-Algorithmus entwickelt, der einen genetischen Algorithmus anwendet, um SRP-Bindung über synonyme Substitutionen zu manipulieren. Veränderungen in der SRP-Besetzung von RNAs könnten sich auch auf Prozesse wie den RNA-Export auswirken, wie in Publikation VII beobachtet wurde, in der exportunterdrückende virale Sequenzelemente analysiert wurden, die vorzugsweise hnRNP-Proteine rekrutieren.

1. Introduction

1.1. Human pre-mRNA splicing

Originally discovered studying the processing of type II adenovirus mRNA transcripts, splicing proved to be a step of precursor mRNA maturation, which affects almost all genes expressed in humans [5]. It describes a cellular process, which removes continuous sequence segments, called introns, from the precursor RNA transcript, while the remaining sequence segments, called exons, are joined together [6]. Variations in intronic boundary selection enables the production of various mature mRNA isoforms, potentially encoding different protein isoforms, originating from the same precursor mRNA [7]. This process, called alternative splicing, can be found in most higher eukaryotes and affects around 95% of human genes. Constitutively spliced exons, which by definition are present in every mRNA isoform of a specific gene, can also be found in most genes [8]. The cellular machinery, that recognizes exon-intron borders and accomplishes splicing is called spliceosome [3].

1.1.1 The spliceosome

The spliceosome, a multi-ribonucleoprotein complex, consists of five uridine-rich small nuclear ribonucleoprotein particles (U snRNPs), called U1, U2, U4/U6 and U5 which already co-transcriptionally assemble at exon-intron borders of the precursor RNA transcript. It recognizes *cis*-acting sequence elements, such as the 5' splice site, or splice donor (SD), and the 3' splice site, or splice acceptor (SA), which define the exon/intron boundaries (see Figure 1) [9].

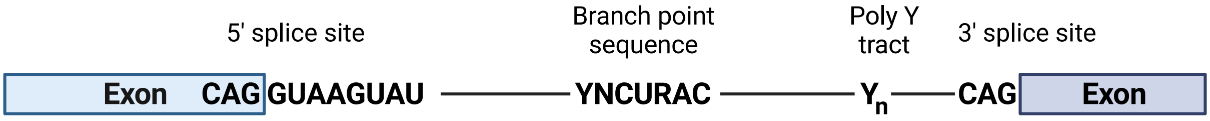


Figure 1: *Cis*-acting sequence elements defining exon/intron boundaries. An intron between two exons requires following structures to be recognized. The 5' splice site, the branch site, the poly-pyrimidine tract (Poly Y tract), and the highly conserved terminal AG-dinucleotide which are integral components of an intron. G=glycine, U=uracil, R=purine, A=adenosine, Y=pyrimidine, N=any nucleotide, C=cytosine.

The splice donor site is characterized by the 11 nucleotides long splice donor sequence, with three nucleotides in the exon and eight nucleotides in the intron, including the invariant GU-dinucleotide at the upstream end of the intronic sequence. The splice acceptor consists of three sequence elements, namely the branch point sequence (BPS), the poly-pyrimidine-tract (PPT) located 15-40nt upstream of the intron 3'-end, the invariant terminal AG-dinucleotide, and three exonic nucleotides downstream of the AG-dinucleotide [10, 11].

The biochemical reaction, catalyzed by the spliceosome, involves two transesterification reactions (Figure 2). First, the oxygen of the 2´ hydroxyl group of the branch adenosine, carries out a nucleophilic attack on the phosphor atom, which attaches the 5´ exon to the intron generating a lariat/3´ exon intermediate. Next, the oxygen atom of the 5´ exon attacks the phosphorous atom that connects the intron and the 3´ exon, releasing the lariat intron and joining the two exons together [3, 11].

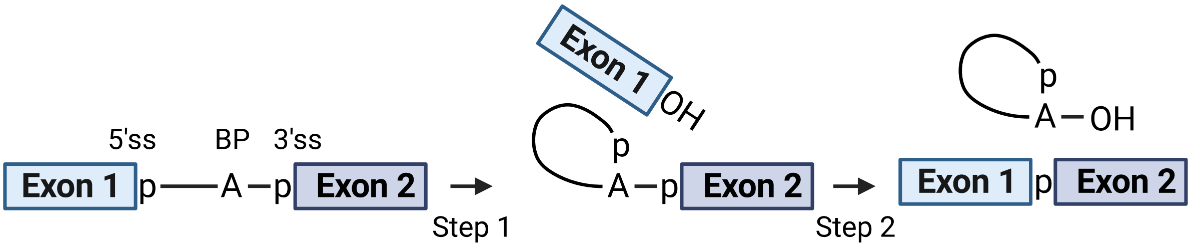


Figure 2: Splicing requires two transesterification reactions, resulting in a lariat structure after the first splicing step and two spliced exons after the second splicing step. Branch point (BP) [3]. Permission to include the figure was kindly granted by Cold Spring Harbor Laboratory Press.

In the first, ATP-independent, step of spliceosome assembly, the 5' splice site sequence is bound by the free 5' end of the U1 snRNA, forming complex E (Figure 3) [12]. This recognition can be positively influenced through binding of the upstream SA by the U2 snRNP, forming the exon definition complex with the U1 snRNP. Next, U2AF65 and U2AF35 are recruited to the PPT and the YAG motif of the downstream 3' splice site, resulting in the binding of the branch point sequence by the U2 snRNP, which interacts with the U1 snRNP, forming the pre-spliceosome or A complex. This process is catalyzed by the DExD/H helicases pre-mRNA-processing 5 (Prp5) and Sub2 and requires ATP.

Subsequently, the pre-assembled tri-snRNP U4/U6/U5 is recruited forming the B complex, which undergoes several rearrangements, releasing U1 and U4 snRNPs [13]. The DEAH-box RNA helicase Prp2 then activates the B complex which catalyzes the first transesterification, resulting in complex C, composed of the first 5' exon and the lariat/3´ exon intermediate. The second step of splicing, catalyzed by the C complex after further structural rearrangements, results in the post-catalytic spliceosome, containing the spliced exons and the lariat intron. Finally, the remaining U-snRNPs are released and recycled for another round of splicing, and the lariat intron is degraded [14].

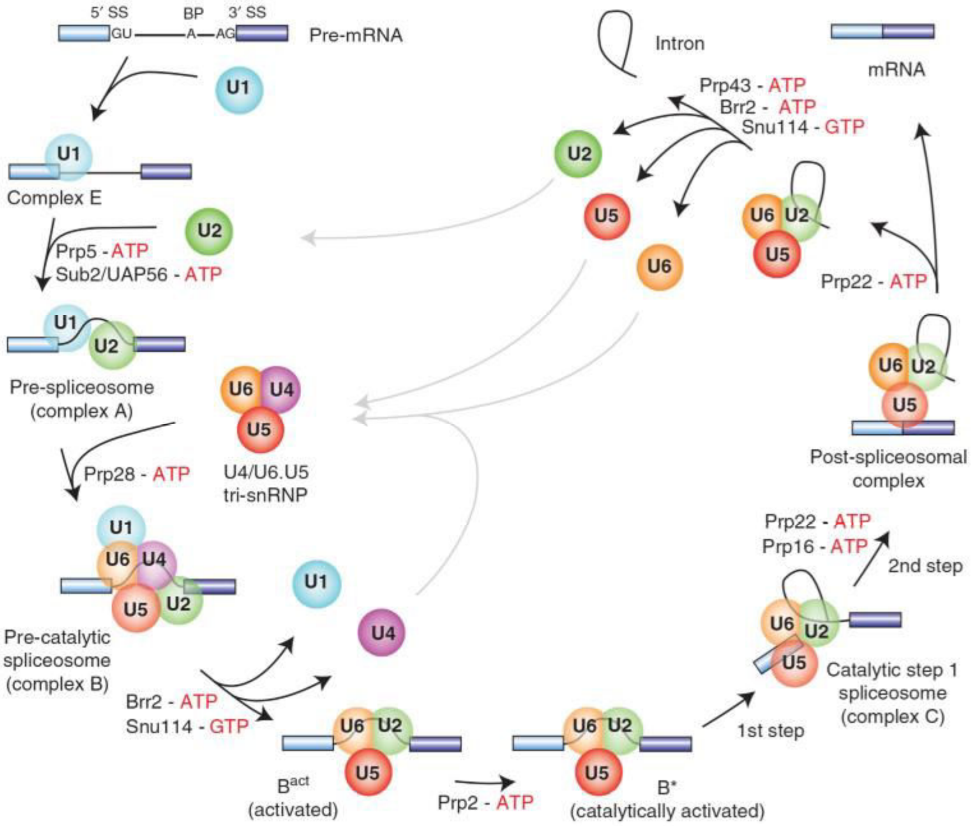


Figure 3: The stepwise assembly of the spliceosome. First the U1 snRNA recognizes the sequence at the 5′ splice site resulting in complex E. Then the U2 snRNP binds to the 3′ splice site forming complex A, followed by the association of U6/U4/U5 tri-snRNP, generating complex B. Subsequently, this pre-catalytic spliceosome is activated by undergoing various molecular rearrangements, releasing U1 and U4 snRNP. The remaining snRNPs, building complex C after the first splicing step, splice the two exons together and dissociate. The free lariat intron becomes degraded [3]. Permission to include the figure was kindly granted and copyright reserved by Cold Spring Harbor Laboratory Press.

The presence of more than one suitable 5′ or 3′ splice site in proximity to each other can result in splice site competition. If e.g. an RNA transcript holds two 5′ splice sites, with equal ability to recruit U1-snRNP, which could result in two different complexes at the upstream end of that intron, two transcript isoforms can subsequently be observed, that, depending on their distance from the upstream splice acceptor, use one of the two 5′ splice sites.

1.1.2 Alternative splicing

The degree of freedom in the selection of splice sites can lead to multiple transcript isoforms of the same gene in a process called alternative splicing (Figure 4). Resulting transcripts can vary in exon length via usage of alternative splice sites. Another possibility is the exclusion of whole exons, called exon skipping, or the retention of introns in the mature RNA. A special form of alternative splicing are mutually exclusive exons, with either one exon or the other being included in the mature RNA transcript.

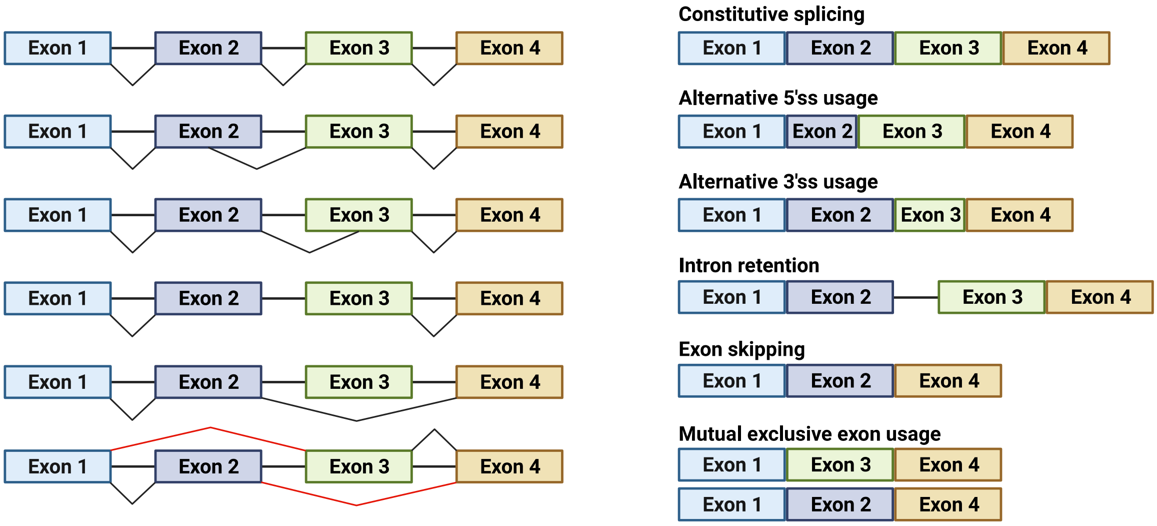


Figure 4: Ways of pre-mRNA alternative splicing. Alternative 5'ss or 3'ss usage can affect exon length. During intron retention, the intron is still part of the mature RNA. Whole exons can be skipped during splicing, resulting in a final transcript missing its sequence. Some exon-pairs were described to be mutually exclusive, with either one or the other being included.

Apart from the core splicing signals, mentioned before, splice site usage is additionally affected by *cis*-acting splicing regulatory elements (SREs) recruiting *trans*-acting splicing regulatory proteins (SRPs), like serine/arginine rich proteins (SR proteins) or heterogeneous nuclear ribonucleoproteins (hnRNPs) [15]. Depending on their binding position, these proteins can enhance or repress splice site usage [16]. SR proteins enhance usage of downstream 5' splice sites and upstream 3' splice sites but repress usage of downstream 3' splice sites and upstream 5' splice sites. hnRNPs act in the opposite manner [17].

1.1.3 Splice site strength

Recognition of the 5' splice site (5'ss or splice donor) during splicing, requires the formation of an RNA duplex between the 11 nucleotides at the free 5' end of the U1 snRNA and the 11 nucleotide long splice donor sequence defining positions -3 to +8 of the exon-intron border

[18]. The higher the complementarity of the two sequences, the more stable the duplex formation [19]. Recognition of the 3’ splice site (3’ ss or splice acceptor) during splicing, requires binding of U2AF65 and U2AF35 to the PPT and the YAG motif of the 3’ ss respectively, resulting in recruitment of the U2 snRNP to the RNA [14].

A widely used algorithm that also considers neighborhood relationships of involved nucleotides to evaluate the strength of splice sites is the MaxEntScan score [20]. The higher the score, the better the predicted splice site recognition and usage.

As an alternative approach to assess 5’ss strength, the HBond score (HBS) is based on the experimentally determined strength of mutated 5’ss taking into account the number and density of HBonds that can potentially be formed between U1 snRNA and 5’ss sequences. It takes all 11 nucleotides of it into account (3 exonic and 8 intronic), in contrast to the MaxEntScan score, which is based on only 9 nucleotides (3 exonic and 6 intronic) [18]. 5’ss sequences showing only the invariant GT dinucleotide at 5’ss position +1/+2 are scored with the lowest HBS of 1.8, whereas the fully complementary 5‘ss sequence is denoted with a HBS of 23.8. However, estimating the intrinsic strength of donor sequences with alternative dinucleotides at splice donor positions +1/+2, like the second most frequent GC-dinucleotide (< 1% of human donor sites) are currently not evaluated [21].

1.1.4 Splicing regulatory proteins

Another layer of splice site recognition relies on the binding pattern of splicing regulatory proteins (SRPs) around splice sites that bind to RNA by sequence elements called splicing regulatory elements (SREs) [22]. Depending on splice site strength and their position to the splice site, they can enhance or repress splice site usage. The two most important protein families, influencing splice site recognition, are serine-arginine rich proteins (SR proteins) and heterogeneous nuclear ribonucleoproteins (hnRNPs). Whereas SR-proteins enhance usage of downstream and repress usage of upstream 5’ss, hnRNPs enhance usage of upstream donors and repress usage of downstream 5’ss. The influence on 3’ss usage is of opposite manner [16].

Splice site usage enhancement can only occur for splice sites of a certain minimal intrinsic strength and becomes less relevant at high intrinsic strengths. RBP mediated repressing of splice site usage, on the other hand, is less efficient for high intrinsic spice site strengths and less relevant for splice sites of weak intrinsic strength [23]. Additionally, expression of SRPs can vary strongly between different cell types, which can make the influence of SREs on splicing dependent on cell type specific RBP expression, to a certain extent [24].

1.1.3.1 SR proteins

Serine/arginine-rich splicing factor (SRSFs, SR proteins) are an evolutionary conserved family of RNA-binding proteins (RBPs), that can be found in all metazoans, plants and some lower eukaryotes [25]. They characteristically show one or two N-terminal RNA recognition motifs (RRMs) followed by a downstream RS domain of at least 50 amino acids, that is rich of serine (S) and arginine (R) amino acids with >40% consecutive RS or SR dipeptide repeats at the C-terminus [26, 27]). RS domains can be found in 160 human proteins, however only 12 of them were classified as classical SR proteins, due to the remaining proteins showing either 1) lack of RRMs, 2) overlapping RRMs and RS domains, 3) additional domains, 4) reverse order of RS domain and RRM or 5) no activity during splicing (Figure 5). The 12 SR proteins (SRSF1-12) range from 20 to 75 kDa in their molecular weight and 164 to 494 amino acids in length.

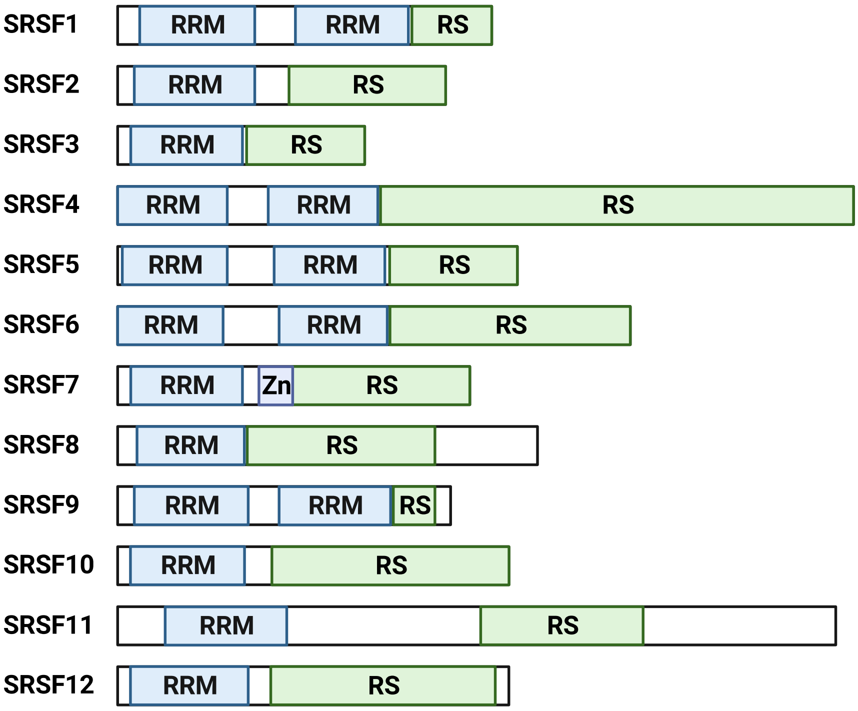


Figure 5. Structure of SR proteins. SR proteins contain one or two RNA recognition motifs (RRM) at the N-terminus and a C-terminal RS domain. SRSF7 additionally contains a zinc finger motif (Zn). This figure was adapted from [4].

The RS domain mostly facilitates protein-protein interactions with other RS domain-containing proteins but can also interact with RNA [28] and serves as a nuclear localization signal [29], whereas the RRM enables protein-RNA interactions. Via the RS domain, SR proteins can interact with the RS domain-containing protein U1-70K, that is associated with the U1 snRNP [30] and U2AF65, that recruits the U2 snRNP together with the U2AF35 [31]. Additional to their role in splicing, biological functions of SR proteins include regulation of transcription [32], RNA stability [33], nuclear export of mRNA [34], mRNA translation [35] and protein degradation [36]. The RS domain of SR proteins additionally enables interaction with polyadenylation factors FIP1 or CPSF6 and the m^{6}A reader YTHDC1, via the RS domain of these proteins [37-39].

The activity of SR proteins is mostly regulated by the phosphorylation status of their RS domain, where serine residues are phosphorylated or dephosphorylated by various protein kinases and phosphatases [40]. Prominent examples include CDC2-like kinases (CLKs) such as CLK1/4 and SRPK1/2, known for their key role regulating proliferation and cell cycle progression [41]. Also, during heat shock or osmotic stress nuclear CLK1/4 kinases are activated, that re-phosphorylate SR proteins [42], whereas the heat-shock-activated serine/arginine phosphatase PP1, was described to dephosphorylate SRSF10 [43]. Altering the phosphorylation status of SR proteins impacts their cellular localization [44], their subnuclear distribution [45], their RNA-binding [46], interaction with the pre-spliceosome [40] and their mRNA export activity [47]. Full phosphorylation of the RS domain is required for spliceosome assembly, whereas dephosphorylation seems to promote splicing catalysis, mRNP packaging and nuclear export [48].

Phosphorylated SR proteins bind RNA transcripts co-transcriptionally mostly downstream of 3’ss and upstream of 5’ss and get hypo-phosphorylated later during splicing [49, 50]. They were shown to position-dependently enhance or repress splice site usage upon binding to the RNA [16]. Upstream of 5’ss, SR protein binding is able to enhance 5’ss usage, via interactions with proteins of the U1 snRNP, like U1-70K [30], U1-C [51], or directly with the U1 snRNP [52, 53]. Downstream of 3’ss, SR protein binding was shown to enhance U2 snRNP recruitment [49, 54, 55]. When bound downstream of 5’ss or upstream of 3’ss, SR protein binding has been described to result in a so-called “dead-end” complex, where stepwise spliceosome assembly seems to be halted during A complex formation, leading to an inhibition of splice site usage [16, 56-59].

Generally, SR proteins can be considered key regulators of human gene expression, allowing adaption of the cellular gene expression program at multiple levels. Correctly predicting binding sites of SR protein, could therefore help to understand the mechanism of a variety of genetic diseases [60, 61]. Although high-affinity binding sites have previously been described for most SR proteins using for instance SELEX (selected evolution of ligands though exponential enrichment) or iCLIP-seq methods, binding of SR proteins seems to be degenerate and redundant to a certain extent [62]. Different SR proteins are often able to bind similar sequence segments or show a general preference of pyrin-rich regions [62-65].

1.1.3.2 hnRNP proteins

Another major family of splicing regulatory proteins are heterogeneous nuclear ribonucleoproteins (hnRNPs). The 20 major members of this protein family are termed hnRNP A-U and range from 34 to 120 kDa in their molecular weight (Figure 6). Although hnRNPs are not as strictly defined by their domain composition, compared to the SR protein family, they were originally defined by their composition of RNA binding domains [66].

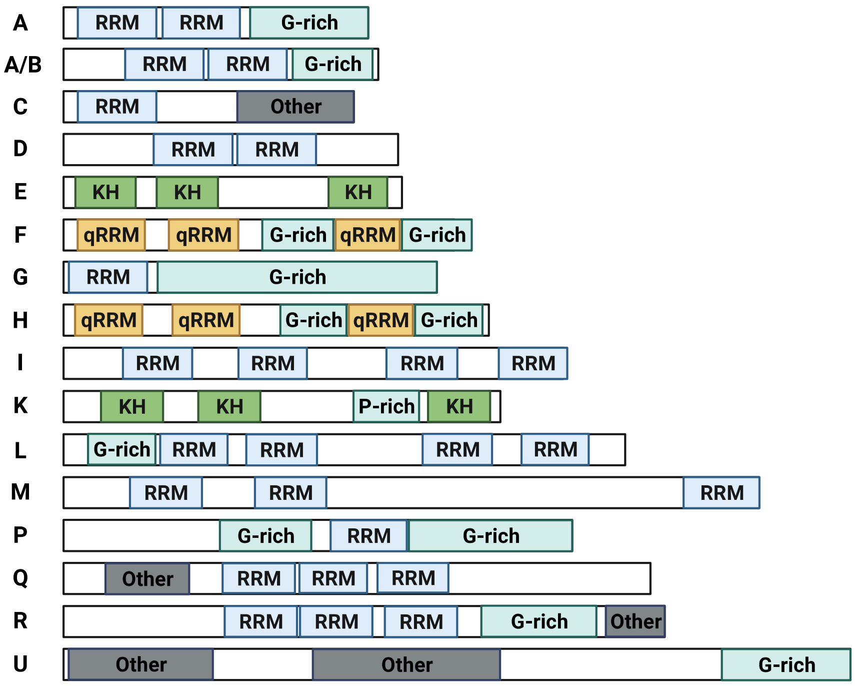


Figure 6: Structure of hnRNPs. The group of hnRNP proteins is highly diverse. Various domains are connected by linker regions. Figure adapted from [1].

Four types of RNA binding domains were found in hnRNP proteins, namely the RNA recognition motif (RRM) found in most hnRNPs, the quasi-RRM, a glycine-rich domain constituting an RGG box and a KH domain [1], as well as auxiliary domains enriched in glycine, arginine, proline, tyrosine, glutamine or asparagine [67, 68]. Various combination of these domains results in high functional diversity of hnRNPs, which includes functions in nuclear export [69, 70], mRNA stability [71], polyadenylation [72], splicing [73, 74], telomer biogenesis [75], translation [76] and RNA editing [77]. Similar to SR proteins, activity of hnRNP proteins was described to be partly regulated via phosphorylation and dephosphorylation, for instance during cell cycle progression [78]. Members of the hnRNPA/B subfamily hnRNPA1 and hnRNPA2 are described to comprise around 60% of all hnRNPs mass and are therefore one of the most studied hnRNP proteins [79].

Although sometimes termed as splicing repressors, they influence splice site usage in an opposite manner, compared to SR proteins. Generally, hnRNPs binding downstream of 3’ss or upstream of 5’ss seem to repress splice site usage but are also able to enhance splice site usage upon biding at the opposite relative position [16, 80]. One way, hnRNPs impact splice site usage is via multimerization at the RNA, sterically hindering either SR protein binding or recruitment of the U1 or U2 snRNP to potential splice sites [81, 82]. For instance, hnRNP C, hnRNP I (or polypyrimidine tract-binding protein, PTB) and hnRNP K have been shown to directly compete with U2AF65 biding at the polypyrimidine tract at 3’ss, repressing its usage [83-85]. Another way, hnRNPs are able to influence splice site usage, is the so-called looping out mechanism, which describes a process, where direct interaction between one hnRNP, that binds the 5’end of an intron, with another hnRNP, that binds the 3’ end of an intron results in looping-out of the intronic sequences in between, supporting to bring both splice sites in close proximity to another. If the same hnRNPs only bind to sequences closer to the middle of the intron this effect is reduced [86-88]. This effect, however, can also result in looping out sequences containing whole exons, inducing exon skipping [89]. As during splicing inhibition via SR proteins, hnRNP binding can than result in the “dead-end” complex of the spliceosome, without reaching its catalytic state [58].

Although binding motifs of hnRNP proteins seem also be somewhat degenerate and redundant, SELEX and iCLIP-seq experiments have shown, that most hnRNP seem to generally prefer to bind pyrimidine-rich sequences, whereas the hnRNP F/H subgroup seems to mostly recognize so-called G runs (DGGGD, with D = A, G or U) [62, 90-92].

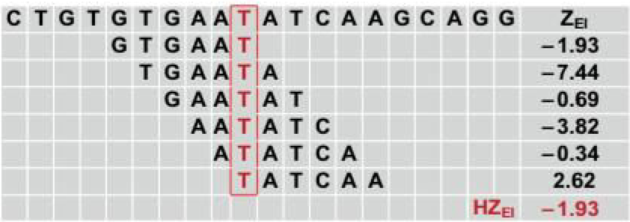
1.2 Predicting splice site usage

Single nucleotide Variants (SNVs) within the coding sequence of a gene, can potentially disrupt the encoded amino acid sequence, causing disease [93]. Approximately 88% of human SNVs associated with disease are, however, not located within the coding sequence of genes, but within intronic and intergenic sequence segments [94]. Regardless of its position, an SNV can additionally disrupt other processes such as, promotor/enhancer activity, splicing and/or binding of RNA binding proteins [95]. Disrupting splicing can influence the encoded amino acid sequence or induce nonsense-mediated decay (NMD) of the RNA transcript. Predicting the effect of SNVs on splicing, could therefore help to identify pathological SNVs as potential risk factors to base recommendations for monitoring or preventive measures against diseases like for instance breast cancer. Correctly estimating, if an SNV influences the splicing pattern, requires robust models of functional splicing. The most important factors to include in these models, that are currently known, are 1) splice donor strength, 2) splice acceptor strength, 3) binding of splicing regulatory proteins (SRPs), and 4) the overall genetic context, like surrounding splice site composition and intron/exon lengths, that can affect intron- and exon-definition [96, 97].

1.2.1 Predicting SREs with the HEXplorer

The typically 4 to 8 nucleotides long SRP binding motifs are degenerate, making prediction of SREs solely based on the nucleotide sequence a challenge. However, due to the position-dependent effect of splicing regulatory proteins, SREs show signature position distributions around splice site sequences, which can be the basis for SRE prediction algorithms [2, 16, 98, 99].

Based on this principle, the HEXplorer applies a RESCUE-type approach on hexamer frequencies around splice sites to estimate their binding potential for splicing regulatory proteins [99]. For this purpose, hexamer frequencies were calculated for up to 100 nucleotides long sequence blocks in the exon upstream or intron downstream of splice donor sites, annotated in the early reference genome (ENSEMBL human chromosomes 6, 7, 9, 10, 13, 14, 20, 22, X). The difference in the frequency of a hexamer either within the exonic or intronic sequence segments was then normalized to a Z-score, the Z_{EI} (exon-intron Z-score), for all 4096 hexamers. Upon Z-score normalization, the mean hexamer frequency ratio between exonic and intronic regions is set to 0 and values differing from the average are indicated as a multiple of the standard derivation of the distribution of all hexamer frequency ratios. The higher the Z-score, the higher the probability to find a certain hexamer sequence upstream of splice donor sites. Enriched presence of a hexamer upstream of 5'ss was interpreted as indication for motifs, recruiting proteins, that enhance usage of downstream splice donor sites, like SR-proteins (and SR-like proteins). Negative Z-scores describe hexamers found predominantly downstream of splice donor sites, thus potentially recruiting proteins that enhance SD usage from a downstream relative position, like hnRNP and hnRNP-like proteins. For a given nucleotide sequence, the HEXplorer calculates a HEXplorer score (HZ_{EI}) per nucleotide by taking the average Z_{EI} score of all six hexamer sequences overlapping an index nucleotide (Figure 7).



C	T	G	T	G	T	G	A	A	T	A	T	C	A	A	G	C	A	G	G	Z_{EI}
				G	T	G	A	A	T											−1.93
					T	G	A	A	T	A										−7.44
						G	A	A	T	A	T									−0.69
							A	A	T	A	T	C								−3.82
								A	T	A	T	C	A							−0.34
									T	A	T	C	A	A						2.62
																			HZ_{EI}	−1.93

Figure 7: HZ_{EI} calculation for an exemplary sequence. The HEXplorer score takes the average Z_{EI} score of all hexamer sequences, overlapping the index nucleotide, coloured in red [2]. Permission to include the figure was kindly granted and copyright reserved by Oxford University Press.

The average of Z_{EI} scores was chosen, to smoothen the HZ_{EI} score variations across the nucleotide sequence and to show how single nucleotide variations can change the SRP binding profile within a given sequence context. The HZ_{EI} profile of nucleotide sequences has been repeatedly shown to be predictive for SRP binding in different experimental settings like mass spectrometric analysis of proteins, isolated by RNA affinity chromatography [23, 100].

1.2.2 The splice site HEXplorer weight (SSHW)

To estimate the influence of splicing regulatory proteins on usage of a given splice site, the HEXplorer scores of the surrounding nucleotides can be used to calculate the so-called splice site HEXplorer weight (SSHW). It summarizes the predicted impact on splice site usage via hnRNP (and hnRNP-like proteins) and SR proteins (and SR protein-like proteins) within an arbitrary sequence window around a splice site. From experience with experimental data, we and others saw that a window of ±50 nucleotides around a given splice site seem to be of special importance concerning its property to recruit splicing regulatory proteins. The SSHW of a splice donor (SD) site is therefore calculated by subtracting the HEXplorer score sum of the 50 nucleotides downstream of the SD from the HEXplorer score sum of the 50 nucleotides upstream to it. For SD sites, this results in a higher score, the higher the amount of predicted upstream SR protein binding and downstream hnRNP binding. To apply the SSHW on splice acceptor (SA) sites, the HEXplorer score sum of the 50 nucleotides downstream of a SA are subtracted from the HEXplorer score sum of the 50 nucleotides upstream to it. The equation is reversed, to reflect the position-dependent effects of SRP binding on splice site usage.

1.2.3 Alternative approaches for SRE prediction

Over the past 20 years, multiple open access tools have been developed to evaluate the interplay between a given splice site and the splicing regulatory proteins bound in its proximity. They can be roughly categorized into tools that are based on 1) the analysis of motif/k-mer frequencies, 2) experimental data, 3) motifs collected from the literature, 4) a combination of different tools or 5) neuronal networks (reviewed in [17]).

1) The HEXplorer, RBPmap [101], Splicing Factor Finder [102] and RESCUE-ESE [99] are based on computational analyzes of nucleotide motifs or k-mer distributions. RESCUE-ESE and the HEXplorer are based on the position-dependent effect of SRPs and the corresponding differences of hexamer frequencies upstream versus downstream of splice donor sites and around intrinsically weak or strong splice sites. RESCUE-ESE scans a genomic sequence for 238 hexamers, that were found to be significantly more frequent upstream of weak splice donor sites than upstream of strong donor sites, indicating potential binding sites of SR proteins. RBPmap and the Splicing Factor Finder predict binding sites of RNA-binding proteins within a

given genomic sequence based on a selection of already described binding motifs and their evolutionary conservation.

2) ESEfinder [103], ESRseq [104] and FAS-ESS [105] are based on observations from experimental data. Like the HEXplorer, ESRseq scores hexamers for their predicted position-dependent enhancing or silencing effect on splice site usage. The ESRseq score is, however, not based on genomic reference sequences around splice donor sites like the HEXplorer, but on the observed effect of a hexamer on splice site usage within a three-exon minigene with a 6-mer library within the central exon. Based on RNA-sequencing data of cells, transfected with the reporter constructs, 2,272 of all 4,096 hexamers were predicted to influence splice site usage ranging from ESRseq score -1.06 (CTTTTA) to +1.03 (AGAAGA), whereas 1,824 hexamers were predicted to have no influence on splicing, i.e. no ESRseq score was assigned. The FAS-ESS and ESEfinder tools, on the other hand, are based in Systematic Evolution of Ligands by EXponential enrichment (SELEX). Whereas FAS-ESS scans a genomic sequence for RNA binding motifs of proteins, that repress downstream splice donor usage and upstream splice acceptor usage, the ESEfinder identifies sequence elements that were experimentally validated to recruit recombinant SR proteins SRSF1, SRSF2, SRSF5 and SRSF6.

3) ATtRACT [106] and SpliceAid [107] are based on previously described motifs curated from literature. SpliceAid and its tissue-specific adaptation SpliceAid2 use a curated set of experimentally validated binding motifs of splicing regulatory proteins that were previously described. Both tools still use historic protein names in their final report, which sometimes requires an additional literature search when searching for binding motifs of a particular RBP. Similar to SpliceAid, ATtRACT also scans sequences for previously described RBP binding motifs, however not distinguishes SR or hnRNP binding visually as SpliceAid does.

4) The Human Splicing Finder [81], EX-SKIP [80] and SROOGLE [82] combine multiple tools for SRE prediction. The Human Splicing Finder, which was recently removed from the public domain, uses RESCUE-ESE and ESEfinder and an additional module to detect SRSF3 and SRSF10 binding motifs. It combines reports from these tools to an overall report about the presence and predictive impact of various sequence elements, important for splicing. Similarly, SROOGLE uses tools like ESE-finder, RESCUE-ESE and FAS-ESS to predict potential binding sites

of splicing regulatory proteins. EX-SKIP, on the other hand, combines RESCUE-ESE, PESE/PESS or FAS-ESS and calculates the ratio of motifs that enhance exon inclusion (ESE) to motifs, that induce exon skipping (ESS). Given a wild type and a mutant version of the same sequence, the tool tries to predict, which sequence has a higher probability to lead to exon skipping.

5) SpliceAI [108], as one of the most recent algorithms, attempts a top-down approach, which does not detect single RBP binding motifs, but utilizes a series of deep residual neuronal networks, that were trained to classify each position of a genomic sequence for its potential to function as splice donor or splice acceptor. The quality of the neural networks is measured by its top-k accuracy, which is the ratio of correctly predicted splice sites at the threshold where the number of predicted sites equals the number of true sites in the dataset. Interestingly, networks working on longer sequence segments of up to 10,000 nt showed significantly higher accuracy than those working on 80 nt: top-k accuracy increased from 0.57 (80 nt) to 0.95 (10,000 nt). This effect might be due to the regulatory effects of neighboring splice sites on splice site selection, opposed to effects of sequence elements in close proximity on the usage of a particular splice site. SpliceAI significantly outperformed GeneSplicer [109], MaxEntScan [20] and NNSplice [110].

1.3 Aberrant splicing in disease

Apart from stress-induced regulation of proteins governing splicing, splicing can also be affected by sequence mutations in critical regions, like 5'/3' splice sites or splicing regulatory elements or even by mutations within snRNAs or the coding sequence of splicing regulatory proteins or proteins of the spliceosome. Most disease-causing mutations, however, are located within splice site sequences or splicing regulatory elements [111, 112]. Depending on where in the genome they occur, mutations can result in recognition of cryptic splice sites, by decreasing the intrinsic strength of the original splice site, affecting SREs or creating a new splice site of sufficient strength at the right position. Alternatively, mutation can lead to exon skipping, intron retention or a general switch in transcript isoform abundance.

One of the first diseases ever described to result from mis-splicing was β-thalassemia, with 1.5% of global population being a carrier of this autosomal recessive hematologic disease

[113]. It is characterized by an altered or missing β-globin chain protein of hemoglobin, which is the primary carrier of oxygen in humans. One common mutation associated with β-thalassemia represents a G to A exchange at the first nucleotide position of the second intron, which disrupts correct splicing by destroying the canonical GT dinucleotide of the 5'ss. This results in expression of two splice-isoforms, that not encode functional β-globin protein [114].

A disease more recently associated with aberrant splicing is Duchenne muscular dystrophy, where non-functional dystrophin protein (DMD), which is essential for muscle contraction and stability, results in a cycle of myofiber (muscle cell fiber) necrosis and regrowth, that eventually leads to muscle waste and death [115, 116]. Again for instance G to A substitutions at the first intron nucleotide position affected the splice site usage and cryptic splicing at multiple introns, sometimes resulting in out-of-frame transcripts, of varying pathological significance [116].

Alternative splicing seems to be more prevalent in the brain compared to other tissues, potentially reflecting its complexity [117, 118]. Many neurodegenerative diseases like Alzheimer disease (AD) [119], frontotemporal dementia (FTD) [120], Parkinson disease (PD) [121] and amyotrophic lateral sclerosis (ALS) [122] are associated with pathological alternative splicing, that disrupts functionality and abundance of associated key proteins (reviewed in [123]). Some Parkinson patients for instance show SNCA transcripts that lack exon 5, removing an important phosphorylation site of the protein, which might drastically enhance SNCA protein aggregation and toxicity [121, 124]. One key player during Alzheimer disease development, PSEN2*,* has been described to be affected by various splice-altering mutations like a deletion of the A of the canonical AG dinucleotide at exon 12, which results in exon 12 skipping and a premature termination codon [119].

Dysregulation of splicing additionally plays an important role during development and progression of various cancers [125]. Mutations affecting expression of oncogenes and tumor suppressors, as well as somatic mutations in components of the spliceosome are abundant for instance in solid brain [126], breast [127], colon [128] or skin cancers [129]. Besides sequence mutations within genes important for cancer, pathological alternative splicing was repeatedly shown to be caused by cancer-induced changes in the expression of splicing regulatory genes.

Overexpression of SRSF5 or SRSF1 has been associates with small-cell lung or glioma oncogenesis respectively [130, 131]. Also increased hnRNP K expression was observed in multiple cancers [132]. Aberrant expression of proteins that regulate splicing changes RNA processing like RNA splicing, but may additionally influence other steps of protein expression, such as RNA export. Shortening the 3′ UTR via alternative splicing seems to be one way during oncogenesis to increase expression of specific proteins [133, 134].

2. Research thesis

2.1 Thesis 1 - Splice site strength should be evaluated together with binding of splicing regulatory proteins

A single RNA transcript can be processed to various different splice-isoforms during RNA maturation increasing potential proteomic diversity of eukaryotic organisms. Regulation of the molecular machinery to accomplish splicing, however, involves dozens of proteins that influence, which splice sites are selected and whether the splicing reaction takes place. One important factor determining splice site usage, is the so-called intrinsic strength of the splice site itself. Among others, it can be measured by the HBond score for the 5′ss or the MaxEntScan score for the 3′ss and provides an estimate of how efficiently the respective splice site is recognized by the spliceosome. However, binding of splicing regulatory proteins in proximity to the splice site add an extra layer of regulation, since they can either repress or enhance splice site usage, especially for splice sites of median intrinsic strength (**Publication I***)*. Understanding the interplay of these two factors could greatly benefit predictive models for splice site selection, which could be important in diagnostics (**Publication II**).

2.1.1 Publication I: Context matters: Regulation of splice donor usage.

Recognition of 5′ splice sites requires RNA duplex formation between the 5′ss sequence and the free 5′ end of the U1 snRNA. The higher the complementarity and the density of HBonds between these sequences, the higher 5′ss recognition. However, other factors additionally influence 5′ss, such as the length of the exon, during exon/intron definition [135], or binding of splicing regulatory proteins (SRPs) in proximity to the 5’ss. Splicing regulatory elements (SREs) are sequences, that recruit SRPs to the RNA, leading to either enhanced or silenced splice site recognition. They have historically been termed exonic splice enhancers (ESE), exonic splice silencers (ESS), intronic splice enhancers (ISE) or intronic splice silencers (ISS), since they were described in most cases in respect to a specific splice site of a gene. Since, however, the activity of splicing regulatory elements is position-dependent, exonic splice enhancers for instance would silence usage of 5’ss upstream to it, making the terminology of these elements redundant in their overall ability to correctly describe SRE influence on splice sites around it. An alternative could potentially be exon-promoting element (EPE) and intron-promoting elements (IPE), which would describe sequences recruiting SR or SR-like proteins and hnRNP or hnRNP-like proteins respectively. Correctly identifying these sequence elements is important, since they are frequently disrupted in various disease. In this work, bioinformatical tools were reviewed, that try to predict human SREs.

The following review was published in Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2019 (doi. 10.1016/j.bbagrm.2019.06.002), by

Johannes Ptok, Lisa Müller, Stephan Theiss and Heiner Schaal

Contribution

JP, LM and HS did the figures. HS wrote parts about the history of splicing regulatory elements, LM wrote parts about the medical relevance of splicing regulatory elements and JP and ST wrote parts about bioinformatic tools to predict splicing regulatory elements. JP did the tool testing and case analysis of the bioinformatics part. Individual contribution of JP at around 35%.

BBA - Gene Regulatory Mechanisms 1862 (2019) 194391

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Context matters: Regulation of splice donor usage☆

Johannes Ptok[a,1], Lisa Müller[a,1], Stephan Theiss[b], Heiner Schaal[a,*]

[a] Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany
[b] Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany

ABSTRACT

Elaborate research on splicing, starting in the late seventies, evolved from the discovery that 5′ splice sites are recognized by their complementarity to U1 snRNA towards the realization that RNA duplex formation cannot be the sole basis for 5'ss selection. Rather, their recognition is highly influenced by a number of context factors including transcript architecture as well as splicing regulatory elements (SREs) in the splice site neighborhood. In particular, proximal binding of splicing regulatory proteins highly influences splicing outcome. The importance of SRE integrity especially becomes evident in the light of human pathogenic mutations where single nucleotide changes in SREs can severely affect the resulting transcripts. Bioinformatics tools nowadays greatly assist in the computational evaluation of 5'ss, their neighborhood and the impact of pathogenic mutations. Although predictions are already quite robust, computational evaluation of the splicing regulatory landscape still faces challenges to increase future reliability. This article is part of a Special Issue entitled: RNA structure and splicing regulation edited by Francisco Baralle, Ravindra Singh and Stefan Stamm.

1. Separating splice site from context

In order to address the impact of sequence context on splicing, we first need to specify what we understand by the “proper splice site” and its “context”. Historically, splice junction consensus sequences were first constrained to the highly conserved GT-AG dinucleotides at both intron ends [1,2], and U1 snRNA was assumed to bind to both intron ends across the excision-ligation point. However, in 1983 Mount and Steitz showed that the 11 nt long free 5′ end of U1 snRNA binds to the 5'ss but not to the 3'ss [3]. A tentative 5′ splice site motif derived from a collection of 139 5'ss showed clear nucleotide preferences in the nine “consensus” positions −3 to +6, but approximately equal probability for all four bases at the terminal intronic positions +7 and +8 of the possible RNA duplex region, leading to a 9 nt long 5'ss consensus motif [4]. A variety of mutational analyses confirmed that overall splicing efficiency is affected by many nucleotide exchanges within this 5′ splice site region [5–8], and that RNA duplex formation between the 5′ splice site and the terminal nucleotides of U1 snRNA is key to 5′ splice site recognition [9].

It was the first decade that paved the way for the concept that 5′ splice sites are recognized by their complementarity to U1 snRNA. The prevailing perception is that an RNA duplex is formed by a linear sequence of hydrogen bonds between U1 snRNA and the 5′ splice site nucleotides in the standard base-pairing register. Alternatively, Roca et al. described different ways of RNA duplex formation with either shifted [10] or bulged base-pairing registers [11], which are statistically hidden in consensus sequence motifs at the exon/intron border. By using massively parallel splicing assays, it was recently shown that 5'ss seem to be predominantly recognized *via* the normal register and that shifted registers might not always be productively recognized for splicing [12].

Since the large majority of 5'ss neither reflect alternative splicing registers nor U1 snRNA complementarity in positions +7 and +8, the 9 nt long 5'ss motif is the base for most scoring algorithms measuring the intrinsic strength of a 5′ splice site (for a review see [13]). Also, when applying tools measuring the information content of 5′ splice sites, information theory-based position weight matrices failed to identify additional nucleotides contributing to 5′ splice site strength [14].

There are several approaches to measuring splice site strength that go beyond independent position weight matrix scores. Derived by applying the statistical maximum entropy concept to separate sets of real and “decoy” splice sites, the 5′ and 3′ splice site MaxEnt scores take non-adjacent nucleotide dependencies into account when scoring 9 nt long sequences for 5'ss and 23 nt for 3'ss. Today, MaxEntScan is the most widely used splice site scoring tool, and it has been included in most online computational tools [15–17].

Using a complementary concept, the SD algorithm implements a dictionary approach measuring 5'ss strength by the logarithm of its sequence frequency in annotated 5'ss of the human genome. It has

☆ This article is part of a Special Issue entitled: RNA structure and splicing regulation edited by Francisco Baralle, Ravindra Singh and Stefan Stamm.

* Corresponding author.

E-mail address: schaal@uni-duesseldorf.de (H. Schaal).

[1] Joint first authors.

https://doi.org/10.1016/j.bbagrm.2019.06.002
Received 26 April 2019; Received in revised form 7 June 2019; Accepted 9 June 2019
Available online 13 June 2019

1874-9399/ © 2019 Elsevier B.V. All rights reserved.

reached 97% sensitivity at 95% specificity on a set of 179 previously reported splicing mutations, analyzed together with 32 minigenes [15–17].

The HBond score (HBS) measures the overall hydrogen bond pattern binding strength in the 11-nt long duplex between the 5'ss and the 5′ end of U1 snRNA. Assigning numerical weights to hydrogen bonds and "mismatches" in individual 5'ss positions, it takes into account interdependencies of up to seven neighboring nucleotides as well as G:U wobble base pairs. As a weighted number of hydrogen bonds, the 5'ss HBond score is dimensionless, and its values range from 1.8 (isolated GT) to 23.8 (full U1 snRNA complementarity CAG GTAAGTAT).

RNA duplex formation of a considerable number of 5′ splice sites [18] can in fact benefit from complementary nucleotides in the 7th and 8th positions of the intron [19–21]. Indeed, 25,289 (8.5%) out of 296,036 annotated human 5′ splices sites (GRCh38.91) contain a U1 snRNA complementary AT dinucleotide in positions +7 and +8, including 5′ splice sites of genes with significant diagnostic relevance, *e.g.* β-globin exon 1, SMN1/2 exon 4, BRCA1 exon 18 and 22, BRCA2 exon 6, 7 and 22. The significance of complementary nucleotides in positions +7 and +8 is supported by *in silico* substituting the AT dinucleotide in these 25,289 5'ss by non-complementary bases. Those 5'ss that exhibited large changes ΔHBS in their HBond scores, and thus were more vulnerable to mutations of positions +7 and +8, were stronger on average (higher MaxEnt and HBond scores) and had intron centered regions of U1 snRNA complementarity. In contrast, less vulnerable 5'ss were weaker on average (lower MaxEnt and HBond scores), and their 11 nucleotides wide motifs showed exon centered U1 snRNA complementarity, indicating a possible compensatory mechanism within the 9 nt long 5'ss consensus motif for mutations in positions +7 and +8 (Fig. 1).

RNA duplex formation alone, however, cannot explain 5′ splice site selection, as numerous sequences within an exon are not used as 5′ splice sites, even though their complementarity to U1 snRNA is even higher than the actually used nearby 5′ splice site.

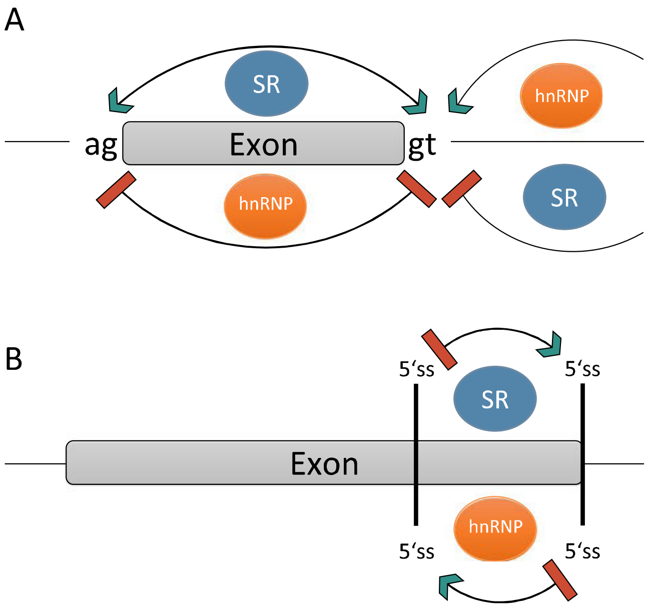


Fig. 2. Schematic drawing of the position dependent regulation of splice sites by SR and hnRNP proteins.
A While SR proteins (blue) enhance the usage of a 5'ss from an upstream position, they repress its recognition from a downstream position. Binding hnRNP proteins (orange) upstream of a 5'ss, on the other hand, repress its usage while they enhance from a downstream position.
B In a situation with two competing 5'ss, the same splicing regulatory element exhibits enhancing features on one splice site while it represses the other. Here, the widely used terms "exonic splicing enhancer (ESE)" and "exonic splicing silencer (ESS)" can be misleading.

2. Splicing regulatory elements: separating true from spurious splice sites

Sequences matching the 5′ and 3′ splice site consensus motifs can be

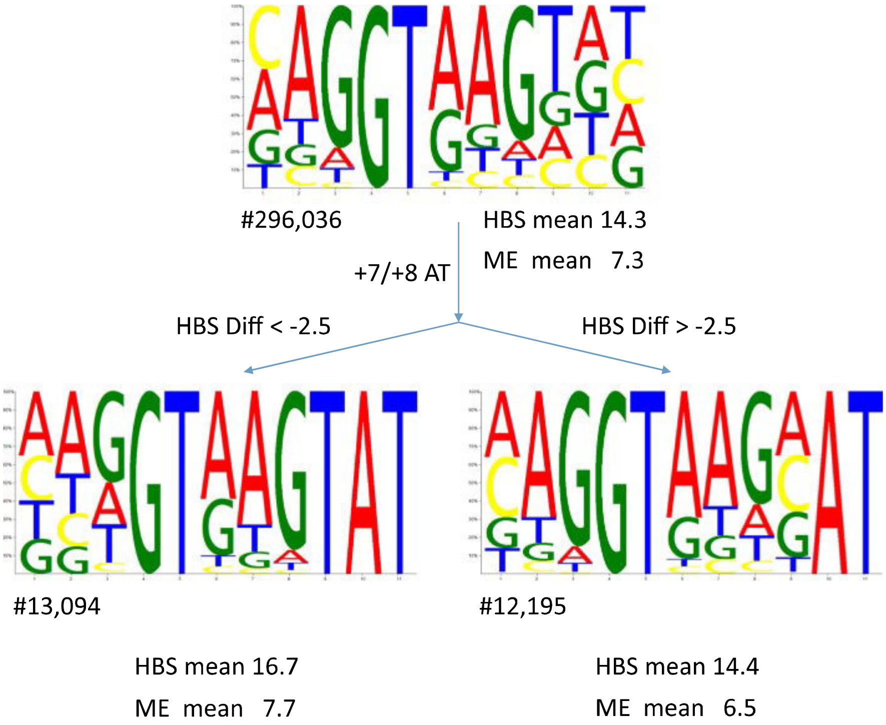


Fig. 1. Different vulnerability to loss of U1 snRNA complementarity in positions +7/+8. 25,289 out of 296,036 canonical human annotated GT 5'ss, with an AT dinucleotide in positions +7 and +8 were split into two groups according to their change in HBS when substituting AT by non-complementary CC. The sequence logos of the wt 5'ss with larger changes (ΔHBS < −2.5, N = 13,094) are depicted on the left side, those with smaller changes (ΔHBS > −2.5, N = 12,195) on the right side. The average MaxEnt scores (ME) and HBond scores (HBS) are shown beneath each sequence logo.

found almost anywhere in the human genome. If all were used as splice sites, transcriptome complexity would increase exponentially. However, a good match to a splice site consensus motif does not imply splice site use, and its recognition is influenced by various context factors: (i) location in first or last *vs.* internal exon [22–25], (ii) exon/intron architecture [26], (iii) regulatory sequences (reviewed in [27]) in close splice site proximity [28], or (iv) generally wider sequence contexts [12].

Frequently, splice site sequence context provides splicing regulatory *cis*-acting elements (SREs) as binding sites for splicing regulatory proteins (SRPs) supporting or inhibiting efficient spliceosomal assembly or (de-)stabilizing the RNA duplex, while interacting with each other and competing for binding sites [29,30] (Fig. 2A). Based on their activity and location in pre-mRNA, these *cis*-acting elements were historically classified as exonic splicing enhancers (ESEs) or silencers (ESSs), and intronic splicing enhancers (ISEs) or silencers (ISSs). However, this classification becomes ambiguous in the presence of several competing potential 5'ss, since the same sequence can be either exonic or intronic dependent on the actually used splice site (Fig. 2B).

2.1. RNA binding proteins and their role in 5'ss selection

RNA binding proteins (RBPs) play important roles in numerous post-transcriptional processes (reviewed in [31]), and they include splicing regulatory proteins (SRPs) that interact with *cis*-acting binding sites in the splicing reaction. Together with the proper splice sites, they contribute highly to the ‘splicing code’, but their mode of action is still not fully unraveled due to various layers of interaction (reviewed in [32]). It is generally understood that two distinct protein families, serine and arginine rich proteins (SR proteins) and heterogeneous nuclear ribonucleoproteins (hnRNP) are main interaction partners of SRE sequences [27,33]. The first group, SR proteins, represent a family of multifunctional RNA binding proteins that are involved in the regulation of constitutive and alternative pre-mRNA splicing [34]. Two distinct structural features characterize this family. First, they share a C-terminal arginine/serine (RS) domain, which is mainly but not exclusively responsible for protein-protein-interactions with the spliceosome, since it is also present in U1-70K, a protein associated with the U1 snRNP [35]. Furthermore, all family members display at least one RNA recognition motif (RRM) at the N-terminus that further provides RNA-binding specificity. The serine residues of the RS domain are targeted for phosphorylation, which in turn highly influences SR protein activity [36,37]. So far, twelve human SR proteins have been identified, and by now, they are uniformly termed serine arginine splicing factor (SRSF) 1–12, while their former names were often derived from their individual molecular weights [38]. Apart from spliceosome interaction, the RS domain is also capable of contacting the pre-mRNA directly either *via* the branch point sequence (BPS) or the 5'ss, which might constitute an alternative way to facilitate spliceosome assembly [39].

As counterparts of SR proteins, hnRNPs are the other key family of proteins involved in splicing regulation and nucleic acid metabolism with additional effects on translation and cellular transport (reviewed in [40,41]). Around 20 major types of hnRNP proteins, which are uniformly termed hnRNP A-U with molecular weights ranging from 34 kDa to 120 kDa, share structural features such as the RNA recognition motif (RRM) and auxiliary domains high in proline, glycine, tyrosine, arginine, glutamine or asparagine [40,42]. Due to their structural diversity, hnRNPs are involved in various cellular processes, though concerning their role in 5'ss selection, they were shown to have adverse behavior compared to SR proteins [28,43].

2.2. Finding a needle in a haystack – defining SRE binding motifs to map the regulatory landscape of 5'ss

Identification of precisely defined binding sites is one of the boxes that needs to be ticked on the way to a full understanding of the splicing code. During the past decades, elaborate approaches have been applied to elucidate the exact interaction sites between RNA and SRPs. In early stages, functional SREs were identified upon their disruption by pathogenic mutations in diseases. Over time, numerous experimental approaches, such as *in vivo* splicing and splicing reporter assays [44–46], and computational methods (*e.g.* [47–50]) were added. More recently, cross-linking and immunoprecipitation (CLIP) and various sub-methods as iClip [51], HITS-CLIP [52] and PAR-CLIP [53] have gained importance and are combined with other large-scale approaches as the identification of alternative splicing events by microarrays and RNA-seq [54]. Both *cis*-acting SRE motifs and *trans*-acting SRP binding domain motifs exhibit high sequence variability, which renders splice site usage prediction difficult. SREs are often degenerate in their sequences and are capable of binding multiple regulatory factors. SRPs, on the other hand, are as well capable of recognizing a wide variety of binding sites, which was shown in several studies [55,56]. While this feature contributes to the tightly regulated splicing process, it is a hurdle in the unraveling of the splicing code.

3. Misguided splicing regulation as the root of disease

Despite SRE motif degeneracy and complexity of splice site context, single nucleotide changes even outside the proper splice site sequence can have dramatic consequences for individuals due to their impact on splicing. For example, when located in SREs, mutations can disrupt binding of SRPs that are crucial for physiological exon/intron border recognition. Hence, mutations can have various effects on the splicing process such as generation of *de novo* splice sites, activation of cryptic splice sites, or decreased use of physiological splice sites [57,58]. In the following, we will give selected examples for these mechanisms:

Duchenne muscular dystrophy (DMD) is an X-linked disorder developing in early childhood and characteristically leaving the early teenage patients wheelchair bound due to muscle weakness caused by lack of dystrophin protein and resulting muscle degeneration. Patients die in their mid-twenties, often due to cardiac or respiratory muscle weakness [59]. One of the disease causing variants identified in DMD is a c.1684C > T mutation in exon 14 of the dystrophin gene that creates a *de novo* GT splice donor site with an HBS of 12.0, which results in a 22 base pair deletion in exon 14. Furthermore, it changes a CAA codon to a TAA stop codon leaving the resulting dystrophin protein dysfunctional [60].

Hutchinson-Gilford Progeria Syndrome (HGPS) is a severe laminopathy and a prominent example for spliceopathies. The main disease associated mutation that causes aberrant splicing of the LMNA gene is the translationally silent c.1824C > T that generates a *de novo* splice site located in exon 11 [61,62]. The mutation increases U1 snRNP complementarity (HBS WT = 12.90, HBS mt = 15.80) which renders this *de novo* splice site active. It generates an alternative lamin A transcript with a deletion of 150 nucleotides that results in a truncated protein (lamin AΔ150) [63].

In a child suffering from neonatal hypotonia, seizures, ataxia and a developmental delay [64], symptoms were related to the activation of a cryptic splice site in the E1α pyruvate dehydrogenase gene (PDH), which plays a key role in energy metabolism [65]. In this particular case, a G-to-A substitution in E1α PDH intron 7-8 was located 26 nucleotides downstream of the physiological splice site. ESEfinder analysis revealed that the mutation generates an SRSF2 protein binding site, that contributes to the use of the cryptic splice site at position 45 in the intron which is naturally inactive [64]. Additionally, the newly created SRE acts silencing on the upstream physiological splice donor which contributes to the aberrant splicing outcome [66].

Especially in a diagnostic setting, silent mutations that do not change the underlying coding potential of a sequence need to be taken into account. Such silent mutations are often ignored in routine diagnostics, since they do not change the resulting amino acids, but they still harbor the potential to severely interfere with splicing regulation.

One well studied example is the silent C > T transition (c.840C > T) in exon 7 of survival motor neuron gene 2 (SMN2) [67]. In patients suffering from spinal muscular atrophy (SMA), loss of survival motor neuron gene 1 (SMN1) could in principle be compensated by the nearly identical SMN2 gene. A critical difference between SMN1 and SMN2, however, lies in the single C > T change in SMN2 disrupting a splicing regulatory element and leading to skipping of SMN2 exon 7, and hence a non-functional SMN2 protein [68,69].

Publicly accessible databases that compile human pathogenic mutations, *e.g.* [70–73], are dominated by protein altering mutations and struggle far more with the inclusion of silent SNVs outside splice site motifs that still have disease-causing potential by mis-regulating splicing. This is due to the overall difficulty of current algorithms to reliably predict the influence of a SNV both on the binding affinity of a specific SRP and on the impact of the altered SRP binding on splice site recognition.

4. Computational evaluation of 5'ss

4.1. Open access tools

During the last two decades, various publicly available tools have been developed which can be used to analyze the interplay of splice site context and intrinsic splice site strength for a given sequence (Table 1). They can broadly be categorized into tools based on (1) a computational analysis of nucleotide motifs or k-mer distributions, (2) individual experimental data, (3) previously described motifs, (4) a combination of multiple tools or (5) neuronal networks (partially reviewed in [74]).

4.1.1. Tools based on computational analysis of nucleotide motifs or k-mer distributions

The HEXplorer [48,66], RBPmap [75], Splicing Factor Finder [76] and RESCUE-ESE [47,77] are based on computational analyses of nucleotide motifs or k-mer distributions. HEXplorer and RESCUE-ESE use the position-dependent effects of splicing regulatory proteins on splice site usage [28] and the corresponding difference in the abundance of hexamers upstream and downstream of splice donors and around intrinsically weak or strong splice sites. Based on a set of hexamer frequency weights from −73.27 (TTTTTT) to +34.35 (GAAGAA), the HEXplorer algorithm [48,66] provides a score profile of a genomic sequence, reflecting its ability to either enhance or silence nearby splice site usage. RESCUE-ESE scans a genomic sequence for the presence of 238 hexamers which were found more frequently within exons with weak 5′ splice sites than in exons with strong 5'ss or introns. These hexamers constitute potential binding sites for SR proteins enhancing downstream splice donor and upstream splice acceptor usage (“ESE”) [47,77]. RBPmap [75] as well as the Splicing Factor Finder [76] predict binding sites of RNA-binding proteins within a given genomic sequence using a collection of well-described binding motifs and their evolutionary conservation.

4.1.2. Tools based on individual experimental data

ESEfinder [49], ESRseq [50] and FAS-ESS [78] are built on grounds of individual experimental data. ESRseq is based on RNA-sequencing data of a three-exon minigene library, resulting in the scoring of 2272 of all 4096 hexamers for their potential position-dependent enhancing/repressing effects on splice site usage. ESRseq hexamer scores range from −1.06 (CTTTTA) to +1.03 (AGAAGA), while no scores are

Table 1
Tools to predict splicing regulatory elements.

Tool	Published	URL	Features
(1) Computational analysis of nucleotide motifs or k-mer distributions			
HEXplorer [48,66]	2014	https://www2.hhu.de/rna/html/hexplorer_score.php	Scores all hexamers for potential position-dependent enhancing/repressing effects on splice site usage
RBPmap [75]	2014	http://rbpmap.technion.ac.il	Calculates potential binding, using well described motifs of the literature and their conservation
Splicing Factor Finder [76]	2009	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2691001	Mapping of splicing factor sites with the help of previously described binding motifs and their evolutionary conservation
RESCUE-ESE [47,77]	2002	http://genes.mit.edu/burgelab/rescue-ese/	Scans for 238 hexamers which are more frequent within exonic sequences of weak splice sites from the reference genome
(2) Based on individual experimental data			
ESRseq [50]	2011	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149502/	Scores all hexamers for potential position-dependent enhancing/repressing effects on splice site usage, based on RNA-sequencing of a minigene library
FAS-ESS [78]	2004	http://genes.mit.edu/fas-ess/	Functional screening for 6-nt SELEX motifs of ESSs
ESEfinder [49]	2003	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home	Scans genomic sequence for functional SELEX motifs of SRSF1, SRSF2, SRSF5 and SRSF6
(3) Based on previously described motifs			
ATtRACT (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823821/)	2016	http://attract.cnic.es	Scanning genomic sequence for hand-curated set of experimentally validated RBP binding motifs
SpliceAid [79]	2009	http://www.introni.it/splicing.html	Scans genomic sequences for validated binding motifs of human splice regulatory proteins
(4) Combining multiple tools			
EX-SKIP [80]	2011	http://ex-skip.img.cas.cz/	Tool predicting exon skipping based on result of multiple SRE prediction tools
Human Splicing Finder [81]	2009	http://www.umd.be/HSF3/	Tool predicting SREs, splice sites or branch sites
SROOGLE [82]	2009	http://sroogle.tau.ac.il	Tool predicting splice sites, SREs, branch sites and Polyadenylation sites
(5) Artificial neural network tool			
SpliceAI [83]	2019	https://github.com/Illumina/SpliceAI	Network predicting splice sites and taking SREs implicitly into account

assigned to 1824 hexamers predicted to have no influence on splicing [50]. The ESEfinder and FAS-ESS tools, on the other hand, are based on Systematic Evolution of Ligands by EXponential enrichment (SELEX) [84]. While ESEfinder scans genomic sequences for the presence of experimentally derived RNA binding motifs of recombinant SR proteins SRSF1, SRSF2, SRSF5 and SRSF6 [49], FAS-ESS correspondingly identifies RNA binding motifs of proteins repressing downstream splice donor and upstream splice acceptor usage ("ESS") [78].

4.1.3. Tools based on previously described motifs

ATtRACT [85] and SpliceAid [79] can be categorized as tools based on previously described motifs. SpliceAid, as well as the tissue-specific adaptation SpliceAid2 [86] are also based on experimentally validated binding motifs of splicing regulatory proteins which were selected from the literature. Similar to SpliceAid, ATtRACT is based on extensive literature search for general RBP binding motifs, thus not directly visually discriminating between SR and hnRNP proteins.

4.1.4. Tools combining multiple tools

The Human Splicing Finder [81], EX-SKIP [80] and SROOGLE [82] combine multiple tools for SRE prediction. Human Splicing Finder combines RESCUE-ESE and ESEfinder as well as an individual module for the detection of SRSF3 and SFRS10 binding motifs [81]. The EX-SKIP tool compares likelihood of exon skipping between exonic wild type and mutant sequences, based on the integration of RESCUE-ESE, PESE/PESS [87] or FAS-ESS [80]. Like Human Splicing Finder, SROOGLE combines tools like ESE-finder, RESCUE-ESE and FAS-ESS to predict potential binding sites of splicing regulatory proteins.

4.1.5. Artificial neural network tool

All above mentioned algorithms for computational identification of SREs follow a bottom-up approach, *i.e.* they collect specific SRE motifs ("atoms"), estimate their individual strength and assemble this information to predict their impact on splice site recognition.

A complementary, top-down approach that is completely agnostic to previously identified motifs has recently been presented by Jaganathan et al., who designed and trained a series of deep residual neuronal networks, SpliceAI, that directly classify each position in a pre-mRNA sequence as either splice donor, splice acceptor, or neither [83]. This type of analysis circumvents the *a-priori* identification of SREs, and rather lets the artificial neural network implicitly learn the splice site recognition rules.

For such a neural network, classification quality is measured by its top-*k* accuracy, which is the fraction of correctly predicted splice sites at the threshold where the number of predicted sites equals the number of true sites in the dataset. Interestingly, networks working on longer sequence segments of up to 10,000 nt exhibited much better splice site prediction than those working on 80 nt: top-*k* accuracy increased from 0.57 (80 nt) to 0.95 (10,000 nt). SpliceAI significantly outperforms GeneSplicer, MaxEntScan and NNSplice, and its source code is publicly hosted at https://github.com/Illumina/SpliceAI.

4.2. Exemplary evaluation of computational tools for assessment of context dependency

In this section, we will briefly demonstrate, on a feature-rich minigene model system, the context-related information that can be gathered from selected online tools of the above list. In a series of minigene reporter experiments, Lu et al. have systematically examined 5'ss context impact on splicing by inserting a set of eight splice sites with varying strengths into two different environments [88]. In detail, an exon of either gene TRIM62 or HMSD, framed by 500 upstream and downstream intronic nucleotides, was inserted between two exons of a minigene splicing reporter, and all combinations of 5'ss and context were assessed (Fig. 3). In the TRIM62 context, decreasing 5'ss strength from MaxEnt score 11 to 4.44 did not significantly disrupt splicing, while in the HMSD context, exon skipping occurred—and gradually increased—below MaxEnt score of 9.6. Thus, the TRIM62 5'ss context clearly supported splicing more effectively than its HMSD counterpart.

For our evaluation of online tools, we analyzed 50 nt up- and downstream of the 5'ss for potential impact on 5'ss usage (Table 2). To measure the "splicing regulatory effect" predicted for this 5'ss neighborhood, we calculated a "splice site enhancer weight" (SSEW) to capture both enhancing and silencing properties [66]: we assigned a weight of +1 or −1 for each exonic enhancer or silencer motif predicted by any of the algorithms, and subtracted the total sum of downstream weights from the sum of upstream weights. Accordingly, we calculated SSEW for ESRseq and HEXplorer as upstream ESRseq and HEXplorer totals minus downstream totals. Since different tools vary in algorithmic principle and value range, it is interesting to compare the different methods.

RESCUE-ESE and the ESEfinder consider subsets of SREs that promote downstream and repress upstream 5'ss usage. In the example above, RESCUE-ESE predicted no difference in 5'ss context, whereas, surprisingly, ESEfinder predicted a stronger 5'ss SRE support within the HMSD gene context, contrary to the experimental observation that the TRIM62 context supported weaker 5'ss more than HMSD.

Similar to the remaining tools, FAS-ESS, which scans sequences for a small set of SREs which repress downstream but enhance upstream 5'ss usage, correctly predicted the stronger TRIM62 5'ss context. RBPmap, HSF3 and EX-SKIP naturally have higher SSEWs than RESCUE-ESE or FAS-ESS, since they comprise predictions of multiple SRE tools. All three tools also graded TRIM62 context significantly more enhancing than HMSD context.

Rather than counting individual motifs, HEXplorer and ESRseq SSEWs total positive and negative hexamer-based scores in genomic regions impacting 5'ss usage. Both ESRseq and HEXplorer correctly assessed the TRIM62 context as more supportive for splicing than HMSD.

In their minigene model system, Lu et al. calculated intrinsic 5'ss strength using the MaxEnt score, based on nine 5'ss nucleotides. Calculating HBond scores of the 11 nucleotide long 5'ss obviously depended on the "context-positions" +7 and +8: within the HMSD context, 5'ss had equal or even slightly higher HBond scores than within the TRIM62 context (Table 3). Thus, in this model system, presence or absence of U1 snRNA complementarity in positions +7 and +8 cannot be responsible for the observed difference between both contexts.

However, for three 5'ss, HBond scores for hypothetically U1 snRNA complementary AT dinucleotides in positions +7/+8 are larger than in the TRIM62 context, so that nucleotide changes in these positions, that are usually included in the context, could in fact modify U1 snRNA complementarity and thus more directly impact splicing (Table 3, *).

Gene architecture could also play a role in differential splicing outcome between TRIM62 and HMSD contexts. The splice acceptor of the shorter 96 nucleotides long TRIM62 exon has a high MaxEnt score of 9.97, whereas the acceptor of the longer 174 nucleotides long HMSD exon is weaker (MaxEnt score 6.67) and thus potentially leads to less efficient exon definition. This is in line with recent findings of Wong et al., comparing the activity of 32,768 unique 5'ss sequences (NNN/GYNNNN) in three different gene contexts (BRCA2 exon 17, SMN1 exon 7, and ELP1 (IKBKAP) exon 20) by massively parallel splicing assays. In these experimental settings, context dependency of 5'ss recognition was strongly determined by the strength of the upstream 3'ss [12].

4.3. Selection of appropriate computational tools for different applications

Generally, specific scientific or diagnostic questions determine which tools are most suitable. RBPmap, RESCUE-ESE, FAS-ESS, SpliceAid, Human Splicing Finder, SROOGLE and especially the commonly used ESEfinder determine locations of highly validated SRP binding motifs within a given RNA sequence. These tools predict individual potential SRP binding sites and can be beneficial in single-case

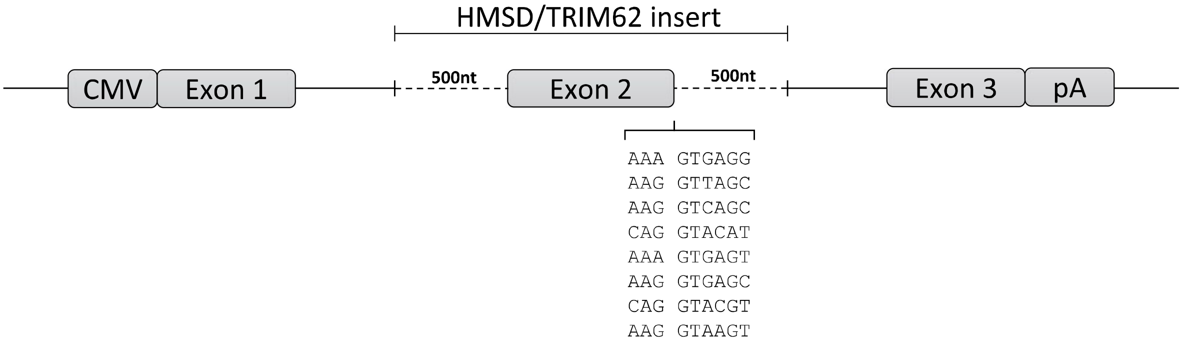


Fig. 3. TRIM62/HMSD minigene reporter. In order to compare two different SD contexts, exon 2 of HMSD and TRIM62 plus 500 nucleotides of the upstream and downstream introns were inserted into a three-exon minigene reporter. The respective wild type splice donor sequences were then replaced (pos. -3 to +6) with eight different splice donor sites, to measure splice donor strength dependent exon skipping in both contexts. CMV: CMV promotor; pA: polyadenylation site.

Table 2
Evaluation of TRIM62 or HMSD context using different tools.

For every tool, splice site enhancer weights were calculated as total upstream predicted SREs minus total downstream SREs, with exonic enhancers counted as +1 and silencers as −1. Higher splice site enhancer weights reflect higher expected enhancing effects on 5'ss usage. The difference between both contexts was calculated by subtracting the TRIM62 SSEW from the HMSD SSEW.

Tool	Splice site enhancer weight		Difference: HMSD-TRIM62
	TRIM62	HMSD	
RESCUE-ESE	−2	−2	0
RBPmap	8	−1	−9
FAS-ESS	3	−2	−5
ESE-finder	−1	7	8
EX-SKIP	51	−13	−64
HSF3	12	4	−8
HEXplorer	504	−92	−596
ESRseq	6,7	−0,5	−7,2

Table 3
Comparison of intrinsic 5'ss strength (HBond score, HBS) including positions +7/+8. Insertion of 9 nt long splice donor sequences into TRIM62 and HMSD contexts leads to different HBond scores depending on the context nucleotides in +7/+8 (CT for TRIM62, AA for HMSD, or AT in a hypothetical context providing U1 snRNA complementarity in positions +7/+8). Large HBond score differences are marked by *.

Inserted SD sequence	HBS: SD in TRIM62 +CT	HBS: SD in HMSD +AA	ΔHBS: HMSD – TRIM62	HBS: SD +AT	ΔHBS: +AT – TRIM62
AAAGTGAGG	10.5	10.7	0.2	12.0	1.5
AAGGTTAGC	13.2	13.4	0.2	14.7	1.5
AAGGTCAGC	13.2	13.4	0.2	14.7	1.5
CAGGTACAT	14.0	14.0	0.0	14.8	0.8
AAAGTGAGT	13.0	14.2	1.2	15.5	2.5*
AAGGTGAGC	15.7	15.9	0.2	17.2	1.5
CAGGTACGT	18.0	19.2	1.2	20.5	2.5*
AAGGTAAGT	19.6	20.8	1.2	22.1	2.5*

analyses of specific SNVs, but they do not explicitly quantify their impact on splicing, which limits their usefulness in a diagnostic setting.

On the other hand, EX-SKIP, the Human Splicing Finder or the state-of-the-art deep-learning based SpliceAI [83] estimate the expected splicing outcome from given sequences in a single, nontransparent "black box" procedure that does not make the incorporation of potential SREs explicit, and—except for HSF3—they do not provide information about the positions of potential SRP binding sites.

For a given sequence, ESRseq and HEXplorer aim at quantifying the potentially enhancing or silencing effects on 5'ss usage. Following a RESCUE-type approach, both algorithms are based on hexamer frequency differences between upstream and downstream sequences, and they reflect position-dependent effects of SREs. Both scores can be used to analyze changes in splice site context on a genome wide scale, but they do not predict individual potential SRP binding motifs like ESE-finder. Comparative studies validated that ESRseq and HEXplorer score changes significantly correlated with changes in splice site usage for known human gene variants as well as in splicing reporter experiments [89,90].

Human Splicing Finder constitutes a well-designed tool providing a clear overview about SREs predicted by various tools like RESCUE-ESE and ESEfinder, in addition to individual motifs for SRSF3 and SRSF10, derived from public data. It may be particularly suited to investigate whether overexpression of an SRP of interest might influence usage of a certain splice site.

5. Challenges in computational 5'ss evaluation

Although there has been a tremendous gain in knowledge on context dependency of splicing, there still remain important questions to be addressed, both locally on a small scale, and involving wider sequence neighborhoods of splice sites on a larger scale.

5.1. Small scale challenges

Generally, RNA secondary structure can impact splice site accessibility, and thus represents another context dependent factor [91–94]. Large scale RBP binding assays showed that although no RBP seems to strictly require a certain RNA secondary structure, some RBPs seem to prefer a binding site within or especially not within a hairpin loop [95,96]. Impact of secondary structure, however, has not been implemented yet in any of the currently available computational tools.

Tools based on SELEX methods [84] lead to the identification of comparatively few high-affinity motifs, leaving potential binding sites of lower affinity aside (*reviewed in* [97]). Additionally, tools mentioned above often predict overlapping or nearby SRP binding motifs which could indicate a protein-RNA binding competition situation, possibly involving steric hindrance. More generally, there is no explicit ("transparent") joint model of 5'ss and SREs addressing 5'ss usage. Quantitatively incorporating/modeling interactions and interdependencies between different SREs as well as between SREs and 5'ss still is a major challenge for computational prediction of splicing.

5.2. Large scale challenges

Gene architecture, *i.e.* genomic distribution of splice sites and SREs as well as exon and intron lengths, is currently not addressed by any of the above tools. Depending on the intrinsic strength and context of nearby splice sites, weakening an index splice site due to a pathogenic mutation can result in different kinds of pathological disruptions of splicing, depending on gene structure [58].

Repression of 5'ss usage can either lead to skipping of the respective exon [98], alternative exon ends due to usage of cryptic splice sites [99], multiple exon skipping due to failed exon definition or changes in the order and dynamics of intron removal [100,101], intron inclusion, or a combination of the above [102]. In most exons below ~250 nt, splice site recognition is well explained by exon definition [103], describing the observation that binding of U1 snRNP at the 3′ end of an exon can enhance recognition of the 5′ exon end [104] and *vice versa* [105].

Mutually enhancing exon end recognition could explain the so-called proximity rule, that in case of competing equally strong 5'ss, the 5'ss closer to the 3'ss of the downstream exon is often predominantly chosen [106,107]. This proximity rule is still applicable after decreasing the proximal 5'ss strength, making it important to not only consider SRP binding sites during 5'ss context evaluation, but also presence and strength of surrounding potential splice site sequences [108]. In line with these findings, SpliceAI neural network classification was much better with up to 10,000 nt long input sequences than with short 80 nt neighborhoods, although SpliceAI does not explicitly identify SREs but implicitly learned their effects.

5.3. Cell-type specific computational prediction of 5'ss and SREs

A fundamental uncertainty of identifying SREs as potential SRP binding sites lies in the unknown availability of the specific SRP binding partners. It is known that expression of SRPs is cell-type specific. A possible approach to address cell-type specificity could be based on RNA-seq data from different cell types: For each cell type, exon junctions and 5'ss predominantly occurring in one cell type but not in others could be collected, and hexamer frequencies could be determined from their respective up- and downstream neighborhoods, along the lines of the RESCUE-ESE concept. From these frequency tables, normalized hexamer *Z*-weights could be calculated and enter into a cell-type specific HEXplorer score. Profiles generated with this HEXplorer score would then hypothetically represent cell-type specific (SRP-availability weighted) splicing enhancing or silencing effects.

6. Conclusion and outlook

Bioinformatics analyses of splice site environment can contribute significantly to understanding splice site usage. With the advent of RNA sequencing techniques, focus has moved to the impact of mutations on RNA processing regulation. This is particularly important in human mutation diagnostics in the vicinity of splice sites [109] as well as in development of therapeutic strategies.

FDA approved drugs have recently used this interventional pathway in SMA and DMD. In SMA patients, the use of antisense oligonucleotides (ASO) to mask an intronic splicing regulatory element has been shown to successfully restore exon inclusion (Nusinersen/Spinraza®) [110]. In DMD patients with loss of exons 49 and 50, the correct reading frame could be reinstated with Eteplirsen (Exondys 51®), leading to skipping of exon 51 through SRE masking [111].

While previously optimal ASO target sequences were typically identified by laborious and costly scanning of large target regions with overlapping ASOs, systematic *in silico* prediction of promising context dependent ASO targets might considerably speed up the development of RNA therapeutic agents.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG) (SCHA909/8-1; TI323/4-1), the Forschungskommission of the Heinrich Heine University Düsseldorf, the Heinz-Ansmann Stiftung für AIDS-Forschung, Düsseldorf (H.S.), and the Jürgen Manchot Stiftung (L.M., H.S.).

Transparency document

The Transparency document associated this article can be found, in online version.

Acknowledgements

We would like to thank all lab members for discussion and critical reading of the manuscript.

Author contributions

J.P., L.M., S.T. and H.S. wrote the manuscript and revised it for important intellectual content.

References

[1] J.F. Catterall, B.W. O'Malley, M.A. Robertson, R. Staden, Y. Tanaka, G.G. Brownlee, Nucleotide sequence homology at 12 intron–exon junctions in the chick ovalbumin gene, Nature 275 (1978) 510–513.

[2] R. Breathnach, C. Benoist, K. O'Hare, F. Gannon, P. Chambon, Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries, Proc. Natl. Acad. Sci. U. S. A. 75 (1978) 4853–4857.

[3] S.M. Mount, I. Pettersson, M. Hinterberger, A. Karmas, J.A. Steitz, The U1 small nuclear RNA-protein complex selectively binds a 5′ splice site in vitro, Cell 33 (1983) 509–518.

[4] S.M. Mount, A catalogue of splice junction sequences, Nucleic Acids Res. 10 (1982) 459–472.

[5] M. Aebi, H. Hornig, C. Weissmann, 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU, Cell 50 (1987) 237–246.

[6] S. Weber, M. Aebi, In vitro splicing of mRNA precursors: 5′ cleavage site can be predicted from the interaction between the 5′ splice region and the 5′ terminus of U1 snRNA, Nucleic Acids Res. 16 (1988) 471–486.

[7] M. Aebi, H. Hornig, R.A. Padgett, J. Reiser, C. Weissmann, Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA, Cell 47 (1986) 555–565.

[8] C. Montell, A.J. Berk, Elimination of mRNA splicing by a point mutation outside the conserved GU at 5′ splice sites, Nucleic Acids Res. 12 (1984) 3821–3827.

[9] Y. Zhuang, A.M. Weiner, A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation, Cell 46 (1986) 827–835.

[10] X. Roca, A.R. Krainer, Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA, Nat. Struct. Mol. Biol. 16 (2009) 176–182, https://doi.org/10.1038/nsmb.1546.

[11] X. Roca, M. Akerman, H. Gaus, A. Berdeja, C.F. Bennett, A.R. Krainer, Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides, Genes Dev. 26 (2012) 1098–1109, https://doi.org/10.1101/gad.190173.112.

[12] M.S. Wong, J.B. Kinney, A.R. Krainer, Quantitative activity profile and context dependence of all human 5′ splice sites, Mol. Cell 71 (2018) 1012–1026 e1013 https://doi.org/10.1016/j.molcel.2018.07.033.

[13] X. Roca, A.R. Krainer, I.C. Eperon, Pick one, but be quick: 5′ splice sites and the problems of too many choices, Genes Dev. 27 (2013) 129–144, https://doi.org/10.1101/gad.209759.112.

[14] N. Caminsky, E.J. Mucaki, P.K. Rogan, Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis, F1000Res 3 (2014) 282, https://doi.org/10.12688/f1000research.5654.1.

[15] G. Yeo, C.B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, J. Comput. Biol. 11 (2004) 377–394, https://doi.org/10.1089/1066527041410418.

[16] K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang, Y. Takeshima, M. Matsuo, G. Sobue, K. Ohno, In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5′ splice sites, Nucleic Acids Res. 35 (2007) 5995–6003, https://doi.org/10.1093/nar/gkm647.

[17] X. Roca, A.J. Olson, A.R. Rao, E. Enerly, V.N. Kristensen, A.L. Borresen-Dale, B.S. Andresen, A.R. Krainer, R. Sachidanandam, Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics, Genome Res. 18 (2008) 77–87, https://doi.org/10.1101/gr.6859308.

[18] L. Hartmann, S. Theiss, D. Niederacher, H. Schaal, Diagnostics of pathogenic splicing mutations: does bioinformatics cover all bases? Front. Biosci. 13 (2008) 3252–3272.

[19] S. Kammler, C. Leurs, M. Freund, J. Krummheuer, K. Seidel, T.O. Tange, M.K. Lund, J. Kjems, A. Scheid, H. Schaal, 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 (2001) 421–434.

[20] M. Freund, M.J. Hicks, C. Konermann, M. Otte, K.J. Hertel, H. Schaal, Extended base pair complementarity between U1 snRNA and the 5′ splice site does not inhibit splicing in higher eukaryotes, but rather increases 5′ splice site recognition,

Nucleic Acids Res. 33 (2005) 5112–5119, https://doi.org/10.1093/nar/gki824.

[21] M. Freund, C. Asang, S. Kammler, C. Konermann, J. Krummheuer, M. Hipp, I. Meyer, W. Gierling, S. Theiss, T. Preuss, D. Schindler, J. Kjems, H. Schaal, A novel approach to describe a U1 snRNA binding site, Nucleic Acids Res. 31 (2003) 6963–6975.

[22] H.V. Colot, F. Stutz, M. Rosbash, The yeast splicing factor Mud13p is a commitment complex component and corresponds to CBP20 the small subunit of the nuclear cap-binding complex, Genes Dev. 10 (1996) 1699–1708, https://doi.org/10.1101/gad.10.13.1699.

[23] J.D. Lewis, E. Izaurralde, A. Jarmolowski, C. McGuigan, I.W. Mattaj, A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5′ splice site, Genes Dev. 10 (1996) 1683–1698.

[24] M.M. Konarska, R.A. Padgett, P.A. Sharp, Recognition of cap structure in splicing in vitro of mRNA precursors, Cell 38 (1984) 731–736.

[25] H.G. Martinson, An active role for splicing in 3′-end formation, Wiley Interdiscip. Rev. RNA 2 (2011) 459–470, https://doi.org/10.1002/wrna.68.

[26] K.L. Fox-Walsh, Y. Dou, B.J. Lam, S.P. Hung, P.F. Baldi, K.J. Hertel, The architecture of pre-mRNAs affects mechanisms of splice-site pairing, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 16176–16181, https://doi.org/10.1073/pnas.0508489102.

[27] A.J. Matlin, F. Clark, C.W. Smith, Understanding alternative splicing: towards a cellular code, Nat. Rev. Mol. Cell Biol. 6 (2005) 386–398, https://doi.org/10.1038/nrm1645.

[28] S. Erkelenz, W.F. Mueller, M.S. Evans, A. Busch, K. Schoneweis, K.J. Hertel, H. Schaal, Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms, RNA 19 (2013) 96–102, https://doi.org/10.1261/rna.037044.112.

[29] J.F. Caceres, S. Stamm, D.M. Helfman, A.R. Krainer, Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors, Science 265 (1994) 1706–1709.

[30] I.C. Eperon, O.V. Makarova, A. Mayeda, S.H. Munroe, J.F. Caceres, D.G. Hayward, A.R. Krainer, Selection of alternative 5′ splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1, Mol. Cell. Biol. 20 (2000) 8303–8318.

[31] E. Dassi, Handshakes and fights: the regulatory interplay of RNA-binding proteins, Front. Mol. Biosci. 4 (2017) 67, , https://doi.org/10.3389/fmolb.2017.00067.

[32] M. Baralle, F.E. Baralle, The splicing code, Biosystems 164 (2018) 39–48, https://doi.org/10.1016/j.biosystems.2017.11.002.

[33] Z. Wang, C.B. Burge, Splicing regulation: from a parts list of regulatory elements to an integrated splicing code, RNA 14 (2008) 802–813, https://doi.org/10.1261/rna.876308.

[34] M.L. Anko, Regulation of gene expression programmes by serine-arginine rich splicing factors, Semin. Cell Dev. Biol. 32 (2014) 11–21, https://doi.org/10.1016/j.semcdb.2014.03.011.

[35] P.J. Shepard, K.J. Hertel, The SR protein family, Genome Biol. 10 (2009) 242, https://doi.org/10.1186/gb-2009-10-10-242.

[36] W. van Der Houven Van Oordt, K. Newton, G.R. Screaton, J.F. Caceres, Role of SR protein modular domains in alternative splicing specificity in vivo, Nucleic Acids Res. 28 (2000) 4822–4831.

[37] V. Botti, F. McNicoll, M.C. Steiner, F.M. Richter, A. Solovyeva, M. Wegener, O.D. Schwich, I. Poser, K. Zarnack, I. Wittig, K.M. Neugebauer, M. Muller-McNicoll, Cellular differentiation state modulates the mRNA export activity of SR proteins, J. Cell Biol. 216 (2017) 1993–2009, https://doi.org/10.1083/jcb.201610051.

[38] J.L. Manley, A.R. Krainer, A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins), Genes Dev. 24 (2010) 1073–1074, https://doi.org/10.1101/gad.1934910.

[39] H. Shen, M.R. Green, A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly, Mol. Cell 16 (2004) 363–373, https://doi.org/10.1016/j.molcel.2004.10.021.

[40] T. Geuens, D. Bouhy, V. Timmerman, The hnRNP family: insights into their role in health and disease, Hum. Genet. 135 (2016) 851–867, https://doi.org/10.1007/s00439-016-1683-5.

[41] J. Jean-Philippe, S. Paz, M. Caputi, hnRNP A1: the Swiss army knife of gene expression, Int. J. Mol. Sci. 14 (2013) 18999–19024, https://doi.org/10.3390/ijms140918999.

[42] G. Dreyfuss, V.N. Kim, N. Kataoka, Messenger-RNA-binding proteins and the messages they carry, Nat. Rev. Mol. Cell Biol. 3 (2002) 195–205, https://doi.org/10.1038/nrm760.

[43] A. Busch, K.J. Hertel, Evolution of SR protein and hnRNP splicing regulatory factors, Wiley Interdiscip. Rev. RNA 3 (2012) 1–12, https://doi.org/10.1002/wrna.100.

[44] H.X. Liu, M. Zhang, A.R. Krainer, Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins, Genes Dev. 12 (1998) 1998–2012.

[45] H. Tian, R. Kole, Selection of novel exon recognition elements from a pool of random sequences, Mol. Cell. Biol. 15 (1995) 6291–6298.

[46] L.R. Coulter, M.A. Landree, T.A. Cooper, Identification of a new class of exonic splicing enhancers by in vivo selection, Mol. Cell. Biol. 17 (1997) 2143–2150.

[47] W.G. Fairbrother, R.F. Yeh, P.A. Sharp, C.B. Burge, Predictive identification of exonic splicing enhancers in human genes, Science 297 (2002) 1007–1013, https://doi.org/10.1126/science.1073774.

[48] S. Erkelenz, S. Theiss, M. Otte, M. Widera, J.O. Peter, H. Schaal, Genomic HEXploring allows landscaping of novel potential splicing regulatory elements, Nucleic Acids Res. 42 (2014) 10681–10697, https://doi.org/10.1093/nar/gku736.

[49] L. Cartegni, J. Wang, Z. Zhu, M.Q. Zhang, A.R. Krainer, ESEfinder: a web resource to identify exonic splicing enhancers, Nucleic Acids Res. 31 (2003) 3568–3571.

[50] S. Ke, S. Shang, S.M. Kalachikov, I. Morozova, L. Yu, J.J. Russo, J. Ju, L.A. Chasin, Quantitative evaluation of all hexamers as exonic splicing elements, Genome Res. 21 (2011) 1360–1374, https://doi.org/10.1101/gr.119628.110.

[51] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule, iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution, Nat. Struct. Mol. Biol. 17 (2010) 909–915, https://doi.org/10.1038/nsmb.1838.

[52] K. Charizanis, K.Y. Lee, R. Batra, M. Goodwin, C. Zhang, Y. Yuan, L. Shiue, M. Cline, M.M. Scotti, G. Xia, A. Kumar, T. Ashizawa, H.B. Clark, T. Kimura, M.P. Takahashi, H. Fujimura, K. Jinnai, H. Yoshikawa, M. Gomes-Pereira, G. Gourdon, N. Sakai, S. Nishino, T.C. Foster, M. Ares Jr., R.B. Darnell, M.S. Swanson, Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy, Neuron 75 (2012) 437–450, https://doi.org/10.1016/j.neuron.2012.05.029.

[53] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (2010) 129–141, https://doi.org/10.1016/j.cell.2010.03.009.

[54] S. De, M. Gorospe, Bioinformatic tools for analysis of CLIP ribonucleoprotein data, Wiley Interdiscip. Rev. RNA 8 (2017), https://doi.org/10.1002/wrna.1404.

[55] Y. Wang, X. Xiao, J. Zhang, R. Choudhury, A. Robertson, K. Li, M. Ma, C.B. Burge, Z. Wang, A complex network of factors with overlapping affinities represses splicing through intronic elements, Nat. Struct. Mol. Biol. 20 (2013) 36–45, https://doi.org/10.1038/nsmb.2459.

[56] Y. Wang, M. Ma, X. Xiao, Z. Wang, Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules, Nat. Struct. Mol. Biol. 19 (2012) 1044–1052, https://doi.org/10.1038/nsmb.2377.

[57] M. Krawczak, N.S. Thomas, B. Hundrieser, M. Mort, M. Wittig, J. Hampe, D.N. Cooper, Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing, Hum. Mutat. 28 (2007) 150–158, https://doi.org/10.1002/humu.20400.

[58] A. Abramowicz, M. Gos, Splicing mutations in human genetic disorders: examples, detection, and confirmation, J. Appl. Genet. 59 (2018) 253–268, https://doi.org/10.1007/s13353-018-0444-7.

[59] M.S. Falzarano, C. Scotton, C. Passarelli, A. Ferlini, Duchenne Muscular Dystrophy: from diagnosis to therapy, Molecules 20 (2015) 18168–18184, https://doi.org/10.3390/molecules201018168.

[60] A. Nishiyama, Y. Takeshima, Z. Zhang, Y. Habara, T.H. Tran, M. Yagi, M. Matsuo, Dystrophin nonsense mutations can generate alternative rescue transcripts in lymphocytes, Ann. Hum. Genet. 72 (2008) 717–724, https://doi.org/10.1111/j.1469-1809.2008.00468.x.

[61] S. Gonzalo, R. Kreienkamp, P. Askjaer, Hutchinson-Gilford Progeria Syndrome: a premature aging disease caused by LMNA gene mutations, Ageing Res. Rev. 33 (2017) 18–29, https://doi.org/10.1016/j.arr.2016.06.007.

[62] P. Scaffidi, T. Misteli, Lamin A-dependent nuclear defects in human aging, Science 312 (2006) 1059–1063, https://doi.org/10.1126/science.1127168.

[63] S. Rodriguez, F. Coppede, H. Sagelius, M. Eriksson, Increased expression of the Hutchinson-Gilford progeria syndrome truncated lamin A transcript during cell aging, Eur. J. Hum. Genet. 17 (2009) 928–937, https://doi.org/10.1038/ejhg.2008.270.

[64] M. Mine, M. Brivet, G. Touati, P. Grabowski, M. Abitbol, C. Marsac, Splicing error in E1alpha pyruvate dehydrogenase mRNA caused by novel intronic mutation responsible for lactic acidosis and mental retardation, J. Biol. Chem. 278 (2003) 11768–11772, https://doi.org/10.1074/jbc.M211106200.

[65] G.K. Brown, L.J. Otero, M. LeGris, R.M. Brown, Pyruvate dehydrogenase deficiency, J. Med. Genet. 31 (1994) 875–879.

[66] A.L. Brillen, K. Schoneweis, L. Walotka, L. Hartmann, L. Muller, J. Ptok, W. Kaisers, G. Poschmann, K. Stuhler, E. Buratti, S. Theiss, H. Schaal, Succession of splicing regulatory elements determines cryptic 5ss functionality, Nucleic Acids Res. 45 (2017) 4202–4216, https://doi.org/10.1093/nar/gkw1317.

[67] C.L. Lorson, E. Hahnen, E.J. Androphy, B. Wirth, A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 6307–6311.

[68] J. Seo, M.D. Howell, N.N. Singh, R.N. Singh, Spinal muscular atrophy: an update on therapeutic progress, Biochim. Biophys. Acta 1832 (2013) 2180–2190, https://doi.org/10.1016/j.bbadis.2013.08.005.

[69] R.N. Singh, N.N. Singh, Mechanism of splicing regulation of Spinal Muscular Atrophy genes, Adv. Neurobiol. 20 (2018) 31–61, https://doi.org/10.1007/978-3-319-89689-2_2.

[70] K. Nakai, H. Sakamoto, Construction of a novel database containing aberrant splicing mutations of mammalian genes, Gene 141 (1994) 171–177.

[71] E. Buratti, M. Chivers, G. Hwang, I. Vorechovsky, DBASS3 and DBASS5: databases of aberrant 3′- and 5′-splice sites, Nucleic Acids Res. 39 (2011) D86–D91, https://doi.org/10.1093/nar/gkq887.

[72] P.D. Stenson, M. Mort, E.V. Ball, K. Shaw, A. Phillips, D.N. Cooper, The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine, Hum. Genet. 133 (2014) 1–9, https://doi.org/10.1007/s00439-013-1358-4.

[73] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (2016) D862–D868, https://doi.org/10.1093/nar/gkv1222.

[74] K. Ohno, J.I. Takeda, A. Masuda, Rules and tools to predict the splicing effects of exonic and intronic mutations, Wiley Interdiscip. Rev. RNA 9 (2018), https://doi.org/10.1002/wrna.1451.

[75] I. Paz, I. Kosti, M. Ares Jr., M. Cline, Y. Mandel-Gutfreund, RBPmap: a web server for mapping binding sites of RNA-binding proteins, Nucleic Acids Res. 42 (2014) W361–W367, https://doi.org/10.1093/nar/gku406.

[76] M. Akerman, H. David-Eden, R.Y. Pinter, Y. Mandel-Gutfreund, A computational approach for genome-wide mapping of splicing factor binding sites, Genome Biol. 10 (2009) R30, https://doi.org/10.1186/gb-2009-10-3-r30.

[77] G. Yeo, S. Hoon, B. Venkatesh, C.B. Burge, Variation in sequence and organization of splicing regulatory elements in vertebrate genes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 15700–15705, https://doi.org/10.1073/pnas.0404901101.

[78] Z. Wang, M.E. Rolish, G. Yeo, V. Tung, M. Mawson, C.B. Burge, Systematic identification and analysis of exonic splicing silencers, Cell 119 (2004) 831–845, https://doi.org/10.1016/j.cell.2004.11.010.

[79] F. Piva, M. Giulietti, L. Nocchi, G. Principato, SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans, Bioinformatics 25 (2009) 1211–1213, https://doi.org/10.1093/bioinformatics/btp124.

[80] M. Raponi, J. Kralovicova, E. Copson, P. Divina, D. Eccles, P. Johnson, D. Baralle, I. Vorechovsky, Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6, Hum. Mutat. 32 (2011) 436–444, https://doi.org/10.1002/humu.21458.

[81] F.O. Desmet, D. Hamroun, M. Lalande, G. Collod-Beroud, M. Claustres, C. Beroud, Human Splicing Finder: an online bioinformatics tool to predict splicing signals, Nucleic Acids Res. 37 (2009) e67, https://doi.org/10.1093/nar/gkp215.

[82] S. Schwartz, E. Hall, G. Ast, SROOGLE: webserver for integrative, user-friendly visualization of splicing signals, Nucleic Acids Res. 37 (2009) W189–W192, https://doi.org/10.1093/nar/gkp320.

[83] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J.F. McRae, S.F. Darbandi, D. Knowles, Y.I. Li, J.A. Kosmicki, J. Arbelaez, W. Cui, G.B. Schwartz, E.D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S.J. Sanders, K.K. Farh, Predicting splicing from primary sequence with deep learning, Cell 176 (2019) 535–548 e524 https://doi.org/10.1016/j.cell.2018.12.015.

[84] C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science 249 (1990) 505–510.

[85] G. Giudice, F. Sanchez-Cabo, C. Torroja, E. Lara-Pezzi, ATtRACT-a Database of RNA-binding Proteins and Associated Motifs, Database: The Journal of Biological Databases and Curation 2016, (2016), https://doi.org/10.1093/database/baw035.

[86] F. Piva, M. Giulietti, A.B. Burini, G. Principato, SpliceAid 2: a database of human splicing factors expression data and RNA target motifs, Hum. Mutat. 33 (2012) 81–85, https://doi.org/10.1002/humu.21609.

[87] X.H. Zhang, L.A. Chasin, Computational definition of sequence motifs governing constitutive exon splicing, Genes Dev. 18 (2004) 1241–1250, https://doi.org/10.1101/gad.1195304.

[88] Z.X. Lu, P. Jiang, J.J. Cai, Y. Xing, Context-dependent robustness to 5′ splice site polymorphisms in human populations, Hum. Mol. Genet. 20 (2011) 1084–1096, https://doi.org/10.1093/hmg/ddq553.

[89] L. Grodecka, E. Buratti, T. Freiberger, Mutations of pre-mRNA splicing regulatory elements: are predictions moving forward to clinical diagnostics? Int. J. Mol. Sci. 18 (2017), https://doi.org/10.3390/ijms18081668.

[90] O. Soukarieh, P. Gaildrat, M. Hamieh, A. Drouet, S. Baert-Desurmont, T. Frebourg, M. Tosi, A. Martins, Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools, PLoS Genet. 12 (2016) e1005756, , https://doi.org/10.1371/journal.pgen.1005756.

[91] T.E. Abbink, B. Berkhout, RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor, J. Virol. 82 (2008) 3090–3098, https://doi.org/10.1128/JVI.01479-07.

[92] D. Zychlinski, S. Erkelenz, V. Melhorn, C. Baum, H. Schaal, J. Bohne, Limited complementarity between U1 snRNA and a retroviral 5′ splice site permits its attenuation via RNA secondary structure, Nucleic Acids Res. 37 (2009) 7429–7440, https://doi.org/10.1093/nar/gkp694.

[93] J. Tan, L. Yang, A.A.L. Ong, J. Shi, Z. Zhong, M.L. Lye, S. Liu, J. Lisowiec-Wachnicka, R. Kierzek, X. Roca, G. Chen, A disease-causing intronic point mutation C19G alters tau exon 10 splicing via RNA secondary structure rearrangement, Biochemistry (Mosc) 58 (2019) 1565–1578, https://doi.org/10.1021/acs.biochem.9b00001.

[94] R. Soemedi, K.J. Cygan, C.L. Rhine, D.T. Glidden, A.J. Taggart, C.L. Lin, A.M. Fredericks, W.G. Fairbrother, The effects of structure on pre-mRNA processing and stability, Methods 125 (2017) 36–44, https://doi.org/10.1016/j.ymeth.2017.06.001.

[95] D. Ray, H. Kazan, K.B. Cook, M.T. Weirauch, H.S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L.H. Matzat, R.K. Dale, S.A. Smith, C.A. Yarosh, S.M. Kelly, B. Nabet, D. Mecenas, W. Li, R.S. Laishram, M. Qiao, H.D. Lipshitz, F. Piano, A.H. Corbett, R.P. Carstens, B.J. Frey, R.A. Anderson, K.W. Lynch, L.O. Penalva, E.P. Lei, A.G. Fraser, B.J. Blencowe, Q.D. Morris, T.R. Hughes, A compendium of RNA-binding motifs for decoding gene regulation, Nature 499 (2013) 172–177, https://doi.org/10.1038/nature12311.

[96] D. Dominguez, P. Freese, M.S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N.J. Lambert, E.L. Van Nostrand, G.A. Pratt, G.W. Yeo, B.R. Graveley, C.B. Burge, Sequence, structure, and context preferences of human RNA binding proteins, Mol. Cell 70 (2018) 854–867 e859 https://doi.org/10.1016/j.molcel.2018.05.001.

[97] K.B. Cook, T.R. Hughes, Q.D. Morris, High-throughput characterization of protein-RNA interactions, Brief. Funct. Genomics 14 (2015) 74–89, https://doi.org/10.1093/bfgp/elu047.

[98] L. Zeng, W. Liu, W. Feng, X. Wang, H. Dang, L. Gao, J. Yao, X. Zhang, A novel donor splice-site mutation of major intrinsic protein gene associated with congenital cataract in a Chinese family, Mol. Vis. 19 (2013) 2244–2249.

[99] Y. Habara, Y. Takeshima, H. Awano, Y. Okizuka, Z. Zhang, K. Saiki, M. Yagi, M. Matsuo, In vitro splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G−>A mutations in introns of the dystrophin gene, J. Med. Genet. 46 (2009) 542–547, https://doi.org/10.1136/jmg.2008.061259.

[100] L.J. Fang, M.J. Simard, D. Vidaud, B. Assouline, B. Lemieux, M. Vidaud, B. Chabot, J.P. Thirion, A novel mutation in the neurofibromatosis type 1 (NF1) gene promotes skipping of two exons by preventing exon definition, J. Mol. Biol. 307 (2001) 1261–1270, https://doi.org/10.1006/jmbi.2001.4561.

[101] T. Hori, T. Fukao, K. Murase, N. Sakaguchi, C.O. Harding, N. Kondo, Molecular basis of two-exon skipping (exons 12 and 13) by c.1248+5g > a in OXCT1 gene: study on intermediates of OXCT1 transcripts in fibroblasts, Hum. Mutat. 34 (2013) 473–480, https://doi.org/10.1002/humu.22258.

[102] K. Takahara, U. Schwarze, Y. Imamura, G.G. Hoffman, H. Toriello, L.T. Smith, P.H. Byers, D.S. Greenspan, Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I, Am. J. Hum. Genet. 71 (2002) 451–465, https://doi.org/10.1086/342099.

[103] B.L. Robberson, G.J. Cote, S.M. Berget, Exon definition may facilitate splice site selection in RNAs with multiple exons, Mol. Cell. Biol. 10 (1990) 84–94, https://doi.org/10.1128/Mcb.10.1.84.

[104] B.E. Hoffman, P.J. Grabowski, U1 snRNP targets an essential splicing factor, U2AF65, to the 3′ splice site by a network of interactions spanning the exon, Genes Dev. 6 (1992) 2554–2568.

[105] S. Ke, L.A. Chasin, Context-dependent splicing regulation: exon definition, co-occurring motif pairs and tissue specificity, RNA Biol. 8 (2011) 384–388.

[106] R. Reed, T. Maniatis, A role for exon sequences and splice-site proximity in splice-site selection, Cell 46 (1986) 681–690.

[107] A.B. Rosenberg, R.P. Patwardhan, J. Shendure, G. Seelig, Learning the sequence determinants of alternative splicing from millions of random sequences, Cell 163 (2015) 698–711, https://doi.org/10.1016/j.cell.2015.09.054.

[108] M.J. Hicks, W.F. Mueller, P.J. Shepard, K.J. Hertel, Competing upstream 5′ splice sites enhance the rate of proximal splicing, Mol. Cell. Biol. 30 (2010) 1878–1886, https://doi.org/10.1128/MCB.01071-09.

[109] D. Baralle, E. Buratti, RNA splicing in human disease and in the clinic, Clin Sci (Lond) 131 (2017) 355–368, https://doi.org/10.1042/CS20160211.

[110] N.N. Singh, M.D. Howell, E.J. Androphy, R.N. Singh, How the discovery of ISS-N1 led to the first medical therapy for spinal muscular atrophy, Gene Ther. 24 (2017) 520–526, https://doi.org/10.1038/gt.2017.34.

[111] E.H. Niks, A. Aartsma-Rus, Exon skipping: a first in class strategy for Duchenne Muscular Dystrophy, Expert. Opin. Biol. Ther. 17 (2017) 225–236, https://doi.org/10.1080/14712598.2017.1271872.

2.1.2 Publication II: Modelling splicing outcome by combining 5’ss strength and splicing regulatory elements

Although various tools exist, that evaluate binding of splicing regulatory proteins (SRPs) on spite site usage, a standardized scoring for the predicted usage of a particular splice site within a given genetic context is still not established, that would reach beyond the above-mentioned intrinsic strength of splice sites. RNA-binding proteins in proximity to the splice site were shown to significantly influence splice site usage and thus should be considered while analyzing the capacity of splice sites to recruit the spliceosome and initiate the assembly of the enzymatic active complex.

In this work, the intrinsic strength of a splice donor (SD) site was combined with the predicted binding of SRPs around it, to estimate the overall capacity of an SD to induce its usage [23]. For every 5’ss annotated in the human genome, every potential GT-site within the respective shortest associated exonic sequence was determined as potential alternative donor sites, to study the differences in the intrinsic strength and the SRP-binding profile, indicated by the surrounding HZ_{EI} score profiles (SSHW), between the annotated SD and the exonic GT site. Generally, the HBond score difference was not a strict discriminator between GT sites and 5’ss and had to be considered together to evaluate GT usage. SSHW distributions of GT sites or annotated 5’ss overlapped to a higher degree than the respective HBS distributions, indicating a stronger influence of HBond score on splice site usage than SSHW. Differences of both HBond score and SSHW from the GT site to those of the annotated 5’ss correlated with GT site usage. Comparing pairs, where the 5’ss showed a weaker HBond score than the GT site, that was either used or not used in our fibroblast RNA-seq data, we detected a significantly higher 5’ss SSHW in pairs, with the unused GT site, than in pairs with the used GT sites. Additionally, the actually used stronger GT sites showed a significantly higher SSHW than the unused stronger GT sites. The ability to better describe 5’ss usage in these potentially competitive situations emphasizes the importance to consider both the intrinsic strength and the surrounding binding landscape of splicing regulatory proteins. Similar observations would be expected with analyzing 3’ss.

The following article was published in Nucleic Acids Res. 2022 Aug 26;50(15):8834-8851.
(doi: 10.1093/nar/gkac663.), by

Lisa Müller*, **Johannes Ptok***, Azlan Nisar, Jennifer Antemann, Ramona Grothmann, Frank Hillebrand, Anna-Lena Brillen, Anastasia Ritchie, Stephan Theiss, Heiner Schaal (*shared first-author)

Contributions

LM, JA, RG, FH, ALB and AR did construction of the reporter constructs and splicing assays. JP, AN, ST and HS did the bioinformatic evaluation of the MPSA and transcriptomic data. ST and HS wrote most of the manuscript. Individual contribution of JP at around 40%.

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

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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-CL-SY-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

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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: 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_{band} <9 reads and n_{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-

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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 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:

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

to the training dataset of 45 165 GT sites and 45 411 annotated 5′ss. The value of the (for γ ≪ 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 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,

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

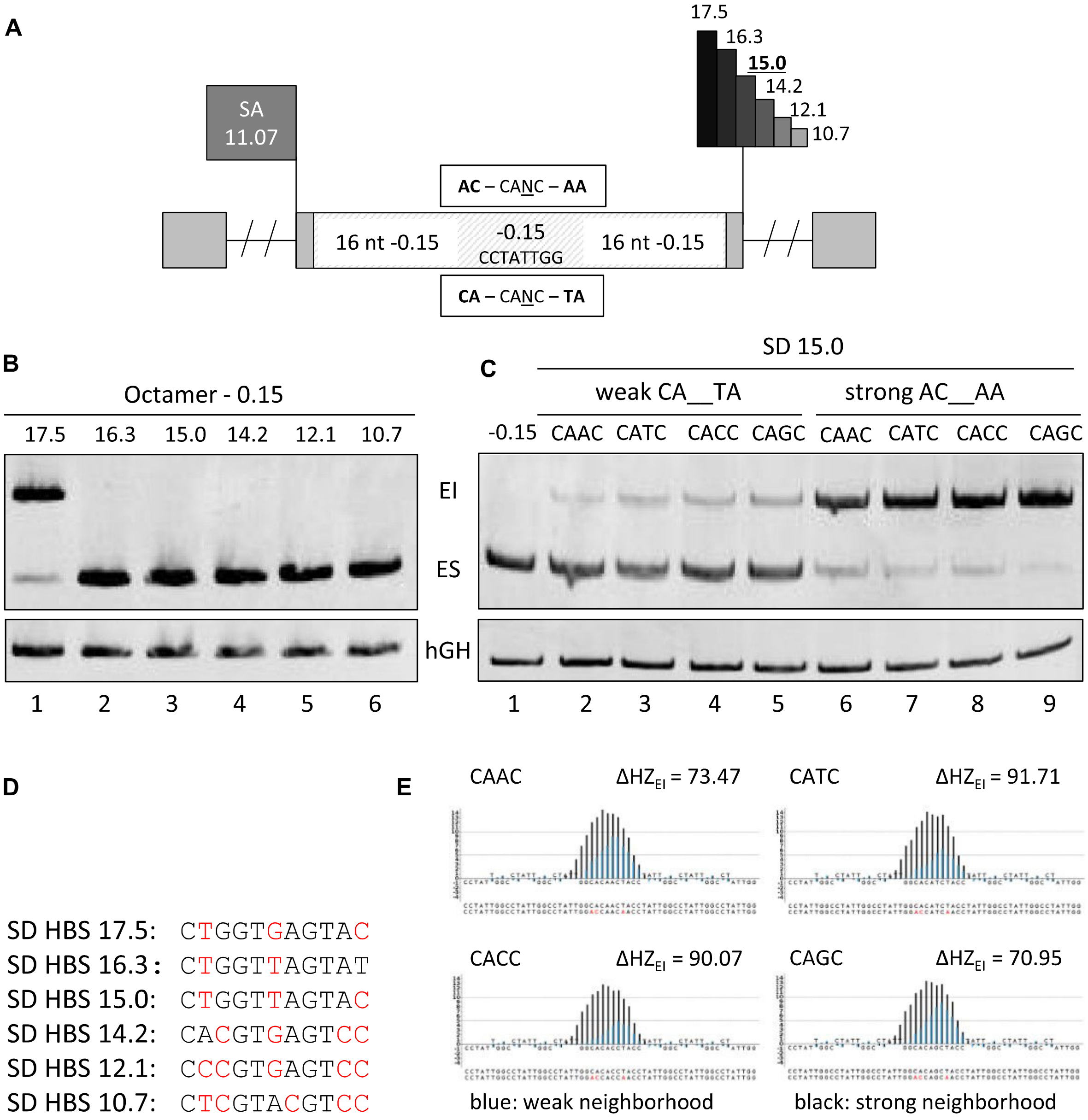


Figure 1. SRSF3 binding motifs CANC mediate exon inclusion in splicing reporter. (**A**) Sketch of the 3-exon minigene reporter plasmid. The middle exon contains an insertion site for different SREs which are flanked by an intrinsically strong splice acceptor (SA MaxEnt 11.07) and splice donor with varying 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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

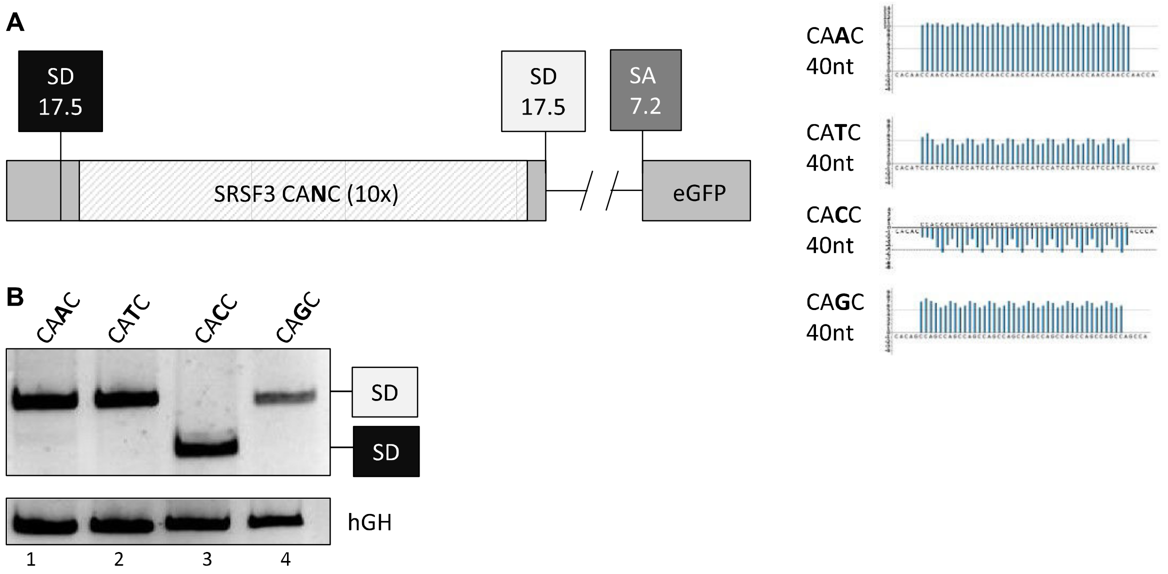


Figure 2. Splicing regulatory effects of concatenated SRSF3 binding sites. (**A**) Schematic drawing of the reporter construct that contains two equally strong splice donors SD with an HBond score of 17.5 (MaxEnt 10.10) and is used to detect up- or downstream enhancing or silencing properties of the 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-

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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 ≥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.

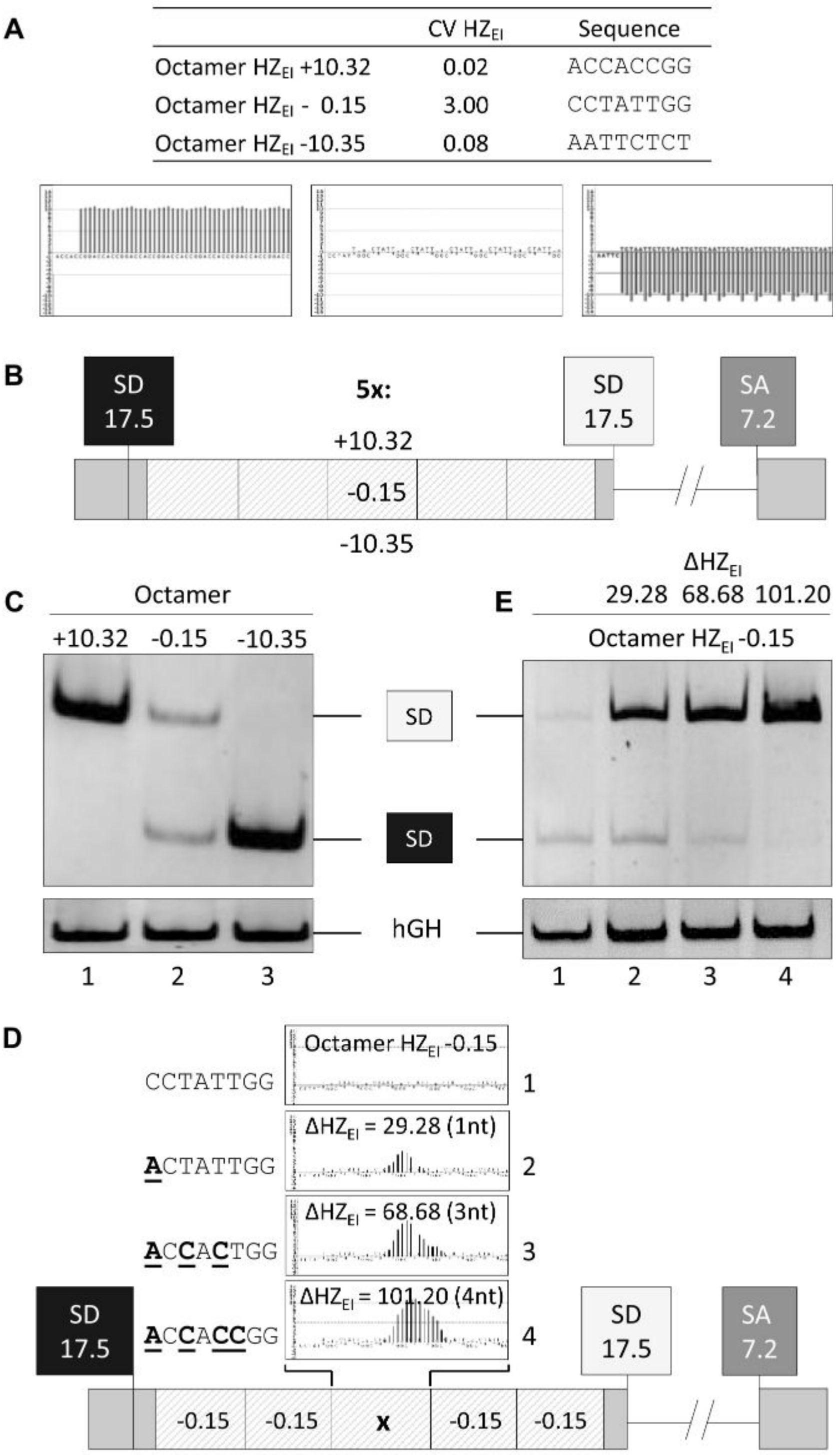
For each of the +10.32 octamers, insertion of repeats resulted in exclusive recognition of the downstream 5′ss, while the use of the upstream donor was completely repressed, confirming their predicted directional splicing regulatory activity (Figure 3C, lane 1; Supplementary Figure S1B, lanes A–G). While 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-

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023



	CV HZ_{EI}	Sequence
Octamer HZ_{EI} +10.32	0.02	ACCACCGG
Octamer HZ_{EI} - 0.15	3.00	CCTATTGG
Octamer HZ_{EI} -10.35	0.08	AATTCTCT

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).

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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 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-

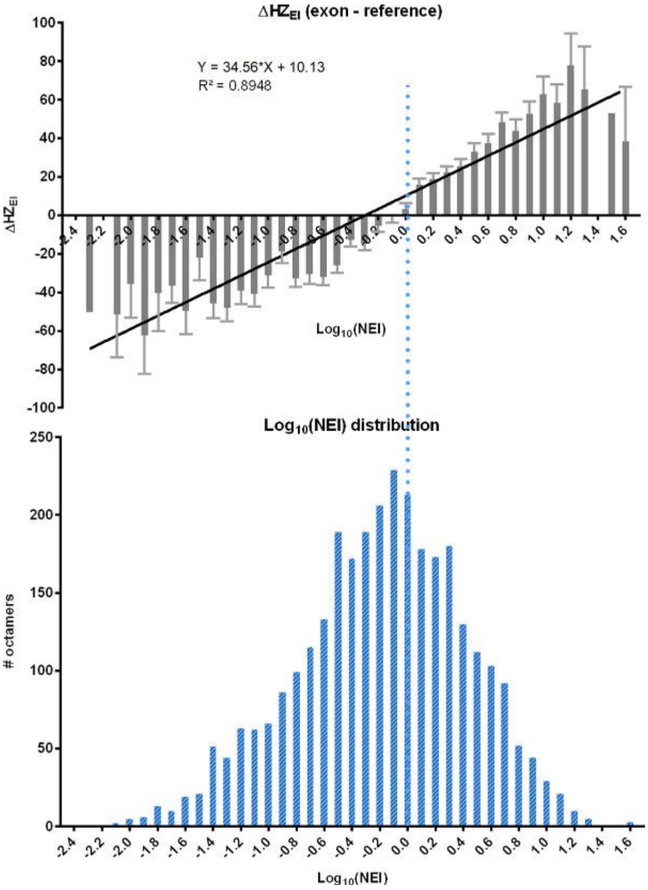


Figure 4. HEXplorer score increases with downstream 5′ss usage in random octamer library assay. Analysis of massively parallel splicing assay with random octamer library inserted in the center of the splicing neutral reference sequence in our splicing competition reporter. Normalized enrichment index (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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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	ΔHBS $\leq$ 0 *weaker* GT-site	ΔHBS > 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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

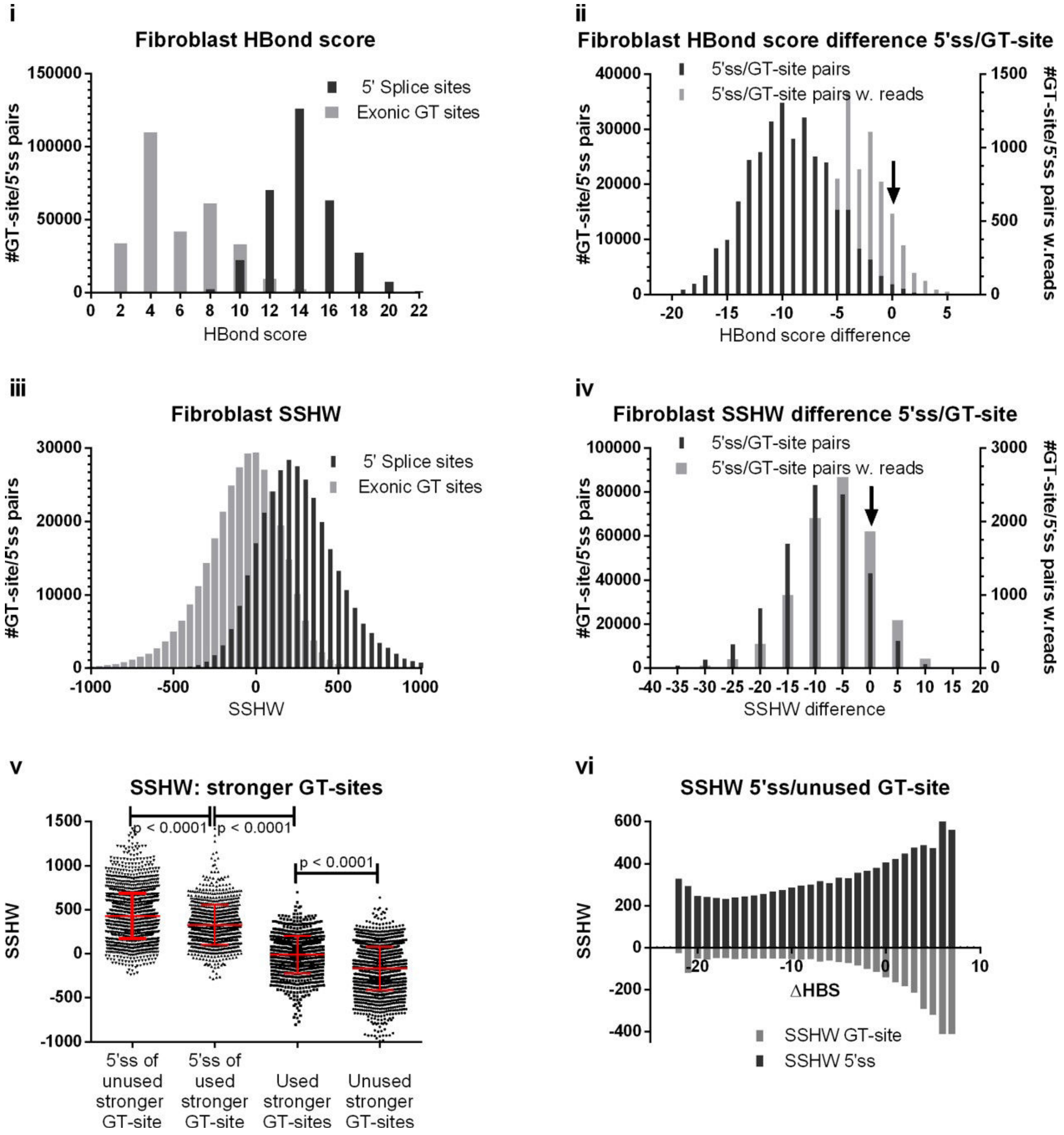


Figure 5. Exonic GT sites have lower U1 snRNA complementarity and weaker SRE support than nearby annotated 5′ss in fibroblast RNA-seq dataset. (**i**) HBond score distributions for 320 601 pairs of high-confidence annotated 5′ss and exonic GT sites closer than 150 nt. (**ii**) HBond score difference $\text{HBS}_{\text{GT}}-\text{HBS}_{5'\text{ss}}$ 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*

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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 ($\Delta HBS > 0$). While for *weaker* GT-sites ($\Delta HBS \leq 0$), GT SSHW was only slightly negative and had little variation, for *stronger* GT-sites ($\Delta HBS > 0$), GT SSHW was increasingly negative, suggesting that SSHW could compensate for $\Delta HBS > 0$ 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 LGNRr $= \log_{10}(\mathrm{GNR}_{\mathrm{GT}}/\mathrm{GNR}_{5'ss})$. 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 ($\mathrm{CV}_{\mathrm{LGNRr}}$ = standard deviation/mean LGNRr) of the LGNRr values in each two-dimensional bin was 0.21 (average $\mathrm{CV}_{\mathrm{LGNRr}} = 0.25$, standard deviation $\mathrm{CV}_{\mathrm{LGNRr}} = 0.18$). 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

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

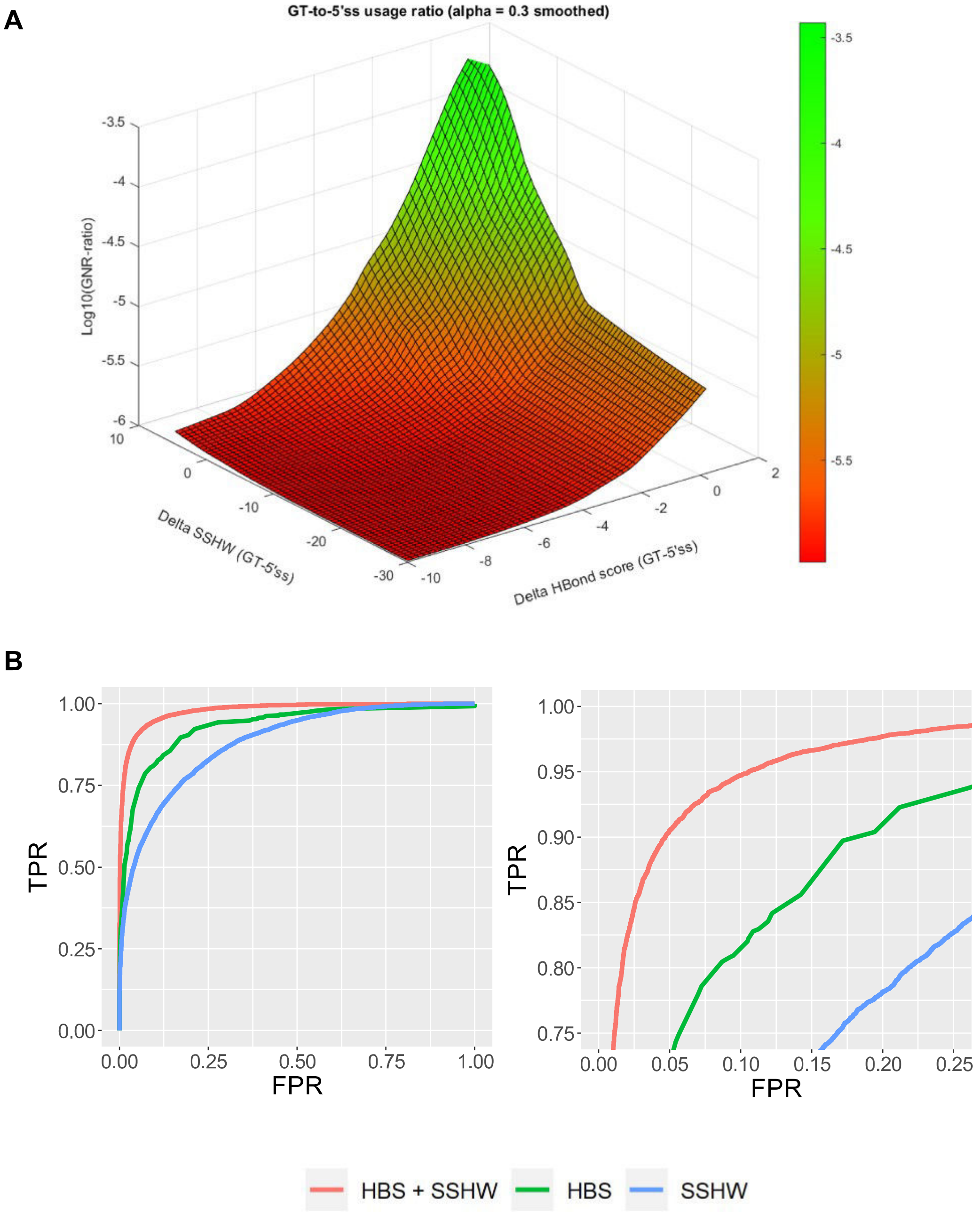


Figure 6. Combination of HBS and SSHW improves classification of GT sites and 5′ss in fibroblast RNA-seq dataset. (**A**) Average LGNRr = $\log_{10}(\mathrm{GNR}_{\mathrm{GT}}/\mathrm{GNR}_{5'\mathrm{ss}})$ as measure of GT-site usage relative to 5′ss (vertical z-axis), plotted as function of HBond score difference $\Delta\mathrm{HBS} = \mathrm{HBS}_{\mathrm{GT}} - \mathrm{HBS}_{5'\mathrm{ss}}$ and splice site HEXplorer weight difference $\Delta\mathrm{SSHW} = \mathrm{SSHW}_{\mathrm{GT}} - \mathrm{SSHW}_{5'\mathrm{ss}}$. 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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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 ($\mathrm{GNRr} = \mathrm{GNR}_{\mathrm{GT}} / \mathrm{GNR}_{5'\mathrm{ss}} < 1\%$) that had medium-to-high U1 snRNA complementarity ($\mathrm{HBS} \geq 10$). 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* ($0 < \mathrm{AUC} < 1$; $\mathrm{AUC} = 0.5$ 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 $\mathrm{AUC} = 0.88$ for SSHW alone, and entering both variables into the model again improved the classification to $\mathrm{AUC} = 0.98$. Thus, in terms of the ROC curves, there is a nearly even AUC spacing of 0.05 each from $\mathrm{SSHW} < \mathrm{HBS} < \mathrm{SSHW} + \mathrm{HBS}$. 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-

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

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.

Downloaded from https://academic.oup.com/nar/article/50/15/8834/6659870 by Universitaets- und Landesbibliothek Duesseldorf user on 31 October 2023

2.2 Thesis 2 – Studying alternative and aberrant splicing in health and cellular stress improves understanding of functional splicing

Studying splice site usage in human transcriptomes shows, that the majority of human transcripts are affected by alternative splicing. Whereas some transcripts show splice-isoforms with variations in exon composition or intron retention events, others show another form of alternative splicing, called noisy splicing, where alternative splice sites in proximity to a highly used splice site are selected. In cancer cells or individual genomes, sequence variations that render the original proximal splice site non-functional, can potentially induce usage of these alternative splice sites with potentially drastic consequences (**Unpublished manuscript I**). Developing tools, that translate sequence variations, that are often based on transcript coordinates rather than genomic coordinates, allows to quickly assess their potential influence on splice site selection (**Publication III**). The amount of alternative splice site usage is additionally regulated during many cellular processes, like reaction to disease or cellular stress. In cardiovascular endothelial cells, alternative splicing affects expression of proteins important during response to oxidative stress, induced by either high-fat diet simulating cell media or lipopolysaccharide (**Publication IV** and **Publication V**).

2.2.1 Unpublished manuscript I: Fully haplotyped genome assemblies of healthy individuals reveal variability in 5’ss strength and support by splicing regulatory proteins

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 [136]. Variant calling from these high-quality haplotype assemblies increases sensitivity and correctly places variants into the right genetic context, which allows to summarize the effect of all relevant SNVs on a given splice site per haplotype. In this work, fully haplotyped genomes of 26 healthy individuals were analyzed to fully assess homozygous and heterozygous sequence variations within the 5'ss and their surrounding regions for their impact on 5’ss strength and predicted binding of splicing regulatory proteins (SRPs) in surrounding sequences. Sequence variations were non-randomly distributed, often inherited, and showed differing tolerance levels in protein-coding and non-coding transcripts. Additionally, we observed a slight balance between sequence variation-induced changes in 5’ss strength and corresponding predicted binding of splicing regulatory proteins in proximity of these 5’ss. Strong 5’ss (HBS > 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 for random variations. Generally, genes analyzed during breast cancer risk assessment seem to only allow HBS reductions of up to 1.1 whereas acceptable changes in predicted SRP binding sites were highly dependent on their respective 5’ss strength.

Article unpublished

Johannes Ptok, Stefan Theiss, Heiner Schaal

Contributions

JP did the bioinformatics analysis and wrote the manuscript, ST and HS supervised the analysis. Individual contribution of JP at around 95%.

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

Abstract

This work presents a comprehensive investigation of sequence variations at human 5' splice sites (5'ss), exploring their impact on splice site strength and potential splicing regulatory protein (SRP) binding sites. Leveraging 26 high-quality genomes, that were recently re-assembled resulting in fully haplotyped assemblies, we were able to fully assess homozygous and heterozygous sequence variations within the 5'ss and their surrounding regions. Sequence variations were non-randomly distributed, often inherited, and showed differing tolerance levels in protein-coding and non-coding transcripts. Additionally, we observed a slight balance between sequence variation-induced changes in 5’ss strength and corresponding predicted binding of splicing regulatory proteins in proximity of these 5’ss. Strong 5’ss (HBS > 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 for random variations. Generally, genes analyzed during breast cancer risk assessment seem to only allow HBS reductions of up to 1.1 whereas acceptable changes in predicted SRP binding sites were highly dependent on their respective 5’ss strength.

1. Introduction

Splicing is a crucial step in 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 snRNPs (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 strength of splice site sequences influences splice site recognition and can be estimated using algorithms such as MaxEntScan, for 5’ and 3’ss [4], or the HBond score (HBS), for 5’ss [5]. While the MaxEntScan for 5'ss only evaluates a 9nt long sequence, the HBS considers all 11 possible nucleotides complementary to the free 5' end of the U1 snRNA. However, such algorithms do not take into account the contribution of SRPs to splice site recognition. SRPs enhance or silence splice site usage depending on splice site strength and their binding position relative to the splice site [6]. The HEXplorer algorithm was developed to predict SRP binding given a genomic sequence and summarizes the overall putative influence of splicing regulatory proteins bound within a +-60nt window with the so called Splice Site HEXplorer Weight (SSHW) [7, 8].

Correctly predicting splice site usage can be very important during diagnosis and the development of treatments for various genetic diseases [9, 10]. Since around 97% of human genes contain intronic sequences, that needs to be removed during pre-mRNA maturation, aberrant splicing can potentially affect translation of almost every human gene [11]. Single nucleotide variants (SNVs) are an important factor that can alter the correct splicing of RNA transcripts. Millions of SNVs directly changing the encoded amino acid sequence via nonsynonymous substitutions, frameshifts or premature stop codons are already described in the literature (ClinVar [12] or SNPdb [13]). However, within an RNA transcript, any SNV, whether so-called silent or not, can alter RNA processing, which is getting more and more attention in diagnostics in recent years. And indeed, approximately 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 [14]. Depending on their genomic location, they can directly alter the sequences of annotated splice sites, alter the SRP-binding landscape or introduce strong cryptic splice sites, potentially leading to splice site competition and aberrant splicing.

Here, we comprehensively analyzed a specific dataset of haplotype genomes from the 1000 Genome Project of 26 healthy individuals to consider the entire individual sequence around a given splice site [15]. The apparent non-pathogenic variability of 5’ss strength or SSHW calculated in this dataset can help us to determine thresholds above which significant changes in splice site usage can be expected.

2. 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 [16]. 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 34 individuals of the 1000 Genome project [17]. Since not all assemblies were generated with the same bioinformatic tools, we first selected genomes from individuals, where both haplotypes were assembled using the flye assembler [18], resulting in fully haplotyped genomes of 26 individuals. To determine variants in proximity to splice sites, we then aligned the assemblies to the human reference genome version GRCh38 and determined variations to it using GSAlign [19]. An R-script then applied functions of our VarCon R package to reconstruct the sequence neighborhood of 5'ss holding the respective sequence variation [20]. R-scripts for the analysis and figures are available on github (https://github.com/caggtaagtat/).

3. Results and Discussion

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

To get a better understanding, when a sequence variation within the immediate region surrounding an exon-intron border, which changes either splice site strength or the binding landscape of splicing regulatory proteins, is sufficiently strong to alter usage of the canonical splice site, we analyzed genomic variations in these regions within a subset of 52 high-quality haplotype assemblies from 26 individuals of the 1000 genome project, that were recently published [17].

After aligning these genome assemblies against the Genome Reference Consortium Human Build 38 (GRCh38), we were able to call the variations to the reference. Based on previously developed methods, we wrote a custom R-script, that integrated sequence variations within a +/-60nt sequence-window around annotated 5’ss into the reference sequence, in order to calculate changes in the 5’ss strength and SSHW, from one to multiple sequence variations in the respective sequence region [20]. Introducing insertion and deletion variations had to be carefully coordinated with the introduction of single nucleotide variations SNVs, since the former naturally change the original genomic coordinates of the respective sequence window (see github.com/caggtaagtat/SNVimpact).

Sometimes, 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 other potentially 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, whether there is a general window of acceptable variation in 5’ss strength, that does not affect 5’ss usage enough to reduce protein production of both gene copies, to be pathogenic. Defining this potential global threshold of variations in 5’ss strength or SSHW could be of clinical importance during identification 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 every 5’ss annotated in transcript isoforms of a gene is equally important for the expression of functional protein—for instance, if the 5’ss is only part of transcript isoforms leading to nuclear retained transcripts and/or transcripts targeted by Nonsense Mediated Decay (NMD). Hence, we generated groups of 5’ss that are either found in (1) exclusively non-coding transcript-isoforms of a gene (“not protein coding”) or (2) in at least one protein coding transcript (“protein coding”). Additionally, the likelihood that an annotated transcript is not described due to technical or biological artifacts can differ across transcripts of a gene. One parameter to assess this “confidence” in a transcript is the transcript support level (TSL, levels ranging from 1 [highest confidence] to 5) provided by the *ensembl* archive. All annotated exon-intron borders of transcript isoforms of TS level 1 (TSL1) are found with the same coordinates in the independent pendant to the *ensembl* archive, called RefSeq, which is provided by the US-American National Center for Biotechnology Information (NCBI) [21]. We therefore additionally labeled 5'ss by whether they were either found in (1) exclusively non-TSL1 transcript-isoforms of a gene ("non-TSL1") or (2) in at least one TSL1 transcript ("TSL1").

3.1.1 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 2,618 (0.84%) showed a homozygous sequence variation across the 23 individuals. 1,433 (54,7%) of the 2,618 5'ss showed a homozygous variation in multiple individuals (from 2 to 22), which additionally indicates that the occurrence of these variations might not be random, but predominantly within the same transcripts and 5'ss. 51.1% of those showed the same HBond score difference across all affected individuals, indicating inheritance rather than random point mutations as the origin.

Grouping these 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 amount of affected 5'ss, relative to the total number of annotated 5'ss in each category (Tab1).

	not TSL1	TSL1
not protein coding	945 (1.1% of 83,910)	170 (1.5% of 11,404)
protein coding	391 (0.9% of 42,920)	1,112 (0.6% of 173,199)

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. 5'ss could either be (1) exclusively part of non-coding transcript isoforms ("not protein coding"), or (2) part of protein coding transcripts ("protein coding"). Additionally, 5'ss were classified whether they were found in 1) transcripts of TSL level lower 1 ("not TSL1"), or 2) 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.

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

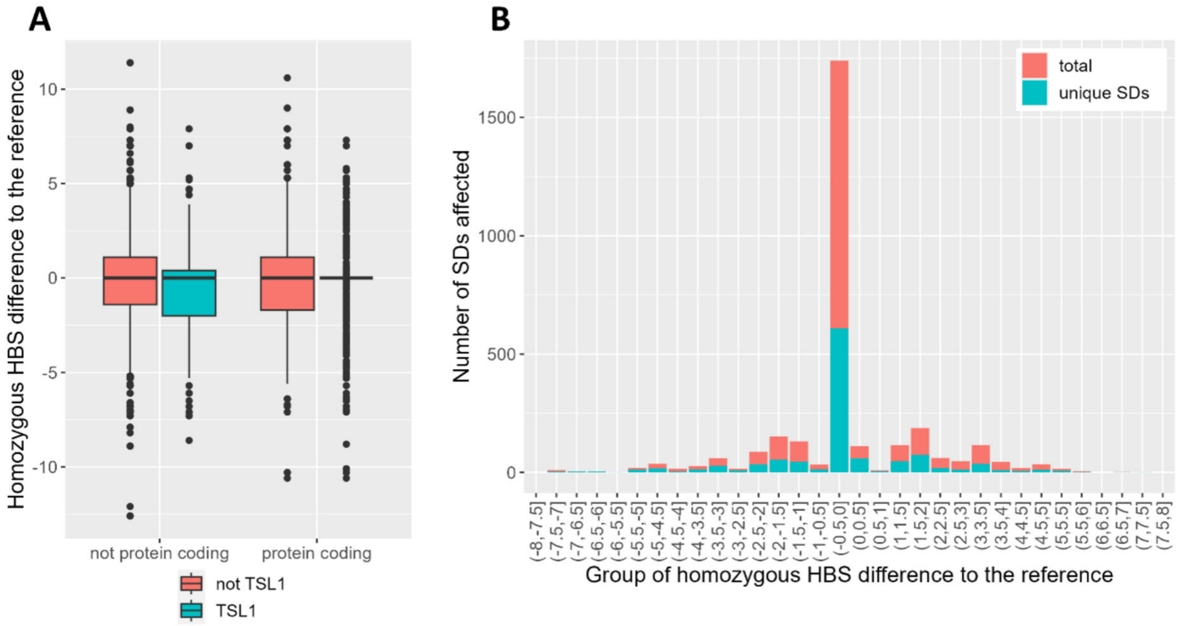


Figure 1. ΔHBS of 5'ss, that are affected by homozygous sequence variations within the 11 nucleotides of the 5'ss sequence by TSL and "protein coding" category. A) Boxplot depicting **ΔHBS** per category. 5'ss were grouped into those, that were 1) exclusively part of non-coding transcript isoforms ("not protein coding"), or 2) part of protein coding transcripts ("protein coding") and those, either found in 1) transcripts of TSL level lower 1 ("not TSL1"), or 2) TSL1 transcripts ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts, that are found with the all 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 in red are the 5'ss counts per ΔHBS group, counting affected 5'ss across all individuals (total = 3,099). Depicted in blue are the counts of unique 5'ss per ΔHBS group (total 1,137).

Comparing homozygous ΔHBS between 5'ss categories, we saw the least deviation from an HBond score difference of 0 for 5'ss that belong to protein coding TSL1 transcripts. Some of the 1,112 5'ss of this category were affected in the genome of multiple individuals, resulting in a total of 3,104 instances. 1,613 of 3,104 (52%) 5'ss of this category showed a ΔHBS of 0 across all individuals. The number of instances with no HBond score 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 TSL1 transcripts showed a slightly lesser amount of ΔHBS of 0 (only 926 of 2,383 instances at 1,506 unique 5'ss, 39%). 5'ss of non-TSL1 transcripts show a much higher ΔHBS deviation around 0, compared to 5'ss of protein coding TSL1 transcripts. However, 5'ss of non-coding TSL1 transcripts seem to show a slight enrichment for HBond score reduction compared to the reference 5'ss (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 transcript support level, TSL1. Around half of affected 5’ss showed an 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 -4 to -1 and 1 to 4 respectively. This could indicate that 5’ss of “highest importance” (only in protein coding TSL1 transcripts) might normally tolerate only a slight HBond score 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 HBond score 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 (Figure 1B). The ΔHBS of heterozygous 5’ss variants showed a similar distribution around zero (data not shown).

Previously we saw that the reduction of 5’ss strength can induce usage of nearby competing 5’ss that are usually not used, independent from the strength of the original 5’ss. This effect was, however, dependent on the HBond score 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).

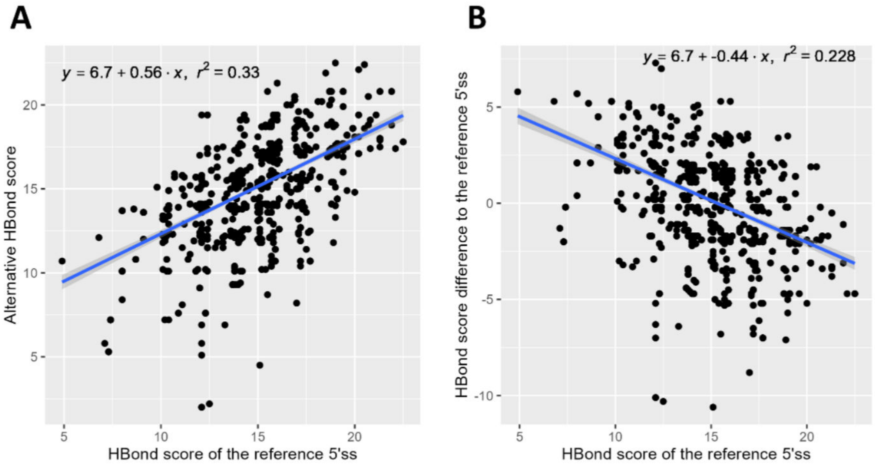


Figure 2: Observed HBond score difference range depends on HBond score of the reference 5'ss. Depicted is the HBond score difference (**B**) or alternative HBond score (**A**), caused by homozygous genetic variations from the reference, within the sequence of annotated 5'ss. 5'ss that were much weaker 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 HBond score compared to the reference. Each regression line is depicted blue, with its formula stated above. The 95% confidence interval of the regression line is marked by a dark grey area. R^2 as a measure how close the single data points are to the regression line is 0.33 (**A**) and 0.23 (**B**) respectively.

We saw a small negative correlation between ΔHBS and the HBond score of the reference 5'ss itself (R^2 of linear regression = 0.228). Homozygous sequence variations within the sequence of coding TSL1 transcript 5'ss more often reduced the HBond score of strong 5'ss. 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 (data not shown). Nevertheless, the detected correlation between reference HBond score and ΔHBS could simply exist, since nucleotide exchanges within high-complementarity 5'ss sequences have a higher chance to decrease the HBond score than within low-complementarity 5'ss sequences by chance.

In order to test, whether the observed sequence variations induce a lower HBond score reduction than expected, we first grouped the 1,112 unique reference 5'ss with homozygous ΔHBS by their HBond score in steps of 1. Subsequently, we removed 5'ss with more than 1 nucleotide differences to the reference, resulting in 1,066 5'ss. Per HBond score 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 expected ΔHBS frequency, which was then compared to the observed real data.

For reference HBond score groups of sufficient size (at least 2% of unique 5’ss in this HBond score group), we compared the theoretically expected ΔHBS frequency of -23.8 to -1 or -1 to 23.8 with the fisher-exact test (two-sided). All HBond score groups from 9.8 to 10.8 until 17.8 to 18.8 showed significantly less negative ΔHBS than expected (p-value < 0.05). The remaining HBond score group of 18.8 to 19.8, however showed no significant shifts from the expected ΔHBS frequency. Stronger 5’ss seem to allow HBond score reduction to a higher extent, because even after moderate HBond score reduction, they can still ensure sufficient “correctly” spliced transcripts and thus functional protein.

To check, whether there is an additional correlation between ΔHBS and ΔSSHW ($SSHW_{alt}$ – $SSHW_{ref}$), we grouped 5’ss in groups of ΔHBS being (i) lower -2, (ii) between -2 and 2, or (iii) greater 2. ΔHBS lower -2 seemed to be slightly compensated by a simultaneous increase of SSHW, whereas a ΔHBS greater 2 seemed to be slightly compensated by a simultaneous decrease of SSHW (Fig. 3).

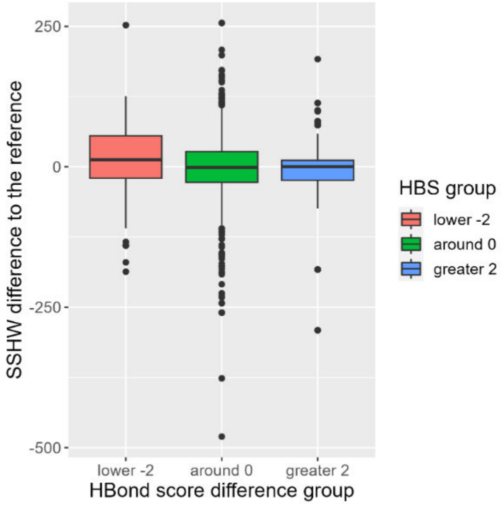


Figure 3. SSHW differences simultaneous to HBond score differences. 5’ss with ΔHBS lower -2 showed a significantly higher simultaneous change in SSHW, than 5’ss with ΔHBS greater 2, which showed the tendency for reduced SSHW of the respective 5’ss.

Next, we analyzed whether these stronger HBond score differences could be found in transcripts that are described to be important factors in breast cancer development. We expected no drastic changes in 5’ss strength for these transcripts, since the haplotyped assemblies came from healthy individuals. Indeed, most 5’ss were either unaffected or only slightly affected in their HBond scores by sequence variations with strongest ΔHBS at -1.1.

However, the 5’ss at exon 9 of the ERI2 gene showed an HBond score reduction from 10 to 7 for one individual, which is expected to very likely reduce usage of this particular 5’ss. Investigating the consequences of reduced 5’ss HBS and thus retention of the respective downstream intron, we found a stop codon around 118 nucleotides into the intronic sequence and the classic poly-A signal motif AATAAA downstream of it. This could indicate the existence of an alternative transcript isoform with lacks the last two exons, already ending at exon 9 with a small alternative amino acid sequence at the end, due to usage of this potential poly-A site. Since this alternative transcript isoform would still contain the encoded ERI1-like Exonuclease at the C-terminus, one could assume that it might still function as the original, thus making a strong increase in its expression not pathogenic. While the resulting transcript could not be found in the *ensembl* reference genome, we could find this predicted shorter transcript isoform in the US-pendant RefSeq.

3.1.2 Heterozygous variations 5’ss sequence variations

The number of 5’ss with heterozygous variations, however, was understandably higher than the number of homozygous variations. From 311,433 5’ss annotated in *ensembl* (version 105), 16,058 (5.2%) showed a heterozygous sequence variation across the 23 individuals. 7,776 (48,4%) of the 16,058 5’ss showed a homozygous variation in multiple individuals (from 2 to 22), which additionally indicates that the occurrence of these variations might also not be random, but predominantly within the same transcripts and 5’ss. 18.8% of those showed the same HBond score difference across all affected individuals, indicating inheritance rather than random point mutations as the origin.

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 six times larger than the number of 5’ss affected by homozygous variations.

	not TSL1	TSL1
not protein coding	6,083 (7.2% of 83,910)	1,088 (9.5% of 11,404)
protein coding	2,411 (5.6% of 42,920)	6,476 (3.7% of 173,199)

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. 5'ss could either be 1) exclusively part of non-coding transcript isoforms ("not protein coding"), or 2) part of protein coding transcripts ("protein coding"). Additionally, 5'ss were classified whether they were found in 1) transcripts of TSL level lower 1 ("not TSL1"), or 2) TSL1 transcripts ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts, that are found with the all same exon-junctions in the ensembl-pendant RefSeq.

As observed for homozygous HBond score differences, 5'ss belonging to protein coding TSL1 transcript isoforms showed the smallest number of affected 5'ss, relative to the total number of annotated 5'ss in each category (Table 2). We again checked the observed HBond score differences per category, splitting the 5'ss into a large set, where one allele is still the same as the reference (16,024 unique 5'ss), and one much smaller set, where both alleles were different from the reference, but not the same (439 unique 5'ss). We then compared the observed HBond score differences of these two datasets (Fig. 4) to the distribution of HBond score differences found in homozygous sequence variations (Fig. 1A).

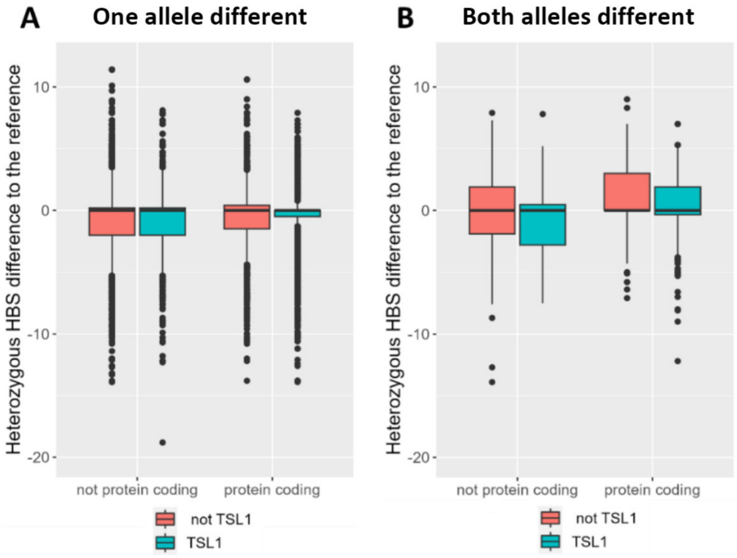


Figure 4. Differences in 5'ss strength to the reference 5'ss of 5'ss that are affected by heterozygous sequence variations within the 11 nucleotides of the 5'ss sequence by TSL and "protein coding" category. Boxplots depicting the HBond score differences to the reference 5'ss per category, where either only one allele (**A**) or both alleles (**B**) are different to the reference 5'ss sequence. 5'ss were grouped into those, that were 1) exclusively part of non-coding transcript isoforms ("not protein coding"), or 2) part of protein coding transcripts ("protein coding") and those, either found in 1) transcripts of TSL level lower 1 ("not TSL1"), or 2) TSL1 transcripts ("TSL1"), with TSL1 transcripts representing high-confident annotated transcripts, that are found with the all same exon-junctions in the ensembl-pendant RefSeq.

5’ss with only one allele different from the reference (Fig. 4A) showed a shift towards weaker HBond scores compared to homozygous situations, although the median ΔHBond still lies at around zero in the heterozygous data. This trend to an HBond score reduction was a little less expressed for protein coding transcripts. The general shift to stronger HBond score reductions than in homozygous variations might potentially be due to the presence of one remaining reference allele, which might compensate for the variation. 5’ss with heterozygous 5’ss differences in both alleles (Fig. 4B) showed a HBond score distribution for 5’ss of non-coding transcripts that was similar to the homozygous variations, whereas 5’ss of protein coding transcripts had a tendency for HBond score increases, although the median was also still at 0.

Next, we again focused in more detail on 5’ss of protein coding TSL1 transcripts with similar HBond score difference distribution compared to homozygous variations (Fig. 5). Single-allele changes again seemed to rather preserve the exact HBond score as the reference, with 52% showing an H-bond score difference of 0, compared to 49% for homozygous variations.

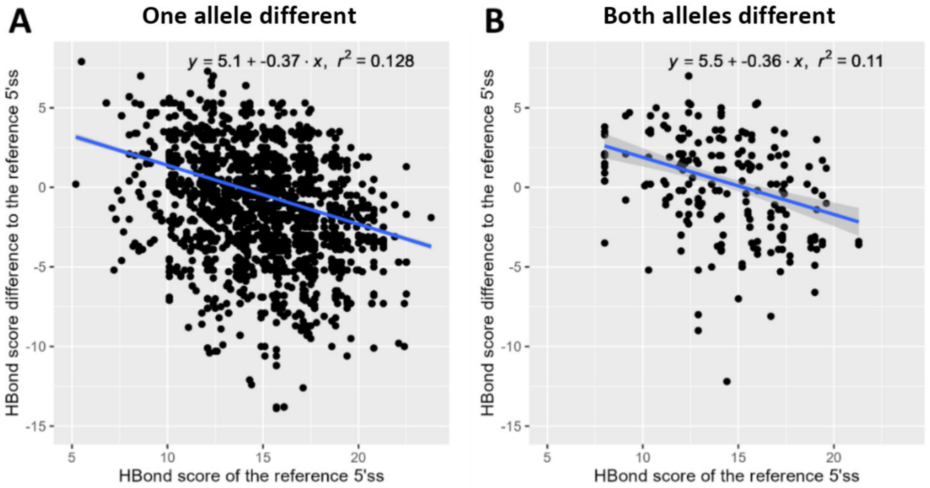


Figure 5. Observed HBond score difference range depends on HBond score of the reference 5’ss. Depicted is the HBond score difference, caused by heterozygous genetic variations from the reference, within the sequence of annotated 5’ss of either one (**A**) or both alleles (**B**). 5’ss that were much 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 HBond score compared to the reference. Each regression line is depicted blue, with its formula stated above. The 95% confidence interval of the regression line is marked by a dark grey 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.

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. Since the SSHW previously showed to be less distinctive for silent or used 5’ss, we next measured SSHW differences to the reference 5’ss, expecting a much greater standard deviation around 0 for SSHW variations of the same 5’ss.

3.1.2 Variations in proximity of annotated 5’ss sequences

Sequence variations were also found in the neighboring sequence segment of ±60 nucleotides around annotated 5’ss. With 127,295 (40.9%) 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 131 nt long sequence (2x60nt + 11nt) with an 11 nt long sequence (Tab. 3).

	Not TSL1	TSL1
Not protein coding	33,206 (39.6% of 83,910)	5,449 (47.8% of 11,404)
protein coding	19,363 (45.1% of 42,920)	69,277 (40.0% of 173,199)

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. 5’ss could either be 1) never part of coding transcripts (“never protein coding”), 2) only part of protein coding transcripts (“only protein coding”) or 3) both part of coding and noncoding transcripts (“partly protein coding”). Additionally, 5’ss were classified into either 1) never being part of TSL1 transcripts, 2) only part of TSL1 transcripts, or 3) both part of TSL1 and not-TSL1 transcripts, with TSL1 representing high-confident annotated transcripts, that are found with the exact same exon-junctions in the ensembl-pendant RefSeq.

With 40% we again saw the smallest fraction of occurrences within 5’ss of the categories protein coding & TSL1 or non-coding and not TLS1, whereas 5’ss of the other categories showed a slightly higher fraction of variations. Overall, variations in the ±60nt window around splice sites affect significantly more unique 5’ss than variations of the splice site sequence itself. This might be due (1) to the larger sequence of interest, which increases the probability to harbor sequence variations and (2) 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. 6A). To analyze variability in 5’ss SSHW, we calculated the SSHW difference (ΔSSHW) by subtracting the SSHW of an affected 5’ss from the

SSHW of the reference 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 approximately 0 (Fig. 6B). 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 ($p_{total} < 0.05$, $p_{unique} < 0.05$). More extreme SSHW differences seem to be overrepresented.

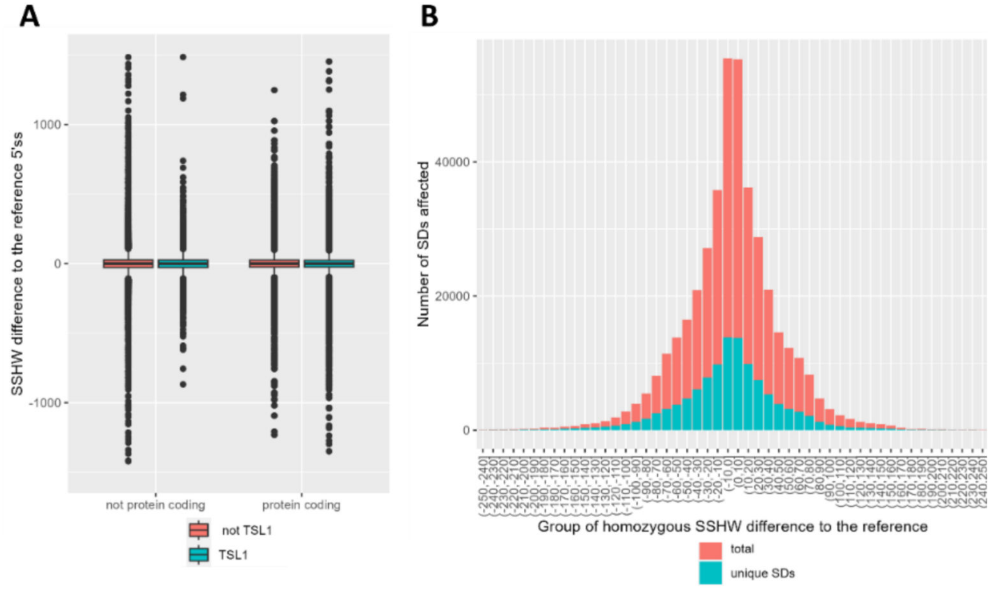


Figure 6. SSHW differences to the reference 5’ss via sequence variations within the +-60 nucleotide window around the 5’ss sequence grouped by TSL and “protein coding” category. A) Boxplot depicting the SSHW differences to the reference 5’ss per category. 5’ss were grouped into those, that were part 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.

Defining extreme 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 69,277 5’ss only found within coding TSL1 transcripts that had variations in the ±60 nt neighborhood, we generated a set of 19,113 5’ss (21%) with extreme ΔSSHW. We subsequently analyzed whether more extreme 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 associated extreme SSHW differences, we binned ΔSSHW into 10 equally large groups (deciles), with lowest ΔSSHW starting in group 1 and ΔSSHW greater 65 starting in group 6 (Fig. 7A).

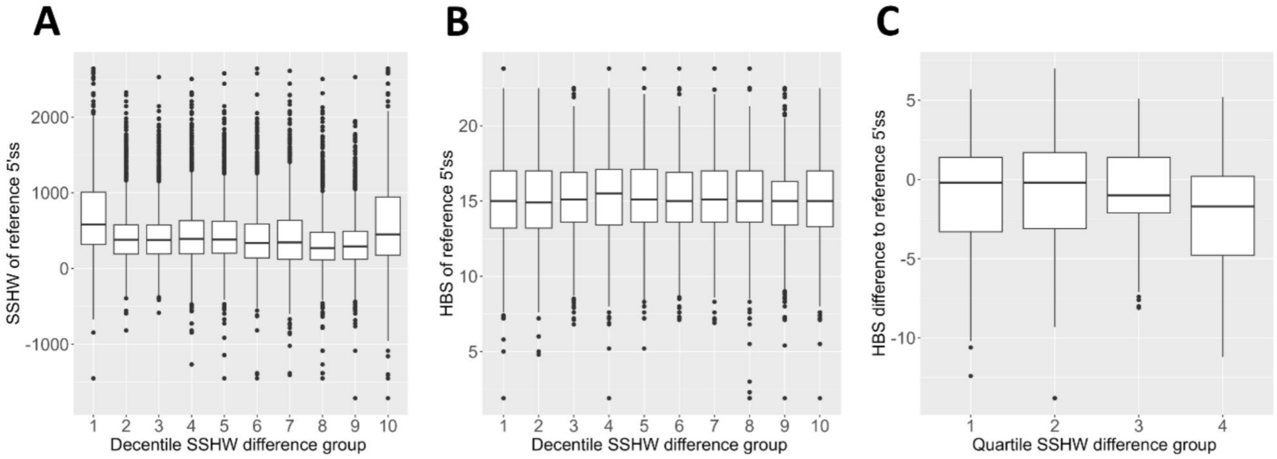


Figure 7. Strong SSHW differences to the 5’ss reference and corresponding factors. A) Boxplot depicting the reference SSHW of 5’ss affected by SSHW differences. SSHW differences were grouped in 10 equally sized decile groups starting from strong negative SSHW differences at decile group 1 to strong positive SSHW differences at decile group 10, 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. **B)** Boxplot depicting the reference HBond-score of 5’ss affected by SSHW differences. SSHW differences were grouped in 10 equally sized decile groups starting from strong negative SSHW differences at decile group 1 to strong positive SSHW differences at decile group 10, showing no clear correlation between these metrics. **C)** Boxplot depicting the HBond-score difference of 5’ss affected by SSHW differences of 5’ss, whose 5’ss sequence was directly affected by nucleotide exchanges. SSHW differences were grouped in 4 equally sized decile groups, due to the very low case numbers (1,816). Starting from strong negative SSHW differences at decile group 1 to strong positive SSHW differences at decile group 4, a tendency of strong negative HBond-score differences can be detected for strong positive SSHW differences, whereas 5’ss with strong negative SSHW differences predominantly showed no change in the HBond-score at all.

Similar to the negative linear correlation between HBond score differences and reference HBond score, we saw that negative ΔSSHW values from -1,500 to -65 (belonging to decile group 1 - 5) seemed to occur more often in 5’ss with a slightly stronger reference SSHW baseline (reference SSHW average around 400), than ΔSSHW greater -60 (reference SSHW average around 300). This observation was even more pronounced analyzing only homozygous SSHW differences (data not shown). Analyzing the allele-specific 5’ss strength per ΔSSHW decile group revealed no observable correlation (Fig. 7B).

Looking at a very small specific group of 5’ss that showed nucleotide exchanges within the ±60 nt neighborhood as well as within the 5’ss sequence, we saw a 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. 7C). It has to be emphasized, that this particular dataset, due to its strict selection, however, only consisted of 1,816 cases, affecting an even lower number of only 764 unique 5'ss.

Finally, we also analysed the breast cancer gene set again, to see, which SSHW changes still seem to be tolerated in essential genes. Since the impact of the SSHW on 5'ss usage is determined by the HBond score of the 5'ss sequence, we analysed the SSHW upon sequence variation (Fig. 8).

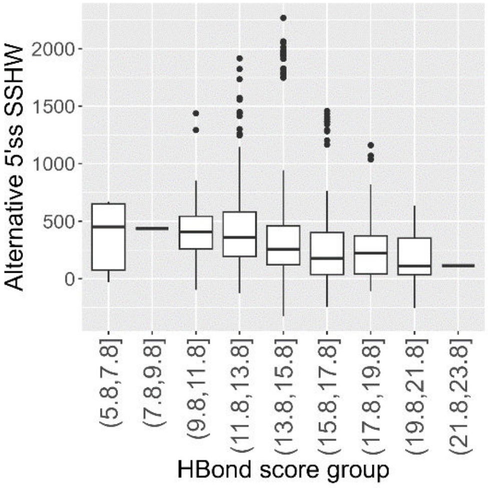


Figure 8. Alternative SSHW per HBond Score group of 5'ss important during breast cancer diagnostics. Depicted are the SSHW distributions per HBond score group for a set of 5'ss, frequently analysed during risk assessment for breast cancer. In 2,479 instances, genomes showed a sequence variation within the genomic surrounding of these 5'ss.

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. Additionally, we observed, that there were no 5'ss (except one) with an HBS score below 12 that showed an SSHW below 0, with a value of 0 indicating equal predicted effects of hnRNPs and SR protein binding. This might indicate that a decrease in SSHW of essential 5'ss should be considered during genetic screening in the future to improve risk assessment for genetic diseases.

4. 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 splicing regulatory proteins, within a curated subset of high-fidelity genomes derived from 26 participants of the 1000 Genome Project [15]. First, we aligned the assembled genomes with the widely recognized Genome Reference Consortium Human Build 38 (GRCh38). Then a custom R-script incorporated sequence variations within a 60-nucleotide window around annotated 5'ss into the reference sequence, allowing us to quantify changes in splice site strength and SSHW due to 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 Transcript Support Level (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 0.8% of annotated 5'ss sites, which reflects previous observations that splice site sequences seemed to be conserved sometimes even across species [23]. Crucially, over half of these sites exhibited variations in multiple individuals, with many 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 5.2% of annotated 5'ss sites. The evolutionary pressure governing the acceptance of heterozygous mutations within the splice sites sequence therefore appeared to 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 half did not result in predicted 5’ss strength. Since for some proteins, a reduced expression is not per se pathogenic, we would have expected this percentage to be significantly lower for heterozygous variations, since here one allele would still remain intact [24].

Grouping 5’ss into categories based on coding potential (protein coding vs. non-protein coding) and transcript support level (TSL1 vs. non-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 the HBond score of the reference 5'ss and the detected HBond score difference due to homozygous sequence variations, we saw that 5’ss with HBS lower 18.8 showed a clear prevalence for sequence variations to increase 5’ss strength, rather than decreasing it, compared with what would be randomly expected. 5’ss of HBond score >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. Similar to the reference H-Bond score, we would have expected the HBond score differences and simultaneous SSHW difference of affected 5'ss to correlate. Indeed, the distribution of SSHW differences was significantly higher for 5’ss with HBond score differences lower -2 than for 5’ss with HBond score differences greater +2. The arbitrary level of ±2 was selected based on the distribution of HBond score differences.

Widening the window around 5’ss, we also analyzed sequence variations, within the ±60nt window around splice sites, excluding the splice site sequence itself. This somewhat arbitrary sequence neighbourhood was repeatedly described to hold an extensive amount of predicted splicing regulatory elements [25, 26]. Sequence variations within these regions were quite frequent, with 40.9% of annotated 5'ss sites affected. Like with HBond score differences, SSHW reductions by sequence variations were slightly enriched in 5’ss with a higher SSHW baseline around 400. SSHW increases on the other hand seemed to correlate with simultaneous HBond score differences, showing almost exclusively 5’ss with HBond score reductions.

Applying these observations on 5’ss of genes, associated with breast cancer development, implied a maximal reduction in 5’ss strength of -1.1 HBS or – SSHW, with 5’ss of HBS < 12 not falling below 0 SSHW. These threshold values could potentially help assessing genetic disease risk and underscore the need to consider SSHW alterations alongside 5’ss strength during genetic screening.

5. Supplements

Supplementary Table 1. Table of genes and associated ensembl transcript IDs, analyzed during genetic screening for breast cancer risk.

Gene name	Transcript ID
ABCA13	ENST00000435803
ACP6	ENST00000369238
ATM	ENST00000278616
BARD1	ENST00000260947
BPIFB4	ENST00000375483
BRCA1	ENST00000357654
BRCA2	ENST00000544455
BRIP1	ENST00000259008
CD1E	ENST00000368160
CDH1	ENST00000261769
CHEK2	ENST00000328354
COL6A2	ENST00000300527
DNMT1	ENST00000359526
EDNRB	ENST00000334286
EPCAM	ENST00000263735
ERCC2	ENST00000391945
ERI2	ENST00000300005
EVC	ENST00000264956
FAM175A	ENST00000321945
FANCA	ENST00000389301
FANCC	ENST00000289081
FANCM	ENST00000267430
FERMT1	ENST00000217289
GABRG3	ENST00000333743
GPRC5A	ENST00000014914
HFE	ENST00000357618
IDS	ENST00000340855
L2HGDH	ENST00000421284
MAP3K1	ENST00000399503
MKI67	ENST00000368653
MLH1	ENST00000231790
MPDZ	ENST00000319217
MRE11A	ENST00000323929
MSH2	ENST00000233146
MSH6	ENST00000234420
MUTYH	ENST00000450313
NBN	ENST00000265433
NF1	ENST00000358273
PALB2	ENST00000261584
PI4KA	ENST00000255882
PIK3CA	ENST00000263967
PMS2	ENST00000265849
PPM1D	ENST00000305921
PTEN	ENST00000371953
RAD50	ENST00000265335
RAD51C	ENST00000337432
RECQL	ENST00000444129
RINT1	ENST00000257700
RPAP1	ENST00000304330
SLC12A6	ENST00000458406
SLX4	ENST00000294008
STK11	ENST00000326873
TP53	ENST00000269305
XRCC2	ENST00000359321

References

1. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proc Natl Acad Sci U S A, 1977. **74**(8): p. 3171-5.
2. Khodor, Y.L., et al., *Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila.* Genes Dev, 2011. **25**(23): p. 2502-12.
3. Aebi, M., et al., *Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA.* Cell, 1986. **47**(4): p. 555-65.
4. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.
5. Freund, M., et al., *A novel approach to describe a U1 snRNA binding site.* Nucleic Acids Res, 2003. **31**(23): p. 6963-75.
6. Wang, Z. and C.B. Burge, *Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.* RNA, 2008. **14**(5): p. 802-13.
7. Erkelenz, S., et al., *Genomic HEXploring allows landscaping of novel potential splicing regulatory elements.* Nucleic Acids Res, 2014. **42**(16): p. 10681-97.
8. Brillen, A.L., et al., *Succession of splicing regulatory elements determines cryptic 5ss functionality.* Nucleic Acids Res, 2017. **45**(7): p. 4202-4216.
9. More, D.A. and A. Kumar, *SRSF3: Newly discovered functions and roles in human health and diseases.* Eur J Cell Biol, 2020. **99**(6): p. 151099.
10. Geuens, T., D. Bouhy, and V. Timmerman, *The hnRNP family: insights into their role in health and disease.* Hum Genet, 2016. **135**(8): p. 851-67.
11. Grzybowska, E.A., *Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing.* Biochem Biophys Res Commun, 2012. **424**(1): p. 1-6.
12. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
13. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
14. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
15. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.
16. Porubsky, D., et al., *Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads.* Nat Biotechnol, 2021. **39**(3): p. 302-308.
17. Ebert, P., et al., *Haplotype-resolved diverse human genomes and integrated analysis of structural variation.* Science, 2021. **372**(6537).
18. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs.* Nat Biotechnol, 2019. **37**(5): p. 540-546.
19. Lin, H.N. and W.L. Hsu, *GSAlign: an efficient sequence alignment tool for intra-species genomes.* BMC Genomics, 2020. **21**(1): p. 182.
20. Ptok, J., S. Theiss, and H. Schaal, *VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV.* Cancer Inform, 2020. **19**: p. 1176935120976399.

21. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
22. Muller, L., et al., *Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements.* Nucleic Acids Res, 2022. **50**(15): p. 8834-8851.
23. Abril, J.F., R. Castelo, and R. Guigo, *Comparison of splice sites in mammals and chicken.* Genome Res, 2005. **15**(1): p. 111-9.
24. Bartha, I., et al., *Human gene essentiality.* Nat Rev Genet, 2018. **19**(1): p. 51-62.
25. Ke, S., et al., *Quantitative evaluation of all hexamers as exonic splicing elements.* Genome Res, 2011. **21**(8): p. 1360-74.
26. Zhang, X.H. and L.A. Chasin, *Computational definition of sequence motifs governing constitutive exon splicing.* Genes Dev, 2004. **18**(11): p. 1241-50.

2.2.2 Publication III: VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV.

Single nucleotide variations (SNVs) within sequences that regulate splicing can result in pathological splicing in the context of various diseases, like cancer. Reporting of SNVs, however, follows the Sequence Variant Nomenclature (http://varnomen.hgvs.org/), sometimes using ambiguous numbering schemes that refer to specific annotated transcript sequences. Calculating the actual genomic position of a given SNV, especially in older literature, can be complicated. However, regarding the impact on of an SNV on nearby splice sites or splicing regulatory elements, the exact genomic position is essential. In this work, the VarCon algorithm was developed, that combines the information of the Ensembl human reference genome and the corresponding transcript table for accurate retrieval of the SNV genomic coordinate. VarCon also shows splice site strengths (scored by HBond and MaxEnt scores) and HEXplorer profiles of an SNV neighborhood, reflecting position-dependent splice enhancing and silencing properties.

Article published in Cancer Informatics. 2020 Nov 24:19:1176935120976399 (doi: 10.1177/1176935120976399), by

Johannes Ptok, Stephan Theiss, Heiner Schaal

Contributions

JP developed the tool. JP, ST and HS wrote the manuscript. Individual contribution of JP at around 90%.

Check for updates

VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV

Cancer Informatics
Volume 19: 1–3
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935120976399
SAGE

Johannes Ptok, Stephan Theiss and Heiner Schaal

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

ABSTRACT: Reporting of a single nucleotide variant (SNV) follows the Sequence Variant Nomenclature (http://varnomen.hgvs.org/), using an unambiguous numbering scheme specific for coding and noncoding DNA. However, the corresponding sequence neighborhood of a given SNV, which is required to assess its impact on splicing regulation, is not easily accessible from this nomenclature. Providing fast and easy access to this neighborhood just from a given SNV reference, the novel tool VarCon combines information of the Ensembl human reference genome and the corresponding transcript table for accurate retrieval. VarCon also displays splice site scores (HBond and MaxEnt scores) and HEXplorer profiles of an SNV neighborhood, reflecting position-dependent splice enhancing and silencing properties.

KEYWORDS: SNPs, alternative splicing, R package, sequence retrieval, HEXplorer score, HBond score

RECEIVED: September 1, 2020. **ACCEPTED:** November 1, 2020.

TYPE: Technical Advances

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Forschungskommission of the Medical Faculty, Heinrich Heine Universität Düsseldorf (2020-12) to H.S.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHORS: Stephan Theiss, Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany.
Email: theiss@uni-duesseldorf.de

Heiner Schaal, Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany.
Email: schaal@uni-duesseldorf.de

Introduction

Comparing genomic DNA sequences of individuals of the same species reveals positions where single nucleotide variations (SNVs) occur. When localized within the coding sequence of a gene, SNVs can, among others, affect which amino acids are encoded by the altered codon, potentially leading to disease. Approximately 88% of human SNVs associated with disease are, however, not located within the coding sequence of genes, but within intronic and intergenic sequence segments.[1] Nevertheless, annotations referring to the coding sequence of a specific transcript are still widely used, for example, c.8754+3G>C (BRCA2 and Ensembl transcript ID ENST00000544455), referring to the third intronic nucleotide downstream of the splice donor (SD) at the position of the 8754th coding nucleotide. Based on its position information referring to the coding sequence (c.) or alternatively to the genomic (g.) position (eg, g.1256234A>G), our tool VarCon retrieves an adjustable SNV sequence neighborhood from the reference genome. To visualize possible effects of SNVs on splice sites or splicing regulatory elements, which play an increasing role in cancer diagnostics and therapy,[2] VarCon additionally calculates HBond scores[3] of SDs and MaxEnt scores[4] of splice acceptor (SA) sites and HEXplorer scores of the retrieved sequences[9].

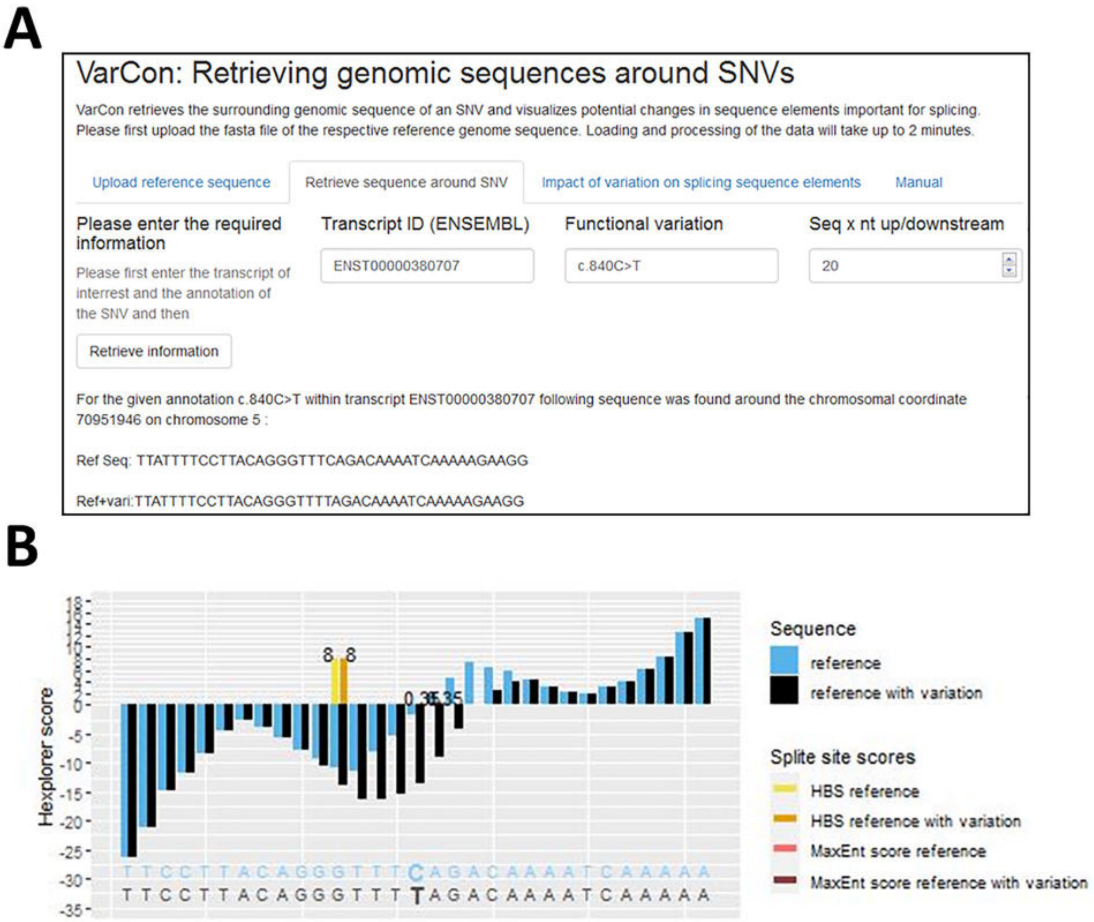
Implementation

VarCon is an R package which can be executed from Windows, Linux, or Mac OS. It executes a Perl script located in its directory and therefore relies on prior installation of some version of Perl (eg, Strawberry Perl). In addition, the human reference genome must be downloaded as fasta file (or zipped fasta.gz) with Ensembl chromosome names (“1” for chromosome 1) and subsequently uploaded into the R working environment, using the function “prepareReferenceFasta” to generate a large DNAStringset (file format of the R package Biostrings). To translate SNV positional information, referring to the coding sequence of a transcript, a transcript table has to be additionally uploaded to the working enviroment. The transcript table has to contain exon and coding sequence coordinates of every transcript from Ensembl. Two zipped transcript table csv-files which either refer to the genome assembly GRCh37 or GRCh38 can be downloaded from https://github.com/caggtaagtat/VarConTables.

As the transcript table with the GRCh38 genomic coordinates (currently from Ensembl version 100) will be updated with further releases, a new transcript table can be downloaded using the Ensembl Biomart interface. Any newly generated transcript table, however, must contain the same columns and column names as described in the documentation of the current transcript tables for correct integration. As, for instance, in cancer research the transcript which is used to refer to genomic positions of SNVs is often the same, a gene-to-transcript conversion table can be used for synonymous usage of certain gene names (or gene IDs) and transcript IDs (Ensembl ID). VarCon deliberately does not rely on Biomart queries using the Biomart R package, as these might be blocked by firewalls.

Due to its structure, the VarCon package can accept any genome and transcript table combination which is available on Ensembl and thus additionally permits usage for any other organism represented in the Ensembl database.[5] The combination of already existing tools like Mutalyzer,[6] SeqTailor,[7] or ensembldb[8] can lead to similar results during the variation conversion and DNA sequence extraction. However, VarCon holds additional benefits, namely, its straightforward usage even on a large-throughput scale, its independence due to the direct data entry, and its instant graphical representation of splicing regulatory elements and intrinsic splice site strength.

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).



VarCon: Retrieving genomic sequences around SNVs

VarCon retrieves the surrounding genomic sequence of an SNV and visualizes potential changes in sequence elements important for splicing.
Please first upload the fasta file of the respective reference genome sequence. Loading and processing of the data will take up to 2 minutes.

Upload reference sequence | Retrieve sequence around SNV | Impact of variation on splicing sequence elements | Manual

Please enter the required information

Please first enter the transcript of interrest and the annotation of the SNV and then

Retrieve information

Transcript ID (ENSEMBL): ENST00000380707

Functional variation: c.840C>T

Seq x nt up/downstream: 20

For the given annotation c.840C>T within transcript ENST00000380707 following sequence was found around the chromosomal coordinate 70951946 on chromosome 5 :

Ref Seq: TTATTTTCCTTACAGGGTTTCAGACAAAATCAAAAAGAAGG

Ref+vari:TTATTTTCCTTACAGGGTTTTAGACAAAATCAAAAAGAAGG

Hexplorer score
18
16
14
12
10
8
6
4
2
0
-5
-10
-15
-20
-25
-30
-35
8
8
0.35
0.35
T T C C T T A C A G G G T T T C A G A C A A A A T C A A A A A
T T C C T T A C A G G G T T T T A G A C A A A A T C A A A A A
Sequence
reference
reference with variation
Splite site scores
HBS reference
HBS reference with variation
MaxEnt score reference
MaxEnt score reference with variation

Figure 1. (A) Exemplary screenshot of VarCon GUI, querying the SNV c.840C>T in gene *SMN1* (transcript ENST00000380707). (B) HEXplorer plot of the sequence neighborhood of the same SNV. Bar plot depicting the HZ_{EI}-score for each nucleotide of the reference sequence in a ±20 nt neighborhood around the position of the variation with (black) or without (blue) the c.840C>T variation. HBond scores of donor sequences within the reference sequence are shown in yellow. HBond scores of donor sequences within the reference sequence with the variation are colored orange. GUI indicates graphical user interface; SNV, single nucleotide variant.

After upload of the human reference genome, selection of the appropriate transcript table and a potential gene-to-transcript conversion table, a transcript ID (or gene name) and an SNV (whose positional information either refers to the coding ["c."] or genomic ["g."] sequence) are requested during the execution of the main function of the package. VarCon then uses the information of the transcripts' exon coordinates to translate the SNV positional information to a genomic coordinate, if needed. Then the genomic sequence around the SNV position is retrieved from the reference genome in the direction of the open reading frame and committed to further analysis, both with and without the SNV.

For analysis of an SNV impact on splicing regulatory elements, VarCon calculates the HZ_{EI} score profile of reference and SNV sequences from the HEXplorer algorithm[9] and visualizes both in a bar plot. The HEXplorer score assesses splicing regulatory properties of genomic sequences, their capacity to recruit splicing regulatory proteins to the pre-mRNA transcript. Highly positive (negative) HZ_{EI} scores indicate sequence segments, which enhance (repress) usage of both downstream 5' splice sites and upstream 3' splice sites.

In addition, intrinsic strengths of SD and SA sites are visualized within the HZ_{EI} score plot. Splice donor strength is calculated by the HBond score, based on hydrogen bonds formed between a potential SD sequence and all 11 nucleotides of the free 5′ end of the U1 snRNA. Splice acceptor strength is calculated by the MaxEnt score, which is essentially based on the observed distribution of SA sequences within the reference genome, while also taking into account dependencies between both non-neighboring and neighboring nucleotide positions.[4]

VarCon can either be executed using integrated R package functions according to the manual on github or with a GUI (graphical user interface) application based on R package shiny with the integrated function "startVarConApp".

Example

The sequence variation c.840C>T within the seventh exon of the *SMN2* gene (Ensembl transcript ID: ENST00000380707) is associated with spinal muscular atrophy. Previous studies have shown that this sequence variation results in a change in splicing regulatory protein binding, increasing skipping of exon 7. Entering this variation and the transcript ID into VarCon (Figure 1A) leads to the following bar plot visualizing this effect with a delta HZ_{EI} of –71.76 (Figure 1B).

Acknowledgements

We would like to thank Gene Yeo for his kind approval to integrate the MaxEnt scoring algorithm into VarCon.

Author Contributions

JP developed the R code of the VarCon package and drafted the manuscript. ST and HS supervised the project and also wrote the manuscript.

Availability

VarCon is available at https://github.com/caggtaagtat/VarCon and released under the MIT License. After installation of the package, an attached shiny app can be started with the integrated function "startVarConApp".

ORCID iDs

Johannes Ptok https://orcid.org/0000-0002-0322-5649
Heiner Schaal https://orcid.org/0000-0002-1636-4365

REFERENCES

1. 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 U S A*. 2009;106:9362-9367.
2. Dong X, Chen R. Understanding aberrant RNA splicing to facilitate cancer diagnosis and therapy. *Oncogene*. 2020;39:2231-2242.
3. Freund M, Asang C, Kammler S, et al. A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res*. 2003;31:6963-6975.
4. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11:377-394.
5. Birney E, Andrews TD, Bevan P, et al. An overview of Ensembl. *Genome Res*. 2004;14:925-928.
6. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat*. 2008;29:6-13.
7. Zhang P, Boisson B, Stenson PD, et al. SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data. *Nucleic Acids Res*. 2019;47:W623-W631.
8. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics*. 2019;35:3151-3153.
9. Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. *Nucleic Acids Res*. 2014;42:10681-10697.

2.2.3 Publication IV: High Concentration of Low-Density Lipoprotein Results in Disturbances in Mitochondrial Transcription and Functionality in Endothelial Cells.

Using alternative splice sites of a given RNA transcript enables a level of protein expression regulation, either by altering regulatory sequence elements, such as the 3’ or 5’ untranslated region, or by altering the resulting protein coding sequence [137-139]. This dynamic regulation of gene expression can change the cellular response to extern stimuli among others and plays therefore an important role in many cellular processes, such as the response to cellular stress [140]. Protein kinases and phosphatases, like serine/arginine phosphatase PP1, that are activated by different types of stress, change the phosphorylation status of SR proteins, affecting their activity and therefore usage of splice sites, whose recognition relies on binding of these proteins to the RNA [40, 43]. Upon stress, some hnRNP and SR proteins seem to localize to the cytoplasm accumulating in discrete phase-dense particles, the cytoplasmic stress granules (SGs) [141, 142]. These non-membrane bound assemblies of proteins and RNA seem to be partly driven by condensation of untranslated RNA transcripts, that were stalled during translation initiation [143]. Since around half of proteins found in these granules are RNA binding proteins, splicing regulatory proteins might be captured within these cytoplasmic bodies to an extent, that impacts their function in the nucleus [144].

In this work, cardiovascular endothelial cells were studied, which are the first barrier between potentially stress-inducing components in the blood stream and other tissues. Treatment with media, simulating a high-fat diet, resulted in significant changes in their transcriptional profile, compared to the control [145]. Among fat catabolism genes, that were expectedly up-regulated, we additionally measured a significant change in splice site usage of the NOS3 gene (or eNOS), which plays an essential role in endothelial stress-induced senescence. Although the overall NOS3 gene expression, did not show significant changes upon treatment with high-fat media on the RNA level, C-terminal splice junctions were significantly upregulated. These specific exon-exon junctions, originated from an alternative transcript isoform, which lacked large regions of the full-length NOS3 transcript, explaining the reduced signal in the immune-blot assay. This dysregulation of NOS3 expression led to an increase in oxidative stress in the cells and a corresponding aberration of ATP-synthesis, due to changes in the expression of mitochondrial expressed genes, that together resulted in increased apoptosis and reduced migratory capacity.

The following article was published in Oxid Med Cell Longev. 2019 Jun 10:2019:7976382 (doi: 10.1155/2019/7976382), by

Stefanie Gonnissen*, **Johannes Ptok***, Christine Goy, Kirsten Jander, Philipp Jakobs, Olaf Eckermann, Wolfgang Kaisers, Florian von Ameln, Jörg Timm, Niloofar Ale-Agha, Judith Haendeler, Heiner Schaal, Joachim Altschmied (* shared first-authors)

Contribution

SG, CG, KJ, PJ, OE and FvA did the migration assays, RT-PCRs and immunoblot. JP, WK, HS did the bioinformatic analysis of the RNA sequencing data (DGE). JH, JA, HS, SG and JP wrote the manuscript. Individual contribution of JP at around 40%.

Hindawi
Oxidative Medicine and Cellular Longevity
Volume 2019, Article ID 7976382, 12 pages
https://doi.org/10.1155/2019/7976382

Research Article

High Concentration of Low-Density Lipoprotein Results in Disturbances in Mitochondrial Transcription and Functionality in Endothelial Cells

Stefanie Gonnissen,[1] Johannes Ptok,[2] Christine Goy,[1] Kirsten Jander,[3] Philipp Jakobs,[1] Olaf Eckermann,[3] Wolfgang Kaisers,[4] Florian von Ameln,[1,3] Jörg Timm,[2] Niloofar Ale-Agha,[1] Judith Haendeler,[1,5] Heiner Schaal,[2] and Joachim Altschmied[3]

[1] *Heisenberg Group-Environmentally-Induced Cardiovascular Degeneration, IUF-Leibniz Research Institute for Environmental Medicine, 40225 Düsseldorf, Germany*
[2] *Institute of Virology, Medical Faculty, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany*
[3] *Core Unit Biosafety Level 2 Laboratory, IUF-Leibniz Research Institute for Environmental Medicine, 40225 Düsseldorf, Germany*
[4] *Department of Anaesthesiology, HELIOS University Hospital Wuppertal, University of Witten/Herdecke, 42283 Wuppertal, Germany*
[5] *Heisenberg Group-Environmentally-Induced Cardiovascular Degeneration, Central Institute of Clinical Chemistry and Laboratory Medicine, Medical Faculty, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany*

Correspondence should be addressed to Judith Haendeler; juhae001@hhu.de

Received 24 January 2019; Revised 30 April 2019; Accepted 14 May 2019; Published 10 June 2019

Guest Editor: Valentina Giorgio

Copyright © 2019 Stefanie Gonnissen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Concentrations of low-density lipoprotein (LDL) above 0.8 mg/ml have been associated with increased risk for cardiovascular diseases and impaired endothelial functionality. Here, we demonstrate that high concentrations of LDL (1 mg/ml) decreased NOS3 protein and RNA levels in primary human endothelial cells. In addition, RNA sequencing data, in particular splice site usage analysis, showed a shift in NOS3 exon-exon junction reads towards those specifically assigned to nonfunctional transcript isoforms further diminishing the functional NOS3 levels. The reduction in NOS3 was accompanied by decreased migratory capacity, which depends on intact mitochondria and ATP formation. In line with these findings, we also observed a reduced ATP content. While mitochondrial mass was unaffected by high LDL, we found an increase in mitochondrial DNA copy number and mitochondrial RNA transcripts but decreased expression of nuclear genes coding for respiratory chain proteins. Therefore, high LDL treatment most likely results in an imbalance between respiratory chain complex proteins encoded in the mitochondria and in the nucleus resulting in impaired respiratory chain function explaining the reduction in ATP content. In conclusion, high LDL treatment leads to a decrease in active NOS3 and dysregulation of mitochondrial transcription, which is entailed by reduced ATP content and migratory capacity and thus, impairment of endothelial cell functionality.

1. Introduction

Diet plays a crucial role in the development and prevention of cardiovascular diseases. A diet high in saturated fat increases the risk of heart disease and stroke. It is estimated to cause about 31% of coronary heart disease and 11% of stroke worldwide. Cholesterol is carried through our blood by particles called lipoproteins: high-density lipoprotein (HDL) and low-density lipoprotein (LDL). HDL cholesterol reduces the risk of cardiovascular disease as it carries cholesterol away from the bloodstream. High levels of LDL cholesterol lead to atherosclerosis increasing the risk of heart attack and ischemic stroke. Already in 2002, Minamino et al. demonstrated that human atherosclerotic lesions contain vascular endothelial cells with senescence-associated phenotypes [1]. We have previously demonstrated for the first time that indeed high LDL is responsible for inducing senescence in human primary endothelial cells by treating them with 1 mg/ml LDL for one week. Those doses induced cellular senescence and loss of endothelial NO synthase (NOS3) [2]. Given the fact that endothelial functionality depends on the production of NO by NOS3 [3], it can be concluded that high doses of LDL lead not only to cellular senescence but also endothelial dysfunction, which is further characterized by endothelial activation, increased apoptosis sensitivity, and decreased migratory capacity [4, 5]. In agreement with our data in human primary endothelial cells, it was recently shown that intravenous injection of LDL in mice resulted in endothelial cell senescence and dysfunction *in vivo*. Interestingly, this dysfunction was accompanied by reduced mitochondrial oxygen consumption in the endothelium [6]. In this context, we already demonstrated that migratory capacity of endothelial cells depends on intact mitochondria. For that purpose, we had generated Rho^0 human primary endothelial cells, which have nonfunctional mitochondria, such that their ATP production solely depends on glycolysis. We found that those cells are unable to migrate underscoring the importance of mitochondria as an energy source [7]. This is further emphasized by the fact that the inhibition of the mitochondrial ATP synthase by oligomycin in endothelial cells reduced the ATP content by approximately 60% [8]. Along the same lines, improvement of mitochondrial functionality with caffeine in endothelial cells increases ATP content and migratory capacity [9]. Thus, endothelial functionality depends on intact mitochondria and ATP produced therein. On the other hand, endothelial senescence and dysfunction seem to be accompanied by reduced mitochondria functionality. However, the underlying mechanisms are poorly understood and not well studied. Therefore, we investigated the effects of high LDL in human primary endothelial cells under conditions leading to endothelial dysfunction on mitochondrial DNA, RNA, protein levels, mass, and functionality.

2. Material and Methods

2.1. Cell Culture. Human primary endothelial cells were cultured in endothelial basal growth media (EBM) from Lonza, supplemented with hydrocortisone (1 μg/ml), bovine brain extract (12 μg/ml), gentamicin (50 μg/ml), human epidermal growth factor (10 ng/ml), and 10% fetal calf serum (EBM complete). The cells were treated with high concentrations of LDL (high LDL) as previously described [2]. In detail, cells were seeded into cell culture dishes and incubated with the cultivation media for two days. After a medium change, the cells were further cultivated in the EBM complete medium (con) or the same medium containing 1 mg/ml LDL (high LDL). The media were changed every two days for 5 to 7 days depending on the experiment.

2.2. Isolation of Total RNA. For isolation of total RNA, cells were lysed using TRIzol (Thermo Fisher Scientific, Schwerte, Germany) and RNA was isolated according to the manufacturer's specifications. RNA was subjected to a second purification step using the RNeasy Mini Kit (Qiagen, Hilden, Germany). RNA concentrations were measured using a NanoDrop 2000c (Thermo Fisher Scientific, Schwerte, Germany), and RNA integrity was determined using a Bioanalyzer (Agilent, Waldbronn, Germany).

2.3. RNA Sequencing Analysis. RNA sequencing data was obtained from quadruplicate total RNA samples. After DNase treatment, a library for sequencing was constructed using the TruSeq® Stranded mRNA Sample Preparation kit (Illumina), according to the Ribo-Zero protocol to remove ribosomal RNA. Subsequently, the libraries were sequenced using HiSeq3000 (Illumina) generating an average of 392 million single-end reads per sample. Library constructions and sequencing were performed at the Genomics and Transcriptomics Laboratory at the Biological Medical Research Centre (BMFZ) of the Heinrich-Heine University Düsseldorf. The quality of the reads was assessed using the tool FASTQC by Andrews (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC [10]. Subsequently, with the help of Trimmomatic version 0.36 [11], reads were trimmed or discarded based on their base calling quality and their adapter content. Afterwards, the extent of rRNA depletion was measured by mapping the reads to rRNA databases using the SortMeRNA algorithm version 2.1b [12]. For alignment and the following analyses, the human genomic reference sequence (GRCh38) and annotation data (release 91) were downloaded from Ensembl [13] and BioMart [14]. After aligning the reads to the human reference using the two-pass mapping protocol of the STAR aligner (2.5.4b) [15], expression of the mitochondrial genes was calculated with the HTSeq python script [16]. Read coverage of gene-to-gene boundaries of mitochondrial transcripts was calculated using the SAMtools software package [17]. For DGE analysis with the R package DESeq2 version 1.18.1 [18], count matrices were generated using the software salmon version 0.9.1 [19].

Scripts used for this work are publicly available at https://github.com/caggtaagtat/Endothelial-mitochondria. FASTQ file preparation and alignment was accomplished using custom BASH shell scripts in the environment of the High Performing Cluster of the Heinrich-Heine University Düsseldorf. Computational support and infrastructure were provided by the "Centre for Information and Media Technology" (ZIM) at the Heinrich-Heine-University Düsseldorf.

2.4. Real-Time PCR Analysis of Transcript Levels. Total RNA was treated with RNAse-free DNase and reversed transcribed using SuperScript IV (Thermo Fisher Scientific, Schwerte, Germany) with random hexamers (pdN$_6$) and oligo dT$_{20}$ as primers. Relative transcript levels were determined by semiquantitative real-time PCR using the nuclear-encoded transcript for the ribosomal protein L32 (RPL32) as reference. The PCR reactions were done in a Rotor-Gene Q instrument (Qiagen, Hilden, Germany) using the SYBR Green qPCR Mastermix (Bimake, Munich, Germany) with the primer pairs listed below. All primer pairs for the analysis of nuclear transcripts were intron-spanning. For quantitation of mitochondrial transcripts, control reactions were performed with mock cDNAs, which were generated in a cDNA synthesis reaction without SuperScript IV. Relative expression was calculated as $2^{-\Delta\mathrm{Ct}\,(\mathrm{Ct\ gene\ of\ interest}-\mathrm{Ct\ RPL32})}$. The following primer pairs were used: MT-ND2: 5′-TCATAGCAGGCAGTTGAGGTG-3′/5′-CGTGGTGCTGGAGTTTAAGTTG-3′, MT-CYB: 5′-CATCGGCATTATCCTCCTGCT-3′/5′-ATCGTGTGAGGGTGGGACTG-3′, MT-CO3: 5′-AGGCATCACCCCGCTAAATC-3′/5′-ACTCTGAGGCTTGTAGGAGG-3′, MT-RNR1: 5′-CAAAACTGCTCGCCAGAACAC-3′/5′-GAGCAAGAGGTGGTGAGGTTG-3′, unprocessed mtRNA precursor transcripts: 5′-CGGACTACAACCACGACCAA-3′/5′-CCAAGGAGTGAGCCGAAGTT-3′ (region 1) and 5′-AGAGGCCTAACCCCTGTCTT-3′/5′-TGCCTAGGACTCCAGCTCAT-3′ (region 2), TFAM: 5′-GATTCACCGCAGGAAAAGCTG-3′/5′-GTGCGACGTAGAAGATCCTTTC-3′, TFB1M: 5′-AGTGGCAGAGAGACTTGCAG-3′/5′-TTCCACCAGCTTGAATGGCT-3′, TFB2M: 5′-GCTGGAAAACCCAAAGCGTA-3′/5′-GTCTATTACAGTGGCGCTGC-3′, and RPL32: 5′-GTGAAGCCCAAGATCGTCAA-3′/5′-TTGTTGCACATCAGCAGCAC-3′.

2.5. Determination of Mitochondrial DNA (mtDNA) Content. Total DNA was isolated using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). DNA concentrations were measured using a NanoDrop 2000c (Thermo Fisher Scientific, Schwerte, Germany). Relative mtDNA levels were determined by semiquantitative real-time PCR using the single copy nuclear nucleoredoxin (NXN) gene as reference. PCR reactions were done in a Rotor-Gene Q instrument (Qiagen, Hilden, Germany) using the SYBR Green qPCR Mastermix (Bimake, Munich, Germany) with the primer pairs listed below. Relative mtDNA content was calculated as $2^{-\Delta\mathrm{Ct}\,(\mathrm{Ct\ mtDNA}-\mathrm{Ct\ NXN})}$. The following primer pairs were used: D-loop: 5′-AGCCACTTTCCACACAGACATCAT-3′/5′-ATCTGGTTAGGCTGGTGTTAGGGT-3′ and NXN: 5′-CCACTCTTGTGTTCTCAGGCAGG-3′/5′-CGTGGGAGCTGTTTGTATGATATGAACC-3′.

2.6. Immunoblotting. Cells were lysed with JNK buffer (10 mM Tris-HCl, pH 7.5,150 mM NaCl, 2.5 mM KCl, 0.5% (*v*/*v*) Triton X-100, 0.5% (*v*/*v*) IGEPAL CA-630), and proteins were separated by SDS-PAGE followed by immunoblotting using primary antibodies against NOS3 (1 : 500, BD, Heidelberg, Germany) and ERK 1/2 (1 : 2000, Cell Signaling Technology, Frankfurt, Germany). All antibodies were diluted in 1% nonfat dry milk. Primary antibodies were incubated overnight at 4°C followed by HRP-coupled secondary antibodies for 2 h at room temperature. For protein detection, the Pierce™ ECL Western Blotting Substrate (Thermo Fisher Scientific, Schwerte, Germany) was used. Semiquantitative analysis from scanned blots was performed by using ImageJ 1.46r [20].

2.7. Cell Migration Assay. Cell migration assay was performed as described previously [9]. Wound closure was determined by setting a small wound and measuring wound width directly afterwards and 2 h later. Pictures were acquired with the Axiovert 200M fluorescent microscope from Carl Zeiss (Jena, Germany).

2.8. ATP Assay. Cells were lysed with JNK buffer, and ATP was measured in total lysates as described previously [9].

2.9. Mitochondrial Mass. Cells were treated with nonyl acridine orange and measured by flow cytometry as previously described [21].

2.10. Statistics. Data were analyzed using unpaired Student's *t*-tests. Differences in gene expression across samples were calculated using the Wald test of the R package DESeq2. *p* values were adjusted to the number of Wald tests following the Benjamini-Hochberg procedure [18]. Adjusted *p* values lower than 0.05 were considered significant.

3. Results

3.1. High LDL Induces Loss of NOS3 Protein and Overall NOS3 mRNA Levels Accompanied by Shifting Alternative Splice Site Use towards Inactive NOS3 Transcript Isoforms. High LDL importantly contributes to the development and progression of cardiovascular diseases. However, the underlying molecular mechanisms how high LDL influences endothelial cell functionality are as yet poorly understood. In particular, detailed transcriptome analyses including the mitochondrial transcripts have not been addressed before. Therefore, we performed RNA deep sequencing of human primary endothelial cells after 7 days of treatment with high LDL. Prior to RNA sequencing, however, we validated that also in the current experimental setting, treatment with high LDL led to a reduction in NOS3 protein levels as described previously by us [2] (Figures 1(a) and 1(b)).

RNA sequencing data showed that the observed decrease in the NOS3 protein levels was accompanied by a decrease in the total amount of NOS3 mRNA reads. In particular, differential gene expression analysis indicated a significant decrease in the NOS3 RNA levels in high LDL-treated cells compared to the healthy control with an adjusted *p* value of 0.06. However, this analysis did not discriminate functional NOS3 mRNA transcripts from nonfunctional NOS3 transcripts generated by alternative splicing. Therefore, we focused on splice site usage within the NOS3 locus, allowing a more detailed view on protein coding and noncoding transcript isoforms. Indeed, splice site usage analysis showed a reduction in gene normalized splice site usage (GNSSR) for those sites involved in the generation of functional NOS3 mRNA transcripts but an increase in the GNSSR contributing to nonfunctional transcript isoforms (Figure 2 and Table 1).

In this analysis, all exon-exon junctions showing a significant decrease in the relative coverage upon LDL treatment were located upstream of exon 18. In contrast, a significant increase was found in the junctions downstream of exon 18 (Table 1). This reciprocal regulation cannot simply be explained with the full-length transcript but might rather be indicative of aberrant transcription from internal promoters generating transcripts not producing functional NOS3 protein. In both regions, the junctions indicative of alternative

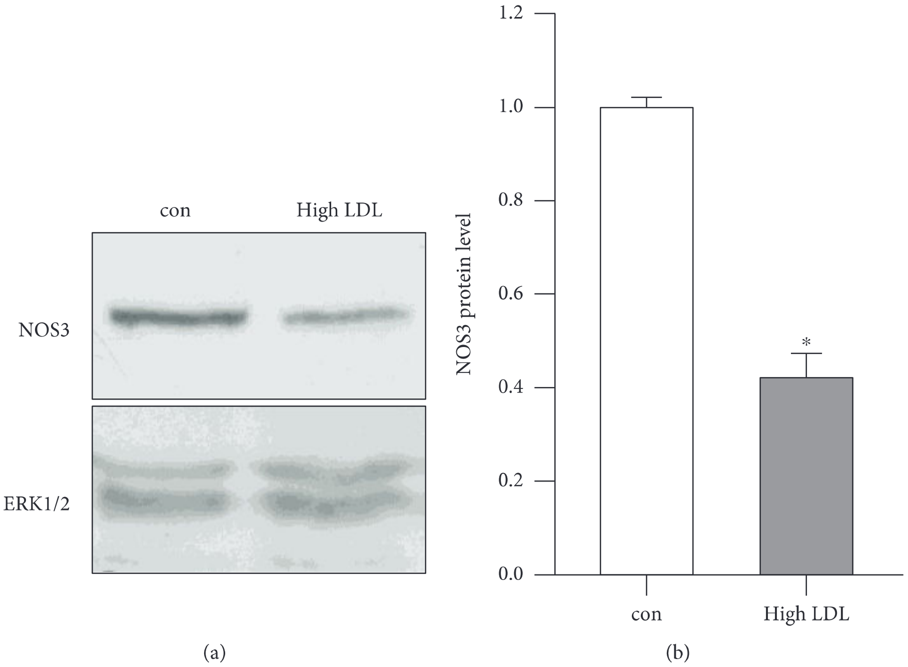


Figure 1: High LDL decreases NOS3 protein levels. Human primary endothelial cells were cultured in the standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Full-length NOS3 protein was detected by immunoblot; ERK1/2 served as a loading control. (a) Representative immunoblots. (b) Semiquantitative analysis. Data are mean ± SEM; $n = 7$; $p < 0.05$ vs. con.

splicing from exon 14 into exons 14A/B/C or resulting in skipping of exon 21 represented only a small fraction of the reads. Thus, in addition to the overall decrease in NOS3 reads, the ratio of gapped reads (exon-exon junction reads) assigned to functional and nonfunctional transcript isoforms is altered towards the latter, which provides an adequate explanation for the reduction of functional NOS3 protein. In agreement with previous observations [22], these results additionally demonstrate that splice site usage can be influenced by high LDL. Differences in the extent of splice site usage between healthy and unhealthy conditions are likely to originate from differential gene expression of genes coding for splicing regulatory proteins (Suppl. Table 1), which mediate splice site usage and consequently RNA transcript isoform levels.

3.2. High LDL Reduces Migratory Capacity, Expression of Genes Associated with Cell Migration, and ATP Content. A reduction in the functional NOS3 levels due to an unhealthy treatment leads to a decrease in the NO levels. Since endothelial cell migration is NO-dependent [23–25], we next investigated endothelial cell migration under high LDL conditions. Endothelial cells treated with high LDL were severely impaired to close a wound (Figure 3).

To address the question whether the expression of genes known to be associated with cell migration capacity could substantiate our finding, we performed differential gene expression analysis. Indeed, the expression of the cell cycle controlling protein CDC42, which is involved in cell migration was reduced by 35% in cells under high LDL conditions ($p < 0.05$). Another factor associated with cell migration, AKT1, was also significantly lower expressed by 17% in high LDL-treated cells ($p < 0.05$). Furthermore, in RNA samples of cells incubated for one week with high LDL, Rho family GTPases RND1 and RND3, but not RND2, showed a decrease in their expression by 55% ($p < 0.05$) and 39% ($p < 0.05$), respectively. RND1/2/3 are known to play a role in cell migration [26]. The reduction in CDC42, AKT1, and RND-transcript levels indicated a negative effect of high LDL on the expression of genes, associated with migration of human endothelial cells. Thus, besides the high LDL mediated decrease in the NO levels, also genes associated with cell migration were significantly decreased in their expression providing a plausible explanation for the observed reduced migratory capacity. Since we have previously demonstrated that the migration of primary human endothelial cells depends on intact mitochondria [7, 9], we next determined the ATP content in those cells. Indeed, the ATP levels in cells treated with high LDL were significantly reduced (Figure 4).

The dependence of migratory capacity on the ATP content was further confirmed by treatment of endothelial cells with oligomycin—a specific inhibitor of the mitochondrial ATP synthase. Both migratory capacity and ATP content were decreased by approximately 60% (Suppl. Figure 1). The latter finding underscores that the mitochondria are one of the main energy sources in endothelial cells.

3.3. Effects of High LDL on Mitochondrial DNA and RNA Levels as well as on Mitochondrial Mass. As we found reduced migratory capacity as well as lower ATP content in primary

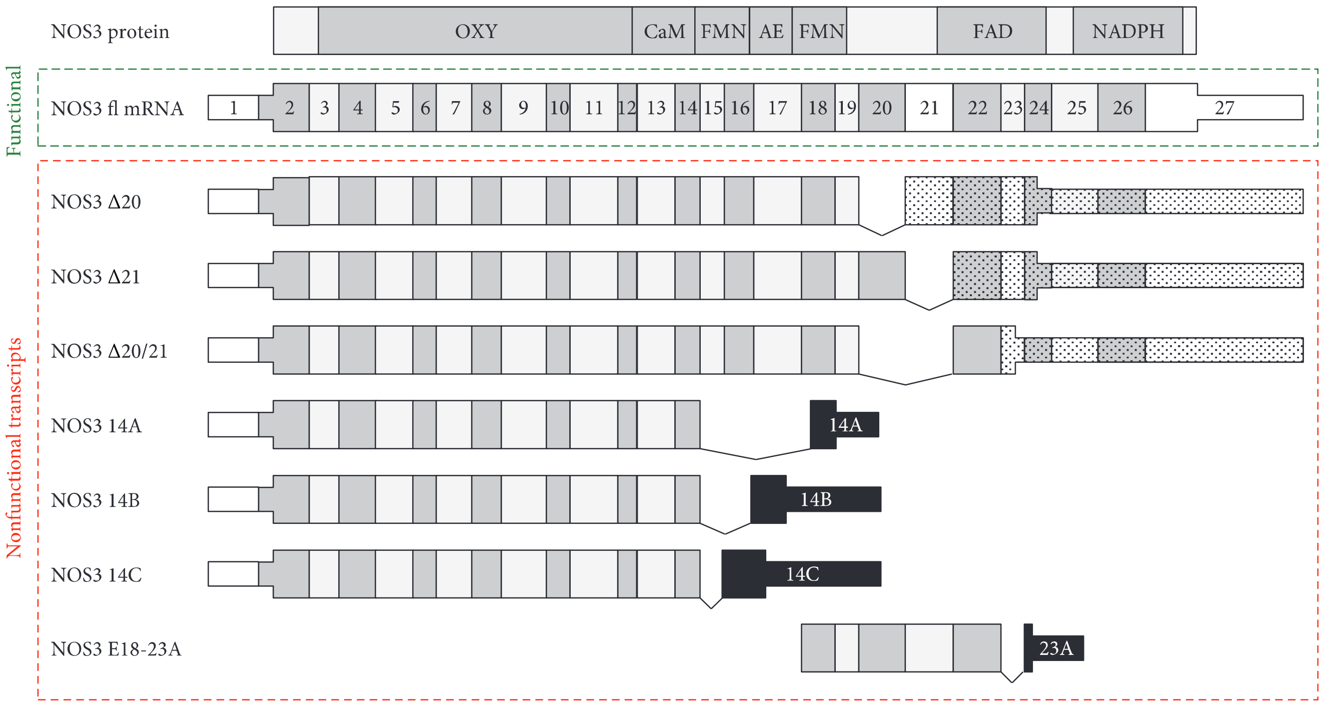


Figure 2: NOS3 transcript isoforms. Depicted at the top is the NOS3 protein (NOS3 protein) with its functional domains (OXY: oxygenase domain; CaM: calmodulin-binding site; FMN: FMN recognition site; AE: autoinhibitory element; FAD: FAD recognition site; NADPH: NADPH recognition site). The corresponding full-length transcript (NOS3 fl mRNA) annotated in Ensembl (v.93) with numbered exons is shown in the green dotted box. The coding region (wide boxes) extends from exon 2 into exon 27. Nonfunctional transcripts, i.e., transcripts not coding for functional NOS3, are shown in the red box below. Skipping of exon 20 (NOS3 *Δ*20), 21 (NOS3 *Δ*21), or both (NOS3 *Δ*20/21) leads to nonfunctional proteins due to frame shifts (dotted boxes) [31]. The variants NOS3 14A/B/C (previously described as NOS3 13A/B/C [32]) originate from splicing events from exon 14 into exons located in the intron of the full-length NOS3 transcript (black boxes). These transcripts terminate in a common polyadenylation signal and encode C-terminally truncated proteins only containing the OXY and CaM domains. The transcript starting at exon 18 (NOS3 E18-23A) lacks the 5′-portion and shows a similar splicing phenomenon as NOS3 14A/B/C at its 3′-end. Thus, it codes for an N- and C-terminally truncated protein.

Table 1: Relative NOS3 exon-exon junction expression. Comparison of relative expression of all exon-exon junctions from exon 1 to exon 18 (1 : 18), exon 18 to exon 27 (18 : 27), and exon-exon junctions indicative of alternative splicing from exon 14 onto exons 14A/B/C (14-14x) or skipping exon 21 (20-22). The exon-exon junction coverage was normalized per sample to the number of gapped reads within the sample and to gene expression; the *p* value was calculated with Student's *t*-test. Exon-exon junctions indicative of skipping exon 20 or exon 20 and 21 simultaneously were not detectable.

Exon-exon junction	Average normalized expression		*p* value
	con	High LDL	
1 : 18	3,740	3,153	0.00
18:27	3,949	4,876	0.02
14-14x	0.009	0.010	0.83
20-22	0.002	0.024	0.16

human endothelial cells upon treatment with high LDL, we next investigated the effects of high LDL on mitochondrial DNA and RNA levels, as well as on mitochondrial mass. Therefore, we first performed an alignment of the sequencing data to the human reference genome. In control samples (con), around 95.5% of the total reads could be mapped to the nuclear and 4.5% to the mitochondrial genome (Table 2). However, in high LDL-treated cells, around 11.2% of the reads could be mapped to mitochondrial transcripts. Thus, high LDL led to a significant, more than two-fold increase in the mitochondrial RNA (mtRNA) levels.

The upregulation of mtRNAs upon high LDL treatment raised the question as to whether this is paralleled by an elevation in mitochondrial DNA (mtDNA) content. Corresponding to the increase in mtRNA content, high LDL-treated cells showed a significantly higher mtDNA content (Figure 5).

The increase in mtRNAs and mtDNA could be indicative of a higher number in mitochondria. Therefore, we determined the expression of genes coding for the translocases of outer (TOMMs) and inner (TIMMs) mitochondrial membrane proteins as surrogate markers. Those transcripts, however, were either not regulated or expressed at lower levels upon high LDL treatment. Corresponding to the non- or downregulated mRNA transcript levels, analysis of the TIMM23 protein levels as a marker for mitochondria showed no significant difference between the two conditions (data not shown). Thus, an increase in overall

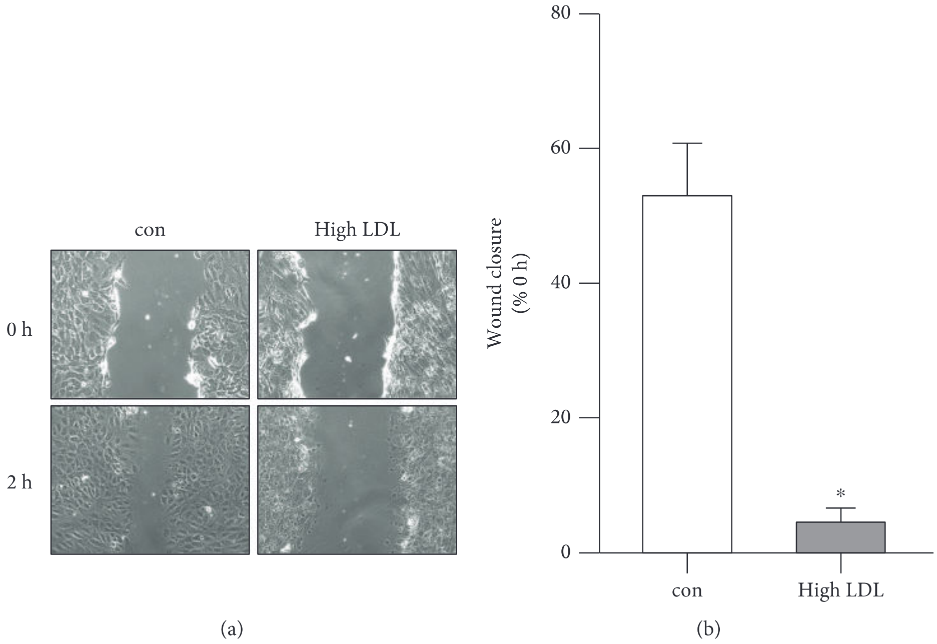


Figure 3: High LDL decreases migratory capacity of endothelial cells. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 5 days. A wound was set, and wound width was determined directly afterwards (0 h) and two hours later (2 h). (a) Representative microscopic pictures. (b) Wound closure relative to the 0 h time point. Data are mean ± SEM; $n = 4$; $p < 0.05$ vs. con.

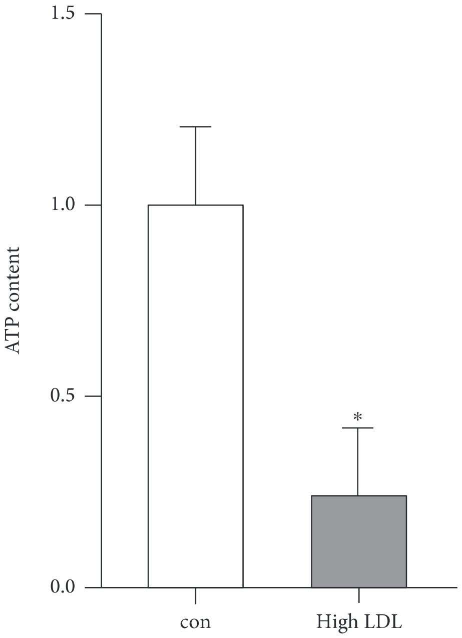


Figure 4: High LDL treatment significantly reduces ATP content in endothelial cells. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 5 days, and ATP content was measured. Data are mean ± SEM; $n = 4$; $p < 0.05$ vs. con.

mitochondrial number induced by high LDL treatment was rather unlikely. We substantiated this by measuring mitochondrial mass using nonyl acridine orange. As shown in Figure 6, high LDL treatment did not result in a change in mitochondrial mass.

As the mtDNA content was increased upon high LDL without a concomitant change in mitochondrial mass, we next investigated the expression of protein coding mtRNA transcripts and mitochondrial ribosomal RNAs. Therefore, the RNA sequencing data were again analyzed for this specific subset of transcripts. High LDL-treated cells displayed an increase in the expression of these mitochondrial transcripts (Table 3).

To validate our RNA sequencing data, the transcript levels of mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 2 (MT-ND2), cytochrome B (MT-CYB), cytochrome C oxidase III (MT-CO3), and the mitochondrial 12S RNA (MT-RNR1) were analyzed by real-time PCR. The first three are subunits of electron transport chain complexes I, III, and IV, respectively. All of the chosen transcripts were significantly increased after treatment with high LDL (Figure 7).

Since the mtRNAs were upregulated, we next investigated whether the nuclear-encoded transcription factors, which are known to be mainly responsible for the transcription of mtDNA, are regulated by high LDL. Expression analysis of the transcripts coding for mitochondrial transcription factor A (TFAM), B1 (TFB1M), and B2 (TFB2M) by real-time PCR, however, did not indicate any significant differences (Figure 8), suggesting that the increase in the mtRNA levels is also not related to an increase in transcription.

Table 2: Percentage of read numbers representing nuclear and mitochondrial transcripts. Shown are total read numbers for all individual biological replicates, i.e., RNAs isolated from human primary endothelial cells cultured in standard medium (con_1-4) or medium containing 1 mg/ml LDL (high_LDL-1-4) and the percentage of reads, which could be mapped to the nuclear or mitochondrial reference genome.

Sample	Nuclear (%)	Mitochondrial (%)	Total # of reads
con_1	96.0	4.0	325,725,044
con_2	95.8	4.2	319,428,178
con_3	93.9	6.1	335,842,395
con_4	96.3	3.7	325,882,327
high_LDL_1	88.9	11.0	328,183,964
high_LDL_2	88.1	11.9	327,355,990
high_LDL_3	89.0	11.0	331,832,849
high_LDL_4	89.0	11.0	331,650,314

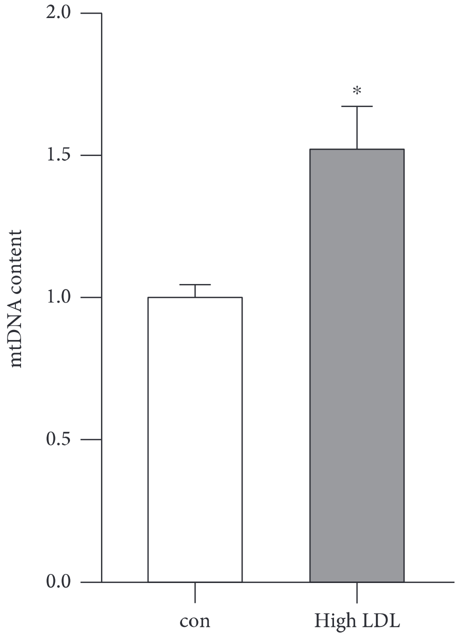


Figure 5: High LDL increases mtDNA content. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Total DNA was isolated, and mtDNA content was measured by semiquantitative real-time PCR using NXN as nuclear reference gene. Data are mean ± SEM; $n = 6$; $p < 0.05$ vs. con.

After transcription, the mtRNA precursor transcripts are cleaved by RNase P and RNase Z at the $5'$- and $3'$ end, respectively, leading to mature mt-mRNA, mt-rRNA, and mt-tRNA transcripts within the mitochondrial matrix. To exclude that high LDL treatment did not impair processing of the precursor mtRNA transcripts, the actual amount of precursor mtRNA transcripts was estimated by counting reads whose alignment covered the border of two neighboring mtRNA transcripts, which get separated after processing by the RNases. Taking the border coordinates of every mtRNA transcript from the human reference genome GRCh38/hg38 (Ensembl version 91), a combination of every transcript border coordinate of neighboring elements of the mitochondrial genome was generated. For every pair, reads were subsequently counted, which simultaneously covered the downstream end of one transcript and the upstream end of the transcript downstream to it, with a minimal overlap of six nucleotides. For comparison across sample groups, the coverage per border pair was normalized to the total number of ungapped reads in a given sample and the relative amount of ungapped reads mapped to the mitochondrial reference sequence. The average border coverage within a sample was then used as an approximation for the relative amount of mtRNA precursor transcripts. Comparing transcript border coverages did not reveal significant differences in mtRNA precursor transcripts (data not shown). For validation, we performed RT PCRs across borders of mature transcripts, thereby detecting the unprocessed precursors. These experiments corroborated the bioinformatic analyses (Figure 9).

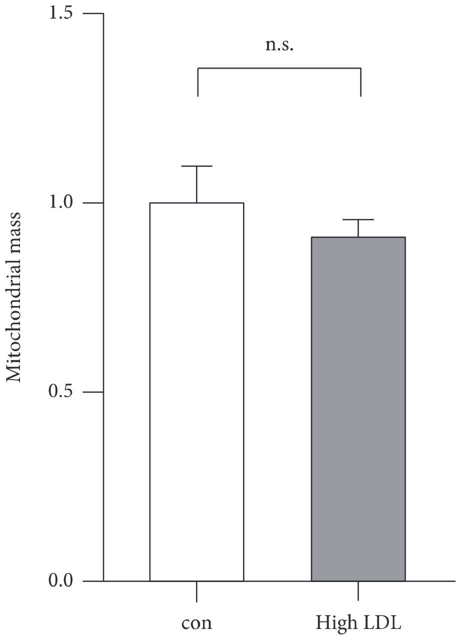


Figure 6: Mitochondrial mass is not altered by high LDL treatment. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Then, cells were incubated with nonyl acridine orange and analyzed by flow cytometry. Data are mean ± SEM; $n = 4$; n.s. = not significant.

Although some proteins of the respiratory chain are encoded on the mitochondrial DNA, most of them are derived from nuclear genes. Undisturbed interplay of mitochondrial and nuclear-encoded proteins ensures efficient respiratory chain complex formation and consequently ATP synthesis [27]. Since treatment with high LDL led to an increase in the expression of mitochondrial encoded proteins, we also analyzed differential expression of nuclear genes for respiratory chain proteins (Suppl. Table 2).

Table 3: Differential gene expression of mitochondrial transcripts after high LDL treatment. Comparison of the expression of mitochondrial protein coding genes and ribosomal RNAs between untreated cells and cells treated with high LDL. The L2FC (log 2-fold change) states the average difference in gene expression between both treatments. Positive L2FC values denote upregulation by high LDL; negative values denote downregulation. A Wald test from DESeq2 was used to calculate the significance of the change in the expression. The adjusted p values take the number of tested genes into account.

Gene name	Ensembl gene ID	L2FC	p value	Adjusted p value
MT-RNR1	ENSG00000211459	1.86	$1.46E-40$	$2.01E-38$
MT-RNR2	ENSG00000210082	1.52	$1.66E-30$	$1.41E-28$
MT-ND6	ENSG00000198695	1.67	$2.98E-24$	$1.68E-22$
MT-ND1	ENSG00000198888	1.34	$3.02E-23$	$1.58E-21$
MT-ND4	ENSG00000198886	1.29	$9.81E-22$	$4.50E-20$
MT-ND3	ENSG00000198840	1.33	$1.48E-20$	$6.17E-19$
MT-CO1	ENSG00000198804	1.34	$3.43E-20$	$1.38E-18$
MT-CO2	ENSG00000198712	1.19	$4.09E-18$	$1.36E-16$
MT-ATP6	ENSG00000198899	1.00	$2.94E-14$	$6.34E-13$
MT-CYB	ENSG00000198727	1.01	$9.95E-13$	$1.76E-11$
MT-ND2	ENSG00000198763	0.90	$1.46E-11$	$2.18E-10$
MT-CO3	ENSG00000198938	0.87	$2.24E-10$	$2.77E-09$
MT-ND4L	ENSG00000212907	1.04	$4.25E-10$	$5.01E-09$
MT-ND5	ENSG00000198786	1.40	$2.67E-08$	$2.31E-07$
MT-ATP8	ENSG00000228253	0.82	$7.33E-08$	$5.85E-07$

In contrast to the overall increased mitochondrial transcript levels following high LDL treatment, differential gene expression analysis for 81 nuclear genes encoding proteins of the respiratory chain revealed that 32% of them were significantly downregulated and only 4% upregulated.

The increase in mitochondrial gene expression and decrease in one third of nuclear genes for respiratory chain proteins could, therefore, restrict efficiency of ATP production due to an imbalance in respiratory chain subunits preventing proper assembly. Thus, one could assume that the reduced ATP content seen in Figure 4 is caused by dysfunctional or not correctly assembled respiratory chain complexes.

4. Discussion

The major findings of our study are that treatment of human primary endothelial cells with 1 mg/ml LDL for seven days decreases the NOS3 protein levels, increases inactive NOS3 splice variants, and reduces mitochondrial functionality in endothelial cells, which results in dramatically reduced migratory capacity and thus, endothelial cell impairment.

A functional endothelial cell layer is important, since it not only regulates vascular tone but as a barrier also regulates the nutrition uptake of the surrounding tissue and protects against pathogens. Endothelial cells are in direct contact with the bloodstream and consequently the first cells affected by LDL. Here, we demonstrate that treatment of endothelial cells with high LDL leads to decreased levels of functional NOS3 protein and mRNA levels. Additionally, RNA sequencing analyses revealed an increase in inactive NOS3 splice variants. This is accompanied by an increase in the expression of all mitochondrially encoded transcripts. However, there was no increase in total mitochondria number, as shown at the RNA level as well as at the protein level. It was previously described that NOS3-deficient mice showed a dysfunctional mitochondrial β-oxidation [28]. This could lead to an increase in reactive oxygen species (ROS) formation and therefore oxidative stress within the cell. The peroxisome proliferator-activated receptor γ (PPARG), which is known to regulate the redox balance, fatty acid oxidation, and mtDNA levels, could therefore be one reason for the high mtDNA levels. Its activation upon oxidative stress could potentially lead to the activation of genes holding a PPAR response element (PPRE) in their promoter resulting in an increase in mtDNA copies [29]. Our RNA sequencing data revealed that cells with low levels of functional NOS3 protein showed an increase in PPARG expression by 110%. The increase in mtRNA transcripts upon high LDL is also in line with findings in mice, which were fed a high-fat, high-sucrose diet for 6 weeks [30] and showed an upregulation of several genes important for mitochondrial biogenesis.

We previously demonstrated that those concentrations of LDL resulted in endothelial cell senescence and increased ROS formation. Thus, we hypothesize that cells try to compensate the increased cellular stress, caused by those unhealthy conditions, by upregulating

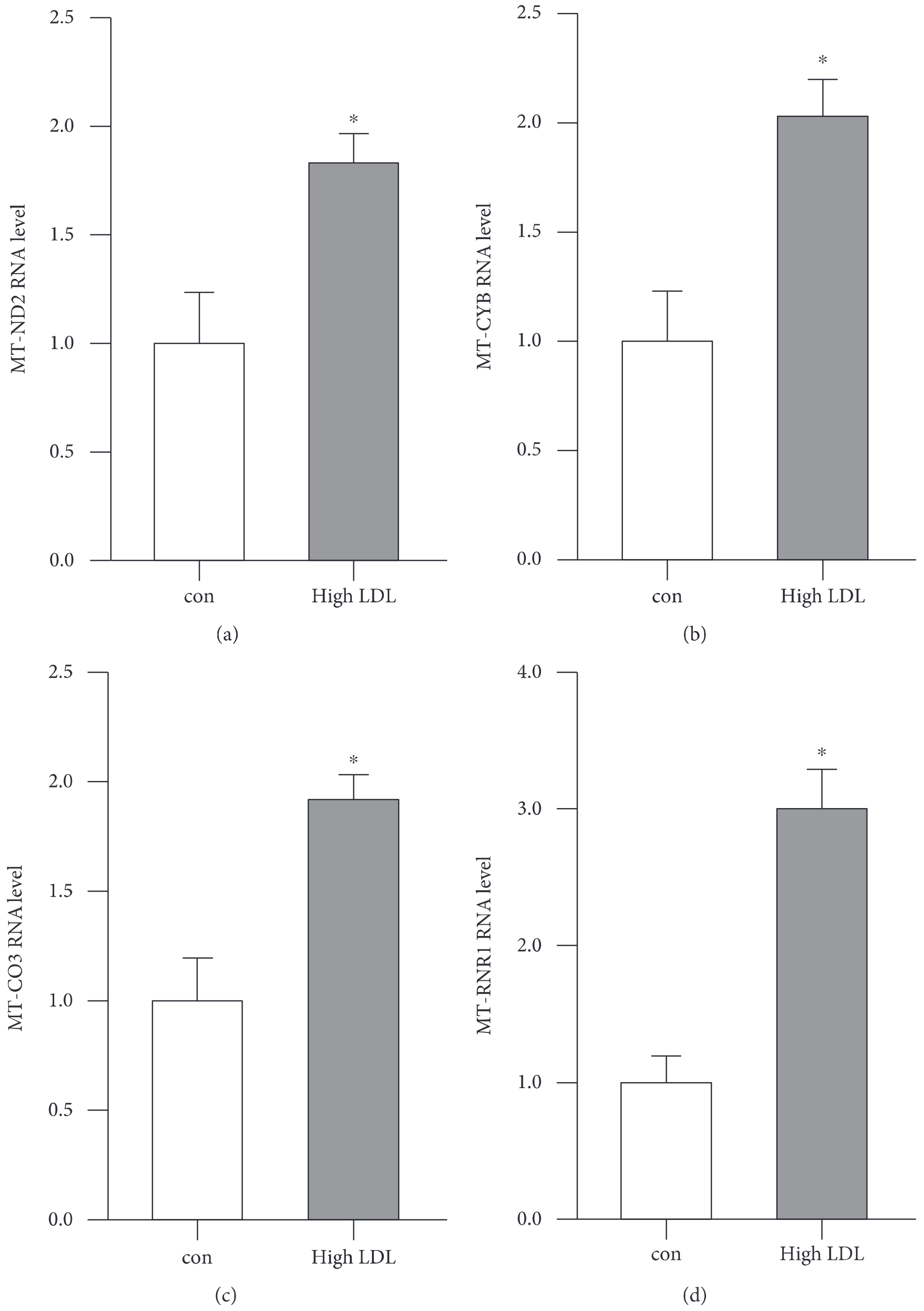


Figure 7: Transcript levels of mitochondrial genes are increased after treatment with high LDL for 7 days. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Semiquantitative real-time PCRs were performed for MT-ND2 (a), MT-CYB (b), MT-CO3 (c), and MT-RNR1 (d) using RPL32 as reference. Data are mean ± SEM; $n = 4$; $p < 0.05$ vs. con.

the expression profile of mitochondrial genes, like MT-ND2 and MT-CO3. Mitochondrial encoded genes are all part of the respiratory chain complexes. Thus, the cells try to cope for energy to handle the unfavorable situation. However, the majority of proteins needed for functional complex formation within the respiratory chain are encoded in the nuclear genome. Our data demonstrate, however, that most of those nuclear-encoded genes are downregulated upon high LDL treatment. Thus, an imbalance in proteins needed for the respiratory chain complexes seems plausible. This in turn would disturb efficient complex formation, resulting in reduced ATP production, which subsequently impairs ATP-dependent processes like endothelial cell migration as we show here.

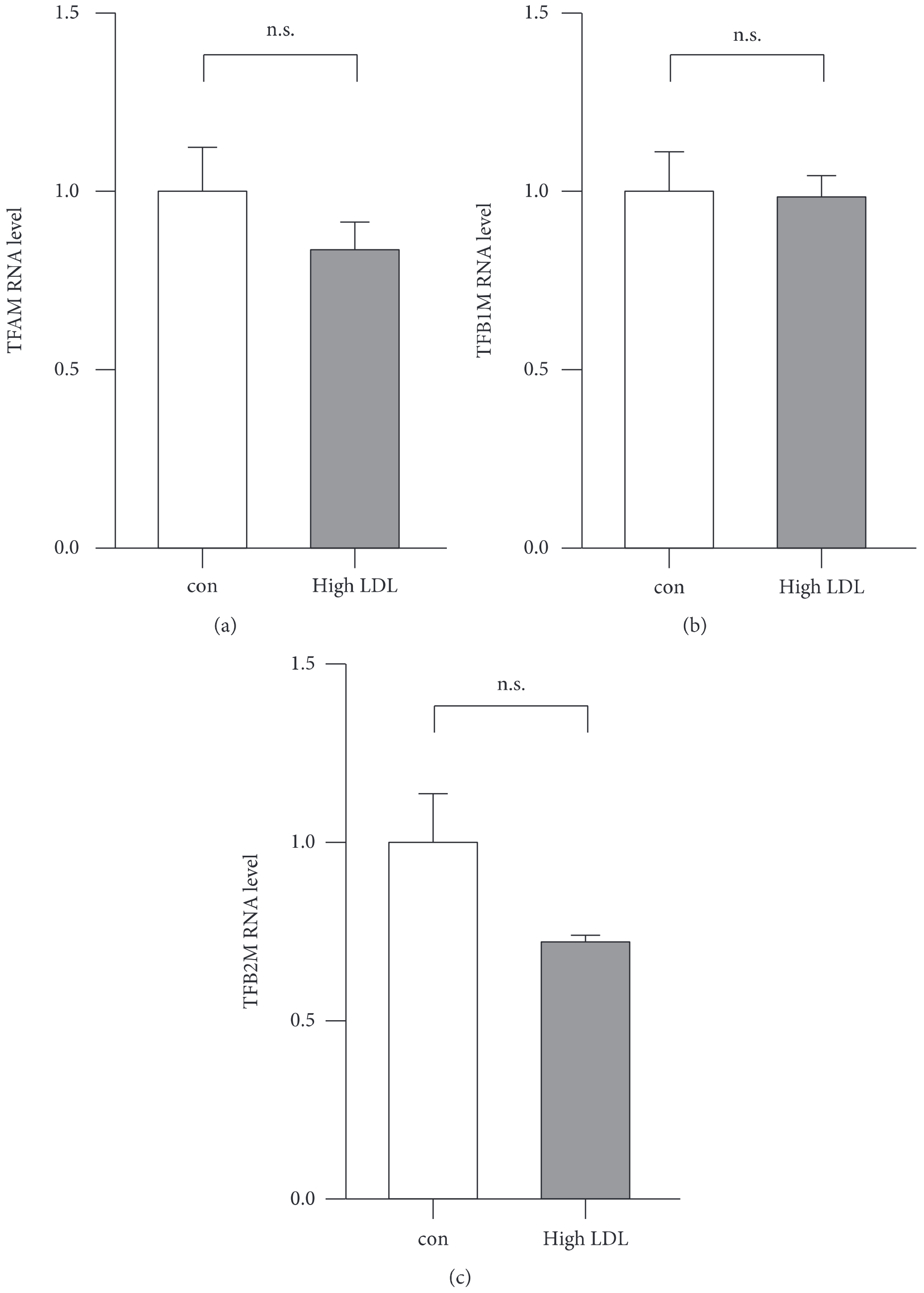


Figure 8: Transcript levels of nuclear-encoded transcription factors of mtDNA transcription are not regulated by high LDL treatment. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Semiquantitative real-time PCRs were performed for TFAM (a), TFB1M (b), and TFB2M (c) using RPL32 as reference. Data are mean ± SEM; $n = 4$; n.s. = not significant.

5. Conclusions

We demonstrate that high LDL concentrations lead to low NOS3 levels in primary human endothelial cells, which is paralleled by mitochondrial dysfunction. We found an increase in mtDNA copy number and mtRNA levels as a potential compensatory mechanism for an unfavorable situation. However, due to an expression imbalance between nuclear and mitochondrial encoded proteins of the respiratory chain, complex formation is most likely impaired resulting in a drastic reduction in ATP levels. Consequently, the migratory capacity of the endothelial cells is reduced, which would negatively affect several cardiovascular diseases.

Data Availability

The RNA sequencing data used to support the findings of this study have been deposited at ArrayExpress under

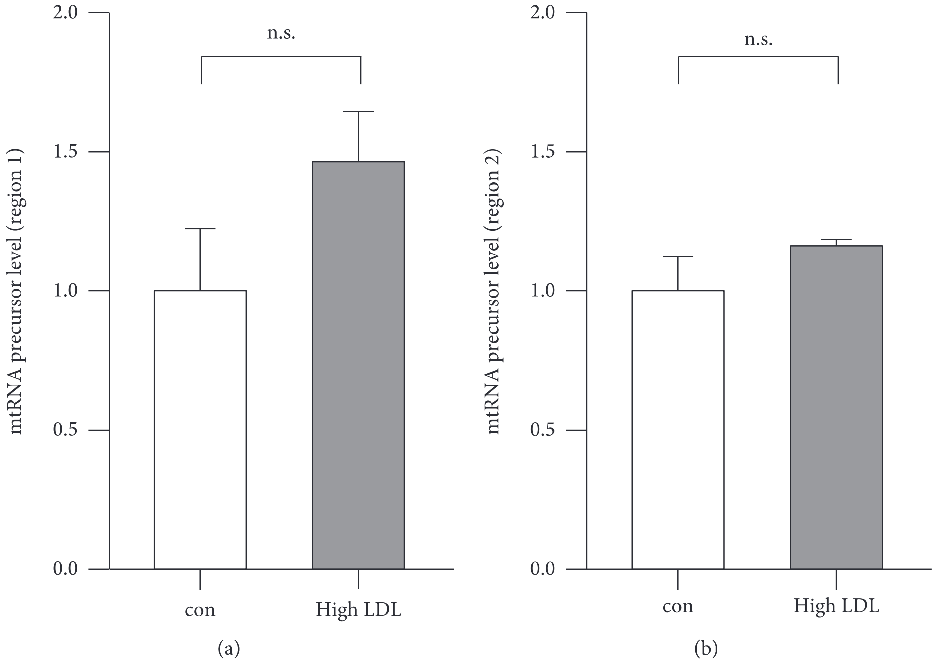


FIGURE 9: mtRNA precursor transcripts are not regulated by high LDL treatment. Human primary endothelial cells were cultured in standard medium (con) or medium containing 1 mg/ml LDL (high LDL) for 7 days. Semiquantitative real-time PCRs were performed for mtRNA precursor transcripts using RPL32 as reference. Data are mean ± SEM; $n = 3 - 4$; n.s. = not significant.

accession number E-MTAB-7647. All other data are available upon request.

Conflicts of Interest

All authors have disclosed that they do not have any conflicts of interest.

Authors' Contributions

Stefanie Gonnissen, Johannes Ptok, Judith Haendeler, Heiner Schaal, and Joachim Altschmied contributed equally to this work.

Acknowledgments

This work was in part supported by the following grants: Deutsche Forschungsgemeinschaft (HA2868/10-2 and HA2868/11-1 to J.H., SCHA909/8-1 to H.S., TI323/4-1 to J.T., and SFB1116-2 A04 to J.H. and J.A.) and Forschungskommission of the Medical Faculty, Heinrich-Heine-Universität Düsseldorf (2016-44) to J.H. and H.S.; S.G. is a former scholarship holder of the IRTG1902, P1; K.J. is a scholarship holder of the IRTG1902, P1; P.J. is a former scholarship holder of the IRTG1902, P2.

Supplementary Materials

Supplementary Figure 1: oligomycin reduces migratory capacity and ATP content in primary human endothelial cells. Primary human endothelial cells were left untreated (con) or treated with 10 μM oligomycin for 24 hours, and migratory capacity (A) and ATP content (B) were measured. Data are mean ± SEM; $n = 4 - 5$; $p < 0.05$ vs. con. Supplementary Table 1: differential gene expression of splicing regulatory genes. DGE calculated using the R package DESeq2 of genes encoding splicing regulatory proteins in samples of untreated cells versus cells treated with high LDL for 7 days. Genes significantly varying in expression are marked in bold. The L2FC (log 2-fold change) states the average difference in gene expression between both treatments. Positive L2FC values denote upregulation by high LDL; negative values denote downregulation. A Wald test from DESeq2 was used to calculate the significance of the change in the expression. The adjusted p values take the number of tested genes into account. Supplementary Table 2: differential gene expression of nuclear-encoded proteins of the mitochondrial electron transport chain. DGE calculated using the R package DESeq2 of genes encoding proteins of the mitochondrial electron transport chain (ETC) in samples of untreated cells versus cells treated with high LDL for 7 days. Genes significantly varying in expression are marked in bold. The L2FC (log 2-fold change) states the average difference in gene expression between both treatments. Positive L2FC values denote upregulation by high LDL, negative values denote downregulation. A Wald test from DESeq2 was used to calculate the significance of the change in the expression. The adjusted p values take the number of tested genes into account. *(Supplementary Materials)*

References

[1] T. Minamino, H. Miyauchi, T. Yoshida, Y. Ishida, H. Yoshida, and I. Komuro, "Endothelial cell senescence in human atherosclerosis: role of telomere in endothelial

dysfunction," *Circulation*, vol. 105, no. 13, pp. 1541–1544, 2002.

[2] N. Buchner, N. Ale-Agha, S. Jakob et al., "Unhealthy diet and ultrafine carbon black particles induce senescence and disease associated phenotypic changes," *Experimental Gerontology*, vol. 48, no. 1, pp. 8–16, 2013.

[3] J. Haendeler, "Nitric oxide and endothelial cell aging," *European Journal of Clinical Pharmacology*, vol. 62, Supplement 1, pp. 137–140, 2006.

[4] C. Heiss, A. Rodriguez-Mateos, and M. Kelm, "Central role of eNOS in the maintenance of endothelial homeostasis," *Antioxidants & Redox Signaling*, vol. 22, no. 14, pp. 1230–1242, 2015.

[5] C. Heiss, I. Spyridopoulos, and J. Haendeler, "Interventions to slow cardiovascular aging: dietary restriction, drugs and novel molecules," *Experimental Gerontology*, vol. 109, pp. 108–118, 2018.

[6] Y. C. Wang, A. S. Lee, L. S. Lu et al., "Human electronegative LDL induces mitochondrial dysfunction and premature senescence of vascular cells in vivo," *Aging Cell*, vol. 17, no. 4, article e12792, 2018.

[7] I. Spyridopoulos, S. Fichtlscherer, R. Popp et al., "Caffeine enhances endothelial repair by an AMPK-dependent mechanism," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 28, no. 11, pp. 1967–1974, 2008.

[8] C. Bergstraesser, S. Hoeger, H. Song et al., "Inhibition of VCAM-1 expression in endothelial cells by CORM-3: the role of the ubiquitin–proteasome system, p38, and mitochondrial respiration," *Free Radical Biology & Medicine*, vol. 52, no. 4, pp. 794–802, 2012.

[9] N. Ale-Agha, C. Goy, P. Jakobs et al., "CDKN1B/p27 is localized in mitochondria and improves respiration-dependent processes in the cardiovascular system—new mode of action for caffeine," *PLoS Biology*, vol. 16, no. 6, article e2004408, 2018.

[10] P. Ewels, M. Magnusson, S. Lundin, and M. Kaller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047-3048, 2016.

[11] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[12] E. Kopylova, L. Noe, and H. Touzet, "SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data," *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, 2012.

[13] F. Cunningham, M. R. Amode, D. Barrell et al., "Ensembl 2015," *Nucleic Acids Research*, vol. 43, no. D1, pp. D662–D669, 2015.

[14] S. Durinck, Y. Moreau, A. Kasprzyk et al., "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, no. 16, pp. 3439-3440, 2005.

[15] A. Dobin, C. A. Davis, F. Schlesinger et al., "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[16] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[17] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078-2079, 2009.

[18] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[19] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, vol. 14, no. 4, pp. 417–419, 2017.

[20] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram, "Image processing with ImageJ," *Biophotonics International*, vol. 11, pp. 36–42, 2004.

[21] J. Haendeler, J. Hoffmann, J. F. Diehl et al., "Antioxidants inhibit nuclear export of telomerase reverse transcriptase and delay replicative senescence of endothelial cells," *Circulation Research*, vol. 94, no. 6, pp. 768–775, 2004.

[22] W. Szeszel-Fedorowicz, I. Talukdar, B. N. Griffith, C. M. Walsh, and L. M. Salati, "An exonic splicing silencer is involved in the regulated splicing of glucose 6-phosphate dehydrogenase mRNA," *Journal of Biological Chemistry*, vol. 281, no. 45, pp. 34146–34158, 2006.

[23] M. Lukosz, A. Mlynek, P. Czypiorski, J. Altschmied, and J. Haendeler, "The transcription factor Grainyhead like 3 (GRHL3) affects endothelial cell apoptosis and migration in a NO-dependent manner," *Biochemical and Biophysical Research Communications*, vol. 412, no. 4, pp. 648–653, 2011.

[24] C. Urbich, E. Dernbach, A. M. Zeiher, and S. Dimmeler, "Double-edged role of statins in angiogenesis signaling," *Circulation Research*, vol. 90, no. 6, pp. 737–744, 2002.

[25] C. Urbich, A. Reissner, E. Chavakis et al., "Dephosphorylation of endothelial nitric oxide synthase contributes to the anti-angiogenic effects of endostatin," *The FASEB Journal*, vol. 16, no. 7, pp. 706–708, 2002.

[26] H. Warner, B. J. Wilson, and P. T. Caswell, "Control of adhesion and protrusion in cell migration by Rho GTPases," *Current Opinion in Cell Biology*, vol. 56, pp. 64–70, 2019.

[27] L. Koch, "Genetic variation: nuclear and mitochondrial genome interplay," *Nature Reviews Genetics*, vol. 17, no. 9, p. 502, 2016.

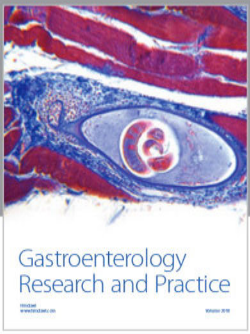
[28] E. L. Gouill, M. Jimenez, C. Binnert et al., "Endothelial nitric oxide synthase (eNOS) knockout mice have defective mitochondrial β-oxidation," *Diabetes*, vol. 56, no. 11, pp. 2690–2696, 2007.

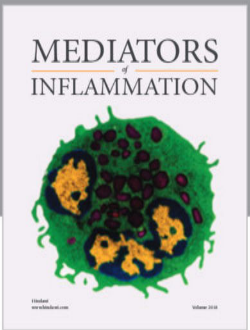
[29] L. Michalik, J. Auwerx, J. P. Berger et al., "International Union of Pharmacology. LXI. Peroxisome proliferator-activated receptors," *Pharmacological Reviews*, vol. 58, no. 4, pp. 726–741, 2006.

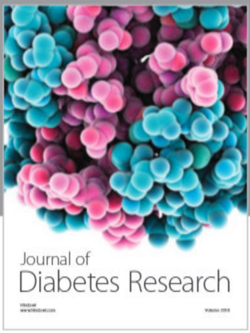
[30] P. W. Wang, H. M. Kuo, H. T. Huang et al., "Biphasic response of mitochondrial biogenesis to oxidative stress in visceral fat of diet-induced obesity mice," *Antioxidants & Redox Signaling*, vol. 20, no. 16, pp. 2572–2588, 2014.

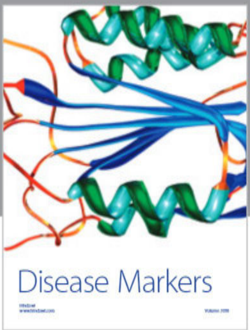
[31] E. Galluccio, L. Cassina, I. Russo et al., "A novel truncated form of eNOS associates with altered vascular function," *Cardiovascular Research*, vol. 101, no. 3, pp. 492–502, 2014.

[32] M. Lorenz, B. Hewing, J. Hui et al., "Alternative splicing in intron 13 of the human eNOS gene: a potential mechanism for regulating eNOS activity," *The FASEB Journal*, vol. 21, no. 7, pp. 1556–1564, 2007.



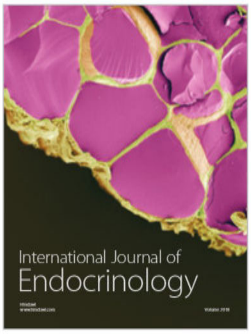


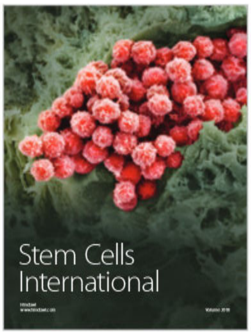


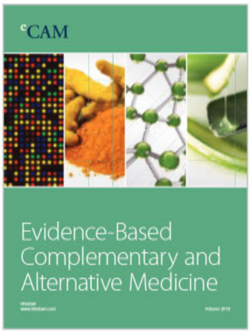


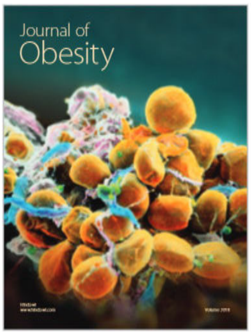
Hindawi

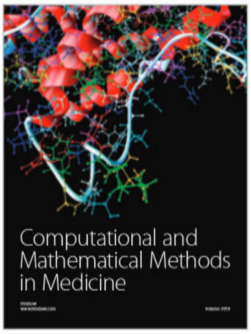
Submit your manuscripts at
www.hindawi.com

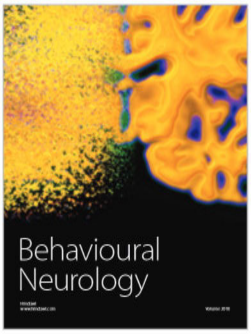


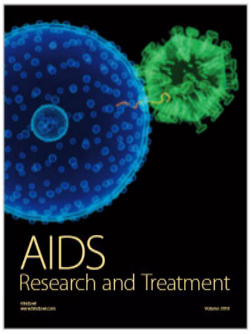


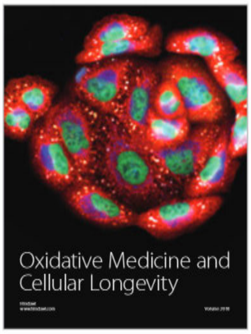












2.2.4 Publication V: Selenoprotein T Protects Endothelial Cells against Lipopolysaccharide-Induced Activation and Apoptosis.

A standard compound to study inflammatory-induced transcriptional changes is Lipopolysaccharide (LPS), which, as part of the cell membrane, gets released into the bloodstream upon lysis of gram-negative bacteria. The response of immune cells to LPS exposure has been shown to include alternative splicing events, which for instance affected genes of the immune response [146]. In this work LPS induced apoptosis in human cardiovascular endothelial cells was analyzed, that results in vascular leakage and subsequent sepsis. Analyzing alternative splicing events in LPS-treated endothelial cells showed for instance an increase in non-functional TRAPPC13 protein (truncated at exon 7). Loss of this protein was previously described to result in less apoptosis, autophagy and endoplasmic stress, which could therefore be a cellular attempt to compensate the LPS-induced cellular stress [147]. We additionally analyzed, whether endothelial apoptosis could be prevented using the first 20 amino acids of the Apurinic/Apyrimidinic Endodeoxyribonuclease 1 (APEX1) [148]. Indeed, APEX1 expression, which resulted in SELENOT upregulation, could successfully reduce LPS-induced endothelial apoptosis.

Article published in Antioxidants (Basel). 2021 Sep 7;10(9):1427 (doi: 10.3390/antiox10091427), by

Dennis Merk*, **Johannes Ptok***, Philipp Jakobs, Florian von Ameln, Jan Greulich, Pia Kluge, Kathrin Semperowitsch, Olaf Eckermann, Heiner Schaal, Niloofar Ale-Agha, Joachim Altschmied, Judith Haendeler (* shared first-author)

<u>Contribution</u>

DM, PJ, FvA, JG, PK, KS and OE did the construction of the reporter constructs, the cell staining and the immunoblot. JP and HS did the bioinformatic analysis of the RNA sequencing data (DGE, GSEA). DM, JP, HS, JH and JA wrote the manuscript. Individual contribution of JP at around 25%.

antioxidants

MDPI

Article

Selenoprotein T Protects Endothelial Cells against Lipopolysaccharide-Induced Activation and Apoptosis

Dennis Merk [1,†], Johannes Ptok [2,†], Philipp Jakobs [1,†], Florian von Ameln [3], Jan Greulich [3], Pia Kluge [1], Kathrin Semperowitsch [1], Olaf Eckermann [1,3], Heiner Schaal [2], Niloofar Ale-Agha [1,*], Joachim Altschmied [1,3,*] and Judith Haendeler [1,*]

[1] Environmentally-Induced Cardiovascular Degeneration, Clinical Chemistry and Laboratory Diagnostics, Medical Faculty, University Hospital and Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany; dennis.merk@hhu.de (D.M.); philipp.jakobs@hhu.de (P.J.); pia.kluge@uni-potsdam.de (P.K.); semperina@gmail.com (K.S.); olaf.eckermann@hhu.de (O.E.)

[2] Institute for Virology, Medical Faculty, University Hospital and Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany; Johannes.ptok@hhu.de (J.P.); schaal@uni-duesseldorf.de (H.S.)

[3] Environmentally-Induced Cardiovascular Degeneration, Clinical Chemistry and Laboratory Diagnostics, Medical Faculty, University Hospital and Heinrich-Heine University Düsseldorf, Germany and IUF-Leibniz Research Institute for Environmental Medicine, 40225 Düsseldorf, Germany; florian.ameln@hhu.de (F.v.A.); jan.greulich@hhu.de (J.G.)

[*] Correspondence: aleagha@hhu.de (N.A.-A.); joalt001@hhu.de (J.A.); juhae001@hhu.de (J.H.); Tel.: +49-211-3389-291 (N.A.-A. & J.A. & J.H.); Fax: +49-211-3389-331 (N.A.-A. & J.A. & J.H.)

[†] D.M., J.P. and P.J. contributed equally to the work.

check for updates

Citation: Merk, D.; Ptok, J.; Jakobs, P.; von Ameln, F.; Greulich, J.; Kluge, P.; Semperowitsch, K.; Eckermann, O.; Schaal, H.; Ale-Agha, N.; et al. Selenoprotein T Protects Endothelial Cells against Lipopolysaccharide-Induced Activation and Apoptosis. *Antioxidants* **2021**, *10*, 1427. https://doi.org/10.3390/antiox10091427

Academic Editor: Stanley Omaye

Received: 17 August 2021
Accepted: 3 September 2021
Published: 7 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Sepsis is an exaggerated immune response upon infection with lipopolysaccharide (LPS) as the main causative agent. LPS-induced activation and apoptosis of endothelial cells (EC) can lead to organ dysfunction and finally organ failure. We previously demonstrated that the first twenty amino acids of the Apurinic/Apyrimidinic Endodeoxyribonuclease 1 (APEX1) are sufficient to inhibit EC apoptosis. To identify genes whose regulation by LPS is affected by this N-terminal APEX1 peptide, EC were transduced with an expression vector for the APEX1 peptide or an empty control vector and treated with LPS. Following RNA deep sequencing, genes upregulated in LPS-treated EC expressing the APEX1 peptide were identified bioinformatically. Selected candidates were validated by semi-quantitative real time PCR, a promising one was Selenoprotein T (SELENOT). For functional analyses, an expression vector for SELENOT was generated. To study the effect of SELENOT expression on LPS-induced EC activation and apoptosis, the SELENOT vector was transfected in EC. Immunostaining showed that SELENOT was expressed and localized in the ER. EC transfected with the SELENOT plasmid showed no activation and reduced apoptosis induced by LPS. SELENOT as well as APEX1(1-20) can protect EC against activation and apoptosis and could provide new therapeutic approaches in the treatment of sepsis.

Keywords: APEX1(1-20); Selenoprotein T; lipopolysaccharide; endothelial cell activation; apoptosis

1. Introduction

Sepsis can best be described as an overwhelming inflammatory condition, in which the body responds to an infection in a hyperactive, dysregulated way, which in turn results in life-threatening organ dysfunction and eventually septic shock. According to an estimate of the World Health Organization (WHO), sepsis affects more than 48 million people every year, potentially leading to 11 million deaths [1]. The basis for the pathophysiological responses in the context of sepsis is multifactorial. Therefore, except for the introduction of vasopressor agents 40 years ago, no new therapeutic principle for the treatment of sepsis has been developed until today.

Lipopolysaccharide (LPS) is an outer membrane component of Gram-negative bacteria. Most bacterial LPS molecules are thermostable and generate a pro-inflammatory stimulus for the immune system in humans. LPS is a serologically reactive bacterial toxin, and 1 to 2 mg in the bloodstream can be lethal. LPS can enter the bloodstream through intestinal absorption of the LPS produced by gut bacteria. Moreover, gut lesions and diet rich in lipids boost the transport across membranes into the systemic circulation [2]. Therefore, at the cellular level, endothelial cells (EC) are directly affected by LPS, which triggers their activation and ultimately apoptosis, leading to vascular leakage. Thus, it is undisputable that the loss of endothelial cell integrity is a mainstay of septic shock [3]. Hence, therapies that could prevent endothelial cell leakage or even restore endothelial cell integrity would be of tremendous value for patients and would address medical needs. EC with LPS affect the endothelial transcriptome by regulating the levels of numerous transcripts, not only of protein coding RNAs, but also of non-coding RNAs such as microRNAs and long non-coding RNAs [4,5]. Having pointed this out, it is a mystery to us that we failed to find any RNA deep sequencing data on LPS-induced transcriptome changes in the endothelium in the established databases such as the Gene expression omnibus (GEO), the European nucleotide archive (ENA), Short Sequence Archive (SRA) or ArrayExpress. However, such an in-depth transcriptome profiling combined with pathway analyses could provide novel targets for the development of new therapeutic principles for the treatment of sepsis, especially for protecting the endothelium. Therefore, one aim of this study was to perform a deep sequencing analysis in LPS-treated primary EC.

Moreover, we have recently shown that the first 20 amino acids of the Apurinic/Apyrimidinic Endodeoxyribonuclease 1 (APEX1) are sufficient to inhibit H_2O_2-induced apoptosis [6]. As the underlying molecular mechanisms initiating apoptosis are independent of the trigger, we hypothesized that this N-terminal APEX1 peptide, APEX1(1-20), could also interfere with LPS-induced apoptosis. Therefore, we included cells expressing the APEX1(1-20) in our deep sequencing analysis to find potential therapeutic targets for sepsis, possibly regulated by this peptide.

2. Materials and Methods

2.1. Cultivation of Primary Human Endothelial Cells and HEK293

Primary human endothelial cells (EC) were obtained from LONZA (Cologne, Germany) and cultured as previously described [7]. In detail, EC were cultured in endothelial basal medium supplemented with 1 μg/mL hydrocortisone, 12 μg/mL bovine brain extract, 50 μg/mL gentamicin, 50 ng/mL amphotericin B, 10 ng/mL epidermal growth factor (LONZA, Cologne, Germany) and 10% fetal bovine serum until the third passage. After detachment with trypsin, cells were grown for at least 20 h. All experiments were performed with EC in passage 3. HEK293 were cultured in DMEM GlutaMAX™ supplemented with 10% heat-inactivated fetal bovine serum and 1% penicillin/streptomycin and then used for the production of lentiviruses.

2.2. Lentiviral Production and Transduction of EC

Generation of VSV-G pseudotyped lentiviral particles and transduction of EC were performed as previously described [8]. Lentiviral titers were determined with the QuickTiter™ Lentivirus Titer kit (Lentivirus-Associated HIV p24) (Biocat, Heidelberg, Germany). EC were transduced with a multiplicity of infection of approximately 20. The day after transduction the cells were washed three times, the medium replaced, and cells cultivated for another day before they were treated with 150 ng/mL LPS for 18 h.

2.3. Isolation of Total Cellular RNA

Cells were lysed using TRIzol® (Thermo Fisher Scientific, Dreieich, Germany) and RNA was isolated according to the manufacturer's instructions. RNAs were subjected to a second purification step using the RNeasy® Mini kit (Qiagen, Hilden, Germany). RNA concentrations were measured using a NanoDrop™ 2000c (Thermo Fisher Scientific, Dreieich, Germany), and RNA integrity and purity were determined by agarose gel electrophoresis.

2.4. RNA Sequencing and Bioinformatic Analysis

RNA sequencing data were obtained from quadruplicate total RNA samples. Total RNAs used for transcriptome analyses were quantified using the Qubit^{TM} RNA HS Assay kit (Thermo Fisher Scientific, Dreieich, Germany) and quality was determined by capillary electrophoresis using the FragmentAnalyzer and the Total RNA Standard Sensitivity Assay (Agilent Technologies, Santa Clara, CA, USA). All samples in this study showed highest RNA Quality Numbers (RQN 10.0). Library construction and sequencing were performed at the Genomics and Transcriptomics Laboratory at the Biological Medical Research Centre (BMFZ) of the Heinrich-Heine University Düsseldorf. Library preparation was performed according to the manufacturer's protocol using the TruSeq Stranded mRNA Assay kit (Illumina, San Diego, CA, USA). Briefly, 500 ng total RNA was used for mRNA capturing, fragmentation, synthesis of cDNA, adapter ligation and library amplification. Bead purified libraries were normalized and finally sequenced on the HiSeq 3000 system (Illumina San Diego, CA, USA) with a read setup of 1 × 150 bp. The bcl2fastq2 (version 2.17.1.4) tool was used to convert the bcl files to fastq files as well for adapter trimming and demultiplexing. GC-content, base-calling quality, adapter content and read length were measured using the tool FASTQC by Andrews (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed on 17 August 2021) and MultiQC [9]. Reads were then trimmed or discarded based on their base calling quality and adapter content with Trimmomatic version 0.36 [10]. Subsequently, with the help of the SortMeRNA algorithm version 2.1b [11], the extent of rRNA depletion was measured by mapping the reads to rRNA databases. For alignment and the following analyses, the human genomic reference sequence (GRCh38) and annotation data (release 101) were downloaded from Ensembl [12] and BioMart [13]. For splice site usage analysis, the reads were then aligned to the human reference genome using the two-pass mapping protocol of the STAR aligner (2.5.4b) [14]. With help of the SAMtools software package [15], uniquely mapped reads were selected for creation of a gap table, listing the coordinates of every gap found in the alignment of the reads and the number of overlapping reads. For DGE analysis with the R package DESeq2 version 1.18.1 [16], count matrices were generated using the software salmon version 0.9.1 [17]. Significantly enriched gene sets were calculated, using the R package GOseq [18]. Scripts used for this work are publicly available at https://github.com/caggtaagtat/SELENOT (accessed on 17 August 2021). FASTQ file preparation and alignment were accomplished using custom BASH shell scripts in the environment of the High Performing Cluster of the Heinrich-Heine University Düsseldorf.

2.5. cDNA Synthesis

Total cellular RNA was reverse transcribed using the QuantiTect Reverse Transcription kit (Qiagen, Hilden Germany) according to the manufacturer's instructions.

2.6. Polymerase Chain Reaction (PCR)

Endpoint PCRs were performed with MyTaq™ HS DNA Polymerase (Biocat, Heidelberg, Germany) according to manufacturer's recommendations in a Bio-Rad T100 Thermal Cycler (Bio-Rad, Feldkirchen Germany). Reaction products were resolved on standard agarose gels.

Relative transcript levels were determined by semi-quantitative real-time PCR using cDNA as a template and the primaQUANT 2x qPCR-SYBR-Green-MasterMix (Steinbrenner, Wiesenbach, Germany), the transcript for the ribosomal protein L32 (RPL32), served as a reference. The PCR reactions were performed in a Rotor-Gene Q instrument (Qiagen, Hilden, Germany). Relative expression was calculated by the ΔC_t method [19].

The sequences of all primer used for PCR are listed in Supplementary Table S1.

2.7. Plasmids

A lentiviral expression vector for the first twenty amino acids of APEX1 was constructed by transferring the coding sequence for APEX1(1-20) with a C-terminal myc-tag from the previously published expression vector [6] into a lentiviral transfer vector, in which the transgene is expressed under the transcriptional control of the cytomegalovirus immediate early promoter/enhancer [8]. To generate an expression vector for human SELENOT with an N-terminal FLAG-tag, the SELENOT coding sequence together with the first 179 bp of the 3′-untranslated region of the human SELENOT gene containing the selenocysteine insertion sequence were amplified from a human EC cDNA using Q5® High-Fidelity DNA Polymerase (New England Biolabs, Frankfurt, Germany). This fragment was inserted into pFLAG-CMV-2 (Sigma-Aldrich, Deisenhofen, Germany) opened with Not I and Xba I using the Gibson Assembly® Cloning kit (New England Biolabs, Frankfurt, Germany) according to the manufacturer's protocol. The construct was verified by DNA sequencing. Cloning details and the complete plasmid sequence are available upon request.

2.8. Transient Transfection of EC

Transient transfections of EC with plasmid DNA were performed using Superfect (Qiagen, Hilden, Germany) as previously described [20,21]. In detail, EC were transfected on 6 cm culture dishes with 3 µg plasmid DNA and 22.5 µL Superfect, or in 6-well plates with 1.2 µg plasmid DNA and 12 µL Superfect per well.

2.9. Immunostaining of EC

EC were fixed and permeabilized as described previously [7]. Afterwards, cells were incubated with an anti-FLAG-tag antibody (1:100, DYKDDDDK Tag Antibody (clone 8H8L17), Abfinity™, Cat. No. 701629, Invitrogen, Darmstadt, Germany). As secondary antibody, a goat anti-rabbit highly cross-adsorbed antibody coupled to Alexa Fluor 594 (1:500, Cat. No. A-11012, Invitrogen, Darmstadt, Germany) was used. For ICAM1 staining, an Alexa Fluor 488-coupled primary antibody (1:50, ICAM1/CD54 (15.2), Cat. No. SC-107 AF488, Santa Cruz Biotechnology, Heidelberg, Germany) was used. The endoplasmic reticulum (ER) was stained with an anti-Calnexin (clone C5C9) Alexa Fluor 488-conjugate (1:25, Cat. No. 38552, Cell Signaling, Technology, Frankfurt, Germany). Nuclei were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) (100 ng/mL, Sigma-Aldrich, Deisenhofen, Germany). Images were taken using Zeiss microscopes (Axio Observer D1 or Axio Imager M2, magnification 400-fold, oil).

2.10. Immunoblotting

Cells were detached from the culture surface with a rubber policeman, centrifuged at 800× *g*, resuspended in radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1% (*v*/*v*) IGEPAL®-CA630, 0.1% (*w*/*v*) SDS and 0.5% (*w*/*v*) Na-Deoxycholate) supplemented with 1/100 volume of a protease inhibitor cocktail (Bimake, Munich, Germany) and lysed for 30 min on ice. The lysates were centrifuged at 18,000× *g* and 4 °C for 15 min and the supernatant was transferred to a fresh tube. Lysate proteins were separated by sodium-dodecyl-sulfate polyacrylamide gel electrophoresis according to standard procedures and transferred onto polyvinylidene difluoride membranes. After blocking with 5% milk powder in TBS (200 mM Tris-HCl pH 8.0, 300 mM NaCl, 100 mM KCl) with 0.1% (*v*/*v*) Tween-20 for 1 h at room temperature, membranes were incubated with an antibody directed against Caspase 3 (1:300 for detection of cleaved Caspase 3, 1:500 for uncleaved Caspase 3, Cat. No. 9662, Cell Signaling Technology, Frankfurt, Germany) and an anti α-Tubulin antibody (clone (DM1A), 1:50,000, Cat. No. T9026, Sigma-Aldrich, Deisenhofen, Germany). Antibodies were incubated overnight at 4 °C. The following day, membranes were incubated with secondary antibodies coupled to horseradish peroxidase (ECL™ Anti-Rabbit or Anti-Mouse IgG, Horseradish Peroxidase linked whole antibody (from sheep), 1:5000; Cat. Nos. NA934V and NA931V, GE healthcare, Solingen, Germany). Detection was performed using ECL substrate (GE healthcare, Solingen, Germany) and X-ray films. Semi-quantitative analyses were performed on scanned X-ray films using Fiji [22].

2.11. Statistics

The number of experiments (n) given in the figure legends represents independent biological replicates, the data shown are mean ± SEM. Normal distribution for all data sets was confirmed by a Shapiro–Wilk test; homogeneity of variances (from means) between groups was verified by Levene's test. Multiple comparisons were performed using one-way ANOVA with post-hoc Tukey LSD test.

3. Results

3.1. APEX1(1-20) Induces Specific Transcriptome Changes in EC in Response to LPS

To identify APEX1(1-20)-mediated transcriptome differences in the response of EC to LPS we performed RNA deep sequencing. For this purpose, primary human EC were transduced with either a lentiviral vector leading to moderate expression of APEX1(1-20) or an empty vector, respectively. Cells were then treated with 150 ng/mL active LPS or detoxified LPS as control. RNA from these cells was used for RNA deep sequencing and analyzed for differential gene expression (DGE). To identify APEX1(1-20)-specific transcriptome changes in response to LPS, we analyzed which genes were specifically regulated by LPS in the APEX1(1-20) expressing cells, but not in the cells transduced with the empty vector.

PCA analysis revealed that all samples from the cells treated with detoxified LPS cluster together, no matter whether the cells expressed APEX1(1-20) or not. The same held true for the LPS-treated cells (Supplementary Figure S1), showing a clear effect of LPS on the cellular transcriptome.

DGE analysis revealed that the APEX1(1-20) transcript derived from the expression vector was only detectable in the cells transduced with this vector, but not in the cells transduced with the empty vector. More importantly, expression of APEX1(1-20) alone did not appear to affect the overall transcriptome as changes in the expression of only a very small number of genes were observed (Supplementary Table S2).

In addition to the DGE analysis, we performed a gene set enrichment analysis (GSEA) focusing on genes that were significantly regulated by LPS exclusively in either the cells transduced with the empty vector or the cells expressing APEX1(1-20). As expected, cells transduced with the empty vector showed a significant enrichment of upregulated genes belonging to gene ontology (GO) terms related to immune responses including the response to bacteria and tumor necrosis factor signaling (Supplementary Table S3). Interestingly, we did not observe these changes in the presence of APEX1(1-20), and, moreover, genes belonging to the GO term cellular response to tumor necrosis factor were significantly downregulated in LPS-treated cells expressing the APEX1 peptide (Supplementary Table S4). These data support the assumption that APEX1(1-20) might provide protection against endothelial cell activation and apoptosis via alteration of the transcriptional responses to LPS treatment.

In the DGE analysis, we found that after LPS treatment, 323 genes were significantly upregulated in cells transduced with the empty vector and 280 were downregulated. In contrast, in the cells expressing APEX1(1-20), only 177 genes were upregulated by LPS and 139 genes were downregulated (Figure 1A,B and Supplementary Tables S5 and S6). Thus, the expression of only roughly half as many genes appeared to be affected by the presence of APEX1(1-20).

Notably, we observed clearly different LPS responses in cells expressing APEX1(1-20) when compared to cells transduced with the empty vector (Figure 1C–F and Supplementary Tables S7–S10).

For functional studies, we focused on genes whose expression is upregulated by LPS only in cells expressing the small APEX1 peptide as the corresponding proteins might evoke APEX1(1-20)-dependent protective effects in EC, which could be of interest in a therapeutic setting.

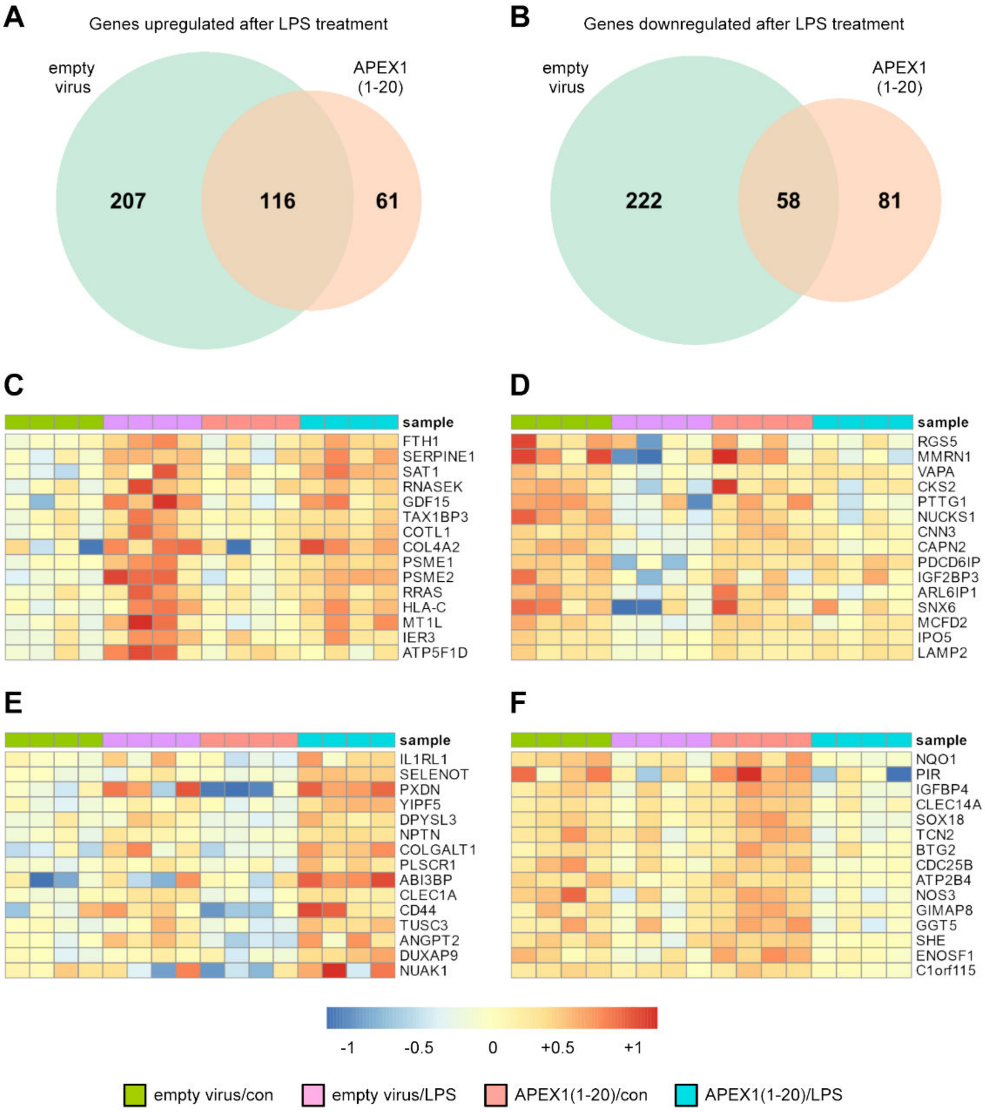


Figure 1. APEX1(1-20) induces specific transcriptome changes in EC in response to LPS. (**A**–**F**) EC were transduced with a lentiviral expression vector for APEX1(1-20) or an empty virus and treated with 150 ng/mL detoxified (con) or active LPS (LPS) for 18 h. RNAs from the transduced cells were subjected to RNA deep sequencing. Differential gene expression was calculated using the R package DESeq2. Wald test from DESeq2 was used to calculate the significance of the change in the expression. (**A**,**B**) Venn-diagrams for genes upregulated (**A**) or downregulated (**B**) after LPS treatment of empty virus transduced cells and cells expressing APEX1(1-20). (**C**–**F**) Heatmaps of genes significantly differentially expressed upon LPS treatment of cells transduced with the empty virus or cells expressing APEX1(1-20). Shown are the 15 top ranked genes from Supplementary Tables S7–S10. The color depicts the normalized expression relative to the respective mean in all samples. (**C**) Genes uniquely upregulated after LPS treatment of empty virus transduced cells. (**D**) Genes uniquely downregulated after LPS treatment of empty virus transduced cells. (**E**) Genes uniquely upregulated after LPS treatment of APEX1(1-20) expressing cells. (**F**) Genes uniquely downregulated after LPS treatment of APEX1(1-20) expressing cells.

3.2. Expression of PXDN and SELENOT Is Specifically Upregulated after LPS Treatment of EC Expressing APEX1(1-20)

As a prerequisite for functional studies, we first validated the regulation of the top-ranked candidates, which, according to the RNA sequencing data, should be expressed to levels allowing reliable detection and quantification.

IL1RL encodes an Interleukin 1 Receptor-like protein, which belongs to a family of ten distinct but structurally related receptors. These proteins serve either as ligand binding or accessory chains and some act as signaling inhibitors. Moreover, two members of this family are orphan receptors [23]. Therefore, IL1RL1 is part of a complex signaling network and one could easily envision that—due to this redundancy—interference with this network might be compensated or evoke unwanted side effects.

Peroxidasin (PXDN), originally described as Vascular Peroxidase 1, is a heme-containing peroxidase, which shows highest expression in the heart and the vascular wall [24]. The protein is rapidly secreted [25] and required for formation of the vascular basement membrane by reinforcing fibrillar network assembly in the extracellular matrix through formation of sulfilimine bonds [26]. It has recently been shown that PXDN promotes angiogenesis [27] and, furthermore, is essential for endothelial cell survival [28].

Selenoprotein T (SELENOT) is a member of the selenoprotein family, whose members are characterized by containing one or more selenocysteine residues, frequently in enzymatically active sites [29]. SELENOT is the most highly conserved selenoprotein throughout evolution [30], suggestive of an essential function, which is underscored by the early embryonic lethality of mice in which the *selenot* gene is constitutively disrupted [31]. SELENOT is one of 7 out of 25 human selenoproteins localized to the ER [32]. The expression of SELENOT, like all other selenoproteins, depends on dietary selenium as shown by a reduced expression in chicken stomach after 55 days on a selenium-deficient diet. Moreover, this regimen resulted in stress injuries [33]. In addition, SELENOT protects kidney cells against cisplatin-induced apoptosis [34]. These observations go along with the notion that ER-resident selenoproteins are critical in cellular stress responses [35].

For the reasons explained above, we did not follow up on IL1RL, but validated the regulation of PXDN and SELENOT by semi-quantitative real-time PCR (Figure 2).

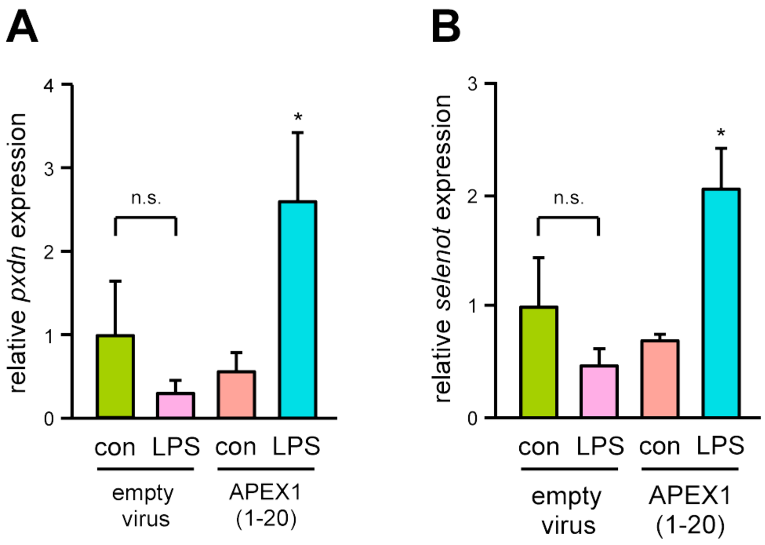


Figure 2. LPS induces upregulation of PXDN and SELENOT expression specifically in EC expressing APEX1(1-20). Transcript levels of PXDN (**A**) and SELENOT (**B**) in EC transduced with an empty virus or the expression vector for APEX1(1-20) and treated with 150 ng/mL detoxified (con) or active LPS (LPS) for 18 h were analyzed by semi-quantitative real-time PCR; RPL32 served as reference (data are mean ± SEM, $n = 4$, * $p < 0.05$ vs. APEX1(1-20)/con, n.s. = not significant, one-way ANOVA with post-hoc Tukey LSD test).

The real-time PCR analysis corroborated the deep sequencing data as for both genes, an upregulation of the transcript level by LPS was only observed in the cells expressing APEX1(1-20). Although PXDN has already been characterized with respect to protective functions in EC [28], these data provide independent proof for the validity of the experimental approach. The second protein, SELENOT, for which no functions in endothelial activation and apoptosis have been described so far, was chosen for functional analyses.

3.3. Generation of a SELENOT Expression Vector and Intracellular Localization of the Overexpressed Protein

To study the impact of SELENOT on endothelial cell functions affected by LPS, we generated an expression vector, which contained a FLAG-epitope tag allowing the identification of the overexpressed protein. For the generation of this expression vector, an aspect unique to selenoproteins had to be taken into account. Selenocysteine (Sec) residues in selenoproteins are not the product of a post-translational modification, but are rather incorporated already during translation by using one of the translation termination codons, namely UGA, for binding of the selenocysteine tRNA (tRNA^{Sec}) to the mRNA. This translational recoding of the UGA codon involves a so-called selenocysteine insertion sequence (SECIS) in the 3′-untranslated region (UTR) of the transcript. The SECIS, which is not highly conserved on the sequence level, forms a stem-loop structure that is required for recruitment of the tRNA^{Sec} to the UGA codon [36]. Consequently, the lack of a SECIS leads to premature translation termination, when the ribosome encounters the first UGA within the open reading frame. Therefore, we included—besides the SELENOT open reading frame—a portion of the SELENOT 3′-UTR including the SECIS in the expression vector.

We first analyzed the expression of FLAG-SELENOT after transient transfection of EC on the RNA level by reverse transcriptase PCR (Figure 3A). We then determined the intracellular localization of the overexpressed FLAG-SELENOT protein by immunofluorescence. As demonstrated by colocalization with the ER-resident protein Calnexin (Figure 3B), FLAG-SELENOT was localized in the ER.

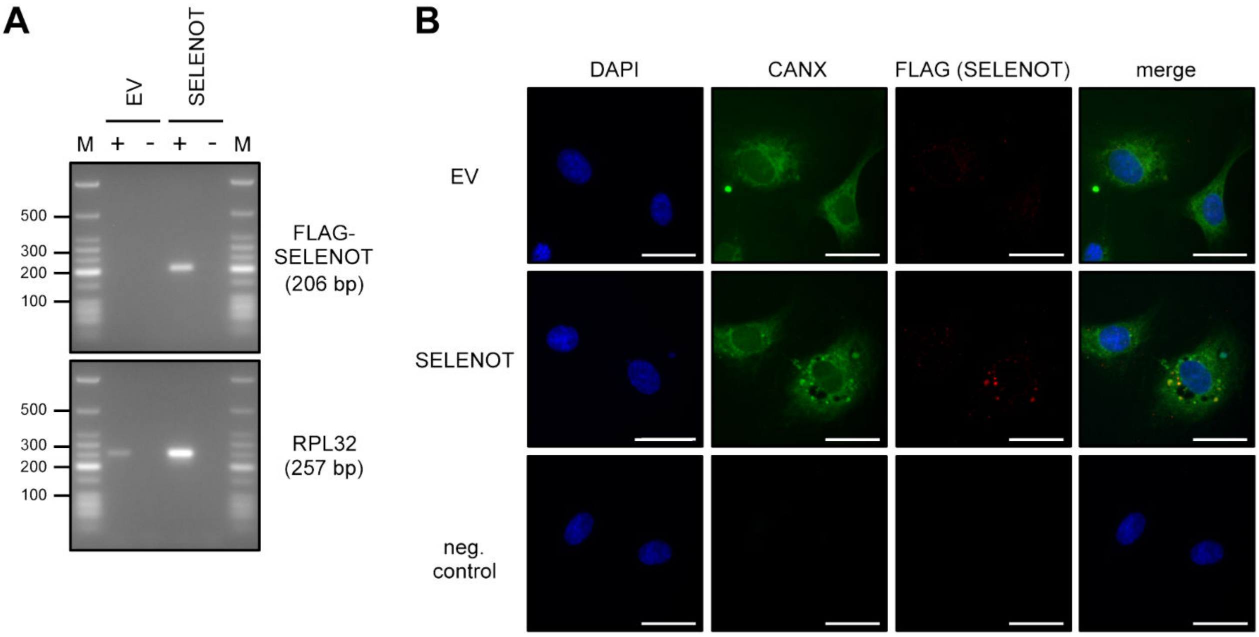


Figure 3. Overexpressed SELENOT is localized in the ER. (**A**,**B**) EC were transfected with the FLAG-SELENOT expression vector (SELENOT) or an empty vector (EV). Expression and localization of exogenously expressed SELENOT was verified on the RNA (**A**) and protein (**B**) level. (**A**) Expression of SELENOT was analyzed by reverse transcription polymerase chain reaction (RT-PCR). Therefore, RNA was isolated from the transfected cells and cDNA was synthesized in the presence (+) or absence (−) of reverse transcriptase. Amplification was performed with primers specifically detecting the FLAG-SELENOT fusion transcript, the housekeeping gene RPL32 served as control. Amplification products were resolved by agarose gel electrophoresis, the expected fragment sizes are specified, numbers on the left indicate selected DNA size markers (M). (**B**) Localization of FLAG-SELENOT was examined by immunostaining and fluorescence microscopy. Cells were stained with an antibody directed against Calnexin (CANX), a marker for the ER (green) and an anti-FLAG antibody (red). Nuclei were counterstained with DAPI (blue); merge is the overlay of all channels (scale bar = 30 μm).

3.4. SELENOT Overexpression Inhibits LPS-Induced Endothelial Cell Activation

Having demonstrated that FLAG-SELENOT is localized in the ER, we next investigated the effect of SELENOT on LPS-induced endothelial cell activation. Therefore, FLAG-SELENOT was expressed in EC as before. After treatment with 150 ng/mL LPS for 18 h, ICAM1—a marker for endothelial cell activation—was detected. As expected, LPS upregulated ICAM1 protein levels in empty vector transfected EC. This upregulation was completely inhibited in cells, in which SELENOT is overexpressed (Figure 4).

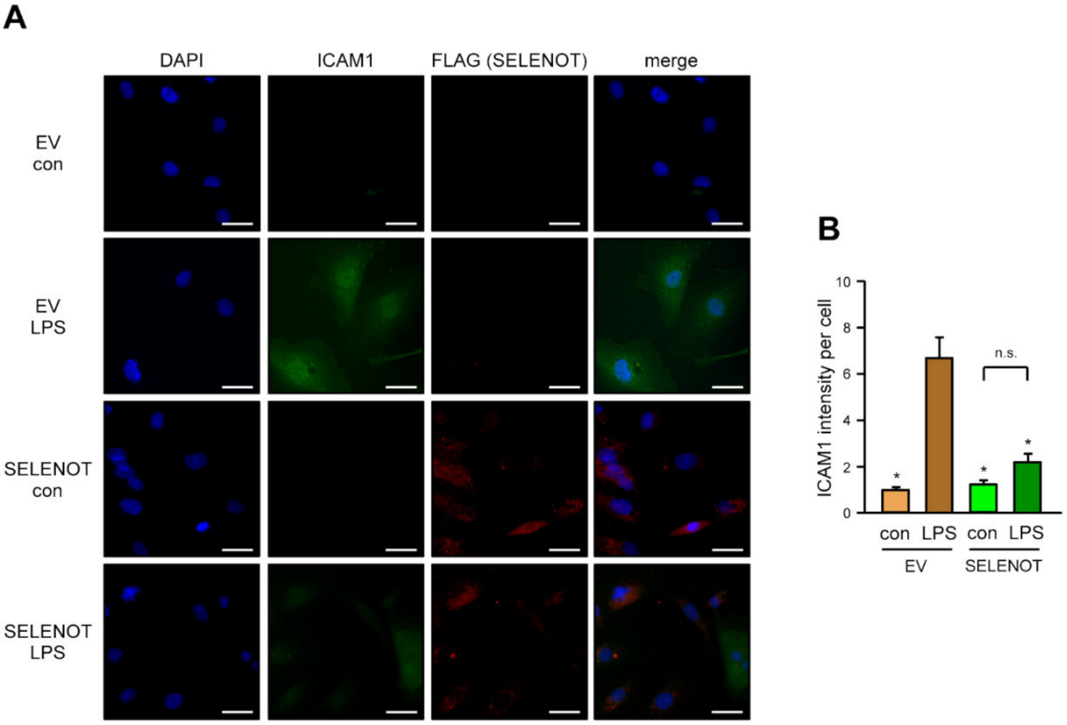


Figure 4. SELENOT suppresses LPS-induced upregulation of ICAM1. (**A**,**B**) EC were transfected with the FLAG-SELENOT expression vector (SELENOT) or an empty vector (EV) and treated with 150 ng/mL detoxified (con) or active LPS (LPS) for 18 h. FLAG-SELENOT and ICAM1 were detected by immunofluorescence. Cells were stained with an antibody directed ICAM1 (green) and an anti-FLAG antibody (red). Nuclei were counterstained with DAPI (blue); merge is the overlay of all channels. (**A**) Representative immunostaining (scale bar = 30 µm). (**B**) Quantitation of ICAM1 levels. The intensity of the green fluorescence per cell was measured using Fiji; in the cells transfected with the SELENOT expression vector, only FLAG-SELENOT positive cells were included (data are mean ± SEM, *n* = 4, * *p* < 0.05 vs. EV/LPS, n.s. = not significant, one-way ANOVA with post-hoc Tukey LSD test).

3.5. SELENOT Overexpression Inhibits LPS-Induced Endothelial Cell Apoptosis

Besides endothelial cell activation, LPS also induces apoptosis of EC [37]. Therefore, we determined Caspase 3 cleavage as a marker for apoptosis in EC. As for ICAM1, LPS increased Caspase 3 cleavage in cells not expressing SELENOT. On the contrary, overexpression of SELENOT completely blunted apoptosis induction by LPS (Figure 5).

In conclusion, SELENOT, which is upregulated by LPS in EC expressing APEX1(1-20), seems to be an important mediator of the protective effects of APEX1(1-20) and could thus be of interest as an adjuvant therapeutic agent in endotoxemia.

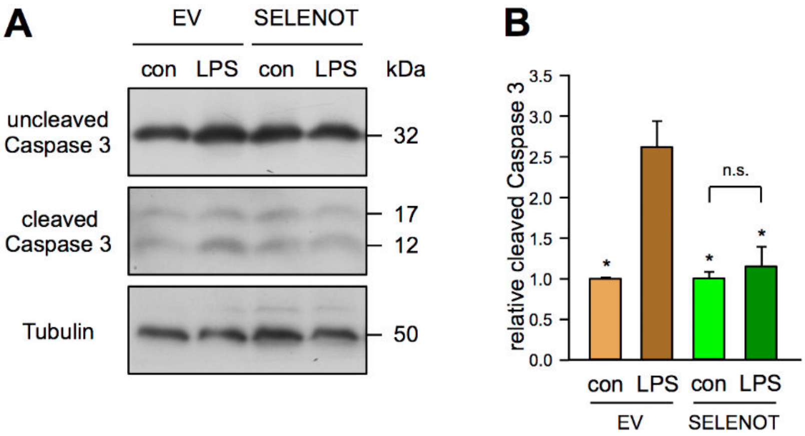


Figure 5. SELENOT suppresses apoptosis induction by LPS. (**A**,**B**) EC were transfected with the FLAG-SELENOT expression vector (SELENOT) or an empty vector (EV) and treated with 150 ng/mL detoxified (con) or active LPS (LPS) for 18 h. Uncleaved and cleaved Caspase 3 were detected by immunoblot, Tubulin served as loading control. (**A**) Representative immunoblot. (**B**) Semi-quantitative analysis of relative amounts of cleaved Caspase 3 (data are mean ± SEM, $n = 4$, * $p < 0.05$ vs. EV/LPS, n.s. = not significant, one-way ANOVA with post-hoc Tukey LSD test).

4. Discussion

The major findings of the present study are the first RNA deep sequencing analysis of LPS-induced changes in primary human EC and the identification of a protective role of APEX1(1-20) and SELENOT in LPS-induced endothelial cell activation and apoptosis.

With respect to the possibility of using an APEX1(1-20) peptide or a related small molecule as a therapeutic agent, it has to be noted that APEX1(1-20) does not change the transcriptome when compared to empty virus transduced cells. Thus, there is no evidence of potential side effects induced by APEX1(1-20) in the endothelium. As expected, LPS treatment induced typical pathways known in sepsis. Those upregulated genes upon LPS treatment in cells not expressing APEX1(1-20) are found, for example, under the GO terms cellular response to tumor necrosis factor, tumor necrosis factor-mediated signaling pathway, and plasma membrane (Supplementary Table S3). It has been known for years that tumor necrosis factor induces endothelial cell activation [38] and apoptosis [39]. Therefore, activation of those pathways is a typical answer of the endothelium to LPS, which in turn leads to loss of endothelial integrity and barrier function. Loss of endothelial cell integrity is a mainstay of septic shock [3], because LPS can enter the systemic circulation destroy endothelial cell integrity, thereby leading to multiple organ failure. Thus, an additional therapy protecting the integrity of the endothelium would be of tremendous interest. Interestingly, APEX1(1-20) leads to reduced responses of the tumor necrosis factor pathways (Supplementary Table S4). Hence, APEX1(1-20) or its downstream targets could be of interest as potential therapeutic options. Therefore, we specifically focused on those targets induced by APEX1(1-20) in the presence of LPS in EC to identify potential candidates. Indeed, we found SELENOT to be upregulated upon APEX1(1-20).

SELENOT is an ER-resident selenoprotein, which is associated with the ER membrane and required to maintain ER redox homeostasis. It is needed to cope with intracellular stress conditions and is one of the most important selenoproteins [30].

As mentioned before, the expression of all selenoproteins depends on selenium. However, there seems to be a hierarchy in the sensitivity of different selenoproteins with respect to selenium levels and SELENOT seems to respond more avidly to selenium depletion than several other proteins of this family [40]. It has been estimated that up to one in seven people worldwide have a low dietary selenium intake [41] and it is clear that proper endothelial functionality depends on an adequate selenium supply [42]. Even more interesting is the observation that selenium serum levels are dramatically reduced in critically ill patients with sepsis [43]. Therefore, selenium supplementation seems to be an obvious supplementary treatment option for sepsis and possibly the protection of the endothelium in this disease. In this context, it is interesting to note that selenium pretreatment or supplementation alleviates some of the deleterious effects of LPS. In the murine macrophage cell line RAW264.7, LPS induced immunological stress as shown by the upregulation of multiple inflammation-related genes. This was accompanied by a reduction in the relative *selenot* mRNA level. Pretreatment with selenium partially rescued this downregulation and had only a very modest effect on the expression of the inflammation-related genes [44]. In mice, LPS-induced myocardial dysfunction, oxidative stress and apoptosis in the heart could be attenuated when the animals were put on a selenium-supplemented diet 2 weeks prior to LPS treatment [45]. Again, this pretreatment did not completely restore heart functionality or prevent oxidative stress and apoptosis induction evoked by LPS. Our experiments did not show a significant downregulation of *selenot* expression in LPS-treated EC, although there seems to be a trend in this direction. On the contrary, the cells expressing APEX1(1-20) showed an upregulation of *selenot* RNA levels of approximately threefold after LPS treatment. This clearly indicates that the small APEX1 peptide can convey a protective outcome, which is much stronger than the effects observed with selenium supplementation or pretreatment.

Up to now, the precise molecular functions of SELENOT have not been elucidated. Nevertheless, a peptide derived from SELENOT has already been used in animal models. Rocca et al. demonstrated that this SELENOT-derived peptide—including the active catalytic site corresponding to the sequence FQICVSUGYR—applied after ischemia and prior to reperfusion is able to protect the heart from ischemia/reperfusion injury. This protection was attributed to a reduction in oxidative stress and inhibition of apoptosis [46]. This is in accordance with our study presented here, in which we demonstrate that SELENOT completely inhibited LPS-induced activation and apoptosis in human primary EC.

The same peptide was applied in a cell-permeable form in a mouse model for Parkinson's disease, where it protected dopaminergic neurons. This effect was also associated with reduced oxidative stress and Caspase 3 activity [47].

Based on the protective effects of this SELENOT peptide in such different organs as the brain and the heart, it is conceivable that it could exert its protective functions also in the vasculature in the setting of sepsis.

Given the high numbers of patients and the up to 11 million deaths per year due to sepsis, a protection of the endothelium as an additional additive therapy could be of tremendous importance. The metabolic response to sepsis entails the rapid breakdown of intracellular reserves of proteins, carbohydrates and fat. This is accompanied by an increase in ER stress. An increase in SELENOT or application of a peptide could dampen this stress and maintain the ER homeostasis, counteracting the overshooting responses of the body to sepsis.

5. Conclusions

In conclusion, our data presented here suggest that APEX1(1-20) and SELENOT are promising therapeutic options for the treatment of sepsis to protect the endothelium and thus, to prevent endothelial cell leakage or even to restore endothelial cell integrity. This would be of tremendous value for patients and would potentially lower the numbers of septic shock, multiple organ failure and deaths.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/antiox10091427/s1, Figure S1: Principal component analysis, Table S1: Primer pairs used for endpoint PCR and semi-quantitative real-time PCR. Primer pairs used for endpoint PCR and semi-quantitative real-time PCR, Table S2: Differential gene expression analysis for genes regulated byexpression of APEX1(1-20), Table S3: Overrepresented GO terms in genes upregulated by LPS exclusively in cells not expressing APEX1(1-20), Table S4: Overrepresented GO terms in genes downregulated by LPS exclusively in cells expressing APEX1(1-20), Table S5: Differentially expressed genes upon LPS treatment of cells not expressing APEX1(1-20), Table S6: Differentially expressed genes upon LPS treatment of cells expressing APEX1(1-20), Table S7: Genes upregulated by LPS exclusively in cells that do not express APEX1(1-20), Table S8: Genes downregulated by LPS exclusively in cells that do not express APEX1(1-20), Table S9: Genes upregulated by LPS exclusively in cells that express APEX1(1-20), Table S10: Genes downregulated by LPS exclusively in cells that express APEX1(1-20).

Author Contributions: Conceptualization, H.S., N.A.-A., J.A. and J.H.; Data curation, H.S., N.A.-A., J.A. and J.H.; Formal analysis, D.M., J.P., P.J. and P.K.; Funding acquisition, H.S. and J.A.; Investigation, D.M., J.P., P.J., F.v.A., J.G., P.K., K.S., O.E. and N.A.-A.; Methodology, D.M., J.P., P.J. and P.K.; Project administration, H.S., J.A. and J.H.; Resources, H.S., J.A. and J.H.; Supervision, P.J., H.S., N.A.-A., J.A. and J.H.; Validation, D.M., J.P., P.J. and P.K.; Visualization, D.M., J.P. and N.A.-A.; Writing—original draft, D.M. and J.P.; Writing—review and editing, H.S., N.A.-A., J.A. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG), IRTG1902-1 and IRTG1902-2 to J.A. and J.H. D.M., J.G. and K.S. are stipend holders of the IRTG1902. This research was also in part funded by Bayer Grant4Target (2019-08-2427) to J.H. and by intramural funding (9772727) of the University Hospital of the Heinrich-Heine University to H.S. and J.H.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article. The RNA sequencing data used to support the findings of this study have been deposited at ArrayExpress under accession number E-MTAB-10936 (https://www.ebi.ac.uk/arrayexpress/09/2021, accessed on 17 August 2021).

Acknowledgments: Computational support and infrastructure were provided by the Center for Information and Media Technology (ZIM) at the Heinrich-Heine University Düsseldorf.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rudd, K.E.; Johnson, S.C.; Agesa, K.M.; Shackelford, K.A.; Tsoi, D.; Kievlan, D.R.; Colombara, D.V.; Ikuta, K.S.; Kissoon, N.; Finfer, S.; et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *Lancet* **2020**, *395*, 200–211. [CrossRef]
2. Boutagy, N.E.; McMillan, R.P.; Frisard, M.I.; Hulver, M.W. Metabolic endotoxemia with obesity: Is it real and is it relevant? *Biochimie* **2016**, *124*, 11–20. [CrossRef]
3. Joffre, J.; Hellman, J.; Ince, C.; Ait-Oufella, H. Endothelial Responses in Sepsis. *Am. J. Respir. Crit. Care Med.* **2020**, *202*, 361–370. [CrossRef]
4. Zhao, B.; Bowden, R.A.; Stavchansky, S.A.; Bowman, P.D. Human endothelial cell response to gram-negative lipopolysaccharide assessed with cDNA microarrays. *Am. J. Physiol. Cell Physiol.* **2001**, *281*, C1587–C1595. [CrossRef]
5. Ho, J.; Chan, H.; Wong, S.H.; Wang, M.H.; Yu, J.; Xiao, Z.; Liu, X.; Choi, G.; Leung, C.C.; Wong, W.T.; et al. The involvement of regulatory non-coding RNAs in sepsis: A systematic review. *Crit. Care* **2016**, *20*, 383. [CrossRef]
6. Dyballa-Rukes, N.; Jakobs, P.; Eckers, A.; Ale-Agha, N.; Serbulea, V.; Aufenvenne, K.; Zschauer, T.C.; Rabanter, L.L.; Jakob, S.; von Ameln, F.; et al. The Anti-Apoptotic Properties of APEX1 in the Endothelium Require the First 20 Amino Acids and Converge on Thioredoxin-1. *Antioxid. Redox Signal.* **2017**, *26*, 616–629. [CrossRef]
7. Ale-Agha, N.; Goy, C.; Jakobs, P.; Spyridopoulos, I.; Gonnissen, S.; Dyballa-Rukes, N.; Aufenvenne, K.; von Ameln, F.; Zurek, M.; Spannbrucker, T.; et al. CDKN1B/p27 is localized in mitochondria and improves respiration-dependent processes in the cardiovascular system-New mode of action for caffeine. *PLoS Biol.* **2018**, *16*, e2004408. [CrossRef]
8. Goy, C.; Czypiorski, P.; Altschmied, J.; Jakob, S.; Rabanter, L.L.; Brewer, A.C.; Ale-Agha, N.; Dyballa-Rukes, N.; Shah, A.M.; Haendeler, J. The imbalanced redox status in senescent endothelial cells is due to dysregulated Thioredoxin-1 and NADPH oxidase 4. *Exp. Gerontol.* **2014**, *56*, 45–52. [CrossRef]
9. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
10. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
11. Kopylova, E.; Noé, L.; Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **2012**, *28*, 3211–3217. [CrossRef]
12. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M.R.; Armean, I.M.; Azov, A.G.; Bennett, R.; et al. Ensembl 2020. *Nucleic Acids Res.* **2020**, *48*, D682–D688. [CrossRef]
13. Durinck, S.; Moreau, Y.; Kasprzyk, A.; Davis, S.; De Moor, B.; Brazma, A.; Huber, W. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **2005**, *21*, 3439–3440. [CrossRef]

14. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
15. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
16. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
17. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef] [PubMed]
18. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef]
19. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **2001**, *29*, e45. [CrossRef]
20. Jakob, S.; Schroeder, P.; Lukosz, M.; Büchner, N.; Spyridopoulos, I.; Altschmied, J.; Haendeler, J. Nuclear protein tyrosine phosphatase Shp-2 is one important negative regulator of nuclear export of telomerase reverse transcriptase. *J. Biol. Chem.* **2008**, *283*, 33155–33161. [CrossRef]
21. Schroeder, P.; Popp, R.; Wiegand, B.; Altschmied, J.; Haendeler, J. Nuclear redox-signaling is essential for apoptosis inhibition in endothelial cells–important role for nuclear thioredoxin-1. *Arterioscler. Thromb. Vasc. Biol.* **2007**, *27*, 2325–2331. [CrossRef] [PubMed]
22. Schindelin, J.; Arganda-Carreras, I.; Frise, E.; Kaynig, V.; Longair, M.; Pietzsch, T.; Preibisch, S.; Rueden, C.; Saalfeld, S.; Schmid, B.; et al. Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **2012**, *9*, 676–682. [CrossRef]
23. Boraschi, D.; Italiani, P.; Weil, S.; Martin, M.U. The family of the interleukin-1 receptors. *Immunol. Rev.* **2018**, *281*, 197–232. [CrossRef]
24. Cheng, G.; Salerno, J.C.; Cao, Z.; Pagano, P.J.; Lambeth, J.D. Identification and characterization of VPO1, a new animal heme-containing peroxidase. *Free Radic. Biol. Med.* **2008**, *45*, 1682–1694. [CrossRef] [PubMed]
25. Cheng, G.; Li, H.; Cao, Z.; Qiu, X.; McCormick, S.; Thannickal, V.J.; Nauseef, W.M. Vascular peroxidase-1 is rapidly secreted, circulates in plasma, and supports dityrosine cross-linking reactions. *Free Radic. Biol. Med.* **2011**, *51*, 1445–1453. [CrossRef]
26. Bhave, G.; Cummings, C.F.; Vanacore, R.M.; Kumagai-Cresse, C.; Ero-Tolliver, I.A.; Rafi, M.; Kang, J.S.; Pedchenko, V.; Fessler, L.I.; Fessler, J.H.; et al. Peroxidasin forms sulfilimine chemical bonds using hypohalous acids in tissue genesis. *Nat. Chem. Biol.* **2012**, *8*, 784–790. [CrossRef]
27. Medfai, H.; Khalil, A.; Rousseau, A.; Nuyens, V.; Paumann-Page, M.; Sevcnikar, B.; Furtmüller, P.G.; Obinger, C.; Moguilevsky, N.; Peulen, O.; et al. Human peroxidasin 1 promotes angiogenesis through ERK1/2, Akt, and FAK pathways. *Cardiovasc. Res.* **2019**, *115*, 463–475. [CrossRef]
28. Lee, S.W.; Kim, H.K.; Naidansuren, P.; Ham, K.A.; Choi, H.S.; Ahn, H.Y.; Kim, M.; Kang, D.H.; Kang, S.W.; Joe, Y.A. Peroxidasin is essential for endothelial cell survival and growth signaling by sulfilimine crosslink-dependent matrix assembly. *FASEB J.* **2020**, *34*, 10228–10241. [CrossRef]
29. Mariotti, M.; Ridge, P.G.; Zhang, Y.; Lobanov, A.V.; Pringle, T.H.; Guigo, R.; Hatfield, D.L.; Gladyshev, V.N. Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS ONE* **2012**, *7*, e33066. [CrossRef]
30. Pothion, H.; Jehan, C.; Tostivint, H.; Cartier, D.; Bucharles, C.; Falluel-Morel, A.; Boukhzar, L.; Anouar, Y.; Lihrmann, I. Selenoprotein T: An Essential Oxidoreductase Serving as a Guardian of Endoplasmic Reticulum Homeostasis. *Antioxid. Redox Signal.* **2020**, *33*, 1257–1275. [CrossRef]
31. Boukhzar, L.; Hamieh, A.; Cartier, D.; Tanguy, Y.; Alsharif, I.; Castex, M.; Arabo, A.; El Hajji, S.; Bonnet, J.J.; Errami, M.; et al. Selenoprotein T Exerts an Essential Oxidoreductase Activity That Protects Dopaminergic Neurons in Mouse Models of Parkinson's Disease. *Antioxid. Redox Signal.* **2016**, *24*, 557–574. [CrossRef] [PubMed]
32. Pitts, M.W.; Hoffmann, P.R. Endoplasmic reticulum-resident selenoproteins as regulators of calcium signaling and homeostasis. *Cell Calcium* **2018**, *70*, 76–86. [CrossRef]
33. Huang, X.; Sun, B.; Zhang, J.; Gao, Y.; Li, G.; Chang, Y. Selenium Deficiency Induced Injury in Chicken Muscular Stomach by Downregulating Selenoproteins. *Biol. Trace Elem. Res.* **2017**, *179*, 277–283. [CrossRef] [PubMed]
34. Huang, J.; Bao, D.; Lei, C.T.; Tang, H.; Zhang, C.Y.; Su, H.; Zhang, C. Selenoprotein T protects against cisplatin-induced acute kidney injury through suppression of oxidative stress and apoptosis. *FASEB J.* **2020**, *34*, 11983–11996. [CrossRef]
35. Addinsall, A.B.; Wright, C.R.; Andrikopoulos, S.; van der Poel, C.; Stupka, N. Emerging roles of endoplasmic reticulum-resident selenoproteins in the regulation of cellular stress responses and the implications for metabolic disease. *Biochem. J.* **2018**, *475*, 1037–1057. [CrossRef]
36. Tujebajeva, R.M.; Copeland, P.R.; Xu, X.M.; Carlson, B.A.; Harney, J.W.; Driscoll, D.M.; Hatfield, D.L.; Berry, M.J. Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO Rep.* **2000**, *1*, 158–163. [CrossRef]
37. Haendeler, J.; Messmer, U.K.; Brüne, B.; Neugebauer, E.; Dimmeler, S. Endotoxic shock leads to apoptosis in vivo and reduces Bcl-2. *Shock* **1996**, *6*, 405–409. [CrossRef]
38. Pober, J.S. Endothelial activation: Intracellular signaling pathways. *Arthritis Res. Ther.* **2002**, *4* (Suppl. 3), S109–S116. [CrossRef]
39. Dimmeler, S.; Haendeler, J.; Rippmann, V.; Nehls, M.; Zeiher, A.M. Shear stress inhibits apoptosis of human endothelial cells. *FEBS Lett.* **1996**, *399*, 71–74. [CrossRef]

40. Carlson, B.A.; Xu, X.M.; Gladyshev, V.N.; Hatfield, D.L. Selective rescue of selenoprotein expression in mice lacking a highly specialized methyl group in selenocysteine tRNA. *J. Biol. Chem.* **2005**, *280*, 5542–5548. [CrossRef]
41. Jones, G.D.; Droz, B.; Greve, P.; Gottschalk, P.; Poffet, D.; McGrath, S.P.; Seneviratne, S.I.; Smith, P.; Winkel, L.H. Selenium deficiency risk predicted to increase under future climate change. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2848–2853. [CrossRef] [PubMed]
42. Lopes Junior, E.; Leite, H.P.; Konstantyner, T. Selenium and selenoproteins: From endothelial cytoprotection to clinical outcomes. *Transl. Res.* **2019**, *208*, 85–104. [CrossRef]
43. Mertens, K.; Lowes, D.A.; Webster, N.R.; Talib, J.; Hall, L.; Davies, M.J.; Beattie, J.H.; Galley, H.F. Low zinc and selenium concentrations in sepsis are associated with oxidative damage and inflammation. *Br. J. Anaesth.* **2015**, *114*, 990–999. [CrossRef]
44. Wang, L.; Jing, J.; Yan, H.; Tang, J.; Jia, G.; Liu, G.; Chen, X.; Tian, G.; Cai, J.; Shang, H.; et al. Selenium Pretreatment Alleviated LPS-Induced Immunological Stress Via Upregulation of Several Selenoprotein Encoding Genes in Murine RAW264.7 Cells. *Biol. Trace Elem. Res.* **2018**, *186*, 505–513. [CrossRef] [PubMed]
45. Wang, X.; Yang, B.; Cao, H.L.; Wang, R.Y.; Lu, Z.Y.; Chi, R.F.; Li, B. Selenium Supplementation Protects Against Lipopolysaccharide-Induced Heart Injury via Sting Pathway in Mice. *Biol. Trace Elem. Res.* **2021**, *199*, 1885–1892. [CrossRef] [PubMed]
46. Rocca, C.; Boukhzar, L.; Granieri, M.C.; Alsharif, I.; Mazza, R.; Lefranc, B.; Tota, B.; Leprince, J.; Cerra, M.C.; Anouar, Y.; et al. A selenoprotein T-derived peptide protects the heart against ischaemia/reperfusion injury through inhibition of apoptosis and oxidative stress. *Acta Physiol.* **2018**, *223*, e13067. [CrossRef] [PubMed]
47. Alsharif, I.; Boukhzar, L.; Lefranc, B.; Godefroy, D.; Aury-Landas, J.; Rego, J.D.; Rego, J.D.; Naudet, F.; Arabo, A.; Chagraoui, A.; et al. Cell-penetrating, antioxidant SELENOT mimetic protects dopaminergic neurons and ameliorates motor dysfunction in Parkinson's disease animal models. *Redox Biol.* **2021**, *40*, 101839. [CrossRef]

2.2 Thesis 3 - The mRNP code can be adjusted to influence splice site usage

Usage of a splice site is not only dependent on its intrinsic strength, but also depends on the binding landscape of splicing regulatory proteins in its proximity. Its usage can be either repressed or enhanced, depending on the protein and its relative position to the splice site. Sequence variations next to splice site positions have been shown to be able to change the binding capacity for splicing regulatory proteins. Using synonymous sequence variations, it has been shown to be possible to specifically manipulate usage of a given splice site within splicing reporters or gene expression vectors, without disrupting the underlying protein coding sequences (**Publication VI**). However, since splicing regulatory proteins also play a role in other biological processes than splicing, these changes in the mRNP code could additionally affect for instance RNA export or RNA stability. One important feature for gene expression would be RNA export, since viral sequence elements, that repress or enhance nuclear export could be distinctly categorized by their binding capacity for splicing regulatory proteins (**Publication VII**).

2.3.1 Publication VI: Modifying splice site usage with ModCon: maintaining the genetic code while changing the underlying mRNP code.

To further deepen our knowledge about the interplay between intrinsic splice site strength and the binding capacity for splicing regulatory proteins, predicted by the HEXplorer based SSHW, the ModCon algorithm was developed to modify the SSHW of a given splice site within a protein coding sequence, without disrupting the encoded amino acid sequence. This allows the generation of expression vectors and/or splicing reporters, with potentially fine-tuned manipulation of the SRP-mediated impact on 5'ss selection. Applying ModCon on our in-house splicing reporter, that controls for transfection-efficiency using a dual-luciferase approach, we successfully initiated the usage of a previously splicing inactive GT-site within the coding sequence of the firefly luciferase [149]. The ModCon algorithm utilizes synonymous sequence variations, keeping the encoded amino acid sequence intact, while modifying the binding capacity of the mRNA transcript for either SR or hnRNP (-like) proteins, predicted by the HEXplorer. To that end, a sliding window approach is combined with a genetic algorithm that uses principles of genetic recombination and natural selection.

To modify the underlying total HEXplorer score of a coding genomic sequence, ModCon first determines the amino acid sequence encoded in the 48 nucleotides upstream and downstream of the selected 5’ss. Then, an initial set of 1,000 sequences (F0), that encode the respective amino acid sequence are generated by random selection of suitable codons per amino acid position. After calculating the total HEXplorer score of each sequence, that estimates its potential to recruit splicing regulatory proteins, a *mating* subset of these sequences (M1) is selected to generate a new generation of sequences via recombination. It is per default assembled from 40% of the fittest sequences of F0 (showing the highest total HEXplorer score, or lowest total HEXplorer score, depending on the intended SSHW manipulation), 20% of F0 using the fitness as probability for selection, and 5% randomly selected parental sequences. Subsequently, M is used to generate the next filial generation of 300 new sequences (F1), using a crossover approach (60%), insertion (30%) or random combination (10%). Next, codons of sequences in the filial generation F1 are randomly mutated to a different codon, still encoding the same amino acid as before, using an experimentally determined mutation rate of 0.01%, to introduce new elements, that potentially were not part of the initial subset. Using the filial generation F1 as the new parental generation, the next mating subset is determined (M2) starting a new generation of the genetic algorithm.

The following article was published in Comput Struct Biotechnol J. 2021 May 21:19:3069-3076 (doi: 10.1016/j.csbj.2021.05.033), by

Johannes Ptok, Lisa Müller, Philipp Niklas Ostermann, Anastasia Ritchie, Alexander T Dilthey, Stephan Theiss, Heiner Schaal

Contribution

JP developed the tool. LM, PNO and AR constructed and tested the reporter construct. JP, HS and ST wrote the manuscript, except for the part about the reporter construct. This was written by LM. ATD helped with the concept of the tool. Individual contribution of JP at around 80%.

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Modifying splice site usage with *ModCon*: Maintaining the *genetic code* while changing the underlying *mRNP code*

Johannes Ptok[a], Lisa Müller[a], Philipp Niklas Ostermann[a], Anastasia Ritchie[a], Alexander T. Dilthey[b,c,d], Stephan Theiss[a,*], Heiner Schaal[a,*]

[a] *Institute of Virology, Medical Faculty, Heinrich Heine University Düsseldorf, D-40225 Düsseldorf, Germany*
[b] *Institute of Medical Statistics and Computational Biology, University of Cologne, Cologne, Germany*
[c] *Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany*
[d] *Institute of Medical Microbiology and Hospital Hygiene, Heinrich Heine University Düsseldorf, Düsseldorf, Germany*

ARTICLE INFO

Article history:
Received 16 February 2021
Received in revised form 14 May 2021
Accepted 20 May 2021
Available online 21 May 2021

Keywords:
pre-mRNA splicing
Splicing reporter
HEXplorer score
HBond score
Splice donor
Splicing regulatory proteins

ABSTRACT

Codon degeneracy of amino acid sequences permits an additional “mRNP code” layer underlying the genetic code that is related to RNA processing. In pre-mRNA splicing, splice site usage is determined by both intrinsic strength and sequence context providing RNA binding sites for splicing regulatory proteins. In this study, we systematically examined modification of splicing regulatory properties in the neighborhood of a GT site, i.e. potential splice site, without altering the encoded amino acids.

We quantified the splicing regulatory properties of the neighborhood around a potential splice site by its *Splice Site HEXplorer Weight* (SSHW) based on the HEXplorer score algorithm. To systematically modify GT site neighborhoods, either minimizing or maximizing their SSHW, we designed the novel stochastic optimization algorithm *ModCon* that applies a genetic algorithm with stochastic crossover, insertion and random mutation elements supplemented by a heuristic sliding window approach.

To assess the achievable range in SSHW in human splice donors without altering the encoded amino acids, we applied ModCon to a set of 1000 randomly selected Ensembl annotated human splice donor sites, achieving substantial and accurate changes in SSHW. Using ModCon optimization, we successfully switched splice donor usage in a splice site competition reporter containing coding sequences from FANCA, FANCB or BRCA2, while retaining their amino acid coding information.

The ModCon algorithm and its R package implementation can assist in reporter design by either introducing novel splice sites, silencing accidental, undesired splice sites, and by generally modifying the entire mRNP code while maintaining the genetic code.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

During splicing, introns of a newly synthetized pre-mRNA strand are mostly co-transcriptionally removed from the transcript, followed by the ligation of the remaining exonic sequence segments [1,2]. The intron excision cellular machinery is called spliceosome and recognizes canonical sequences with a GT dinucleotide at the upstream end of an intron, the 5′ splice site (5′ss) or splice donor (SD), and with an invariant AG at the downstream end of an intron, the 3′ splice site (3′ss) or splice acceptor (SA) [3,4].

Recognition of splice donor sites during spliceosome formation is accomplished through RNA duplex formation with 11 nucleotides of the free 5′end of the U1 snRNA [5], while splice acceptor sites are bound by U2 auxiliary factors (U2AF). A higher U1 snRNA complementarity is beneficial for 5′ splice site recognition and usage [6]. Several algorithms are available for scoring splice site strength: e.g. maximum entropy algorithms providing maxent scores for both 5′ss and 3′ss [7], and the HBond score reflecting 5′ss complementarity to U1 snRNA [5].

Abbreviations: A, adenine; eGFP, enhanced green fluorescent protein; F1, filial sequence 1; G, guanine; GA, genetic algorithm; hnRNP, heterogeneous nuclear ribonucleoproteins; HBS, HBond score; HZ_{EI}, HEXplorer score; nt, nucleotides; P1, parental sequence 1; pre-mRNA, precursor messenger RNA; SA, splice acceptor; SD, splice donor; snRNA, small nuclear RNA; SR proteins, serine- and arginine-rich proteins; SRP, splicing regulatory protein; SSHW, splice site HEXplorer weight; SW, sliding window; T, thymine.

* Corresponding authors.

https://doi.org/10.1016/j.csbj.2021.05.033
2001-0370/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Beyond the proper splice site consensus sequences, splice site recognition has been shown to greatly depend on proximal binding of splicing regulatory proteins (SRPs) [8-11], which can significantly enhance or repress splice site usage. Splicing regulatory proteins can be divided into two major families, differing in their position-dependent effect on splice site usage [12]. Serine- and arginine-rich proteins (SR proteins) enhance usage of upstream splice acceptors and downstream splice donors, but repress usage of upstream splice donors and downstream splice acceptors. Heterogeneous nuclear ribonucleoproteins (hnRNP) on the other hand have an opposite effect on splice site usage. The proximal splice donor context beneficial for its usage therefore consists of upstream binding motifs of SR proteins and downstream binding motifs of hnRNP proteins (reviewed in [11]).

In general, different nucleotide sequences coding for the same amino acid sequence can contain different splicing regulatory elements—binding sites for SR- or hnRNP proteins. For any genomic sequence, its splicing regulatory properties are reflected by its HEXplorer score ($\mathrm{HZ_{EI}}$) profile, calculated for every nucleotide [13]. Here, we examine the possible variation in the total HEXplorer score while preserving the encoded amino acid sequence for a given reading frame. To this end, we designed an algorithm to maximize or minimize total $\mathrm{HZ_{EI}}$ by variation of admissible codons. With an average codon degeneracy of three (range 1–6), the number of alternatively admissible nucleotide sequences for a given sequence of N amino acids is $\sim 3^N$ and grows exponentially with the number of codons, N. An exhaustive search in the space of all alternative sequences is therefore very time-consuming and not feasible in practice.

Stochastic optimization algorithms like Monte Carlo or evolutionary algorithms are particularly well suited for optimizing an objective function—total $\mathrm{HZ_{EI}}$—in an exponentially large configuration space ($\sim 3^N$) under a set of constraints (amino acids). Here, we designed a genetic algorithm with recombination between mating configuration populations using crossover, insertion and random mutations, and combined it with a heuristic sliding window approach. This *ModCon* algorithm (Modulator of Context) permits enhancing or silencing splice site usage by manipulating their sequence neighborhoods while preserving the encoded amino acids. As a proof of principle, we applied ModCon to sequences within a splice donor competition reporter and additionally demonstrated the impact of a change in HEXplorer score for a naturally occurring GT site within a common luciferase expression reporter system. We tested the scope of HEXplorer score manipulation with ModCon on a set of 1000 randomly selected human SD sites.

2. Material and methods

2.1. HEXplorer algorithm

The HEXplorer score is based on hexamer frequency differences in 100 nt long neighborhoods upstream compared to downstream of splice donor sites [13], resulting in a $\mathrm{Z_{EI}}$-score for each hexamer. Hexamers predominantly found upstream of splice donor sites have positive $\mathrm{Z_{EI}}$-scores, and they frequently overlap SR protein binding sites, while negative $\mathrm{Z_{EI}}$-score hexamers often relate to hnRNP binding sites. Proceeding from a single hexamer-based quantification to a score for each nucleotide of a genomic sequence, we calculated the HEXplorer score ($\mathrm{HZ_{EI}}$) as the average $\mathrm{Z_{EI}}$-score of all six hexamers overlapping an index nucleotide: $\mathrm{HZ_{EI}} = \sum Z_{EI}/6$. The total $\mathrm{HZ_{EI}}$ of a sequence stretch, e.g. a splice site neighborhood, indicates its overall splicing regulatory property, likely due to hnRNP or SR protein binding sites. Changes in HEXplorer score induced by mutations have been shown to correlate well with the mutation's impact on nearby splice site usage [14].

For any splice donor site, the overall SRP-mediated impact of its sequence neighborhood on splice site usage is then captured by its *Splice Site HEXplorer Weight* (SSHW), the total upstream minus downstream $\mathrm{HZ_{EI}}$ [14]: $\mathrm{SSHW} = \sum_{up} HZ_{EI} - \sum_{dn} HZ_{EI}$. The higher the SSHW, the higher the predicted SRP binding potential of a splice site sequence neighborhood, potentially enhancing its usage.

2.2. The ModCon algorithm

To optimize a splice donor site's SSHW, ModCon combines a genetic algorithm applying principles of natural selection and sexual recombination with a sliding window approach, and separately addresses up- and downstream sequences of the given splice site. Driven by the optimization algorithm, ModCon varies the sequence neighborhood of the splice donor site under the constraint of preserving the encoded amino acids and calculates the HEXplorer score of each alternative neighborhood as well as the SSHW.

As input, ModCon takes (1) a coding sequence, (2) the position of the first nucleotide of the "index" GT site within the coding sequence and (3) either maximization or minimization of the target function SSHW. ModCon outputs a coding sequence with a SSHW-optimized alternative neighborhood (16 codons upstream and downstream by default) for the GT site. The graphical abstract provides a structural overview of the ModCon algorithm.

2.2.1. Maximizing or minimizing the total HEXplorer score of a coding sequence

To maximize the SSHW of a splice donor site, ModCon maximizes the total upstream HEXplorer score and minimizes the total downstream HEXplorer score, and *vice versa*.

By default, ModCon considers a sequence window of ±48 nucleotides around the selected index GT site in frame, excluding codons which would overlap with the 11 nucleotides of the GT donor sequence. Synonymous substitutions are then applied to 16 codons ($\widehat{=}$ 48 nt) upstream and downstream of the GT site to increase or decrease the underlying total $\mathrm{HZ_{EI}}$, possibly regulating GT site usage through introduction or modification of splicing regulatory elements.

Ideally, the total HEXplorer score $\mathrm{HZ_{EI}}$ of all sequences encoding the same amino acid sequence would be calculated to determine the highest or lowest total $\mathrm{HZ_{EI}}$. However, a sequence of 16 amino acids can potentially be encoded by up to 6^{16} or 2.8 trillion different nucleotide sequences, because some amino acids can be encoded by up to 6 different codons. Since the HEXplorer score computation of all 2.8 trillion eligible sequences, however, would require extensive time and memory resources, we developed an evolutionary algorithm supplemented by a sliding window approach.

2.2.2. Genetic algorithm

Genetic algorithms can be used to approach optimization tasks with trillions of potential solutions for combinatorial problems by applying principles of genetic recombination and natural selection [15].

Here, a genetic algorithm is developed to combine distinct sets of nucleotide sequences depending on their "fitness", defined by their total $\mathrm{HZ_{EI}}$. It applies a cyclic iterative optimization process that consists of (1) generating an initial sequence population and calculating its fitness, (2) selecting a suitable mating population, (3) creating a new filial generation from it and (4) introducing random mutations (see flowchart Fig. 1).

First, by randomly selecting eligible codons, an initial parental population **F0** of 1000 sequences is generated, all encoding the

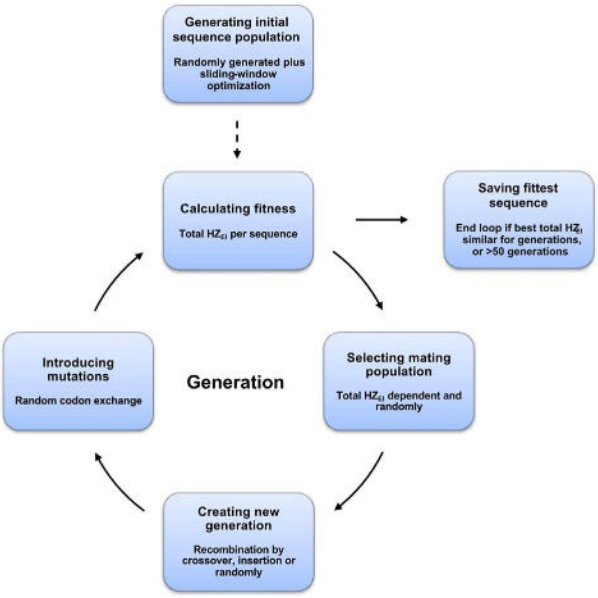

Generating initial sequence population
Randomly generated plus sliding-window optimization
Calculating fitness
Total HZ_{EI} per sequence
Saving fittest sequence
End loop if best total HZ_{EI} similar for generations, or >50 generations
Generation
Introducing mutations
Random codon exchange
Selecting mating population
Total HZ_{EI} dependent and randomly
Creating new generation
Recombination by crossover, insertion or randomly

Fig. 1. Schematic of the genetic algorithm. After generation of the initial sequence population, the "fitness" of the sequences, defined by the total HZ_{EI}, is calculated. Next, a defined number of sequences are selected for recombination, either randomly, by crossover, or sequence insertion. Afterwards, every codon within a sequence undergoes random codon exchange with a certain probability. The fitness of every newly generated filial generation of sequences is again determined by total HZ_{EI} calculation. This cycle of generations is continued, while determining the best fitness of every generation, until no further significant increase in total HZ_{EI} can be measured anymore. Finally, the sequence with the highest or lowest total HZ_{EI} is returned. A dashed arrow represents a one-time action, whereas a solid arrow represents an action repeated in every generation.

same initial amino acid sequence. In order to improve convergence speed, this sequence set can optionally be supplemented with typically 100 (10%) sequences previously calculated by a sliding-window approach. This step significantly accelerates the process of the genetic algorithm by directing it towards an optimal solution.

In the first part of the genetic algorithm loop, the fitness of every sequence of the parental population **F0** is calculated as its respective total HZ_{EI}. For each generation, the sequence with the highest or lowest total HZ_{EI} is stored for later reporting on efficiency, and to check if the overall scores could be significantly improved during the last 30 generations.

A new set of sequences, the *mating* population **M**, is then generated through recombination from a subset of the parental population **F0**. This subset is per default assembled from 40% of the fittest sequences of **F0**, 20% of **F0** using the fitness as probability for selection, and 5% randomly selected parental sequences.

Next, **M** is used to generate a set of recombined sequences (300 per default). The resulting *filial* population **F1** is created through random combination of sequence blocks from sequences of the mating population. Every newly generated codon sequence is generated by recombination of two randomly selected codon sequences from the mating population, using three distinct modes of recombination (Fig. 2).

Usage frequency of the three methods of recombination is based on the extent to which the three modes preserve continuous parental codon sequence stretches within the resulting filial sequences. Therefore, 60% of filial sequences originate from "crossover" recombination, where a filial sequence is made from two continuous sequence stretches coming from one parental sequence each. With 30%, the second most applied mode of recombination is "insertion", where filial sequences consist of the sequence of one parental sequence, which holds a random-sized insertion from a second parental sequence in-between. The least used recombination method is the random selection of codons from either one parental sequence. It is used for the remaining 10% of filial sequences.

An important step of evolutionary algorithms is the introduction of mutations after generation of the filial population **F1**, since a carefully selected mutation rate increases the probability to escape potential local maxima or minima during the search for the global peak in the fitness function. From a series of preliminary experiments, we identified an optimal mutation rate of 10^{-4} or 0.01%, meaning that one in 10,000 codons is randomly exchanged with another codon encoding the same amino acid.

The introduction of mutations marks the last step during the cycle of generations. Afterwards, the total HZ_{EI} is again calculated for each sequence, to determine the likeliness to further contribute to the following generations of filial sequences. Then, again, a subset of sequences is selected from **F1** based on their fitness to constitute the next mating sequence population, to generate the second filial population **F2** through recombination.

The generation of newly combined sequences is repeated, until the total HZ_{EI} holds approximately the same level for at least 30 generations or the maximal number of generations (50 generations by default) is reached.

2.2.3. Sliding-window algorithm

In order to keep the computational effort manageable, we furthermore applied a stepwise optimization of codon-quadruplets to optimize the total HZ_{EI} of a 16-codon long sequence. In contrast to the up to 2.8 trillion different nucleotide sequences for a stretch of 16 codons, a four amino acids long sequence can only be encoded by up to $6^4 = 1296$ distinct nucleotide sequences, which enables more efficient total HZ_{EI} calculation.

To optimize the total HZ_{EI} of a sequence, the sliding window algorithm first makes a list of every potential nucleotide sequence encoding the most upstream stretch of four amino acids (codons 1–4). Then, total HZ_{EI} is calculated and the most downstream hexamer of each nucleotide sequence is saved. For every unique hexamer within the sequence pool, the maximal associated total HZ_{EI} is determined. Since the HEXplorer score of each nucleotide is calculated from all six overlapping hexamers, a hexamer between two nucleotides constitutes a barrier in the HZ_{EI} score dependencies. In particular, a sequence downstream of a hexamer can be changed without affecting the HZ_{EI} upstream of that hexamer. The algorithm then proceeds with the optimal nucleotide sequences in each hexamer group, reducing the number of sequences drastically (up to $6^2 = 36$ sequences, in case the last hexamer encodes amino acids with a codon degeneracy of 6).

Subsequently, the algorithm makes a list of every potential nucleotide sequence encoding the next four codons (codons 5–8) and combines every new nucleotide sequence with every previously determined one. Since always those sequences with the highest total HZ_{EI} are selected, the first four codons are now the same in every sequence. To decrease computation time, we can save them for the output and remove them from our sequence list, reducing the sequence length to four codons. This process is repeated until the end of the 16 codons is reached.

While the sliding window algorithm enables fast calculation (taking only a few seconds per run on a standard machine), its fast convergence skips sequences with an intermediate HZ_{EI} increase. The sliding window algorithm is therefore primarily used to quickly obtain a few near-optimal sequences, in particular as supplementary sequences for the genetic algorithm.

2.2.4. Additional sequence processing

A sequence found to maximize or minimize total HZ_{EI} may still contain GT or AG dinucleotides that may accidentally correspond to strong splice donor or acceptor sites. To reduce coincidental

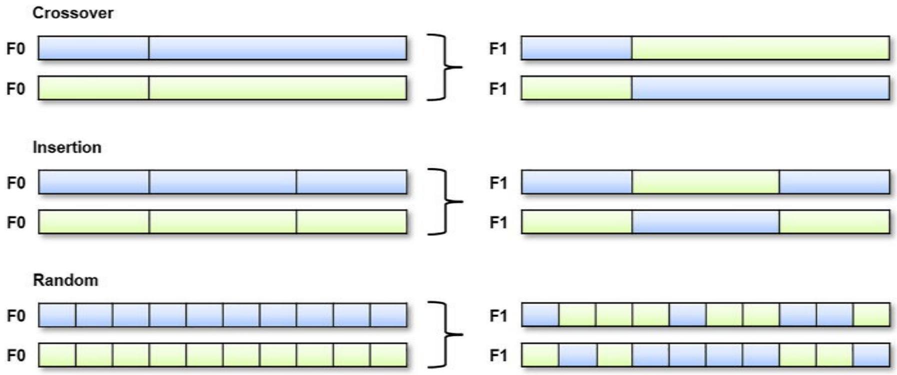


Fig. 2. Modes of sequence recombination. Mating of two codon sequences (F0, F0) of the parental sequence population can lead to various resulting potential filial sequences (F1, F1) depending on the modus of recombination. Filial sequences coming from crossover combinations are constituted of two continuous sequence stretches, coming from one parental codon sequence each. During insertion recombination, the filial sequences consist of one of the parental sequences, holding an insertion from the other one. Random recombination describes the random mixture of codons from both parental sequences.

introduction of such undesired splice sites, ModCon proceeds to degrade splice sites exceeding a threshold of HBS >10 for splice donors and Maxent score >4 for splice acceptors, while preserving the encoded amino acids.

For that purpose, codons overlapping these GT or AG sites are exchanged by alternative codons leading to no or much weaker sites, while keeping total HZ_{EI} close to the identified optimum. Degrading or increasing the HBond score of a specific index GT site of interest can be performed if needed, using the respective R-functions.

2.2.5. HBond, maxent and HEXplorer data sources

Required data for HBond scores of 5′ splice sites [5] and HZ_{EI} scores of splice site neighborhoods [13] were taken from previous work of this group (cf. http://www.uni-duesseldorf.de/rna). The maxent score for human splice sites has been integrated for the evaluation of 3′ splice site strength [7] with kind approval of Gene Yeo. Data for the calculation of 3′ splice site maxent scores were adopted from the website http://hollywood.mit.edu/burgelab/maxent/download/fordownload/.

2.3. The luciferase reporter

A dual luciferase reporter was used to monitor differences in the splicing outcome upon using ModCon to render an unused 5′ss functional. The reporter construct consists of renilla- and firefly-luciferase transcription units under the control of an SV40 promoter which are terminated by an SV40 polyadenylation site. Renilla luciferase expression was used for normalization of mRNA abundancies in a PCR based readout. For cloning, synonymous mutations were placed in the firefly luciferase coding sequence to create an EcoRV restriction site upstream of an unused splice donor with an HBS of 14. Downstream of the firefly luciferase stop codon, the HIV derived 3′ss $SA7^{opt}$ and an artificial exon (99 bp) were placed [16]. Algorithm amended sequences were inserted as a gene strand synthesized by Eurofins Genomics, Germany (Eurofins Gene Strand #11106560588).

To analyze the splicing pattern, HeLa cells cultivated in Dulbecco's high-glucose modified Eagle's medium (Invitrogen) supplemented with 10% fetal calf serum and 50 µg/ml penicillin and streptomycin each (Invitrogen) were used for transient-transfection experiments. For that, $2.5x10^5$ cells were plated in six-well plates and transfected with 1 µg of the reporter plasmid using TransIT®-LT1 transfection reagent (Mirus Bio LLC US) according to the manufacturer's instructions. Total RNA was isolated 24 h post-transfection by using acid guanidinium thiocyanate-phenol-chloroform. For semiquantitative RT-PCR analyses, RNA was reverse transcribed by using Superscript III reverse transcriptase (Invitrogen) and oligo(dT) primers (Invitrogen) and amplified using the primer pair #6575/#6381, as well as #6167/6168 for the renilla luciferase internal control. Splicing patterns were visualized via a non-denaturing 10% polyacrylamide gel. Primer sequences for the RT-PCR:

#6575 FW (modified) firefly luciferase GTGTTGTTCCATTCCATCACG
#6381 REV firefly luciferase CAGCTGTTCTCCAGCTGT
#6167 FW renilla luciferase GCGTTGATCAAATCTGAAGAAGG
#6168 REV renilla luciferase TTGGACGACGAACTTCACCT

2.4. Splice donor competition reporter

In order to experimentally test the splicing behavior of ModCon designed sequences, 40 nt stretches of either wild type or modified sequence were inserted between two identical copies of a strong 5′ss sequence with an HBond score of 17.5. These two competing 5′ss define the 3′ end of the first exon of an HIV-based two-exon splicing reporter. 40 nt long sequences between the competing 5′ss were derived from FANCA, FANCB and BRCA2 (Suppl. Table 3). All sequences can be obtained upon request.

To analyze the splicing pattern, transient-transfection experiments were carried out as described above. To monitor transfection efficiency, 1 µg of pXGH5 expression plasmid (hGH) was co-transfected. For semiquantitative RT-PCR analyses, RNA was reverse transcribed by using Superscript III reverse transcriptase (Invitrogen) and oligo(dT) primers (Invitrogen) and amplified using the primer pair #3210/#3211, as well as #1224/#1225 for hGH. Splicing patterns were visualized via a non-denaturing 10% polyacrylamide gel. Primer sequences for the RT-PCR:

#3210 TGAGGAGGCTTTTTTGGAGG
#3211 TTCACTAATCGAATGGATCTGTC
#1224 TCTTCCAGCCTCCCATCAGCGTTTGG
#1225 CAACAGAAATCCAACCTAGAGCTGCT

3. Results

3.1. Similar SSHW ranges obtained by GA and SW for 1000 human TSL1 SD sites

In order to determine the achievable range of SSHW optimization, we extracted 185,190 unique splice donors annotated in Ensembl transcripts (version 101) with the highest transcript support level of 1 (TSL1, Suppl. Table 1). After removing 1% of extremely high and low SSHW values, the remaining 183,339 wild type donor sites (99%) showed SSHW values ranging from around −300 to 1000 (average SSHW 235).

For a random sample of 1000 splice donors drawn from this set (Suppl. Table 2), we then minimized and maximized SSHW using both the sliding window algorithm (SW) and the genetic algorithm with the results from the SW added to the initial population (GA). Table 1 presents the average and standard deviation SSHW difference for the four combinations of algorithm (GA, SW) and optimization (SSHW min/max).

Fig. 3 shows the distributions of minimal and maximal SSHW difference (optimized—wild type) obtained by the GA and SW algorithm. Note that the resulting distributions for the GA and SW algorithm practically coincide, while the maximal SSHW distribution is ~17% narrower and higher compared to the minimal SSHW distribution.

Comparing the distribution of SSHW values achieved with the two algorithms showed no significant differences during SSHW minimization or maximization. However, for around a third of the 1000 SD sites, one of the two approaches performed marginally better. During SSHW minimization and maximization, the GA with the input from the SW algorithm exceeded the achieved SSHW of the SW algorithm alone in 26% of the cases and underperformed for 8% of the SD sites. However, although both algorithms obtained similar extreme values for SSHW, the GA also provides a wide range of intermediate SSHW, and thus permits fine adjustment of potential binding sites for SRPs.

3.2. Faster convergence of GA if SW results are added to initial population

While the sliding window algorithm is deterministic in nature, the genetic algorithm is stochastic and progressively converges to a sequence with optimized SSHW. Convergence speed depends on the choice of mating populations and filial generations, but also on the initial sequence population chosen. Here, we in particular examined the impact of adding 10% sliding-window optimized sequences to the initial generation of the genetic algorithm during total HZ_{EI} maximization. We generated scatterplots of the total HZ_{EI} values of all 300 sequences in each generation both with (Fig. 4B) and without these additional sequences (Fig. 4A). Upon adding 100 SW-optimized sequences generated from the initial WT sequence, the convergence of the genetic algorithm was faster and reached sequences with optimal total HZ_{EI} values with fewer iterations, significantly reducing the algorithm's runtime. These observations also held true for total HZ_{EI} minimizations (data not shown).

In contrast to the sliding window algorithm, the genetic algorithm has the benefit of approaching the total HZ_{EI} maximum or minimum with many slightly different intermediate sequences, and thus avoids getting trapped in local maxima or minima during optimization. This effect could be nicely observed with the example of Fig. 4, where the maximal total HZ_{EI} of the sliding window algorithm was even exceeded after the third generation of the genetic algorithm.

Naturally, the GA with input from the SW algorithm required more CPU time per SSHW minimization or maximization than the SW algorithm alone. During SSHW adjustment of the 1000 human donor sites from Suppl. Table 2 on a machine with 4 CPUs and 8 GB RAM, the SW algorithm took an average of 16.0 s per SD, whereas the GA with input from the SW algorithm took an average of 54.3 s per SD, making the former 3.4 times faster. Similar running time was measured during SSHW minimizations. ModCon, however, also runs with only 1 CPU and a few Mb of RAM available. ModCon per default applies the SW algorithm for SSHW optimizations to save running time, while still achieving a similarly high or low SSHW than with the GA. Alternatively, the combined algorithm can still be applied setting the parameter “optiRate” of the R function “ModCon” to any value other than 100. Setting optiRate to a value >100 results in ModCon using the combined approach to optimize the SSHW of a given GT site. A value lower than 100 triggers the same, but also reports an alternative sequence neighborhood for the donor, which shows optiRate % of the maximal SSHW increase or decrease, enabling fine adjustment of SSHW values.

3.3. Applying ModCon to reporter constructs

To experimentally test the splicing regulatory effect of ModCon designed nucleotide sequences, we selected three 40 nucleotides long wildtype sequences from FANCA, FANCB and BRCA2 with negative HZ_{EI} scores, and positioned them between two identical strong splice donors (HBS 17.5) in an HIV-based two-exon splice site competition reporter. Different sequences placed between these two donor sites can lead to recruitment of splicing regulatory proteins, whose impact on donor usage is position-dependent (Fig. 5A). Whereas hnRNP protein binding enhances upstream donor usage and represses downstream donor usage, SR proteins act in the opposite way.

We specifically selected wildtype sequences with negative HZ_{EI} regions at different levels in order to observe the gradual switch of splice site usage between the competing splice donors. With a HZ_{EI}/nt of −3.28, the wild type FANCB sequence segment induced usage of both donor sites, while a slightly reduced HZ_{EI}/nt of −5.59 in the wild type FANCA sequence segment led to exclusive upstream donor usage, further confirmed by the wild type BRCA2 sequence with HZ_{EI}/nt of −5.90.

Maximizing the total HZ_{EI} of the sequence segments while retaining the amino acid coding information (Fig. 5C) yielded positive HZ_{EI} regions and completely switched splice site usage to the downstream splice donor in all three sequences (Fig. 5B).

In a last step, we examined ModCon on a longer coding sequence with a highly variable HEXplorer profile. Firefly luciferase as a widely used standard expression vector provides a simple experimental read-out. Aiming at turning the firefly luciferase into a splicing reporter, we attempted to switch on usage of a moderately strong GT site (HBS 14 ≈ median HBS = 15 in all human annotated 5′ss, MaxEnt score of 7.33) deep in the coding region that is unused in the wild type luciferase (Fig. 6A).

Using ModCon, we were able to induce usage of the internal, unused GT site by modifying its SSHW from −63.7 to 782.3 and additionally shifting the HEXplorer profile of the sequence seg-

Table 1
SSHW difference, applying the sliding window algorithm (SW) and the genetic algorithm (GA) with the results from the SW added to the initial population. ΔSSHW was calculated subtracting the wild type SSHW from the algorithm achieved SSHW.

	GA ΔSSHW min	SW ΔSSHW min	GA ΔSSHW max.	SW ΔSSHW max
Average SSHW	−1109.6	−1111.8	701.2	701.3
St. Dev. SSHW	227.3	220.8	181.2	181.3

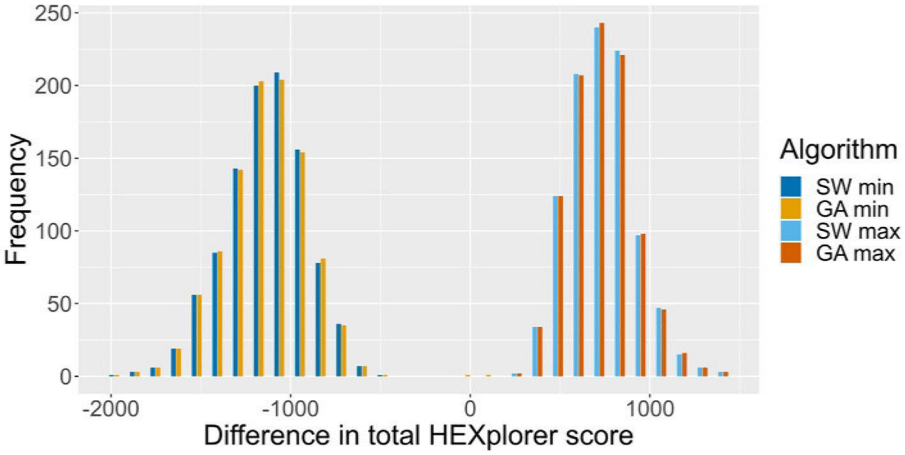


Fig. 3. Bar plot depicting the distribution for the SSHW difference of 1000 human TSL1 SD sites applying different settings of ModCon. SSHW difference (optimized—wild type) is shown on the horizontal axis. Note that bars for GA max and SW max, as well as GA min and SW min lie right next to each other.

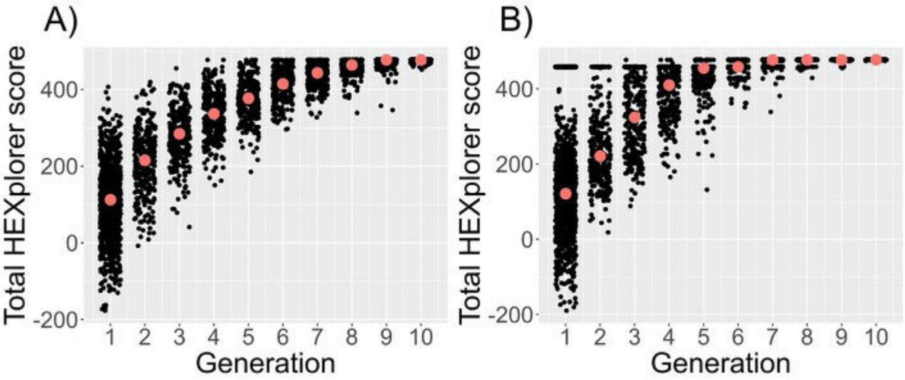


Fig. 4. Total HEXplorer score per generation of the genetic algorithm. Depicted is the total HEXplorer score of every sequence generated during the first 10 generations of the genetic algorithm applied to an exemplary 48 nucleotide long sequence. The first run without spike-in sequences of the sliding window approach shown in (A) needs more generations to reach the maximal total HEXplorer score than with the additional input, shown in (B). For each generation, the median total HEXplorer score is shown in red. Running time on a machine with 4 CPUs and 8 GB RAM for A) 15.3 s and B) 15.7 s. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ment between the GT site and the acceptor site from exon-like to intron-like (Fig. 6B, C). We thus successfully applied ModCon to a much longer and diverse coding sequence.

4. Discussion

Codon degeneracy of amino acid sequences permits an additional "mRNP code" layer underlying the genetic code that is related to RNA processing like pre-mRNA splicing, RNA stability, RNA secondary and tertiary structure, nuclear retention or export [17]. In this study, we addressed pre-mRNA splicing regulation to promote or repress GT site usage without altering encoded amino acids and GT site sequences. Depending on the fine balance between its intrinsic strength and sequence context, GT site usage can be enhanced or repressed by proximal binding of splicing regulatory proteins. Computationally, the neighboring SRP binding landscape is reflected by the SSHW of a given GT site based on its neighborhood's HEXplorer score profile [14]. To systematically modify GT site neighborhoods with respect to their predicted splicing regulatory properties without altering the encoded amino acids, we developed the stochastic optimization algorithm ModCon. We experimentally verified the ModCon algorithm in a splice site competition reporter using wild type sequences from FANCA, FANCB and BRCA2 genes, as well as in a common reporter system of firefly luciferase.

In particular, moderately strong splice sites are most susceptible to regulation by SRP binding. Splicing regulatory proteins binding within a ~50 nt neighborhood of splice donor sites are generally assumed to potentially impact splice site recognition [18,19]. In the evaluation of the ModCon GA and SW algorithm, we therefore used 48 nt wide neighborhoods close to this estimate. The SW algorithm performed equally well as the combined GA during SSHW manipulation, with a 3-times shorter running time. However, the GA allows a much finer SSHW tuning at intermediate levels and avoids local maxima or minima much better than the SW algorithm. Therefore, for SSHW maximization and minimization, ModCon per default applies the SW algorithm and for precise SSHW adjustments, ModCon applies the GA.

Stochastic Monte Carlo or evolutionary algorithms are particularly suited to optimization tasks in large configuration spaces growing exponentially with sequence length, like ~3^N for N amino acids in our case. Here, we chose a genetic algorithm with stochastic crossover, insertion and random mutation elements [15]. We successfully reduced computational effort (time and memory demands) by supplementing this genetic algorithm by a heuristic sliding window approach. On a set of 1000 human splice donor sites, both approaches were equally able to significantly optimize SSHW in both directions. In a splice site competition reporter with two identical competing splice donors, we demonstrated that it is possible to induce a switch in SD usage by designing nucleotide

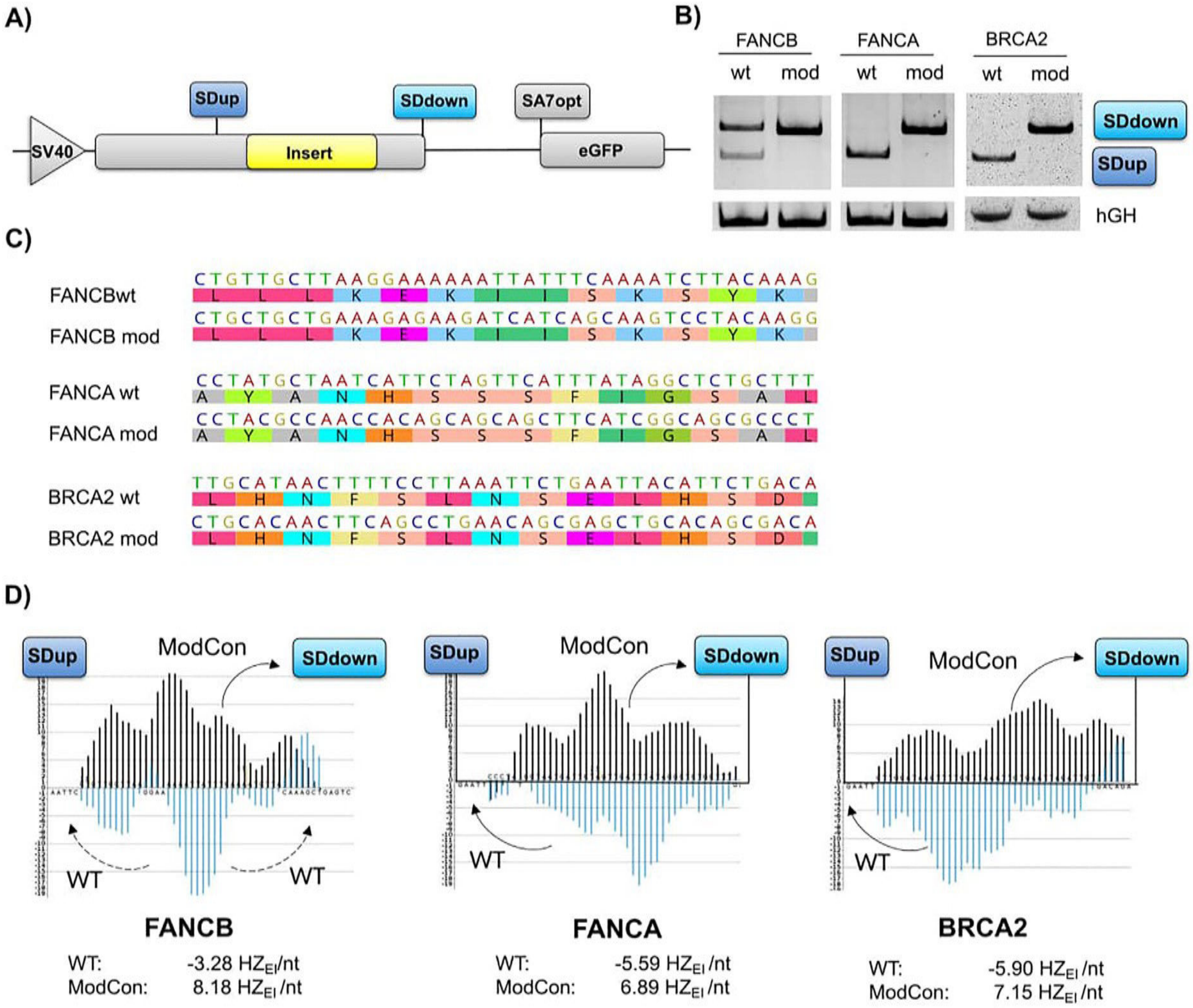

A)
SV40
SDup
Insert
SDdown
SA7opt
eGFP
B)
FANCB
FANCA
BRCA2
wt
mod
wt
mod
wt
mod
SDdown
SDup
hGH
C)
FANCBwt
CTGTTGCTTAAGGAAAAAATTATTTCAAAATCTTACAAAG
L L L K E K I I S K S Y K
FANCB mod
CTGCTGCTGAAAGAGAAGATCATCAGCAAGTCCTACAAGG
L L L K E K I I S K S Y K
FANCA wt
CCTATGCTAATCATTCTAGTTCATTTATAGGCTCTGCTTT
A Y A N H S S S F I G S A L
FANCA mod
CCTACGCCAACCACAGCAGCAGCTTCATCGGCAGCGCCCT
A Y A N H S S S F I G S A L
BRCA2 wt
TTGCATAACTTTTCCTTAAATTCTGAATTACATTCTGACA
L H N F S L N S E L H S D
BRCA2 mod
CTGCACAACTTCAGCCTGAACAGCGAGCTGCACAGCGACA
L H N F S L N S E L H S D
D)
SDup
ModCon
SDdown
WT
WT
SDup
ModCon
SDdown
WT
SDup
ModCon
SDdown
WT
FANCB
WT: -3.28 HZ_{EI}/nt
ModCon: 8.18 HZ_{EI}/nt
FANCA
WT: -5.59 HZ_{EI}/nt
ModCon: 6.89 HZ_{EI}/nt
BRCA2
WT: -5.90 HZ_{EI}/nt
ModCon: 7.15 HZ_{EI}/nt

Fig. 5. Switch in splice donor usage by nucleotide sequences encoding the same amino acids. A) Splice donor competition reporter system with the SV40 promoter, the SV40 poly-A site and a strong splice acceptor site. Between the two identical, strong splice donors (SDup and SDdown), any sequence can be inserted and studied regarding its effect on splice donor selection. B) RT-PCR analysis showing a switch in donor usage upon increasing the total HZ_{EI} of the sequence in between, while keeping the coded amino acids, observed for exemplary sequences from FANCB, FANCA or BRCA2. C) Encoded amino acid sequence of the wild type (wt) and the ModCon (mod) generated alternative sequences. D) HEXplorer profiles of the respective tested sequences, with the wild type sequences depicted in blue and the ModCon generated sequences depicted in black. The total HZ_{EI}/nt of the sequences is shown below. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

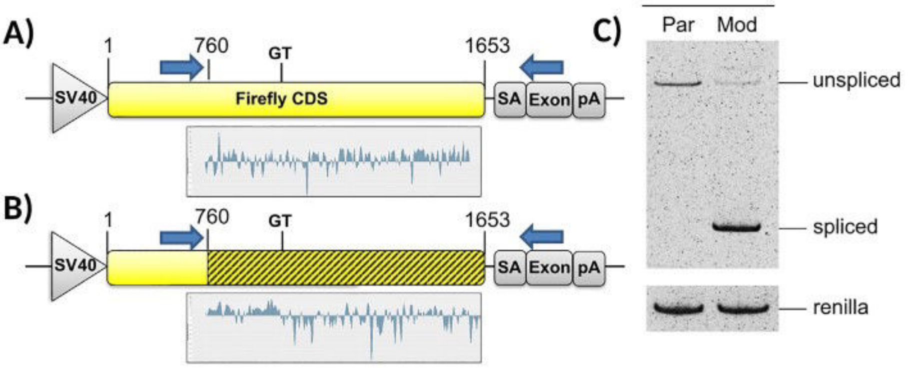


Fig. 6. Luciferase reporter construct encoding identical amino acids and associated splicing pattern. Depicted is the reporter system including the SV40 promoter (**SV40**), a strong splice acceptor site (**SA**), an artificial exon sequence (**Exon**) and the SV40 polyadenylation signal (**pA**). (A) Parental coding sequence of the firefly luciferase, holding an unused GT site (**GT**) at position 1001. (B) ModCon optimized luciferase CDS (hatched), encoding identical amino acids, but containing maximized GT site (**GT**) SSHW and shifted HEXplorer profile between GT site and SA. The 11 nt of the GT site (CAG/GTATCAGG) were not modified. The HEXplorer profiles are shown below the CDS. Primer positions are indicated by blue arrows. (C) RT-PCR analysis showing activation of the internal donor site after modifying the firefly luciferase CDS of the parental construct (**Par**) with ModCon to increase its SSHW (**Mod**). Sequence positions refer to the first nucleotide of the luciferase CDS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sequences with "opposite sign" HZ_{EI} scores without altering the encoded amino acid sequence. Adjusting the SSHW of an unused GT site within the luciferase reporter enabled activation of this sequence as splice site. To increase the possibility of a splicing event within the CDS and due to position of restriction sites, we modified the 240 nt sequence upstream and additionally adjusted the total HZ_{EI} of the 594 nt long sequence downstream of the GT site to mimic intron-like HZ_{EI} profiles, using functions of the ModCon R-package. All adjustments were done exclusively based on synonymous mutations, maintaining the encoded amino acid sequence. Additional embedded functions allow degradation of the intrinsic strength of cryptic splice sites within a given nucleotide sequence.

The ModCon algorithm and its R package implementation thus open the perspective to conveniently assist in reporter design by either introducing novel splice sites, silencing accidental, undesired splice sites, and by generally modifying the entire mRNP code while maintaining the genetic code.

Availability and implementation

The ModCon R-script is an open source R package available with all needed data in the GitHub repository (https://github.com/caggtaagtat/ModCon). It was uploaded to the Bioconductor R package library.

Author statement

All authors have seen and approved the final version of the manuscript being submitted. They warrant that the article is the authors' original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Funding

This work was supported by the Forschungskommission of the Medical Faculty, Heinrich-Heine-Universität Düsseldorf (H.S.), the Heinz-Ansmann Stiftung für AIDS-Forschung (H.S.) and the Jürgen Manchot Stiftung (L.M., P.O., A.R., H.S.).

CRediT authorship contribution statement

Johannes Ptok: Conceptualization, Software, Methodology, Investigation, Validation, Visualization, Writing - original draft. **Lisa Müller:** Validation, Visualization, Investigation, Writing - original draft. **Philipp Niklas Ostermann:** Validation, Visualization, Investigation. **Anastasia Richie:** Validation, Visualization, Investigation. **Alexander T. Dilthey:** Methodology. **Stephan Theiss:** Supervision, Writing - review & editing, Writing - original draft. **Heiner Schaal:** Conceptualization, Supervision, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Aljoscha Tersteegen for cloning assistance and Gene Yeo for his friendly approval to integrate the MaxEntScan scoring algorithm into our software.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.05.033.

References

[1] Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 1977;74(8):3171–5.
[2] Khodor YL, Rodriguez J, Abruzzi KC, Tang C-H- A, Marr MT, Rosbash M. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev 2011;25(23):2502–12.
[3] Aebi M, Hornig H, Padgett RA, Reiser J, Weissmann C. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell 1986;47(4):555–65.
[4] Matera AG, Wang Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol 2014;15(2):108–21.
[5] Freund M et al. A novel approach to describe a U1 snRNA binding site. Nucleic Acids Res 2003;31(23):6963–75.
[6] Freund M et al. Extended base pair complementarity between U1 snRNA and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. Nucleic Acids Res 2005;33(16):5112–9.
[7] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 2004;11(2-3):377–94.
[8] Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 2005;6(5):386–98.
[9] Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 2008;14(5):802–13.
[10] Baralle M, Baralle FE. The splicing code. Biosystems 2018;164:39–48.
[11] Ptok J, Müller L, Theiss S, Schaal H. Context matters: Regulation of splice donor usage. Biochim Biophys Acta Gene Regul Mech 2019;1862(11-12):194391. https://doi.org/10.1016/j.bbagrm.2019.06.002.
[12] Erkelenz S, Mueller WF, Evans MS, Busch A, Schoneweis K, Hertel KJ, et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. RNA 2013;19(1):96–102.
[13] Erkelenz, S., et al., Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res, 2014. 42(16): p. 10681-97.
[14] Brillen AL et al. Succession of splicing regulatory elements determines cryptic 5ss functionality. Nucleic Acids Res 2017;45(7):4202–16.
[15] Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control. and artificial intelligence 1992.
[16] Kammler S et al. The strength of the HIV-1 3' splice sites affects Rev function. Retrovirology 2006;3:89.
[17] Gehring NH, Wahle E, Fischer U. Deciphering the mRNP code: RNA-bound determinants of post-transcriptional gene regulation. Trends Biochem Sci 2017;42(5):369–82.
[18] Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell 2019;176(3):535–548.e24.
[19] Zhang XH et al. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res 2003;13(12):2637–50.

2.3.2 Publication VII: Let it go: HIV-1 cis-acting repressive sequences.

Splicing regulatory proteins facilitate many different biological functions apart from regulation of splice site usage, such as alternative polyadenylation, RNA stability and nuclear export. This was, for instance, demonstrated by analyzing *cis*-acting repressive sequence elements (CRS), that repress export of transcripts of human immunodeficiency virus type 1 (HIV-1) [150]. Analyzing previously described CRS of the HIV-1 reference genome for their binding capacity of splicing regulatory proteins showed, that sequence elements, that were described to repress nuclear export were predicted to introduce HEXplorer-predicted hnRNP binding motifs. hnRNP proteins are predominantly described to reduce RNA nuclear export upon binding and RNA transcript, with a few exceptions. Generally, viral export elements could be grouped into export supporting or export repressing function by their overall total HEXplorer score, which implies, that their function could be predominantly facilitated by splicing regulatory proteins. It is therefore important to note, that any manipulation or change in the binding landscape of splicing regulatory proteins might not only alter splice site usage, but additionally alter further attributes of the resulting RNA transcripts, based on the change in SRP composition.

The following article was published in J Virol.2021 Jul 12;95(15):e0034221 (doi: 10.1128/JVI.00342-21), by

Philipp Niklas Ostermann, Anastasia Ritchie, **Johannes Ptok**, Heiner Schaal

Contribution

PNO, AR and HS wrote the manuscript and created the figures. JP helped with the bioinformatic analysis of the sequence elements. Individual contribution of JP at around 10%.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

MINIREVIEW

Let It Go: HIV-1 *cis*-Acting Repressive Sequences

Philipp Niklas Ostermann,[a] Anastasia Ritchie,[a] Johannes Ptok,[a] Heiner Schaal[a]

[a]Institute of Virology, University Hospital Düsseldorf, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Philipp Niklas Ostermann and Anastasia Ritchie contributed equally to this work. Author order was determined alphabetically.

ABSTRACT After human immunodeficiency virus type 1 (HIV-1) was identified in the early 1980s, intensive work began to understand the molecular basis of HIV-1 gene expression. Subgenomic HIV-1 RNA regions, spread throughout the viral genome, were described to have a negative impact on the nuclear export of some viral transcripts. Those studies revealed an intrinsic RNA code as a new form of nuclear export regulation. Since such regulatory regions were later also identified in other viruses, as well as in cellular genes, it can be assumed that, during evolution, viruses took advantage of them to achieve more sophisticated replication mechanisms. Here, we review HIV-1 *cis*-acting repressive sequences that have been identified, and we discuss their possible underlying mechanisms and importance. Additionally, we show how current bioinformatic tools might allow more predictive approaches to identify and investigate them.

KEYWORDS HIV-1, *cis*-acting repressive sequences, instability or inhibitory elements, nuclear retention, bioinformatic analysis, HEXplorer, hnRNPs

Viruses are instrumental in advancing our understanding of cellular mechanisms. Because viruses are obligate parasites, viral replication strategies must be well adapted to host cell machinery. Regardless of genome structure or replication mechanism, all viruses use cellular processes in order to propagate. Several viruses replicate in the nucleus to gain access to certain host factors required for viral replication. Research into replication mechanisms of these viruses, such as human immunodeficiency virus type 1 (HIV-1), has yielded important insights into the exploited host cell mechanisms, thus contributing to our current understanding of the splicing machinery and nuclear RNA export (1).

The complex replication cycle of HIV-1, the underlying cause of AIDS, has been extensively studied (2). HIV-1 viral particles harbor two identical positive-sense single-stranded RNA (ssRNA) copies that are reverse transcribed during infection, resulting in proviral DNA that is integrated into the host cell genome, which also contributes to integration site-specific effects on host cells (3–7). HIV-1 then uses the host cell machinery to express a single viral pre-mRNA variant. Using at least four splice donor sites and seven splice acceptor sites, HIV-1 generates more than 50 different viral mRNA transcripts through alternative splicing (8–10).

Traditionally, human transcripts must be fully matured before being exported from the nucleus (11). This includes splicing, which results in exon junction complexes (EJCs) being deposited on the transcript. TREX has been shown to interact with the most 5′ EJC and recruit nuclear export factors NXF1 and NXT1, leading to the nuclear export of these mature transcripts (12). HIV-1 takes advantage of host export pathways to export the intronless viral transcripts coding for Tat, Rev, and Nef (Fig. 1, 2 kb), which characterize its early gene expression (13). Intronless refers here to already processed HIV-1 mRNAs in which no further sequence removal via splicing using HIV-1 splice sites is possible.

In contrast to intronless transcripts, intron-containing transcripts still harbor at least one sequence that is bordered by functional splice sites that were not used for splicing

Citation Ostermann PN, Ritchie A, Ptok J, Schaal H. 2021. Let it go: HIV-1 *cis*-acting repressive sequences. J Virol 95:e00342-21. https://doi.org/10.1128/JVI.00342-21.

Editor Britt A. Glaunsinger, University of California, Berkeley

Copyright © 2021 American Society for Microbiology. All Rights Reserved.

Address correspondence to Heiner Schaal, schaal@uni-duesseldorf.de.

Accepted manuscript posted online 12 May 2021

Published 12 July 2021

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

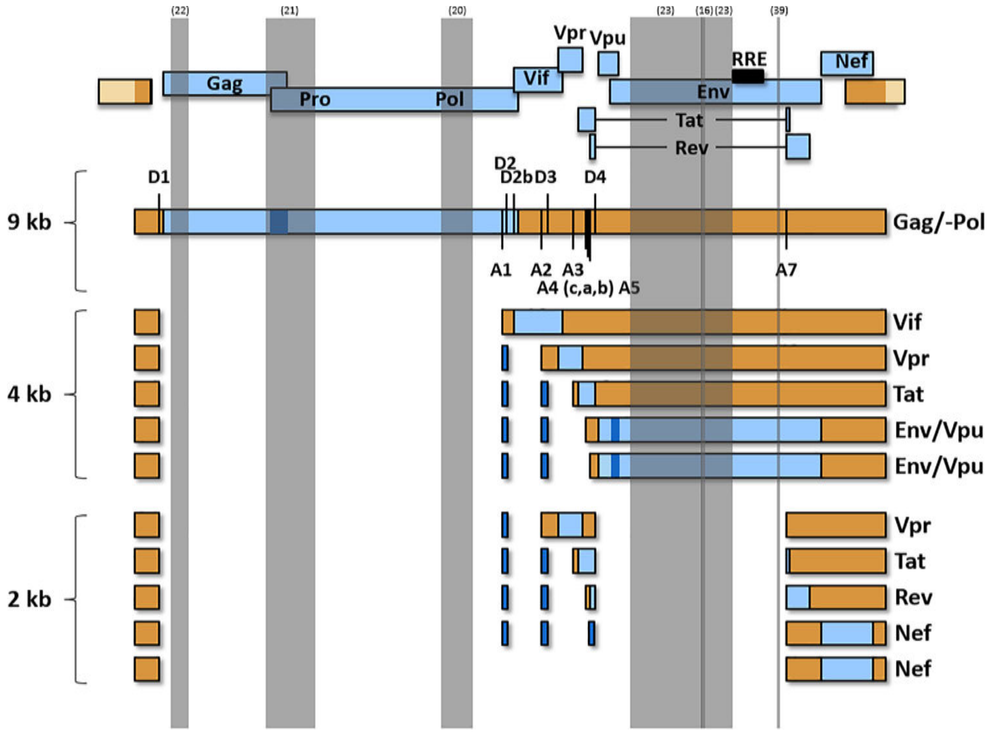


FIG 1 Schematic illustration of HIV-1 CRS within unspliced and intron-containing viral transcripts. The HIV-1 provirus with viral ORFs is shown in blue. Untranslated regions are shown in brown, and overlapping open reading frames are in dark blue. Gray shaded bands highlight positions of HIV-1 CRS that inhibit cytoplasmic RNA accumulation within the full genome, as well as spliced isoforms. The references describing these CRS are noted at the top of the image. 9 kb, viral pre-mRNA (RNA genome, *gag/pol*) with splice sites; 4 kb, intron-containing viral transcripts coding for Vif, Vpr, Tat, Env, and Vpu; 2 kb, intronless viral transcripts coding for Tat, Rev, and Nef. Illustrated are CRS that were well defined during the underlying experimental studies and whose nucleotide positions could be reconstructed. D1 to D4, 5′ splice (donor) sites; A1 to A7, 3′ splice (acceptor) sites.

during processing of this particular HIV-1 mRNA. The terms exon and intron *per se* are to be understood as gene specific. For example, the subgenomic region limited by the HIV-1 splice sites SD4 and SA7 is intronic for the genes *tat*, *rev*, and *nef* but exonic for *env*.

For HIV-1 replication, nuclear export of unspliced and intron-containing transcripts is essential. However, transcripts that harbor introns are typically trapped in the nucleus (Fig. 1, 9 kb and 4 kb). Therefore, export of unspliced and intron-containing viral transcripts requires interaction between Rev and the Rev response element (RRE), a viral RNA element that forms a secondary structure located within the HIV-1 *env* coding region (14–17). Thus, the expression of HIV-1 structural and enzymatic precursor proteins Gag, Pol, and Env is dependent on nuclear-cytoplasmic shuttling of Rev and the RRE in *cis* (3, 4, 17–19).

While investigating HIV-1 Rev-RRE interaction, viral sequences were identified that prevented cytoplasmic RNA accumulation in the absence of Rev (14, 16, 18). Although in the following years many of these *cis*-acting repressive sequences (CRS) were identified throughout the HIV-1 genome, no uniformly recognizable sequence commonality was detected, and underlying mechanisms remain unclear (20–26).

In contrast to well-characterized viral elements that promote nuclear export, like the HIV-1 RRE, whose different aspects have been reviewed regularly since its discovery (for reviews, see, for example, references 27–29), we spotlight studies identifying sequences that regulate gene expression by preventing cytoplasmic accumulation of viral RNA. Although the focus of recent studies has shifted away from CRS to other features of the complex life cycle of HIV-1, the HIV-1 CRS mode of action still needs to be described (27). Interestingly, CRS were also identified in other viruses, as well as in cellular transcripts. Therefore, in this review, we provide a comprehensive overview of CRS found in HIV-1 with respect to potential common mechanisms underlying CRS

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

function. In the context of these CRS, we show that the bioinformatic HEXplorer algorithm for prediction and landscaping of splicing regulatory elements may also be suitable to predict and to analyze RNA sequences that prevent nuclear export (30).

HIV-1 CRS THROUGHOUT THE VIRAL GENOME: HISTORICAL RETROSPECTIVE

The first viral sequences shown to inhibit nuclear export, later described as CRS, were found in the HIV-1 *env* gene (16). In that study, chloramphenicol acetyltransferase (CAT) reporter constructs containing different fragments of the HIV-1 *env* gene showed decreased expression levels. Expression was partially rescued by coexpression of the HIV-1 protein Rev, suggesting that these fragments contain an element enabling Rev-mediated export, namely, the RRE, but also sequences that reduce the expression. The CRS acted independently of any splice sites, which also have an inhibitory effect on nuclear export, hinting at an unknown mechanism for nuclear retention (31–34). Deletion mutants revealed an RNA sequence of only 60 nucleotides that repressed CAT expression even with Rev coexpression. These findings revealed the existence of CRS not only within the RRE but also in regions of HIV-1 *env* not involved in Rev-mediated nuclear export. Although the underlying mechanism was not investigated further, Rosen and coworkers were the first to describe viral sequences responsible for nuclear retention of intron-containing viral transcripts (16).

While investigating Rev-mediated expression of HIV-1 structural proteins, additional CRS within the HIV-1 *gag* and *env* genes were suggested (35). All were localized in intronic sequences (Fig. 1) and contributed to low levels of unspliced and intron-containing HIV-1 RNA in the cytoplasm by a Rev-independent mechanism that is overridden by the Rev-RRE system (35).

After the identification of CRS in HIV-1 *env* and the first hints of their presence in *gag*, the necessity of Rev for expression of all viral structural genes led to the identification of specific CRS within the *gag* and *gag/pol* genes. Following the same approach as for the analysis of CRS within the *env* gene, joining viral *gag* and *gag/pol* sequences 3′ terminally to the CAT gene inhibited expression of CAT protein (20). In particular, the *pol* fragment decreased CAT expression comparably to the previously identified *env* CRS. This *pol* CRS-induced inhibition of CAT gene expression was rescued only by addition of HIV-1 RRE in *cis* and Rev in *trans*. Deletion mutants indicated again that splice sites within the CRS did not play any role in the inhibition of CAT expression. Additionally, the use of cellular fractionation strengthened the hypothesis that CRS function by trapping transcripts in the nucleus in the absence of Rev (20).

In a subsequent study, a 1,295-nucleotide-long *cis*-acting inhibitory region (IR) within the *gag* gene (IR-1) and a 1,932-nucleotide-long IR within the *pol* gene (IR-2) were shown to inhibit expression in distinct reporter systems (21). Although deletion analysis indicated the existence of independent CRS within these regions, the complete regions exerted the strongest inhibitory effect. Focusing on the complete regions, cellular fractionation assays with a CAT gene expression system demonstrated an inhibitory effect of IR-1 on nuclear export, while IR-2 was not analyzed in that assay (21). Due to the lengths of IR-1 and IR-2, there were no conclusions about any motif- or sequence-specific effects. However, the experiments conducted clearly demonstrated an inhibitory effect of regions within the HIV-1 *gag* gene on nuclear export (21).

A deletion-based investigation was then able to narrow a CRS in *gag* down to a 218-nucleotide-long fragment at the 5′ end of the open reading frame (ORF) that profoundly inhibited CAT gene expression in an HIV-1 *tat*-based CAT expression system, in a splicing-independent manner (22). CAT expression was partially rescued by the HIV-1 RRE-Rev system, showing that this repressive sequence is important for Rev-regulated gene expression. Intriguingly, the experiments indicated decreased amounts of repressive-sequence-containing RNAs in the nucleus as well as the cytoplasm, suggesting that this novel element affected mRNA stability rather than nuclear export (22). Therefore, this repressive sequence was introduced as instability or inhibitory element 1 (INS-1), although the possibility that it influences expression via nuclear retention

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

could not be ruled out. Deletion of a long stretch of AU-rich regions within INS-1, which shows a generally high AU content, diminished its inhibitory effect (22).

Based on these findings, it was possible to reach a high level of Rev-independent *gag* expression by mutational inactivation of INS-1 without deleting parts of its sequence or altering the resulting amino acid sequence (36). Insertion of synonymous mutations along the INS-1 region of an HIV-1 p17^{gag} expression vector showed increased expression levels, compared to the wild-type construct. This confirmed the existence of a sequence element with a strong negative effect on protein expression within the 5′ region of HIV-1 *gag* that might affect RNA stability or nuclear export and whose inhibitory effect could be abrogated by synonymous mutations. Since most of the nucleotides that were mutated were found to be highly conserved in different HIV-1 and HIV-2 isolates, the inhibitory effect of INS-1 was assumed to be important for regulation of viral gene expression (36). Together, these studies identified INS-1 in the 5′ end of HIV-1 p17^{gag}, which might explain the early and presumably stronger impact on a transcript's further processing exerted by RNA elements that are located near the 5′ end (22, 36) and which might mean that the fate of the mRNA is decided before the RNA polymerase reaches the 3′ end of the region to be transcribed.

The nonmutated INS-1 element was found to bind poly(A)-binding protein 1 (PABPC1), whereas the mutant p17^{gag} version did not (37). Furthermore, transfection of different cell lines with the p17^{gag} expression vector and analysis of intracellular PABPC1 protein levels indicated a positive correlation between the inhibitory effect of INS-1 and PABPC1 expression. These results suggested a role for PABPC1 as an inhibitory factor binding to HIV-1 CRS, thus preventing viral protein expression (37).

While PABPC1 was described as a binding factor of HIV-1 INS-1 with a potential role in regulating viral protein expression, additional CRS were identified within the *gag/pol* gene (26). Here, a comprehensive approach analyzing several RNA regions within the HIV-*1 gag/pol* gene identified sequence elements INS-2 in the p24 capsid and INS-3 in the viral protease coding region. Additionally, other sequences with a presumed function as INS elements were found in HIV-1 reverse transcriptase and integrase coding regions. Mutational inactivation by insertion of silent mutations rescued protein expression in reporter constructs in the absence of HIV-1 Rev. Measuring the amount of the respective reporter mRNAs in the cytoplasm as well as nucleus supported a model in which such INS elements affect RNA half-life rather than nuclear export, as shown previously for other repressive sequences. Although some of the previously discussed INS elements overlap with the characterized CRS in HIV-1 *pol*, the authors concluded that INS elements exert different effects and are therefore distinct from CRS (26).

These INS-like sequences were also described within HIV-1 *env* transcripts (23). Analysis of different *env* fragments in a p37^{gag} expression vector decreased p24 capsid protein expression for single fragments, and the combination of some fragments increased this inhibitory effect. As shown for other CRS and INS elements, the observed inhibitory effects were not dependent on splice sites, viral factors in *trans*, or the RRE in *cis*. Thus, interaction with cellular proteins was suggested as the mechanism behind posttranscriptional regulation of gene expression and counteraction of CRS by RRE-Rev-mediated export (23).

At the same time, an independent study identified a CRS overlapping the RRE that caused nuclear retention of reporter mRNA in transfected *Drosophila* cells (24). It was shown that a 240-nucleotide-long deletion also spanning the RRE led to nuclear export in the context of the HIV-1 *env* expression vector used. Since deletion of this presumed CRS in the gp41^{env} coding frame alone reversed nuclear retention, it was suggested that the CRS is not only sufficient but also necessary to prevent nuclear export of *env* and *gag/pol* transcripts in *Drosophila* cells. Furthermore, the mutation of the 3′ *tat/rev* splice acceptor site did not impair the inhibitory effect on nuclear export, again ruling out that splice signals are responsible for nuclear retention of unspliced and intron-containing viral transcripts (24).

The CRS overlapping the RRE in *Drosophila* cells was also found to function in primate and human cells, including the CD4^{+} Jurkat T-cell line (25). As expected, when

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

employing the HIV-1 *env* expression vectors adapted for mammalian cells, nuclear retention was demonstrated in the absence of Rev for all constructs except for that missing the 240-nucleotide CRS overlapping the RRE. In that study, CRS were found only within the gp41^{env} coding region and not within the gp120^{env} coding frame. This result clearly contradicted the previously described findings, which also demonstrated CRS within gp120^{env}, thus challenging the conclusion that the newly identified gp41^{env} CRS on its own is sufficient for nuclear retention in an HIV-1 infection context (23, 25). Rather, the study found an independent CRS further unravelling the convoluted system of such sequences found throughout the HIV-1 genome, which seem to act by a mechanism conserved in *Drosophila* and mammalian cells (24, 25).

Next to their role as mere inhibitory signals, the use of different β-globin constructs revealed the requirement for elements such as the p17^{gag} INS-1 element as well as the *pol* CRS for efficient Rev-RRE-mediated nuclear export in the context of efficiently spliced pre-mRNA constructs (20, 22, 36, 38). Therefore, the authors suggested a spatial separation of INS- and CRS-containing transcripts from the splicing apparatus, not only preventing splicing but also enabling Rev-mediated export (38). Furthermore, the experiments provided evidence that INS elements might not lead to a reduced RNA half-life but inhibit nuclear export like CRS. That comprehensive study not only found INS or CRS dependence of Rev-mediated export by nuclear separation of transcripts from the splicing apparatus but also suggested that INS elements have the same underlying mechanism as CRS (38).

Like PABPC1, heterogeneous nuclear ribonucleoprotein C (hnRNPC) was found to bind specifically to a newly identified CRS within the HIV-1 *env* gene (39). Deletion mapping in different reporter constructs narrowed this strong CRS down to only 48 nucleotides. While the deletion of this CRS alone did not result in cytoplasmic accumulation of full-length *env*, it confirmed the existence of additional synergistic inhibitory elements within the *env* coding region (39).

Regarding the interplay and importance of several CRS within HIV-1 structural genes, it was shown that certain point mutations proximal to HIV-1 splice acceptor site 7 and within the *env* ORF enable Rev-independent Env but not Gag protein expression (40). This observation, combined with a missing exon-junction complex on genomic RNA that might be necessary for nuclear export of genomic RNA, was suggested as the result of CRS inactivation that led to *env* RNA export (40). However, excessive CRS located within the *gag/pol* ORF might offer a different explanation for the observed nuclear retention of genomic but not *env* RNA.

In conclusion, all of these CRS were found within coding regions of structural genes *gag* and *env*, as well as within the *pol* coding region of the unspliced HIV-1 *gag/pol* transcript. Hence, cytoplasmic accumulation of these viral transcripts occurs only in the presence of viral protein Rev, overriding nuclear retention. Therefore, the identified CRS were suggested to play a role in HIV-1 gene regulation by inhibiting expression of structural proteins required for the generation of infectious particles (16, 35). Furthermore, the finding that Rev-mediated export depends on CRS in efficiently spliced transcripts not only revealed a situation in which Rev-mediated export overrides nuclear retention but also indicated that Rev contributes to rapid export of unspliced and intron-containing viral RNA, which by default do not have a license for nuclear export (38).

HIV-1 CRS-EXPLOITING CELLULAR PROTEINS FOR VIRAL STRATEGIES

Several proteins were found to specifically bind the identified HIV-1 CRS or to play a role in nuclear retention of HIV-1 RNA. Among them were RNA-binding proteins PABPC1, NONO, SFPQ, hnRNPA2, hnRNPA1, and hnRNPC (37, 39, 41–43).

As a member of the hnRNP family, hnRNPC not only was shown to bind AU clusters within cellular transcripts but also contains a nuclear retention signal (NRS) of approximately 78 amino acids. As opposed to most other hnRNPs, hnRNPC is found predominately, if not exclusively, in the nucleus (44–46). In addition to its critical role protecting the human transcriptome from U2AF65 binding to Alu elements and thus masking the

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

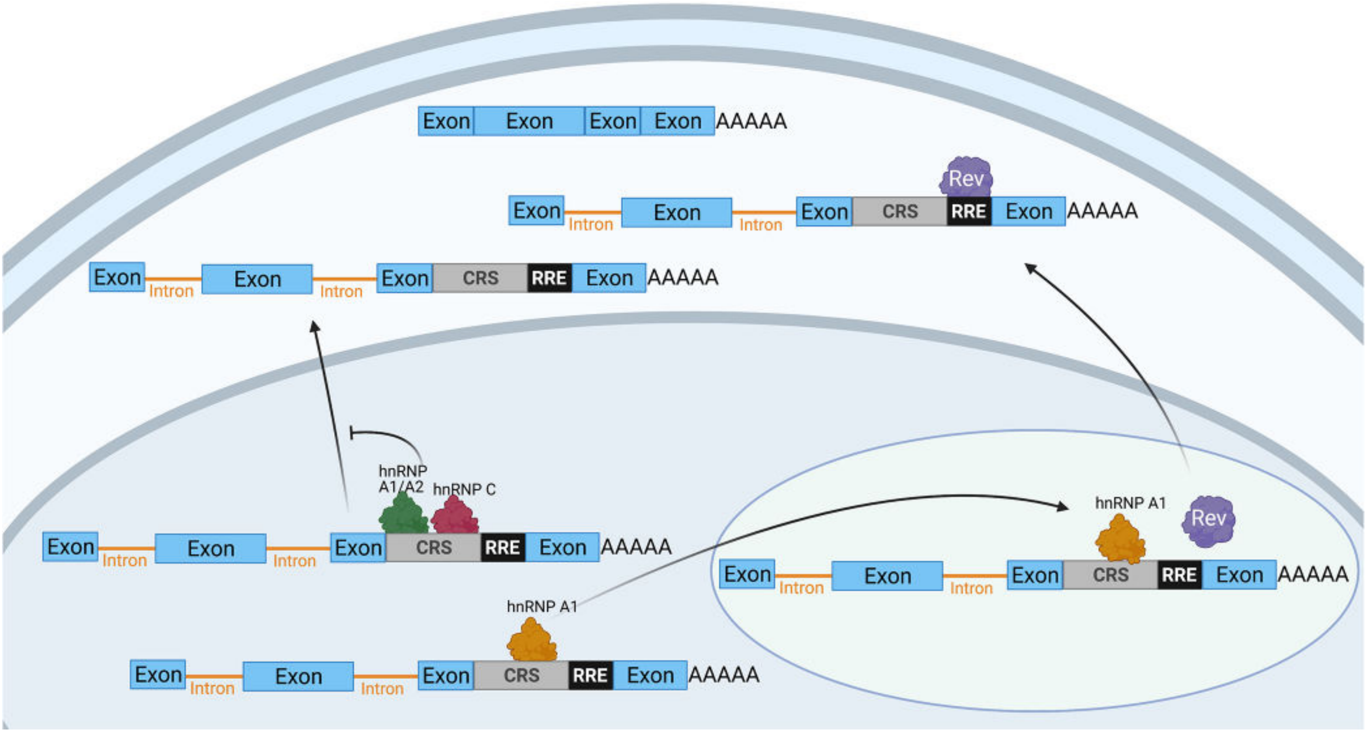

Exon Exon Exon Exon AAAAA
Rev
Exon Intron Exon Intron Exon CRS RRE Exon AAAAA
Exon Intron Exon Intron Exon CRS RRE Exon AAAAA
hnRNP A1/A2
hnRNP C
Exon Intron Exon Intron Exon CRS RRE Exon AAAAA
hnRNP A1
Rev
Exon Intron Exon Intron Exon CRS RRE Exon AAAAA
hnRNP A1
Exon Intron Exon Intron Exon CRS RRE Exon AAAAA

FIG 2 Model for the possible role of the hnRNP family in nuclear export. The hnRNPC NRS overrides the NES found in hnRNPA1 and other shuttling hnRNPs (42). hnRNPA1 collaborates with Rev to export unspliced RNA from the nucleus (40). The image was created with BioRender.

potential use of several hundred thousand cryptic splice sites, hnRNPC is likely a key player in nuclear retention of viral RNA (47). In agreement with this, hnRNPC was also found to play a role in the nuclear retention of human papillomavirus type 1 (HPV-1) RNA (48, 49).

Another member of the hnRNP family, hnRNPA1, was found to bind to a short fragment within the HIV-1 INS-1 element (22, 43). Further, small interfering RNA (siRNA)-mediated knockdown of several hnRNPs to screen potential effects on HIV-1 gene expression revealed that hnRNPA1 depletion led to increased expression of viral structural proteins without altering RNA levels (50). This observation substantiates a possible direct role for hnRNPA1 in nuclear retention of unspliced and intron-containing HIV-1 RNA. In another retrovirus, human T-cell leukemia virus 2 (HTLV-2), hnRNPA1 also binds repressive sequences to inhibit nuclear export. hnRNPA1 expression levels corresponded to HTLV-2 repressive sequence-induced inhibition of gene expression (51). In the context of cellular genes, injection of hnRNPA1 into the nucleus and cytoplasm of oocytes inhibited mRNA export (52). Although both hnRNPC and hnRNPA1 bind AU-containing sequences (44), hnRNPA1 is a shuttling protein that contains a nuclear localization signal as well as an export signal and participates in RNA processing in both the nucleus and cytoplasm (46, 53). In accordance with its shuttling nature, hnRNPA1 has been shown to facilitate the nuclear export of mRNAs (54) (Fig. 2). However, several studies indicate that hnRNPA1 is also involved in the nuclear retention of transcripts (50–52). The mechanism by which hnRNPA1 decides to retain or export transcripts is still unclear and needs more elucidation. HIV-1 seems to induce hnRNPA1 expression, and during late HIV-1 infection nuclear import of hnRNPA1 is inhibited, leading to cytoplasmic accumulation (53). This accumulation may be dependent on the nuclear export of the unspliced viral RNA, since hnRNPA1 still interacts with the RNA once in the cytoplasm (53). Additionally, hnRNPA1 binding stimulates Rev-mediated nuclear export of reporter RNA (38, 43). The concept that the CRS mechanism is separate from the splicing apparatus is reinforced by the finding that hnRNPA/hnRNPB proteins, including hnRNPA1, inhibit HIV-1 pre-mRNA splicing (55). Thus, hnRNPA1 plays different roles in HIV-1 propagation beyond sole retention of unspliced and intron-containing HIV-1 transcripts.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

Within the diverse family of hnRNPs, hnRNPA2 also specifically binds certain elements (A2RE-1 and A2RE-2) in the HIV-1 genome (56). hnRNPA2 is an important factor for viral RNA trafficking in late HIV-1 gene expression. Knockdown of hnRNPA2 along with hnRNPB1 in the absence of viral Rev resulted in cytoplasmic localization of the otherwise nucleus-retained HIV-1 RNA genome (41, 56, 57). Furthermore, the knockdown of hnRNPA2 resulted in increased expression of viral structural proteins, confirming hnRNPA2-dependent nuclear retention of HIV-1 transcripts (50). Although the specific hnRNPA2 binding sites were not initially described as CRS, they, together with hnRNPA2 as binding partner, might play a role in HIV-1 RNA export regulation.

As a protein found solely in the nucleus, hnRNPC is a predestined candidate for retaining HIV-1 unspliced and intron-containing RNA in the nucleus (45, 46). In addition to their roles in nuclear retention of HIV-1 RNA, hnRNPA1 and hnRNPA2 are also well known for their roles in other steps of RNA processing, particularly splicing (41, 43, 44, 56–65). They all seem to function as key regulators of HIV-1 gene expression by nuclear retention of viral transcripts in the absence of Rev, either directly or indirectly by occupying the RNA binding site for an export factor or an adaptor for an export molecule such as an SR protein (66–73). Although hnRNPA1 and hnRNPA2 shuttle between the cytoplasm and the nucleus, hnRNPC and other nucleus-retained hnRNPs contain a NRS that can override the nuclear export signals (NES) found in shuttling hnRNPs (45, 46, 74).

While the exact mechanisms and pathways by which hnRNPC, hnRNPA1, and hnRNPA2 regulate the viral transcriptome are still not completely understood, the studies discussed contributed to a better understanding of their effects on different levels of HIV-1 gene expression and gave crucial insights into the regulation of nuclear RNA export in viruses and eukaryotes.

In addition to hnRNPs, other proteins bind to HIV-1 CRS and participate in HIV-1 RNA nuclear retention. A recent study identified additional proteins implicated in the nuclear retention of HIV-1 RNA using a comprehensive genome-wide CRISPR-Cas9 knockout library-based approach (75–77). In addition to proteins that contribute to HIV-1 splicing regulation, that study identified three putative proteins (CRNKL1, DHX38, and BUD31) that regulated viral gene expression via nuclear retention. CRNKL1 showed the strongest effect on nuclear retention of cellular genes, including intron-containing transcripts (77).

Therefore, it would be conceivable that simply a certain number of retention proteins must bind the RNA in order to effectively retain it in the nucleus. Inactivation of individual but not necessarily adjacent RNA binding sites would independently lead to a reduction in the overall binding of these retention proteins, thus weakening nuclear retention. If sufficient RNA binding sites, which could be distributed over a larger area, are destroyed, then these transcripts could eventually accumulate in the cytoplasm in a Rev-independent manner based on cellular export factors.

INHIBITION OF CYTOPLASMIC mRNA ACCUMULATION BY CRS IS FOUND THROUGHOUT DIFFERENT VIRAL FAMILIES

Like HIV-1, other members of the *Retroviridae* family make use of CRS. For instance, equine infectious anemia virus (EIAV) contains several CRS that were suggested to inhibit nuclear export of viral transcripts (32). Again, using a heterologous CAT reporter, it was shown that these CRS could contribute to Rev dependence and that the inhibitory effect exerted was due to binding of cellular proteins, rather than viral proteins.

In addition to the *Retroviridae* family, CRS were found to be important in other viral families (32, 48, 78–80). Research into human herpesvirus 8 (HHV-8) showed that herpesviruses also contain CRS. Here, it was found that the polyadenylated nuclear RNA (PAN) contains a 79-nucleotide-long sequence integral to its nuclear entrapment (81). Later, this element was characterized as an RNA secondary structure constituting a stability element rather than a CRS, called the element for nuclear expression (ENE), which can sequester its own poly(A) tail (82). Although the PAN-ENE might not be a classic CRS, its high nuclear abundance and early description as a NRS indicates a function similar to

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

that of CRS. In this context, the striking pyrimidine/purine (Y/R) bias found in ENEs is reminiscent of described CRS elements, contributing to the CRS-like function (83).

Another interesting aspect related to the nuclear export of HHV-8 is the viral protein ORF57. Expression of ORF57 leads to nuclear export of viral transcripts otherwise trapped and degraded within the nucleus (84). Until now, no specific ORF57 response element was found to contribute to RNA export, like in the case of the HIV-1 Rev-RRE system (85). ORF57 rescues nuclear export of those HHV-8 RNAs that exhibit a high AT content and also was able to rescue nuclear export of an HIV-1 *gag* reporter construct or a "deoptimized" enhanced green fluorescent protein (eGFP) transcript, both of which contained comparable high AT contents (84). Since these transcripts have a higher density of hnRNP binding sites than do transcripts with lower AT contents, ORF57 was suggested to exert its nuclear export-promoting function via binding to cellular proteins mainly involved in nuclear retention (84). Therefore, ORF57-mediated nuclear export illustrates another way viruses not only make use of but override the cellular nuclear RNA retention machinery.

As a member of the *Hepadnaviridae* family, hepatitis B virus (HBV) was shown to harbor sequences that inhibit nuclear export in an additive manner. Polypyrimidine tract-binding protein (PTB) contributes to this additive effect by binding separate repressive sequences (86).

Another important virus family that is dependent on CRS for regulated gene expression is the *Papillomaviridae* family. Studies on bovine papillomavirus type 1 (BPV-1) identified a short fragment of about 50 nucleotides that could inhibit CAT expression when inserted only in the sense orientation into the 3′ untranslated region (UTR) of the expression vector employed (78, 79). Additionally, deletion of this fragment from a BPV-1 L1 expression vector led to higher levels of the expressed mRNA (78). Although this short sequence exerted the strongest negative effect, neighboring sequences were also found to show inhibitory potential (78). In the identified fragment, four overlapping motifs with homology to the consensus 5′ splice site were found to be responsible for nuclear retention (79). In addition to its protective function on transcript integrity by suppressing cryptic polyadenylation signals, binding of U1 small nuclear ribonucleoprotein (snRNP) likely leads to the observed nuclear retention (87). Together, these results further support the idea that not one specific element alone, but rather distinct elements or even clusters of still poorly defined sequences with binding affinity for retention proteins, like hnRNPC or U1 snRNP, have the ability to prevent cytoplasmic mRNA accumulation.

Other members of the *Papillomaviridae* family, HPV-1 and HPV-16, were specifically shown to use AU-rich sequences to contribute to decreased cytoplasmic mRNA levels to regulate gene expression (80, 88, 89). Successive studies revealed that the cellular factors hnRNPC1 and hnRNPC2 bind to the negative regulatory elements within HPV-1 transcripts (48, 49). That work also demonstrated that the specific pentameric motifs AUUUA and UUUUU found within a short AU-rich inhibitor sequence in the HPV-1 3′ UTR were functionally important for nuclear retention (48). These motifs are also found primarily in HIV-1 introns (Fig. 3), where they may contribute to the retention of intron-containing transcripts.

A BIOINFORMATIC TOOL, THE HEXplorer ALGORITHM, BORROWED FROM SPLICE ANALYSES TO INSPECT CRS

The HEXplorer algorithm allows landscaping of genomic sequences, revealing potential splicing regulatory elements, probably SR and hnRNP binding sites (30, 90). It follows a RESCUE (Relative Enhancer and Silencer Classification by Unanimous Enrichment) sequence analysis approach, resulting in a HEXplorer score per nucleotide (30, 91). Based on 6-mer distributions around human splice donor sites, the HEXplorer score is the average Z-score of all six hexamers that overlap with one single nucleotide. Positive Z-scores are associated with 6-mers that were found more frequently upstream of splice donor sites, whereas negative Z-scores are associated with 6-mers that were found more frequently downstream of splice donor sites. Proximal binding

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

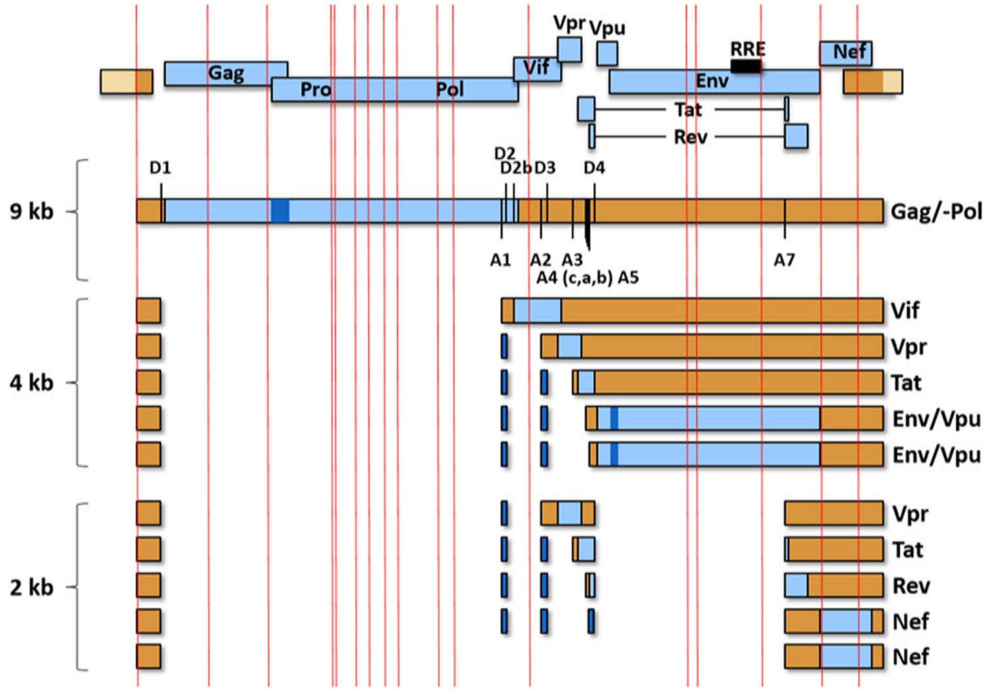


FIG 3 Schematic illustration of HIV-1 AUUUA/UUUUU sequences within unspliced and intron-containing viral transcripts. HIV-1 provirus with viral ORFs is shown in blue. Untranslated regions are shown in brown, and overlapping open reading frames are in dark blue. Red vertical lines indicate AUUUA/UUUUU sequences within the HIV genome. 9 kb, viral pre-mRNA (RNA genome, *gag/pol*) with splice sites; 4 kb, intron-containing viral transcripts coding for Vif, Vpr, Tat, Env, and Vpu; 2 kb, intronless viral transcripts coding for Tat, Rev, and Nef; D1 to D4, 5′ splice (donor) sites; A1 to A7, 3′ splice (acceptor) sites.

of SR proteins upstream or binding of hnRNPs downstream of a particular splice donor site enhances its usage by interaction with compounds of the molecular machinery regulating splicing. Thus, sequences enriched upstream of splice donor sites (positive HEXplorer scores) often correlate with binding sites of SR proteins, and sequences enriched downstream of splice donor sites (negative HEXplorer scores) often correlate with hnRNP binding sites. Therefore, HEXplorer score profiles indicate a sequence's potential to either enhance or silence nearby splice sites through the binding of SR or hnRNPs. Since an accumulation of different hnRNPs at CRS was detected (see above), the HEXplorer algorithm may also be suitable for the prediction of sequences involved in nuclear retention (67, 69, 70, 72).

Based on the somewhat more length-defined viral CRS described in this review for which nucleotide sequences could be reconstructed, we calculated the HEXplorer score per nucleotide, to allow comparison of sequence elements of different lengths (Table 1). Consistent with preferential binding of hnRNPs to viral CRS as a possible cause of nuclear RNA retention, the analyzed sequences exhibited low HEXplorer scores per nucleotide, from 0.441 to −6.146, which are significantly lower than those of natural RNA elements that are supportive for nuclear export (Fig. 4A).

Mutational inactivation of the inhibitory effect of the INS-1 element (Table 1, positions 867 to 1084) within the HIV-1 *gag* ORF, enabling Rev-independent *gag* expression (22, 36), went along with an increase in the HEXplorer score from −0.048 to +3.63, suggesting reduced hnRNP binding. Furthermore, there is a positive correlation between the length of described repressive sequences and their HEXplorer score per nucleotide (Fig. 4B). This observation could be due to only approximately defined element coordinates or the more-or-less alternating RNA binding sites for RNA-binding proteins, such as hnRNPs that can retain RNA in the nucleus or SR proteins that can serve as adapters for nuclear export factors (67, 69–72).

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

TABLE 1 Viral RNA elements regulating gene expression by inhibiting or supporting cytoplasmic RNA accumulation

Virus	Effect on expression	HEXplorer score per nucleotide	Nucleotide positions	Length (nt)	GenBank accession no. for reference genome	Source
HIV-1	Inhibitory	−4.690	7315–7374	60	HIVHXB2CG	16
	Inhibitory	+0.441	4157–4647	390	HIVHXB2CG	20
	Inhibitory	+0.252	2006–2619	614	HIVHXB2CG	21
	Inhibitory	−0.048	867–1084	218	HIVHXB2CG	22
	Inhibitory	−1.835	6458–6889	432	HIVHXB2CG	23
	Inhibitory	−0.362	6889–7332	444	HIVHXB2CG	23
	Inhibitory	−1.546	7332–7720	389	HIVHXB2CG	23
	Inhibitory	−6.146	8241–8288	48	HIVHXB2CG	39
BPV-1	Inhibitory	−5.766	7094–7146	53	NC_001522	79
EIAV	Inhibitory	−0.880	4479–5170	692	NC_001450.1	32
HPV-16	Inhibitory	+0.056	7008–7226	219	NC_001526	80
HHV-8	Inhibitory	−2.328	29706–29784	79	NC_009333	81
Moloney murine leukemia virus	Supporting	+1.873	725–1037	313	NC_001501	120
HBV	Supporting	+1.904	1154–1413	262	AY128092	124
HBV	Supporting	+1.453	1352–1684	332	AY128092	124
Murine leukemia virus	Supporting	+2.943	2918–3016	99	NC_001362	125
Borna disease virus	Supporting	+2.020	4070–4269	200	NC_030692	92

Taken with the underlying hexamer distribution of the HEXplorer algorithm, there is no HEXplorer score per nucleotide for well-defined repressive pentamers. However, taking the possible neighborhood into account, i.e., adding one missing nucleotide at either site of the respective pentamers, allows calculation of an average Z-score from Z-scores of every hexamer containing the respective pentamer. This way, especially the two described papillomavirus CRS pentameric motifs AUUUA and UUUUU show strong average negative scores of −16.8 and −42.9, irrespective of where they are located in the transcript. Hence, like with the CRS of the HIV-1 genome, the well-defined papillomavirus CRS motifs contribute to a more-negative HEXplorer score per nucleotide of RNA regions harboring them.

In conclusion, algorithms profiling certain protein RNA binding sites such as the HEXplorer algorithm also might exhibit predictive power for identifying CRS. In addition to CRS identification, these algorithms could also be used to characterize CRS by evaluating their hnRNP-binding capacity.

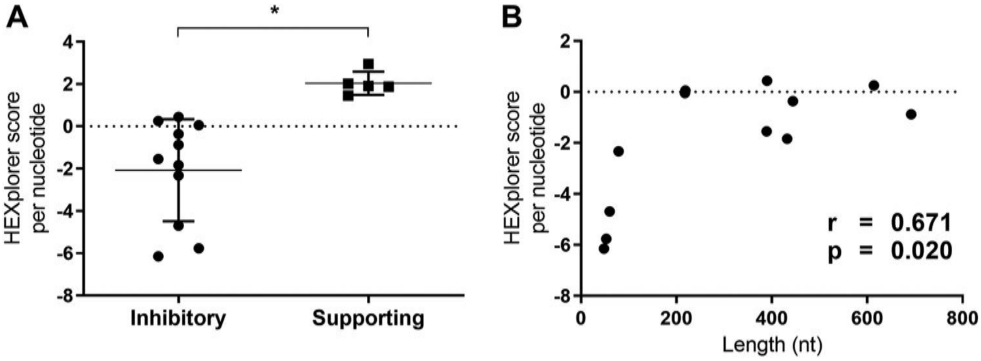


FIG 4 HEXplorer algorithm-based statistical analysis of viral RNA elements inhibiting and supporting cytoplasmic RNA accumulation. (A) HEXplorer scores for inhibitory and supporting elements. Based on the RNA elements listed in Table 1, HEXplorer scores per nucleotide for RNA elements with inhibitory effects on cytoplasmic accumulation were compared to HEXplorer scores per nucleotide for RNA elements that were shown to support nuclear export. The graph shows individual data points and mean values ± standard deviations. Statistical significance was determined by unpaired two-tailed *t* test. *, $P < 0.05$. (B) Correlation between the lengths of identified repressive sequences and the respective HEXplorer scores per nucleotide. Statistical significance was determined by nonparametric Spearman two-tailed correlation. In panels A and B, statistical analysis and graphical illustration were performed with GraphPad Prism.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

CELLULAR TRANSCRIPTS CONTAINING CRS

The comprehensive work on *Papillomaviridae* strains found that viral CRS share their function with those in cellular genes. Such CRS and other inhibitory sequences are localized either within the ORF, e.g., in the case of β-globin mRNA (82) and the α-chain of collagen type 1 (COL1A1) (92), or within the 3′ UTR, e.g., in the case of c-*fos* mRNA (46) and *p14/robld3* (93). The β-globin mRNA contains a CRS within its ORF whose inhibitory function was disrupted when the transcript was extended or spliced or elements that promote export were added (94).

Besides such physiological examples in ORFs or in the 3′ UTR, CRS have also been described in introns that remain in the transcript due to splice site mutations. The inclusion of intron 26 (HEXplorer score of −2.671) in COL1A1, caused by such a human pathogenic mutation, leads to the nuclear retention of the transcript by impeding its exit from the SC-35 domain. The COL1A1 deficiency causes osteogenesis imperfecta (92). A similar effect of intron retention can be observed in transcripts with detained introns, in which slow excision of intronic sequences results in nuclear retention rather than export from the nucleus (95, 96). In the *Clk1* transcript, intron 3 (HEXplorer score of −6.22) and intron 4 (HEXplorer score of −10.19) are detained in several tissues, although only intron 4 contains a previously identified CRS. The intron-detaining transcript is retained in the nucleus until exposed to TG003 or stress, when it is then spliced into fully mature mRNA and exported to the cytoplasm (97).

The previously discussed AUUUA and UUUUU motifs found in HPV-1 were also identified in the 3′ UTR instability element of the protooncogene c-*fos*. Three poly(U)-binding proteins bind to these elements, resulting in the inhibition of c-*fos* expression (46). Additionally, with respect to inhibitory effects of splice sites on nuclear export, several studies found that the consensus sequence for the 5′ splice site inhibits export and localizes the transcript to nuclear speckles when found in the 3′ terminal exon or in unspliced mRNA (31–34, 98–101). This motif was found primarily in long noncoding RNAs (lncRNAs), which are retained in the nucleus by specific motifs (98). In fact, a mutation that creates a 5′ splice site in the 3′ UTR of the *p14/robld3* transcript inhibits poly(A) tail formation through splicing-inactive U1 snRNP binding (93). Additionally, the 5′ splice site consensus sequence was shown to have an inhibitory effect on nuclear export (98). Due to the absence of a poly(A) tail and the nuclear export factors it binds, the transcript is retained in the nucleus and detained, causing a complex immunodeficiency syndrome (102).

While most mRNAs are predominately cytoplasmic, lncRNA are typically located in the nucleus. Short sequences derived from Alu sequences and C-rich motifs were found to bind hnRNPK and to promote nuclear accumulation of common lncRNAs (103). One particular lncRNA, MALAT-1, has two separate regions that are vital for nuclear localization, E and M. Region M was found to bind to RNPS1, a protein that localizes to nuclear speckles, and deletion of either of these regions leads to nuclear export of MALAT-1 transcripts (104, 105). Recently reviewed with respect to pre-mRNA splicing, nuclear speckles are nuclear compartments formed by liquid-liquid separation and enriched in RNA-binding proteins (99, 106–108). Although a role for nuclear speckles in inhibition of RNA export by CRS was not specifically investigated, nuclear compartmentalization might contribute to decreased RNA export, as previously indicated by HIV-1 CRS-containing transcripts (38, 109, 110).

BORG, another strictly nuclear lncRNA, was shown to contain at least three independent NRSs, and it was suggested that they contain a short RNA motif (AGCCC) that might contribute to nuclear retention (111). Compared to the three independent BORG retention signals (HEXplorer scores per nucleotide of 0.7, −0.6, and −2.6), the AGCCC motif, if calculated (as described above) with a sixth nucleotide extended, exhibits an average positive HEXplorer score per nucleotide of +1.8, which suggests that other RNA-binding proteins may also play a role here. Apart from different CRS found within cellular transcripts described here, many other sequences and specific binding proteins that contribute to nuclear retention of cellular transcripts were

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

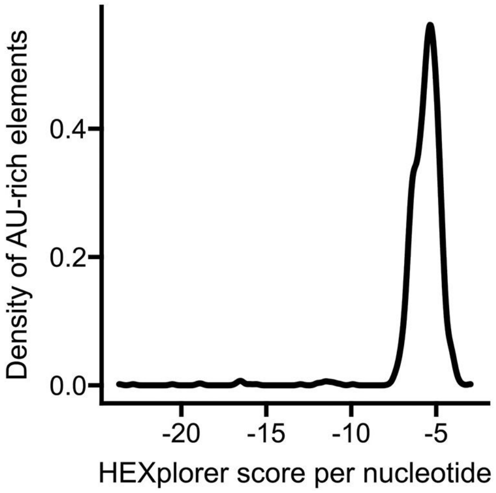


FIG 5 Density function describing the distribution of HEXplorer score per nucleotide values of 1,405 unique human AREs found within 3′ UTRs of cellular transcripts (114). The *x* axis shows the HEXplorer scores per nucleotide, and the *y* axis shows the relative density of AREs per HEXplorer score per nucleotide.

identified (reviewed in references 112 and 113). For further examples of cellular CRS, please refer to the recently updated AU-Rich Element (ARE) Database (ARED-Plus; https://brp.kfshrc.edu.sa/ared), which comprises sequences involved in posttranscriptional expression regulation of 63% of all human genes and lists 1,405 unique human AREs found within the 3′ UTR of cellular transcripts. Since, e.g., hnRNPD, also known as ARE RNA-binding protein 1 (AUF1), influences RNA stability via ARE binding, we calculated the HEXplorer scores per nucleotide of ARE sequences from ARED-Plus to see whether ARE-hnRNP interactions are reflected in the HEXplorer scores of these elements (Fig. 5) (114, 115). With an average HEXplorer score per nucleotide of −5.7, AREs within the 3′ UTR in general showed negative HEXplorer scores per nucleotide, indicating that many AREs theoretically may be able to facilitate hnRNP binding. Since ARE-containing sequences are predominantly suggested to regulate RNA stability, like HIV-1 INS elements, their effect on nuclear retention often is dismissed as the absence of nuclear export-promoting sequences (22, 113, 114). However, AREs may also act via nuclear retention through recruitment of hnRNPs. Thus, bioinformatic tools predicting hnRNP binding sites might pave the way for more directed analyses.

CONCLUSION AND OUTLOOK

Nuclear export of viral transcripts is another step at which viral gene expression is regulated. HIV-1 has evolved a sophisticated system to regulate the export of its RNA. Investigating the regulation of HIV-1 gene expression led to the discovery of CRS throughout its viral RNA genome. These sequences, together with their binding factors, inhibit nuclear export of unspliced and intron-containing transcripts in the absence of HIV-1 Rev, a mechanism that foremost represses early expression of HIV-1 structural genes *gag*, *gag/pol*, and *env*.

In particular, members of the hnRNP family seem to be important binding factors of HIV-1 CRS. Different members of the hnRNP family can lead to distinct effects on nuclear retention and stability upon binding, perhaps leading to the original classification as CRS or INS elements.

The high AT content of HIV-1 CRS seems to indicate that the host restriction factor APOBEC3G might have played a role in retroviral CRS formation. APOBEC3G is a polynucleotide cytidine deaminase that restricts HIV-1 replication by dC-to-dU deamination in the minus-strand DNA, thus leading to G-to-A substitutions in the plus-strand DNA during reverse transcription (116). APOBEC3G was previously proposed to enhance

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

HIV-1 evolution. However, evidence for this came mainly from *in vitro* experiments, whereas computational analysis based on *in vivo* data could not confirm a role of APOBEC3G-induced G-to-A substitution as a trigger for HIV-1 evolution (117–119). Because other viruses that are not restricted by APOBEC3G also harbor CRS, it is questionable whether this restriction factor contributed to the HIV-1 CRS landscape (see above). Although there is currently no identified factor contributing to CRS formation, nuclear RNA retention is essential for regulating viral gene expression.

Although the growing body of CRS found within viral and cellular transcripts consists of rather undefined regions of several hundred nucleotides or specific motifs of only a few nucleotides bound by distinct cellular proteins, there might be a common mechanism based on RNA occupation by certain retention proteins, probably concomitant with intranuclear allocation (32, 49, 79–81, 103, 109, 110, 112, 113, 120, 121). Nevertheless, all CRS found in both viral and cellular transcripts are key players in regulating subcellular localization, and the disruption of CRS can lead to illness or, in the case of viruses, altered replication (93, 101, 104).

Identification of sequences with potential effects on nuclear retention often proves experimentally difficult because of their diverse sequence motifs. Thus, bioinformatic tools, such as the HEXplorer algorithm (30), RBPmap (122), or ESRseq (123), may be promising tools to uncover and to analyze CRS in a targeted manner.

Further investigation into sequence-mediated nuclear retention may both offer insights into the mechanisms of nuclear RNA retention to elucidate cellular pathomechanisms and identify new targets for antiviral therapy.

ACKNOWLEDGMENTS

This work was funded by the Jürgen Manchot Stiftung (P.N.O., A.R., and H.S.), by the Stiftung für AIDS-Forschung, Düsseldorf (H.S.), and by the Forschungskommission (FoKo) of the Heinrich Heine University (J.P. and H.S.).

P.N.O., A.R., and H.S. discussed the structure and scope of the review. P.N.O. designed the first draft. P.N.O. and J.P. performed the bioinformatic analyses. All authors edited the manuscript and approved the final version.

We declare no conflicts of interest.

REFERENCES

1. Ramdas P, Sahu AK, Mishra T, Bhardwaj V, Chande A. 2020. From entry to egress: strategic exploitation of the cellular processes by HIV-1. Front Microbiol 11:559792. https://doi.org/10.3389/fmicb.2020.559792.
2. Barre-Sinoussi F, Ross AL, Delfraissy JF. 2013. Past, present and future: 30 years of HIV research. Nat Rev Microbiol 11:877–883. https://doi.org/10.1038/nrmicro3132.
3. Joshi S, Joshi RL. 1996. Molecular biology of human immunodeficiency virus type-1. Transfus Sci 17:351–378. https://doi.org/10.1016/0955-3886(96)00004-5.
4. Frankel AD, Young JA. 1998. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem 67:1–25. https://doi.org/10.1146/annurev.biochem.67.1.1.
5. Yoon JK, Holloway JR, Wells DW, Kaku M, Jetton D, Brown R, Coffin JM. 2020. HIV proviral DNA integration can drive T cell growth ex vivo. Proc Natl Acad Sci U S A 117:32880–32882. https://doi.org/10.1073/pnas.2013194117.
6. Krupkin M, Jackson LN, Ha B, Puglisi EV. 2020. Advances in understanding the initiation of HIV-1 reverse transcription. Curr Opin Struct Biol 65:175–183. https://doi.org/10.1016/j.sbi.2020.07.005.
7. Bedwell GJ, Engelman AN. 2021. Factors that mold the nuclear landscape of HIV-1 integration. Nucleic Acids Res 49:621–635. https://doi.org/10.1093/nar/gkaa1207.
8. Nguyen Quang N, Goudey S, Segeral E, Mohammad A, Lemoine S, Blugeon C, Versapuech M, Paillart JC, Berlioz-Torrent C, Emiliani S, Gallois-Montbrun S. 2020. Dynamic Nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. Retrovirology 17:25. https://doi.org/10.1186/s12977-020-00533-1.
9. Emery A, Swanstrom R. 2021. HIV-1: to splice or not to splice, that is the question. Viruses 13:181. https://doi.org/10.3390/v13020181.
10. Sertznig H, Hillebrand F, Erkelenz S, Schaal H, Widera M. 2018. Behind the scenes of HIV-1 replication: alternative splicing as the dependency factor on the quiet. Virology 516:176–188. https://doi.org/10.1016/j.virol.2018.01.011.
11. Stewart M. 2019. Polyadenylation and nuclear export of mRNAs. J Biol Chem 294:2977–2987. https://doi.org/10.1074/jbc.REV118.005594.
12. Cheng H, Dufu K, Lee CS, Hsu JL, Dias A, Reed R. 2006. Human mRNA export machinery recruited to the 5′ end of mRNA. Cell 127:1389–1400. https://doi.org/10.1016/j.cell.2006.10.044.
13. Taniguchi I, Mabuchi N, Ohno M. 2014. HIV-1 Rev protein specifies the viral RNA export pathway by suppressing TAP/NXF1 recruitment. Nucleic Acids Res 42:6645–6658. https://doi.org/10.1093/nar/gku304.
14. Truman CT, Järvelin A, Davis I, Castello A. 2020. HIV Rev-isited. Open Biol 10:200320. https://doi.org/10.1098/rsob.200320.
15. Hoffmann D, Schwarck D, Banning C, Brenner M, Mariyanna L, Krepstakies M, Schindler M, Millar DP, Hauber J. 2012. Formation of *trans*-activation competent HIV-1 Rev:RRE complexes requires the recruitment of multiple protein activation domains. PLoS One 7:e38305. https://doi.org/10.1371/journal.pone.0038305.
16. Rosen CA, Terwilliger E, Dayton A, Sodroski JG, Haseltine WA. 1988. Intragenic *cis*-acting *art* gene-responsive sequences of the human immunodeficiency virus. Proc Natl Acad Sci U S A 85:2071–2075. https://doi.org/10.1073/pnas.85.7.2071.
17. Knight DM, Flomerfelt FA, Ghrayeb J. 1987. Expression of the art/trs protein of HIV and study of its role in viral envelope synthesis. Science 236:837–840. https://doi.org/10.1126/science.3033827.
18. Sodroski J, Goh WC, Rosen C, Dayton A, Terwilliger E, Haseltine W. 1986. A second post-transcriptional *trans*-activator gene required for HTLV-III replication. Nature 321:412–417. https://doi.org/10.1038/321412a0.
19. Feinberg MB, Jarrett RF, Aldovini A, Gallo RC, Wong-Staal F. 1986. HTLV-III expression and production involve complex regulation at the levels of

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

splicing and translation of viral RNA. Cell 46:807–817. https://doi.org/10.1016/0092-8674(86)90062-0.

20. Cochrane AW, Jones KS, Beidas S, Dillon PJ, Skalka AM, Rosen CA. 1991. Identification and characterization of intragenic sequences which repress human immunodeficiency virus structural gene expression. J Virol 65:5305–5313. https://doi.org/10.1128/JVI.65.10.5305-5313.1991.
21. Maldarelli F, Martin MA, Strebel K. 1991. Identification of posttranscriptionally active inhibitory sequences in human immunodeficiency virus type 1 RNA: novel level of gene regulation. J Virol 65:5732–5743. https://doi.org/10.1128/JVI.65.11.5732-5743.1991.
22. Schwartz S, Felber BK, Pavlakis GN. 1992. Distinct RNA sequences in the *gag* region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of Rev protein. J Virol 66:150–159. https://doi.org/10.1128/JVI.66.1.150-159.1992.
23. Nasioulas G, Zolotukhin AS, Tabernero C, Solomin L, Cunningham CP, Pavlakis GN, Felber BK. 1994. Elements distinct from human immunodeficiency virus type 1 splice sites are responsible for the Rev dependence of *env* mRNA. J Virol 68:2986–2993. https://doi.org/10.1128/JVI.68.5.2986-2993.1994.
24. Brighty DW, Rosenberg M. 1994. A *cis*-acting repressive sequence that overlaps the Rev-responsive element of human immunodeficiency virus type 1 regulates nuclear retention of *env* mRNAs independently of known splice signals. Proc Natl Acad Sci U S A 91:8314–8318. https://doi.org/10.1073/pnas.91.18.8314.
25. Churchill MJ, Moore JL, Rosenberg M, Brighty DW. 1996. The Rev-responsive element negatively regulates human immunodeficiency virus type 1 *env* mRNA expression in primate cells. J Virol 70:5786–5790. https://doi.org/10.1128/JVI.70.9.5786-5790.1996.
26. Schneider R, Campbell M, Nasioulas G, Felber BK, Pavlakis GN. 1997. Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation. J Virol 71:4892–4903. https://doi.org/10.1128/JVI.71.7.4892-4903.1997.
27. Gales JP, Kubina J, Geldreich A, Dimitrova M. 2020. Strength in diversity: nuclear export of viral RNAs. Viruses 12:1014. https://doi.org/10.3390/v12091014.
28. Sherpa C, Le Grice SFJ. 2020. Structural fluidity of the human immunodeficiency virus Rev response element. Viruses 12:86. https://doi.org/10.3390/v12010086.
29. Rausch JW, Le Grice SF. 2015. HIV Rev assembly on the Rev response element (RRE): a structural perspective. Viruses 7:3053–3075. https://doi.org/10.3390/v7062760.
30. Erkelenz S, Theiss S, Otte M, Widera M, Peter JO, Schaal H. 2014. Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. Nucleic Acids Res 42:10681–10697. https://doi.org/10.1093/nar/gku736.
31. Takemura R, Takeiwa T, Taniguchi I, McCloskey A, Ohno M. 2011. Multiple factors in the early splicing complex are involved in the nuclear retention of pre-mRNAs in mammalian cells. Genes Cells 16:1035–1049. https://doi.org/10.1111/j.1365-2443.2011.01548.x.
32. Rosin-Arbesfeld R, Yaniv A, Gazit A. 2000. Suboptimal splice sites of equine infectious anaemia virus control Rev responsiveness. J Gen Virol 81:1265–1272. https://doi.org/10.1099/0022-1317-81-5-1265.
33. Huang Y, Carmichael GG. 1996. A suboptimal 5′ splice site is a *cis*-acting determinant of nuclear export of polyomavirus late mRNAs. Mol Cell Biol 16:6046–6054. https://doi.org/10.1128/mcb.16.11.6046.
34. Chang DD, Sharp PA. 1989. Regulation by HIV Rev depends upon recognition of splice sites. Cell 59:789–795. https://doi.org/10.1016/0092-8674(89)90602-8.
35. Hadzopoulou-Cladaras M, Felber BK, Cladaras C, Athanassopoulos A, Tse A, Pavlakis GN. 1989. The *rev* (*trs*/*art*) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a *cis*-acting sequence in the *env* region. J Virol 63:1265–1274. https://doi.org/10.1128/JVI.63.3.1265-1274.1989.
36. Schwartz S, Campbell M, Nasioulas G, Harrison J, Felber BK, Pavlakis GN. 1992. Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent *gag* expression. J Virol 66:7176–7182. https://doi.org/10.1128/JVI.66.12.7176-7182.1992.
37. Afonina E, Neumann M, Pavlakis GN. 1997. Preferential binding of poly(A)-binding protein 1 to an inhibitory RNA element in the human immunodeficiency virus type 1 *gag* mRNA. J Biol Chem 272:2307–2311. https://doi.org/10.1074/jbc.272.4.2307.
38. Mikaelian I, Krieg M, Gait MJ, Karn J. 1996. Interactions of INS (CRS) elements and the splicing machinery regulate the production of Rev-responsive mRNAs. J Mol Biol 257:246–264. https://doi.org/10.1006/jmbi.1996.0160.
39. Suh D, Seguin B, Atkinson S, Ozdamar B, Staffa A, Emili A, Mouland A, Cochrane A. 2003. Mapping of determinants required for the function of the HIV-1 *env* nuclear retention sequence. Virology 310:85–99. https://doi.org/10.1016/s0042-6822(03)00073-4.
40. Pfeiffer T, Erkelenz S, Widera M, Schaal H, Bosch V. 2013. Mutational analysis of the internal membrane proximal domain of the HIV glycoprotein C-terminus. Virology 440:31–40. https://doi.org/10.1016/j.virol.2013.01.025.
41. Gordon H, Ajamian L, Valiente-Echeverria F, Levesque K, Rigby WF, Mouland AJ. 2013. Depletion of hnRNP A2/B1 overrides the nuclear retention of the HIV-1 genomic RNA. RNA Biol 10:1714–1725. https://doi.org/10.4161/rna.26542.
42. Zolotukhin AS, Michalowski D, Bear J, Smulevitch SV, Traish AM, Peng R, Patton J, Shatsky IN, Felber BK. 2003. PSF acts through the human immunodeficiency virus type 1 mRNA instability elements to regulate virus expression. Mol Cell Biol 23:6618–6630. https://doi.org/10.1128/mcb.23.18.6618-6630.2003.
43. Najera I, Krieg M, Karn J. 1999. Synergistic stimulation of HIV-1 rev-dependent export of unspliced mRNA to the cytoplasm by hnRNP A1. J Mol Biol 285:1951–1964. https://doi.org/10.1006/jmbi.1998.2473.
44. Hamilton BJ, Nagy E, Malter JS, Arrick BA, Rigby WF. 1993. Association of heterogeneous nuclear ribonucleoprotein A1 and C proteins with reiterated AUUUA sequences. J Biol Chem 268:8881–8887. https://doi.org/10.1016/S0021-9258(18)52955-0.
45. Nakielny S, Dreyfuss G. 1996. The hnRNP C proteins contain a nuclear retention sequence that can override nuclear export signals. J Cell Biol 134:1365–1373. https://doi.org/10.1083/jcb.134.6.1365.
46. Pinol-Roma S. 1997. HnRNP proteins and the nuclear export of mRNA. Semin Cell Dev Biol 8:57–63. https://doi.org/10.1006/scdb.1996.0122.
47. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell 152:453–466. https://doi.org/10.1016/j.cell.2012.12.023.
48. Sokolowski M, Schwartz S. 2001. Heterogeneous nuclear ribonucleoprotein C binds exclusively to the functionally important UUUUU-motifs in the human papillomavirus type-1 AU-rich inhibitory element. Virus Res 73:163–175. https://doi.org/10.1016/s0168-1702(00)00238-0.
49. Sokolowski M, Zhao C, Tan W, Schwartz S. 1997. AU-rich mRNA instability elements on human papillomavirus type 1 late mRNAs and c-*fos* mRNAs interact with the same cellular factors. Oncogene 15:2303–2319. https://doi.org/10.1038/sj.onc.1201415.
50. Lund N, Milev MP, Wong R, Sanmuganantham T, Woolaway K, Chabot B, Abou Elela S, Mouland AJ, Cochrane A. 2012. Differential effects of hnRNP D/AUF1 isoforms on HIV-1 gene expression. Nucleic Acids Res 40:3663–3675. https://doi.org/10.1093/nar/gkr1238.
51. Black AC, Luo J, Watanabe C, Chun S, Bakker A, Fraser JK, Morgan JP, Rosenblatt JD. 1995. Polypyrimidine tract-binding protein and heterogeneous nuclear ribonucleoprotein A1 bind to human T-cell leukemia virus type 2 RNA regulatory elements. J Virol 69:6852–6858. https://doi.org/10.1128/JVI.69.11.6852-6858.1995.
52. Izaurralde E, Jarmolowski A, Beisel C, Mattaj IW, Dreyfuss G, Fischer U. 1997. A role for the M9 transport signal of hnRNP A1 in mRNA nuclear export. J Cell Biol 137:27–35. https://doi.org/10.1083/jcb.137.1.27.
53. Monette A, Ajamian L, López-Lastra M, Mouland AJ. 2009. Human immunodeficiency virus type 1 (HIV-1) induces the cytoplasmic retention of heterogeneous nuclear ribonucleoprotein A1 by disrupting nuclear import: implications for HIV-1 gene expression. J Biol Chem 284:31350–31362. https://doi.org/10.1074/jbc.M109.048736.
54. Roy R, Durie D, Li H, Liu BQ, Skehel JM, Mauri F, Cuorvo LV, Barbareschi M, Guo L, Holcik M, Seckl MJ, Pardo OE. 2014. hnRNPA1 couples nuclear export and translation of specific mRNAs downstream of FGF-2/S6K2 signalling. Nucleic Acids Res 42:12483–12497. https://doi.org/10.1093/nar/gku953.
55. Caputi M, Mayeda A, Krainer AR, Zahler AM. 1999. hnRNP A/B proteins are required for inhibition of HIV-1 pre-mRNA splicing. EMBO J 18:4060–4067. https://doi.org/10.1093/emboj/18.14.4060.
56. Mouland AJ, Xu H, Cui H, Krueger W, Munro TP, Prasol M, Mercier J, Rekosh D, Smith R, Barbarese E, Cohen EA, Carson JH. 2001. RNA trafficking signals in human immunodeficiency virus type 1. Mol Cell Biol 21:2133–2143. https://doi.org/10.1128/MCB.21.6.2133-2143.2001.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

57. Beriault V, Clement JF, Levesque K, Lebel C, Yong X, Chabot B, Cohen EA, Cochrane AW, Rigby WF, Mouland AJ. 2004. A late role for the association of hnRNP A2 with the HIV-1 hnRNP A2 response elements in genomic RNA, Gag, and Vpr localization. J Biol Chem 279:44141–44153. https://doi.org/10.1074/jbc.M404691200.
58. Jain N, Morgan CE, Rife BD, Salemi M, Tolbert BS. 2016. Solution structure of the HIV-1 intron splicing silencer and its interactions with the UP1 domain of heterogeneous nuclear ribonucleoprotein (hnRNP) A1. J Biol Chem 291:2331–2344. https://doi.org/10.1074/jbc.M115.674564.
59. Jean-Philippe J, Paz S, Lu ML, Caputi M. 2014. A truncated hnRNP A1 isoform, lacking the RGG-box RNA binding domain, can efficiently regulate HIV-1 splicing and replication. Biochim Biophys Acta 1839:251–258. https://doi.org/10.1016/j.bbagrm.2014.02.002.
60. Hillebrand F, Peter JO, Brillen AL, Otte M, Schaal H, Erkelenz S. 2017. Differential hnRNP D isoform incorporation may confer plasticity to the ESSV-mediated repressive state across HIV-1 exon 3. Biochim Biophys Acta Gene Regul Mech 1860:205–217. https://doi.org/10.1016/j.bbagrm.2016.12.001.
61. Bilodeau PS, Domsic JK, Mayeda A, Krainer AR, Stoltzfus CM. 2001. RNA splicing at human immunodeficiency virus type 1 3′ splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element. J Virol 75:8487–8497. https://doi.org/10.1128/jvi.75.18.8487-8497.2001.
62. Domsic JK, Wang Y, Mayeda A, Krainer AR, Stoltzfus CM. 2003. Human immunodeficiency virus type 1 hnRNP A/B-dependent exonic splicing silencer ESSV antagonizes binding of U2AF65 to viral polypyrimidine tracts. Mol Cell Biol 23:8762–8772. https://doi.org/10.1128/mcb.23.23.8762-8772.2003.
63. Madsen JM, Stoltzfus CM. 2005. An exonic splicing silencer downstream of the 3′ splice site A2 is required for efficient human immunodeficiency virus type 1 replication. J Virol 79:10478–10486. https://doi.org/10.1128/JVI.79.16.10478-10486.2005.
64. Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, Ares M, Yeo GW. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep 1:167–178. https://doi.org/10.1016/j.celrep.2012.02.001.
65. Bruun GH, Doktor TK, Borch-Jensen J, Masuda A, Krainer AR, Ohno K, Andresen BS. 2016. Global identification of hnRNP A1 binding sites for SSO-based splicing modulation. BMC Biol 14:54. https://doi.org/10.1186/s12915-016-0279-9.
66. Escudero-Paunetto L, Li L, Hernandez FP, Sandri-Goldin RM. 2010. SR proteins SRp20 and 9G8 contribute to efficient export of herpes simplex virus 1 mRNAs. Virology 401:155–164. https://doi.org/10.1016/j.virol.2010.02.023.
67. Hargous Y, Hautbergue GM, Tintaru AM, Skrisovska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH. 2006. Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. EMBO J 25:5126–5137. https://doi.org/10.1038/sj.emboj.7601385.
68. Huang Y, Gattoni R, Stévenin J, Steitz JA. 2003. SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. Mol Cell 11:837–843. https://doi.org/10.1016/s1097-2765(03)00089-3.
69. Lai MC, Tarn WY. 2004. Hypophosphorylated ASF/SF2 binds TAP and is present in messenger ribonucleoproteins. J Biol Chem 279:31745–31749. https://doi.org/10.1074/jbc.C400173200.
70. Muller-McNicoll M, Botti V, de Jesus Domingues AM, Brandl H, Schwich OD, Steiner MC, Curk T, Poser I, Zarnack K, Neugebauer KM. 2016. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. Genes Dev 30:553–566. https://doi.org/10.1101/gad.276477.115.
71. Tintaru AM, Hautbergue GM, Hounslow AM, Hung ML, Lian LY, Craven CJ, Wilson SA. 2007. Structural and functional analysis of RNA and TAP binding to SF2/ASF. EMBO Rep 8:756–762. https://doi.org/10.1038/sj.embor.7401031.
72. Huang Y, Steitz JA. 2005. SRprises along a messenger's journey. Mol Cell 17:613–615. https://doi.org/10.1016/j.molcel.2005.02.020.
73. Lee ES, Wolf EJ, Ihn SSJ, Smith HW, Emili A, Palazzo AF. 2020. TPR is required for the efficient nuclear export of mRNAs and lncRNAs from short and intron-poor genes. Nucleic Acids Res 48:11645–11663. https://doi.org/10.1093/nar/gkaa919.
74. Michael WM, Choi M, Dreyfuss G. 1995. A nuclear export signal in hnRNP A1: a signal-mediated, temperature-dependent nuclear protein export pathway. Cell 83:415–422. https://doi.org/10.1016/0092-8674(95)90119-1.
75. Sanjana NE, Shalem O, Zhang F. 2014. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods 11:783–784. https://doi.org/10.1038/nmeth.3047.
76. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343:84–87. https://doi.org/10.1126/science.1247005.
77. Xiao H, Wyler E, Milek M, Grewe B, Kirchner P, Ekici A, Silva ABOV, Jungnickl D, Full F, Thomas M, Landthaler M, Ensser A, Überla K. 2021. CRNKL1 is a highly selective regulator of intron-retaining HIV-1 and cellular mRNAs. mBio 12:e02525-20. https://doi.org/10.1128/mBio.02525-20.
78. Furth PA, Baker CC. 1991. An element in the bovine papillomavirus late 3′ untranslated region reduces polyadenylated cytoplasmic RNA levels. J Virol 65:5806–5812. https://doi.org/10.1128/JVI.65.11.5806-5812.1991.
79. Furth PA, Choe WT, Rex JH, Byrne JC, Baker CC. 1994. Sequences homologous to 5′ splice sites are required for the inhibitory activity of papillomavirus late 3′ untranslated regions. Mol Cell Biol 14:5278–5289. https://doi.org/10.1128/mcb.14.8.5278.
80. Kennedy IM, Haddow JK, Clements JB. 1990. Analysis of human papillomavirus type 16 late mRNA 3′ processing signals in vitro and in vivo. J Virol 64:1825–1829. https://doi.org/10.1128/JVI.64.4.1825-1829.1990.
81. Conrad NK, Steitz JA. 2005. A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts. EMBO J 24:1831–1841. https://doi.org/10.1038/sj.emboj.7600662.
82. Mitton-Fry RM, DeGregorio SJ, Wang J, Steitz TA, Steitz JA. 2010. Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. Science 330:1244–1247. https://doi.org/10.1126/science.1195858.
83. Torabi SF, Vaidya AT, Tycowski KT, DeGregorio SJ, Wang J, Shu MD, Steitz TA, Steitz JA. 2021. RNA stabilization by a poly(A) tail 3′-end binding pocket and other modes of poly(A)-RNA interaction. Science 371: eabe6523. https://doi.org/10.1126/science.abe6523.
84. Vogt C, Hackmann C, Rabner A, Koste L, Santag S, Kati S, Mandel-Gutfreund Y, Schulz TF, Bohne J. 2015. ORF57 overcomes the detrimental sequence bias of Kaposi's sarcoma-associated herpesvirus lytic genes. J Virol 89:5097–5109. https://doi.org/10.1128/JVI.03264-14.
85. Verma D, Li DJ, Krueger B, Renne R, Swaminathan S. 2015. Identification of the physiological gene targets of the essential lytic replicative Kaposi's sarcoma-associated herpesvirus ORF57 protein. J Virol 89:1688–1702. https://doi.org/10.1128/JVI.02663-14.
86. Roy D, Bhanja Chowdhury J, Ghosh S. 2013. Polypyrimidine tract binding protein (PTB) associates with intronic and exonic domains to squelch nuclear export of unspliced RNA. FEBS Lett 587:3802–3807. https://doi.org/10.1016/j.febslet.2013.10.005.
87. Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468:664–668. https://doi.org/10.1038/nature09479.
88. Tan W, Felber BK, Zolotukhin AS, Pavlakis GN, Schwartz S. 1995. Efficient expression of the human papillomavirus type 16 L1 protein in epithelial cells by using Rev and the Rev-responsive element of human immunodeficiency virus or the *cis*-acting transactivation element of simian retrovirus type 1. J Virol 69:5607–5620. https://doi.org/10.1128/JVI.69.9.5607-5620.1995.
89. Tan W, Schwartz S. 1995. The Rev protein of human immunodeficiency virus type 1 counteracts the effect of an AU-rich negative element in the human papillomavirus type 1 late 3′ untranslated region. J Virol 69:2932–2945. https://doi.org/10.1128/JVI.69.5.2932-2945.1995.
90. Brillen AL, Schöneweis K, Walotka L, Hartmann L, Müller L, Ptok J, Kaisers W, Poschmann G, Stühler K, Buratti E, Theiss S, Schaal H. 2017. Succession of splicing regulatory elements determines cryptic 5′ss functionality. Nucleic Acids Res 45:4202–4216. https://doi.org/10.1093/nar/gkw1317.
91. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297:1007–1013. https://doi.org/10.1126/science.1073774.
92. Schneider PA, Schwemmle M, Lipkin WI. 1997. Implication of a *cis*-acting element in the cytoplasmic accumulation of unspliced Borna disease virus RNAs. J Virol 71:8940–8945. https://doi.org/10.1128/JVI.71.11.8940-8945.1997.
93. Langemeier J, Schrom EM, Rabner A, Radtke M, Zychlinski D, Saborowski A, Bohn G, Mandel-Gutfreund Y, Bodem J, Klein C, Bohne J. 2012. A complex immunodeficiency is based on U1 snRNP-mediated poly(A) site suppression. EMBO J 31:4035–4044. https://doi.org/10.1038/emboj.2012.252.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

94. Akef A, Lee ES, Palazzo AF. 2015. Splicing promotes the nuclear export of β-globin mRNA by overcoming nuclear retention elements. RNA 21:1908–1920. https://doi.org/10.1261/rna.051987.115.

95. Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev 29:63–80. https://doi.org/10.1101/gad.247361.114.

96. Pendleton KE, Park SK, Hunter OV, Bresson SM, Conrad NK. 2018. Balance between MAT2A intron detention and splicing is determined cotranscriptionally. RNA 24:778–786. https://doi.org/10.1261/rna.064899.117.

97. Ninomiya K, Kataoka N, Hagiwara M. 2011. Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor. J Cell Biol 195:27–40. https://doi.org/10.1083/jcb.201107093.

98. Lee ES, Akef A, Mahadevan K, Palazzo AF. 2015. The consensus 5′ splice site motif inhibits mRNA nuclear export. PLoS One 10:e0122743. https://doi.org/10.1371/journal.pone.0122743.

99. Gordon JM, Phizicky DV, Neugebauer KM. 2021. Nuclear mechanisms of gene expression control: pre-mRNA splicing as a life or death decision. Curr Opin Genet Dev 67:67–76. https://doi.org/10.1016/j.gde.2020.11.002.

100. Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev 26:1209–1223. https://doi.org/10.1101/gad.188037.112.

101. Palazzo AF, Lee ES. 2018. Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. Front Genet 9:440. https://doi.org/10.3389/fgene.2018.00440.

102. Fuke H, Ohno M. 2008. Role of poly(A) tail as an identity element for mRNA nuclear export. Nucleic Acids Res 36:1037–1049. https://doi.org/10.1093/nar/gkm1120.

103. Lubelsky Y, Ulitsky I. 2018. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature 555:107–111. https://doi.org/10.1038/nature25757.

104. Guo CJ, Xu G, Chen LL. 2020. Mechanisms of long noncoding RNA nuclear retention. Trends Biochem Sci 45:947–960. https://doi.org/10.1016/j.tibs.2020.07.001.

105. Miyagawa R, Tano K, Mizuno R, Nakamura Y, Ijiri K, Rakwal R, Shibato J, Masuo Y, Mayeda A, Hirose T, Akimitsu N. 2012. Identification of *cis*- and *trans*-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. RNA 18:738–751. https://doi.org/10.1261/rna.028639.111.

106. Guo YE, Manteiga JC, Henninger JE, Sabari BR, Dall'Agnese A, Hannett NM, Spille JH, Afeyan LK, Zamudio AV, Shrinivas K, Abraham BJ, Boija A, Decker TM, Rimel JK, Fant CB, Lee TI, Cisse II, Sharp PA, Taatjes DJ, Young RA. 2019. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature 572:543–548. https://doi.org/10.1038/s41586-019-1464-0.

107. Kim J, Han KY, Khanna N, Ha T, Belmont AS. 2019. Nuclear speckle fusion via long-range directional motion regulates speckle morphology after transcriptional inhibition. J Cell Sci 132:jcs226563. https://doi.org/10.1242/jcs.226563.

108. Saitoh N, Spahr CS, Patterson SD, Bubulya P, Neuwald AF, Spector DL. 2004. Proteomic analysis of interchromatin granule clusters. Mol Biol Cell 15:3876–3890. https://doi.org/10.1091/mbc.e04-03-0253.

109. Berthold E, Maldarelli F. 1996. *Cis*-acting elements in human immunodeficiency virus type 1 RNAs direct viral transcripts to distinct intranuclear locations. J Virol 70:4667–4682. https://doi.org/10.1128/JVI.70.7.4667-4682.1996.

110. Séguin B, Staffa A, Cochrane A. 1998. Control of human immunodeficiency virus type 1 RNA metabolism: role of splice sites and intron sequences in unspliced viral RNA subcellular distribution. J Virol 72:9503–9513. https://doi.org/10.1128/JVI.72.12.9503-9513.1998.

111. Zhang B, Gunawardane L, Niazi F, Jahanbani F, Chen X, Valadkhan S. 2014. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. Mol Cell Biol 34:2318–2329. https://doi.org/10.1128/MCB.01673-13.

112. Wegener M, Muller-McNicoll M. 2018. Nuclear retention of mRNAs: quality control, gene regulation and human disease. Semin Cell Dev Biol 79:131–142. https://doi.org/10.1016/j.semcdb.2017.11.001.

113. Barreau C, Paillard L, Osborne HB. 2005. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res 33:7138–7150. https://doi.org/10.1093/nar/gki1012.

114. Bakheet T, Hitti E, Khabar KSA. 2018. ARED-Plus: an updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. Nucleic Acids Res 46:D218–D220. https://doi.org/10.1093/nar/gkx975.

115. Loflin P, Chen CY, Shyu AB. 1999. Unraveling a cytoplasmic role for hnRNP D in the in vivo mRNA destabilization directed by the AU-rich element. Genes Dev 13:1884–1897. https://doi.org/10.1101/gad.13.14.1884.

116. Delviks-Frankenberry KA, Desimmie BA, Pathak VK. 2020. Structural Insights into APOBEC3-mediated lentiviral restriction. Viruses 12:587. https://doi.org/10.3390/v12060587.

117. Sadler HA, Stenglein MD, Harris RS, Mansky LM. 2010. APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. J Virol 84:7396–7404. https://doi.org/10.1128/JVI.00056-10.

118. Kim EY, Bhattacharya T, Kunstman K, Swantek P, Koning FA, Malim MH, Wolinsky SM. 2010. Human APOBEC3G-mediated editing can promote HIV-1 sequence diversification and accelerate adaptation to selective pressure. J Virol 84:10402–10405. https://doi.org/10.1128/JVI.01223-10.

119. Deforche K, Camacho R, Laethem KV, Shapiro B, Moreau Y, Rambaut A, Vandamme AM, Lemey P. 2007. Estimating the relative contribution of dNTP pool imbalance and APOBEC3G/3F editing to HIV evolution in vivo. J Comput Biol 14:1105–1114. https://doi.org/10.1089/cmb.2007.0073.

120. King JA, Bridger JM, Gounari F, Lichter P, Schulz TF, Schirrmacher V, Khazaie K. 1998. The extended packaging sequence of MoMLV contains a constitutive mRNA nuclear export function. FEBS Lett 434:367–371. https://doi.org/10.1016/s0014-5793(98)00948-x.

121. Reddy TR, Kraus G, Suhasini M, Leavitt MC, Wong-Staal F. 1995. Identification and mapping of inhibitory sequences in the human immunodeficiency virus type 2 *vif* gene. J Virol 69:5167–5170. https://doi.org/10.1128/JVI.69.8.5167-5170.1995.

122. Paz I, Kosti I, Ares M, Cline M, Mandel-Gutfreund Y. 2014. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 42:W361–W367. https://doi.org/10.1093/nar/gku406.

123. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res 21:1360–1374. https://doi.org/10.1101/gr.119628.110.

124. Donello JE, Beeche AA, Smith GJ, Lucero GR, Hope TJ. 1996. The hepatitis B virus posttranscriptional regulatory element is composed of two subelements. J Virol 70:4345–4351. https://doi.org/10.1128/JVI.70.7.4345-4351.1996.

125. Bartels H, Luban J. 2014. Gammaretroviral *pol* sequences act *in cis* to direct polysome loading and NXF1/NXT-dependent protein production by *gag*-encoded RNA. Retrovirology 11:73. https://doi.org/10.1186/s12977-014-0073-0.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

Philipp Niklas Ostermann did his bachelor's and master's studies in Molecular Biomedicine at the Heinrich Heine University (Germany). He then received a scholarship from the Jürgen Manchot Foundation to perform his doctoral studies on HIV-1 gene expression at the Institute of Virology (University Hospital Düsseldorf) in the laboratory of Prof. Heiner Schaal. He was awarded two independent fellowships from the German Society for Virology and a Volkswagen Foundation Scholarship to attend a workshop and conference on emerging RNA viruses in Tofo, Mozambique, as mentee of Prof. Paul Young (University of Queensland). Afterwards, he joined Prof. Ali Mirazimi's laboratory (Karolinska Institute) to work on Crimean-Congo hemorrhagic fever virus (CCHFV) reporter systems and Hazara virus replication as a visiting researcher under the supervision of Dr. Vanessa Monteil at the Public Health Agency of Sweden. He currently works on replication strategies of HIV-1 and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) back in the laboratory of Prof. Schaal to finish his Ph.D.

Anastasia Ritchie received her Bachelor of Science in Chemistry from Michigan State University (USA) as a part of the Alumni Distinguished Scholarship cohort. While completing her bachelor's degree, she worked in Prof. Robert Hausinger's laboratory, studying the kinetics and structure of ethylene-forming enzyme (EFE). She went on to complete her master's degree in Biomedical Sciences at Florida Atlantic University (USA) in the laboratory of Prof. Massimo Caputi, where she studied transcriptional regulation of the tumor necrosis factor (TNF) inflammatory pathway. She is currently doing her Ph.D. work with Prof. Heiner Schaal at Heinrich Heine University (Germany) as a part of the MOI IV Graduate School, where she is investigating the regulation of nuclear export of intron-containing RNA.

Johannes Ptok acquired his Bachelor and Master of Science degrees from the Heinrich Heine University in Germany with a major in Bioinformatics and Quantitative Biology, focusing on functional pre-mRNA splicing and the analysis of RNA sequencing data. Before starting his Ph.D. work in the work group of Prof. Heiner Schaal, he joined the laboratory of Prof. Dr. Hertel at the University of California, Irvine, expanding his bioinformatic skills. Besides development of software tools in R, he now evaluates RNA sequencing data for differential gene expression and splicing in studies to further understand cellular and viral spicing.

Heiner Schaal obtained his diploma and Ph.D. degrees from the University of Cologne (Germany), focusing on differential gene expression in the central nervous system of mice at the Institute of Genetics. With a postdoctoral fellowship from the Boehringer Ingelheim Fonds, he moved to the Neurological Clinic of Heinrich Heine University, Dusseldorf (Germany), before he joined the laboratory of Prof. Andreas Scheid at the Institute of Virology, Dusseldorf (Germany). He has pursued HIV research for more than 30 years. In his studies to elucidate the replication cycle of HIV, he has inevitably turned to general questions of pre-mRNA processing and nuclear export, particularly of intron-containing RNA.

Downloaded from https://journals.asm.org/journal/jvi on 31 October 2023 by 134.99.174.106.

3. General discussion

The major sequence elements determining splicing are splice site sequences, including the branch point sequence and the poly-pyrimidine tract at the 3’ss and splicing regulatory elements, recruiting splicing regulatory proteins (SRPs) to the RNA. In this thesis, the combinatorial influence of these factors on splice site recognition haven been studied in human cells during health and during cellular stress. Estimation of splice site strength by the MaxEntScan score and the HBond score can be complemented by prediction of splicing regulatory proteins (**Thesis 1**). In this work, SRP binding was predicted by the HEXplorer algorithm. Although the HEXplorer algorithm provides a numeric value estimating the overall splicing regulating property of a given genomic sequence, it is not suitable to predict binding of specific SRPs like alternative tools. The HEXplorer algorithm, as most alternatives, is based on sequences of the human reference genome. Alternative splicing, however, is quite diverse across cell types, depending on cell type specific expression of SRPs [117, 151] (**Publication I**). The estimated overall impact of SREs on 5’ss recognition using the Splice Site HEXplorer Weight (SSHW), is calculated, subtracting the HEXplorer score sum of the 50 nucleotides downstream of a 5’ss from the HEXplorer score sum of the 50 nucleotides upstream of a 5’ss. The window of 50 nucleotides is somewhat arbitrary, since putative exonic splicing enhancer and silencer octamer frequencies, as well as the top 400 ESEseqs and ESSseqs, showed relatively little variation across 100nt long exons [104, 152]. Since also the hexamer weights, used in the HEXplorer algorithm, highly correlated between 100 nt and 30 nt wide 5’ss sequence surroundings, the 50 nucleotides window was expected to capture relevant SREs. Multimerization, observed with some splicing regulatory proteins, would however potentially not be fully reflected by the RESCUE-type approach of the HEXplorer algorithm.

To analyze the additional benefit combining the SSHW with the intrinsic strength to model 5’ss recognition, thousands of SSHW for a 5’ss at the first exon of a two-exon minigene splicing reporter were tested (**Publication II**). The 5’ss was accompanied by a second equally strong 5’ss with an intrinsic strength of 17.5 HBS. For the massively parallel splicing assay (MPSA), a random library of 3,127 octamers was inserted in between the competing 5’ss. Resulting 5’ss usage was measured comparing the frequency of octamer sequences in the sequenced reporter plasmid with the frequency in sequenced cDNA of the processed RNA transcript.

Since octamer sequences were not associated with barcode sequences positioned to remain within the sequenced RNA after splicing, usage of the downstream 5’ss was only indirectly measured, opposite to previous comparison studies of 5’ss intrinsic strengths [153]. To find similar 5’ss competitions in the human genome, a large RNA sequencing data set of 57 human fibroblasts samples was analyzed. For every annotated 5’ss used in this data set, the next shortest respective exon was determined. Next, every position in the exonic sequence with a GT dinucleotide was determined, as the minimal requirement for a potential canonical 5’ss. Forming pairs of every GT site with its respective annotated 5’ss, the differences in intrinsic strength, SSHW and 5’ss usage were calculated. With decreasing differences in SSHW and intrinsic strength (HBond score) an increase in relative GT usage was detected. Splitting GT sites and annotated 5’ss intro two data sets served as basis for calculation of a logistic 5’ss usage prediction model, that best discriminated between unused GT sites und annotated highly used 5’ss, when combining SSHW and HBond score in the model than using one of the factors separately. This observation might, however, be less relevant for non-coding transcripts, that also undergo splicing, since they show significantly less interaction with SR proteins [154]. Studying mRNA sequencing data instead of total RNA sequencing data, potential 5’ss competitions situations within non-coding transcripts are excluded from the analysis. Additionally, 5’ss usage inducing non-sense mediated mRNA decay (NMD) are degraded before sequencing [155], since no NMD-inhibitors, like *cycloheximide* (CHX), were applied in this study. 5’ss showing no or low usage in our mRNA sequencing data might therefore be not due to differences in SSHW or intrinsic strength, but due to initiation of NMD. Generally, the SSHW still seems to be a metric, that allows to better estimate functional splice site strength, combined with the intrinsic strength estimated by for instance the HBond score. Correctly estimating changes in splice site recognition upon mutations within the splice site sequence or its immediate surrounding, however, additionally partly relies on the presence and functional strength of alternative potentially competing splice sites in proximity and general gene context. The best 5’ss with no competing splice sites in the transcript will not be used, if there is no 3’ss. This might be one reason why tools like SpliceAI, that trained neuronal networks to predict sequence variant induced changes in splice site usage, worked better using longer sequence segments (up to 10,000 nt) than shorter (80 nt) [108]. Another reason might be the inclusion of upstream and downstream splice site strengths. 5’ss recognized by the special U12-type spliceosome were not included in the analysis of GT/5’ss pairs, since they make up only around 0.4% of human 5’ss and are recognized not by the U1 snRNP of the major spliceosome, but the U11 snRNP of the minor spliceosome [156-158]. Exonic 5’ss showing a GC dinucleotide instead of the canonical GT were also excluded in my analysis, since they only make up less than 1% of human 5’ss [159], as well as other even less frequent noncanonical 5’ss.

Studying noncanonical splicing, how intrinsic strength and SREs are affected by mutations and alternative splicing in health and disease can further deepen our knowledge about functional splicing (**Thesis 2**). Sequence variations are sometimes located within the 5’ss sequence itself. Depending on the variation, the intrinsic strength of the 5’ss might change, potentially drastically affecting 5’ss usage. In this work, homo- and heterozygous sequence variations of high-resolution genomes from the 1000 Genomes Project were analyzed, that were located directly within or in proximity to annotated 5’ss (**Unpublished manuscript I**). Sequence variations that altering the 5’ss sequence were relatively rare, affecting around 0.8% of human annotated splice sites. This was expected, since splice site sequences are described to be conserved regions of the genome [160]. Heterozygous variations were found at 5.2% of human 5’ss, indicating that changes that affect only one allele might be less strictly inhibited because in some cases functional protein expression is still sufficient with one unaffected allele. In half of all homo- and heterozygous sequence variations, the variation did not result in changes in intrinsic strength and 5’ss with intrinsic strength lower 18.8 HBond score showed less HBond score reduction, than statistically randomly expected. Additionally, we observed a slight balance between sequence variation-induced changes in 5’ss strength and corresponding predicted binding of splicing regulatory proteins in its proximity. Generally, genes analyzed during breast cancer risk assessment, only allow reductions in intrinsic 5’ss strength of up to 1.1 HBS whereas acceptable changes in predicted SRP binding were highly dependent on the respective 5’ss strength. The analysis was done on 26 genomes, since they were part of a recent attempt to sequence the whole genome with all its repetitive sequences or recently still uncovered regions [161], using more recent methods, that might increase the amount of detected sequence variants per genome, compared to before. The nomenclature of sequence variations, however, sometimes only refers to positions within specific process RNA transcripts, making projection on the genomic context complicated. For this purpose, the

VarCon tool was developed (**Publication III**), that converts the variant position referring to coding sequences to positions on the human reference genome.

Another aspect to study splicing regulation is studying how splicing is altered during for instance cellular stress. Protein kinases and phosphatases, like nuclear kinase CLK1/4, are described to change the phosphorylation state of SR or hnRNP proteins upon heat shock or osmotic stress. This is able to alter for instance cellular localization [44], their subnuclear distribution [45], their RNA-binding [46], interaction with the pre-spliceosome [40] and their mRNA export activity [47]. Splice sites normally enhance or silenced by the affected splicing regulatory proteins could show drastic changes in splice site recognition. In this work, transcriptional changes in primary cardiovascular endothelial cells were analyzed upon treatment with high-fat diet, introducing high concentrations of low-density lipoprotein (LDL) (**Publication VI**). It was previously shown that high-fat diet results in loss of functional NOS3 protein in endothelial cells, impairing their function [162]. Upon sequencing of the mRNA, a strong decrease in NOS3 splice junction usage at the N-terminus and an increase of splice junctions at the C-terminus was observed. Apparently, these changes in splice site usage could not be attributed to regulation of NOS3 splicing, but rather activation of an internal secondary promotor, resulting in expression of a truncated NOS3 protein, lacking essential domains for full functionality. Interestingly, the reduction of NOS3 protein expression was not captured by classical differential gene expression (DGE) analysis, since here all reads within the gene locus are summarized, emphasizing the benefit of analysis of splice site usage additionally to DGE analysis. Non-functional NOS3 was previously described to result in defective mitochondrial beta-oxidation, which leads to increased levels of reactive oxygen species (ROS) and oxidative stress [163]. One important player regulating the redox balance and fatty acid oxidation, the peroxisome proliferator-activated receptor *γ* (PPARG), was also described to regulate mtDNA levels [164]. Indeed, LDL treatment additionally led to increased expression of PPARG and increased levels of mtDNA, with no increase in absolute mitochondrial mass, resulting in increased expression of mitochondrial encoded genes, whereas most nuclear encoded genes that are transported to mitochondria were downregulated. This imbalance might further disrupt mitochondrial functionality resulting in endothelial senescence and apoptosis. Endothelial apoptosis results in vascular leakage and subsequent sepsis. In this work, a promising compound was analyzed, that seemed to reduced apoptosis, induced by one of the most relevant sepsis-inducing factors lipopolysaccharide (LPS) (**Publication V**). Endothelial apoptosis could be prevented using the first 20 amino acids of the Apurinic/Apyrimidinic Endodeoxyribonuclease 1 (APEX1) [148], probably via Selenoprotein T (SELENOT) upregulation, which was also shown to reduce endothelial apoptosis in vitro. SELENOT ca be found in high concentrations in embryonic structures and decreases progressively during development. It is also described to stimulate liver regeneration [165]. Studying alternative splicing upon SELENOT overexpression might therefore provide precious insights into pathways of tissue regeneration.

Expression vectors, like for SELENOT overexpression or during designing of splicing reporter, sometimes require careful manipulation of splice site usage. In some cases, usage of splice sites within a coding sequence has to be reduced to increase protein expression. Splicing reporter sometimes require careful manipulation of splice site strength or SREs to generate the desired outcome. For this purpose, the ModCon algorithm was developed, that allows to adjust the Splice Site HEXplorer weight (SSHW) of 5’ss or to drastically decrease the intrinsic strength of splice sites within a coding sequence, using synonymous substitutions (**Publication VI**). This allows to keep the encoded amino acid the same, while changing the underlying mRNP code (**Thesis 3**) [166]. The mRNP code describes the observation, that the sequence of RNA transcripts drastically influences the fate of the RNA, by recruitment of various RNA-binding proteins, depending on its nucleotide sequence. Changing the mRNP code, therefore not only potentially changes splicing, but also RNA modification, RNA export or translation. Applying ModCon on a 5’ss in the middle of the firefly luciferase gene, resulted in activation of the previously unused 5’ss of only medium-strong intrinsic strength (HBond score 14). The default window of manipulation of the algorithm is 50 nucleotides upstream and downstream, where the total HEXplorer is either up- or downregulated via synonymous substitutions to either increase or decrease the SSHW. However, the firefly reporter system only provided a limited set of restriction sites, that restricted the pool of insertion sites during cloning of the construct. Therefore, a segment of 241 nucleotides upstream of the 5’ss was altered to increase SR-protein binding and an additional segment of 652 nucleotides downstream of the 5’ss position was altered, to resemble a typical SRP binding profile of intronic sequences (relatively negative total HEXplorer score). In other systems, the similar adjustments would probably be necessary, however, this greatly influences not only SRP mediated splicing regulation but also the total mRNP code potentially changing the fate of RNA transcripts in unintended ways. The SRP binding profile of intronic sequences is enriched with hnRNP binding sites. Since some hnRNP were shown to repress RNA export from the nucleus, introducing significant amounts of hnRNP binding sites could alter nuclear export of the respective transcript [69, 70]. The reverse was previously frequently done by researchers studying *Cis*-acting Repressive Sequences (CRS) in viral transcripts, that repress nuclear RNA export. The export-inhibiting element INS-1 within the *gag* ORF (positions 867 to 1084) of human immunodeficiency virus type 1 (HIV-1), for instance, could be deactivated by synonymous mutations, that increased the total HEXplorer score from -0.05 to +3.63, indicating reduced hnRNP binding [167, 168]. Reviewing further viral export elements revealed, that those facilitating RNA retention in the nucleus were enriched of hnRNP binding sites, whereas elements supporting RNA export from the nucleus were enriched of SR protein binding sites (**Publication IX**).

Future endeavors to improve models predicting splice site usage would probably greatly benefit from systematically including predicted SRP binding, taking cell-type specific SRP expression into account. Studying cell-type specific splice site usage in ultra deep RNA-sequencing data, potentially upon inhibition of RNA degradation like NMD, might capture the cell-type specific impact of SREs. Additionally considering splice site competition at various genetic contexts, like varying 3'ss strength or exon length, could further improve prediction of splice site usage in the future.

4. Curriculum Vitae

Education

01/2019 – 11/2023	**Doctoral researcher** Heinrich-Heine-Universität, Düsseldorf
07/2016 – 12/2018	**Master of Science in Biology** Heinrich-Heine-Universität, Düsseldorf *Major: Bioinformatics and quantitative Biology*
10/2012 – 06/2016	**Bachelor of Science in Biology** Heinrich-Heine-Universität, Düsseldorf
04/2018	**Advanced RNA-Seq analysis** EMBL-EBI Cambridge
10/2019	**Computing Skills In Python For Reproducible Research** EMBL Heidelberg

Publications

1. Magvan B, Kloeble AA, **Ptok J**, Hoffmann D, Habermann D, Gantumur A, Paluschinski M, Enebish G, Balz V, Fischer JC, Chimeddorj B, Walker A, Timm J. Sequence diversity of hepatitis D virus in Mongolia. *Front Med (Lausanne)*. 2023 Mar 13;10:1108543. doi: 10.3389/fmed.2023.1108543. PMID: 37035318; PMCID: PMC10077969.
2. Radulovic I, Schündeln MM, Müller L, **Ptok** J, Honisch E, Niederacher D, Wiek C, Scheckenbach K, Leblanc T, Larcher L, Soulier J, Reinhardt D, Schaal H, Andreassen PR, Hanenberg H. A novel cancer risk prediction score for the natural course of FA patients with biallelic BRCA2/FANCD1 mutations. *Hum Mol Genet.* 2023 Jan 31:ddad017. doi: 10.1093/hmg/ddad017.
3. Müller L, Andrée M, Moskorz W, Drexler I, Hauka S, **Ptok** J, Walotka L, Grothmann R, Hillebrandt J, Ritchie A, Peter L, Walker A, Timm J, Adams O, Schaal H. Adjusted COVID-19 booster schedules balance age-dependent differences in antibody titers benefitting risk populations. *Front Aging.* 2022 Oct 12;3:1027885. doi: 10.3389/fragi.2022.1027885.
4. Müller L*, **Ptok** J*, Nisar A, Antemann J, Grothmann R, Hillebrand F, Brillen AL, Ritchie A, Theiss S, Schaal H. Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements. *Nucleic Acids Res*. 2022 Aug 26;50(15):8834-8851. doi: 10.1093/nar/gkac663.

5. Müller L, Moskorz W, Brillen AL, Hillebrand F, Ostermann PN, Kiel N, Walotka L, **Ptok J**, Timm J, Lübke N, Schaal H. Altered HIV-1 mRNA Splicing Due to Drug-Resistance-Associated Mutations in Exon 2/2b. *Int J Mol Sci.* 2021 Dec 23;23(1):156. doi: 10.3390/ijms23010156. PMID: 35008581; PMCID: PMC8745674.
6. **Merk** D*, **Ptok J***, Jakobs P, Ameln P, Greulich J, Kluge P, Semperowitsch K, Eckermann O, Schaal H, Ale-Agha N, Altschmied J, Haendeler J. *Selenoprotein T Protects Endothelial Cells against Lipopolysaccharide-Induced Activation and Apoptosis.* Antioxidance. 2021 Sep 7; 10(9), 1427; doi:10.3390/antiox10091427
7. Goranci-Buzhala G, Mariappan A, Ricci-Vitiani L, Josipovic N, Pacioni S, Gottardo M, **Ptok J**, Schaal H, Callaini G, Rajalingam K, Dynlacht B, Hadian K, Papantonis A, Pallini R, Gopalakrishnan J. *Cilium induction triggers differentiation of glioma stem cells.* Cell Rep. 2021 Sep 7;36(10):109656. doi: 10.1016/j.celrep.2021.109656. PMID: 34496239.
8. Ostermann PN, Ritchie A, **Ptok J**, Schaal H. *Let It Go: HIV-1 cis-Acting Repressive Sequences*. J Virol. 2021 Jul 12;95(15):e0034221. doi: 10.1128/JVI.00342-21. Epub 2021 Jul 12. PMID: 33980600
9. **Ptok J**, Müller L, Ostermann PN, Ritchie A, Dilthey AT, Theiss S, Schaal H. Modifying splice site usage with ModCon: Maintaining the genetic code while changing the underlying mRNP code. Comput Struct Biotechnol J. 2021 May 21;19:3069-3076. doi: 10.1016/j.csbj.2021.05.033. PMID: 34136105; PMCID: PMC8178101.
10. Müller L, Andrée M, Moskorz W, Drexler I, Walotka L, Grothmann R, **Ptok J**, Hillebrandt J, Ritchie A, Rabl D, Ostermann PN, Robitzsch R, Hauka S, Walker A, Menne C, Grutza R, Timm J, Adams O, Schaal H. Age-dependent immune response to the Biontech/Pfizer BNT162b2 COVID-19 vaccination. Clin Infect Dis. 2021 Apr 27:ciab381. doi: 10.1093/cid/ciab381. Epub ahead of print. PMID: 33906236; PMCID: PMC8135422.
11. Erkelenz S, Poschmann G, **Ptok J**, Müller L, Schaal H. Profiling of cis- and trans-acting factors supporting noncanonical splice site activation. RNA Biol. 2021 Jan;18(1):118-130. doi: 10.1080/15476286.2020.1798111. Epub 2020 Aug 5. PMID: 32693676; PMCID: PMC7834088.
12. **Ptok J**, Theiss S, Schaal H. VarCon: An R Package for Retrieving Neighboring Nucleotides of an SNV. Cancer Inform. 2020 Nov 24;19:1176935120976399. doi: 10.1177/1176935120976399. PMID: 33281441; PMCID: PMC7691889.
13. **Ptok J***, Müller L*, Theiss S, Schaal H. Context matters: Regulation of splice donor usage. Biochim Biophys Acta Gene Regul Mech. 2019 Nov-Dec;1862(11-12):194391. doi: 10.1016/j.bbagrm.2019.06.002. Epub 2019 Jun 13. PMID: 31202784.
14. Gonnissen S*, **Ptok J***, Goy C, Jander K, Jakobs P, Eckermann O, Kaisers W, von Ameln F, Timm J, Ale-Agha N, Haendeler J, Schaal H, Altschmied J. High Concentration of Low-Density Lipoprotein Results in Disturbances in Mitochondrial Transcription and Functionality in Endothelial Cells. Oxid Med Cell Longev. 2019 Jun 10;2019:7976382. doi: 10.1155/2019/7976382. PMID: 31281593; PMCID: PMC6590621.
15. Erkelenz S, Theiss S, Kaisers W, **Ptok J**, Walotka L, Müller L, Hillebrand F, Brillen AL, Sladek M, Schaal H. Ranking noncanonical 5' splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. Genome Res. 2018 Dec;28(12):1826-1840.

doi: 10.1101/gr.235861.118. Epub 2018 Oct 24. PMID: 30355602; PMCID: PMC6280755.

16. Kaisers W, **Ptok J**, Schwender H, Schaal H. Validation of Splicing Events in Transcriptome Sequencing Data. Int J Mol Sci. 2017 May 23;18(6):1110. doi: 10.3390/ijms18061110. PMID: 28545234; PMCID: PMC5485934.
17. Brillen AL, Schöneweis K, Walotka L, Hartmann L, Müller L, **Ptok J**, Kaisers W, Poschmann G, Stühler K, Buratti E, Theiss S, Schaal H. Succession of splicing regulatory elements determines cryptic 5′ss functionality. Nucleic Acids Res. 2017 Apr 20;45(7):4202-4216. doi: 10.1093/nar/gkw1317. PMID: 28039323; PMCID: PMC5397162.

Talks

Ptok J*, Müller L*, Nisar A, Antemann J, Grothmann R, Hillebrand F, Brillen AL, Ritchie A, Theiss S, Schaal H. Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements. 5th International Caparica Conference in SPLICING 2023. 06/2023

Ptok J*, Müller L*, Nisar A, Antemann J, Grothmann R, Hillebrand F, Brillen AL, Ritchie A, Theiss S, Schaal H. Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements. BMFZ meeting Duesseldorf. 09/2022

Ptok J*, Gonnissen S*, Goy C, Jander K, Jakobs P, Eckermann O, Kaisers W, von Ameln F, Timm J, Ale-Agha N, Haendeler J, Schaal H, Altschmied J. High concentration of low-density lipoprotein results in disturbances in mitochondrial transcription and functionality in endothelial cells. Bioinformatics and Aging 2019

Poster

Ptok J*, Müller L*, Nisar A, Antemann J, Grothmann R, Hillebrand F, Brillen AL, Ritchie A, Theiss S, Schaal H. Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements. 5th International Caparica Conference in SPLICING 2023. 06/2023

Ptok J*, Theiss S, Schaal H. Reconstruction of exon composition using RNA-seq data. Advanced RNA-Seq analysis course EMBL-EBI Cambridge 04/2018

5. Acknowledgements

Danken möchte ich an dieser Stelle:

Professor Dr. Heiner Schaal für die Hilfestellungen, die Motivation und die angeregten wissenschaftlichen Diskussionen.

Prof. Dr. Michael Feldbrügge für die bereitwillige Übernahme des Co-Referats.

Dr. Stephan Theiss für konstruktive Diskussionen und Anregungen.

Des Weiteren möchte ich mich bei Prof. Dr. Jörg Timm dafür bedanken, dass ich diese Arbeit am Institut für Virologie fertigstellen konnte.

Vielen Dank auch an Alex Dilthey für die Aufnahme in die bioinformatische Arbeitsgruppe und Betreuung der Institut-übergreifenden Projekte.

Ganz besonders danken möchte ich der AG Schaal für die Zusammenarbeit und schöne Zeit am Institut.

6. Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Johannes Ptok

Düsseldorf, den

7. References

1. Geuens, T., D. Bouhy, and V. Timmerman, *The hnRNP family: insights into their role in health and disease.* Hum Genet, 2016. **135**(8): p. 851-67.
2. Erkelenz, S., et al., *Genomic HEXploring allows landscaping of novel potential splicing regulatory elements.* Nucleic Acids Res, 2014. **42**(16): p. 10681-97.
3. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
4. Twyffels, L., C. Gueydan, and V. Kruys, *Shuttling SR proteins: more than splicing factors.* FEBS J, 2011. **278**(18): p. 3246-55.
5. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proc Natl Acad Sci U S A, 1977. **74**(8): p. 3171-5.
6. Khodor, Y.L., et al., *Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila.* Genes Dev, 2011. **25**(23): p. 2502-12.
7. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
8. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
9. Aebi, M., et al., *Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA.* Cell, 1986. **47**(4): p. 555-65.
10. Moore, M.J., *Intron recognition comes of AGe.* Nat Struct Biol, 2000. **7**(1): p. 14-6.
11. Matera, A.G. and Z. Wang, *A day in the life of the spliceosome.* Nat Rev Mol Cell Biol, 2014. **15**(2): p. 108-21.
12. Michaud, S. and R. Reed, *An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway.* Genes Dev, 1991. **5**(12B): p. 2534-46.
13. Raghunathan, P.L. and C. Guthrie, *RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2.* Curr Biol, 1998. **8**(15): p. 847-55.
14. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.
15. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.
16. Erkelenz, S., et al., *Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms.* RNA, 2013. **19**(1): p. 96-102.
17. Ptok, J., et al., *Context matters: Regulation of splice donor usage.* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2019.
18. Freund, M., et al., *A novel approach to describe a U1 snRNA binding site.* Nucleic Acids Res, 2003. **31**(23): p. 6963-75.
19. Zhuang, Y. and A.M. Weiner, *A compensatory base change in U1 snRNA suppresses a 5' splice site mutation.* Cell, 1986. **46**(6): p. 827-35.
20. Yeo, G. and C.B. Burge, *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* J Comput Biol, 2004. **11**(2-3): p. 377-94.
21. Burset, M., I.A. Seledtsov, and V.V. Solovyev, *Analysis of canonical and non-canonical splice sites in mammalian genomes.* Nucleic Acids Res, 2000. **28**(21): p. 4364-75.
22. Wang, Z. and C.B. Burge, *Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.* RNA, 2008. **14**(5): p. 802-13.
23. Muller, L., et al., *Modeling splicing outcome by combining 5'ss strength and splicing regulatory elements.* Nucleic Acids Res, 2022. **50**(15): p. 8834-8851.

24. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nat Rev Mol Cell Biol, 2017. **18**(7): p. 437-451.
25. Busch, A. and K.J. Hertel, *Evolution of SR protein and hnRNP splicing regulatory factors.* Wiley Interdiscip Rev RNA, 2012. **3**(1): p. 1-12.
26. Wegener, M. and M. Muller-McNicoll, *View from an mRNP: The Roles of SR Proteins in Assembly, Maturation and Turnover.* Adv Exp Med Biol, 2019. **1203**: p. 83-112.
27. Manley, J.L. and A.R. Krainer, *A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins).* Genes Dev, 2010. **24**(11): p. 1073-4.
28. Anko, M.L., *Regulation of gene expression programmes by serine-arginine rich splicing factors.* Semin Cell Dev Biol, 2014. **32**: p. 11-21.
29. Caceres, J.F., et al., *Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity.* J Cell Biol, 1997. **138**(2): p. 225-38.
30. Shepard, P.J. and K.J. Hertel, *The SR protein family.* Genome Biol, 2009. **10**(10): p. 242.
31. Graveley, B.R., K.J. Hertel, and T. Maniatis, *The role of U2AF35 and U2AF65 in enhancer-dependent splicing.* RNA, 2001. **7**(6): p. 806-18.
32. Legartova, S., et al., *The SC-35 Splicing Factor Interacts with RNA Pol II and A-Type Lamin Depletion Weakens This Interaction.* Cells, 2021. **10**(2).
33. Lemaire, R., et al., *Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins.* Genes Dev, 2002. **16**(5): p. 594-607.
34. Muller-McNicoll, M., et al., *SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export.* Genes Dev, 2016. **30**(5): p. 553-66.
35. Sanford, J.R., et al., *A novel role for shuttling SR proteins in mRNA translation.* Genes Dev, 2004. **18**(7): p. 755-68.
36. Pelisch, F., et al., *The serine/arginine-rich protein SF2/ASF regulates protein sumoylation.* Proc Natl Acad Sci U S A, 2010. **107**(37): p. 16119-24.
37. Schwich, O.D., et al., *SRSF3 and SRSF7 modulate 3'UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels.* Genome Biol, 2021. **22**(1): p. 82.
38. Xiao, W., et al., *Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing.* Mol Cell, 2016. **61**(4): p. 507-519.
39. Zhu, Y., et al., *Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation.* Mol Cell, 2018. **69**(1): p. 62-74 e4.
40. Zhou, Z. and X.D. Fu, *Regulation of splicing by SR proteins and SR protein-specific kinases.* Chromosoma, 2013. **122**(3): p. 191-207.
41. Blackie, A.C. and D.J. Foley, *Exploring the roles of the Cdc2-like kinases in cancers.* Bioorg Med Chem, 2022. **70**: p. 116914.
42. Ninomiya, K., N. Kataoka, and M. Hagiwara, *Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor.* J Cell Biol, 2011. **195**(1): p. 27-40.
43. Shi, Y. and J.L. Manley, *A complex signaling pathway regulates SRp38 phosphorylation and pre-mRNA splicing in response to heat shock.* Mol Cell, 2007. **28**(1): p. 79-90.
44. Lai, M.C., R.I. Lin, and W.Y. Tarn, *Transportin-SR2 mediates nuclear import of phosphorylated SR proteins.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10154-9.
45. Misteli, T. and D.L. Spector, *Serine/threonine phosphatase 1 modulates the subnuclear distribution of pre-mRNA splicing factors.* Mol Biol Cell, 1996. **7**(10): p. 1559-72.
46. Tacke, R., Y. Chen, and J.L. Manley, *Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer.* Proc Natl Acad Sci U S A, 1997. **94**(4): p. 1148-53.

47. Botti, V., et al., *Cellular differentiation state modulates the mRNA export activity of SR proteins.* J Cell Biol, 2017. **216**(7): p. 1993-2009.
48. Howard, J.M. and J.R. Sanford, *The RNAissance family: SR proteins as multifaceted regulators of gene expression.* Wiley Interdiscip Rev RNA, 2015. **6**(1): p. 93-110.
49. Graveley, B.R., *Sorting out the complexity of SR protein functions.* RNA, 2000. **6**(9): p. 1197-211.
50. Bradley, T., M.E. Cook, and M. Blanchette, *SR proteins control a complex network of RNA-processing events.* RNA, 2015. **21**(1): p. 75-92.
51. Jamison, S.F., et al., *U1 snRNP-ASF/SF2 interaction and 5' splice site recognition: characterization of required elements.* Nucleic Acids Res, 1995. **23**(16): p. 3260-7.
52. Shen, H. and M.R. Green, *A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly.* Mol Cell, 2004. **16**(3): p. 363-73.
53. Shen, H. and M.R. Green, *RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans.* Genes Dev, 2006. **20**(13): p. 1755-65.
54. Zuo, P. and T. Maniatis, *The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing.* Genes Dev, 1996. **10**(11): p. 1356-68.
55. Wang, Z., H.M. Hoffmann, and P.J. Grabowski, *Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity.* RNA, 1995. **1**(1): p. 21-35.
56. Kanopka, A., O. Muhlemann, and G. Akusjarvi, *Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA.* Nature, 1996. **381**(6582): p. 535-8.
57. Shen, M. and W. Mattox, *Activation and repression functions of an SR splicing regulator depend on exonic versus intronic-binding position.* Nucleic Acids Res, 2012. **40**(1): p. 428-37.
58. Sharma, S., et al., *Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome.* Nat Struct Mol Biol, 2008. **15**(2): p. 183-91.
59. Sharma, S., et al., *U1 snRNA directly interacts with polypyrimidine tract-binding protein during splicing repression.* Mol Cell, 2011. **41**(5): p. 579-88.
60. More, D.A. and A. Kumar, *SRSF3: Newly discovered functions and roles in human health and diseases.* Eur J Cell Biol, 2020. **99**(6): p. 151099.
61. Ortiz-Sanchez, P., et al., *Loss of SRSF3 in Cardiomyocytes Leads to Decapping of Contraction-Related mRNAs and Severe Systolic Dysfunction.* Circ Res, 2019. **125**(2): p. 170-183.
62. Van Nostrand, E.L., et al., *A large-scale binding and functional map of human RNA-binding proteins.* Nature, 2020. **583**(7818): p. 711-719.
63. Tacke, R. and J.L. Manley, *Determinants of SR protein specificity.* Curr Opin Cell Biol, 1999. **11**(3): p. 358-62.
64. Liu, H.X., M. Zhang, and A.R. Krainer, *Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins.* Genes Dev, 1998. **12**(13): p. 1998-2012.
65. Sliskovic, I., H. Eich, and M. Muller-McNicoll, *Exploring the multifunctionality of SR proteins.* Biochem Soc Trans, 2022. **50**(1): p. 187-198.
66. Dreyfuss, G., et al., *hnRNP proteins and the biogenesis of mRNA.* Annu Rev Biochem, 1993. **62**: p. 289-321.

67. Hoffman, D.W., et al., *RNA-binding domain of the A protein component of the U1 small nuclear ribonucleoprotein analyzed by NMR spectroscopy is structurally similar to ribosomal proteins.* Proc Natl Acad Sci U S A, 1991. **88**(6): p. 2495-9.
68. Han, S.P., Y.H. Tang, and R. Smith, *Functional diversity of the hnRNPs: past, present and perspectives.* Biochem J, 2010. **430**(3): p. 379-92.
69. Pinol-Roma, S., *HnRNP proteins and the nuclear export of mRNA.* Semin Cell Dev Biol, 1997. **8**(1): p. 57-63.
70. Shen, E.C., et al., *Arginine methylation facilitates the nuclear export of hnRNP proteins.* Genes Dev, 1998. **12**(5): p. 679-91.
71. Khan, M.I., J. Zhang, and Q. Liu, *HnRNP F and hnRNP H1 regulate mRNA stability of amyloid precursor protein.* Neuroreport, 2021. **32**(9): p. 824-832.
72. Kajitani, N., et al., *hnRNP L controls HPV16 RNA polyadenylation and splicing in an Akt kinase-dependent manner.* Nucleic Acids Res, 2017. **45**(16): p. 9654-9678.
73. Martinez-Contreras, R., et al., *hnRNP proteins and splicing control.* Adv Exp Med Biol, 2007. **623**: p. 123-47.
74. Mayeda, A. and A.R. Krainer, *Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2.* Cell, 1992. **68**(2): p. 365-75.
75. Zhang, Q.S., et al., *hnRNP A1 associates with telomere ends and stimulates telomerase activity.* RNA, 2006. **12**(6): p. 1116-28.
76. Lee, E.K., et al., *hnRNP C promotes APP translation by competing with FMRP for APP mRNA recruitment to P bodies.* Nat Struct Mol Biol, 2010. **17**(6): p. 732-9.
77. Quinones-Valdez, G., et al., *Regulation of RNA editing by RNA-binding proteins in human cells.* Commun Biol, 2019. **2**: p. 19.
78. Fung, P.A., R. Labrecque, and T. Pederson, *RNA-dependent phosphorylation of a nuclear RNA binding protein.* Proc Natl Acad Sci U S A, 1997. **94**(4): p. 1064-8.
79. Jean-Philippe, J., S. Paz, and M. Caputi, *hnRNP A1: the Swiss army knife of gene expression.* Int J Mol Sci, 2013. **14**(9): p. 18999-9024.
80. Wang, Y., et al., *A complex network of factors with overlapping affinities represses splicing through intronic elements.* Nat Struct Mol Biol, 2013. **20**(1): p. 36-45.
81. Okunola, H.L. and A.R. Krainer, *Cooperative-binding and splicing-repressive properties of hnRNP A1.* Mol Cell Biol, 2009. **29**(20): p. 5620-31.
82. Zhu, J., A. Mayeda, and A.R. Krainer, *Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins.* Mol Cell, 2001. **8**(6): p. 1351-61.
83. Zarnack, K., et al., *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.* Cell, 2013. **152**(3): p. 453-66.
84. Cao, W., et al., *Control of alternative splicing by forskolin through hnRNP K during neuronal differentiation.* Nucleic Acids Res, 2012. **40**(16): p. 8059-71.
85. Patton, J.G., et al., *Characterization and molecular cloning of polypyrimidine tract-binding protein: a component of a complex necessary for pre-mRNA splicing.* Genes Dev, 1991. **5**(7): p. 1237-51.
86. Robinson, R., *Looping out introns to help splicing.* PLoS Biol, 2006. **4**(2): p. e41.
87. Martinez-Contreras, R., et al., *Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing.* PLoS Biol, 2006. **4**(2): p. e21.
88. Fisette, J.F., et al., *hnRNP A1 and hnRNP H can collaborate to modulate 5' splice site selection.* RNA, 2010. **16**(1): p. 228-38.

89. Nasim, F.U., et al., *High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism.* RNA, 2002. **8**(8): p. 1078-89.

90. Li, S., et al., *Identification of an aptamer targeting hnRNP A1 by tissue slide-based SELEX.* J Pathol, 2009. **218**(3): p. 327-36.

91. Han, K., et al., *A combinatorial code for splicing silencing: UAGG and GGGG motifs.* PLoS Biol, 2005. **3**(5): p. e158.

92. Caputi, M. and A.M. Zahler, *Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family.* J Biol Chem, 2001. **276**(47): p. 43850-9.

93. Freedman, M.L., et al., *Principles for the post-GWAS functional characterization of cancer risk loci.* Nat Genet, 2011. **43**(6): p. 513-8.

94. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

95. He, B., et al., *Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers.* Sci Adv, 2020. **6**(30): p. eaba3064.

96. Fox-Walsh, K.L., et al., *The architecture of pre-mRNAs affects mechanisms of splice-site pairing.* Proc Natl Acad Sci U S A, 2005. **102**(45): p. 16176-81.

97. Pai, A.A., et al., *The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture.* Elife, 2017. **6**.

98. Lim, K.H., et al., *Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes.* Proc Natl Acad Sci U S A, 2011. **108**(27): p. 11093-8.

99. Fairbrother, W.G., et al., *Predictive identification of exonic splicing enhancers in human genes.* Science, 2002. **297**(5583): p. 1007-13.

100. Brillen, A.L., et al., *Succession of splicing regulatory elements determines cryptic 5ss functionality.* Nucleic Acids Res, 2017. **45**(7): p. 4202-4216.

101. Paz, I., et al., *RBPmap: a web server for mapping binding sites of RNA-binding proteins.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W361-7.

102. Akerman, M., et al., *A computational approach for genome-wide mapping of splicing factor binding sites.* Genome Biol, 2009. **10**(3): p. R30.

103. Cartegni, L., et al., *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res, 2003. **31**(13): p. 3568-71.

104. Ke, S., et al., *Quantitative evaluation of all hexamers as exonic splicing elements.* Genome Res, 2011. **21**(8): p. 1360-74.

105. Wang, Z., et al., *Systematic identification and analysis of exonic splicing silencers.* Cell, 2004. **119**(6): p. 831-45.

106. Giudice, G., et al., *ATtRACT-a database of RNA-binding proteins and associated motifs.* Database (Oxford), 2016. **2016**.

107. Piva, F., et al., *SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans.* Bioinformatics, 2009. **25**(9): p. 1211-3.

108. Jaganathan, K., et al., *Predicting Splicing from Primary Sequence with Deep Learning.* Cell, 2019. **176**(3): p. 535-548 e24.

109. Pertea, M., X. Lin, and S.L. Salzberg, *GeneSplicer: a new computational method for splice site prediction.* Nucleic Acids Res, 2001. **29**(5): p. 1185-90.

110. Reese, M.G., et al., *Improved splice site detection in Genie.* J Comput Biol, 1997. **4**(3): p. 311-23.

111. Scotti, M.M. and M.S. Swanson, *RNA mis-splicing in disease.* Nat Rev Genet, 2016. **17**(1): p. 19-32.
112. Singh, R.K. and T.A. Cooper, *Pre-mRNA splicing in disease and therapeutics.* Trends Mol Med, 2012. **18**(8): p. 472-82.
113. Cao, A. and R. Galanello, *Beta-thalassemia.* Genet Med, 2010. **12**(2): p. 61-76.
114. Treisman, R., et al., *A single-base change at a splice site in a beta 0-thalassemic gene causes abnormal RNA splicing.* Cell, 1982. **29**(3): p. 903-11.
115. Fletcher, S., et al., *Antisense suppression of donor splice site mutations in the dystrophin gene transcript.* Mol Genet Genomic Med, 2013. **1**(3): p. 162-73.
116. Takeshima, Y., et al., *Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center.* J Hum Genet, 2010. **55**(6): p. 379-88.
117. Yeo, G., et al., *Variation in alternative splicing across human tissues.* Genome Biol, 2004. **5**(10): p. R74.
118. Su, C.H., D. D, and W.Y. Tarn, *Alternative Splicing in Neurogenesis and Brain Development.* Front Mol Biosci, 2018. **5**: p. 12.
119. Braggin, J.E., et al., *Alternative splicing in a presenilin 2 variant associated with Alzheimer disease.* Ann Clin Transl Neurol, 2019. **6**(4): p. 762-777.
120. Kar, A., et al., *Tau alternative splicing and frontotemporal dementia.* Alzheimer Dis Assoc Disord, 2005. **19 Suppl 1**(Suppl 1): p. S29-36.
121. Tseng, E., et al., *The Landscape of SNCA Transcripts Across Synucleinopathies: New Insights From Long Reads Sequencing Analysis.* Front Genet, 2019. **10**: p. 584.
122. Perrone, B., et al., *Alternative Splicing of ALS Genes: Misregulation and Potential Therapies.* Cell Mol Neurobiol, 2020. **40**(1): p. 1-14.
123. Nikom, D. and S. Zheng, *Alternative splicing in neurodegenerative disease and the promise of RNA therapies.* Nat Rev Neurosci, 2023. **24**(8): p. 457-473.
124. Oueslati, A., *Implication of Alpha-Synuclein Phosphorylation at S129 in Synucleinopathies: What Have We Learned in the Last Decade?* J Parkinsons Dis, 2016. **6**(1): p. 39-51.
125. Ouyang, J., et al., *The role of alternative splicing in human cancer progression.* Am J Cancer Res, 2021. **11**(10): p. 4642-4667.
126. Babic, I., et al., *EGFR mutation-induced alternative splicing of Max contributes to growth of glycolytic tumors in brain cancer.* Cell Metab, 2013. **17**(6): p. 1000-1008.
127. Anczukow, O., et al., *The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation.* Nat Struct Mol Biol, 2012. **19**(2): p. 220-8.
128. Zhou, X., et al., *BCLAF1 and its splicing regulator SRSF10 regulate the tumorigenic potential of colon cancer cells.* Nat Commun, 2014. **5**: p. 4581.
129. Jensen, M.A., J.E. Wilkinson, and A.R. Krainer, *Splicing factor SRSF6 promotes hyperplasia of sensitized skin.* Nat Struct Mol Biol, 2014. **21**(2): p. 189-97.
130. Kim, H.R., et al., *SRSF5: a novel marker for small-cell lung cancer and pleural metastatic cancer.* Lung Cancer, 2016. **99**: p. 57-65.
131. Zhou, X., et al., *The RNA-binding protein SRSF1 is a key cell cycle regulator via stabilizing NEAT1 in glioma.* Int J Biochem Cell Biol, 2019. **113**: p. 75-86.
132. Gallardo, M., et al., *Aberrant hnRNP K expression: All roads lead to cancer.* Cell Cycle, 2016. **15**(12): p. 1552-7.
133. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.* Science, 2008. **320**(5883): p. 1643-7.

134. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.

135. Robberson, B.L., G.J. Cote, and S.M. Berget, *Exon definition may facilitate splice site selection in RNAs with multiple exons.* Mol Cell Biol, 1990. **10**(1): p. 84-94.

136. Porubsky, D., et al., *Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads.* Nat Biotechnol, 2021. **39**(3): p. 302-308.

137. Chan, J.J., et al., *Pan-cancer pervasive upregulation of 3' UTR splicing drives tumourigenesis.* Nat Cell Biol, 2022. **24**(6): p. 928-939.

138. Ciolli Mattioli, C., et al., *Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments.* Nucleic Acids Res, 2019. **47**(5): p. 2560-2573.

139. Graveley, B.R., *Alternative splicing: increasing diversity in the proteomic world.* Trends Genet, 2001. **17**(2): p. 100-7.

140. Dutertre, M., et al., *The emerging role of pre-messenger RNA splicing in stress responses: sending alternative messages and silent messengers.* RNA Biol, 2011. **8**(5): p. 740-7.

141. Shattuck, J.E., et al., *Sky1: at the intersection of prion-like proteins and stress granule regulation.* Curr Genet, 2020. **66**(3): p. 463-468.

142. Allemand, E., et al., *Regulation of heterogenous nuclear ribonucleoprotein A1 transport by phosphorylation in cells stressed by osmotic shock.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3605-10.

143. Buchan, J.R. and R. Parker, *Eukaryotic stress granules: the ins and outs of translation.* Mol Cell, 2009. **36**(6): p. 932-41.

144. Jain, S., et al., *ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure.* Cell, 2016. **164**(3): p. 487-98.

145. Gonnissen, S., et al., *High Concentration of Low-Density Lipoprotein Results in Disturbances in Mitochondrial Transcription and Functionality in Endothelial Cells.* Oxid Med Cell Longev, 2019. **2019**: p. 7976382.

146. Janssen, W.J., et al., *Inflammation-Induced Alternative Pre-mRNA Splicing in Mouse Alveolar Macrophages.* G3 (Bethesda), 2020. **10**(2): p. 555-567.

147. Ramirez-Peinado, S., et al., *TRAPPC13 modulates autophagy and the response to Golgi stress.* J Cell Sci, 2017. **130**(14): p. 2251-2265.

148. Merk, D., et al., *Selenoprotein T Protects Endothelial Cells against Lipopolysaccharide-Induced Activation and Apoptosis.* Antioxidants (Basel), 2021. **10**(9).

149. Ptok, J., et al., *Modifying splice site usage with ModCon: Maintaining the genetic code while changing the underlying mRNP code.* Comput Struct Biotechnol J, 2021. **19**: p. 3069-3076.

150. Ostermann, P.N., et al., *Let It Go: HIV-1 cis-Acting Repressive Sequences.* J Virol, 2021. **95**(15): p. e0034221.

151. Grosso, A.R., et al., *Tissue-specific splicing factor gene expression signatures.* Nucleic Acids Res, 2008. **36**(15): p. 4823-32.

152. Zhang, X.H. and L.A. Chasin, *Computational definition of sequence motifs governing constitutive exon splicing.* Genes Dev, 2004. **18**(11): p. 1241-50.

153. Wong, M.S., J.B. Kinney, and A.R. Krainer, *Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites.* Mol Cell, 2018. **71**(6): p. 1012-1026 e3.

154. Krchnakova, Z., et al., *Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins.* Nucleic Acids Res, 2019. **47**(2): p. 911-928.

155. Zhang, Z., et al., *Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay.* BMC Biol, 2009. **7**: p. 23.
156. Hall, S.L. and R.A. Padgett, *Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites.* J Mol Biol, 1994. **239**(3): p. 357-65.
157. Turunen, J.J., et al., *The significant other: splicing by the minor spliceosome.* Wiley Interdiscip Rev RNA, 2013. **4**(1): p. 61-76.
158. Alioto, T.S., *U12DB: a database of orthologous U12-type spliceosomal introns.* Nucleic Acids Res, 2007. **35**(Database issue): p. D110-5.
159. Sheth, N., et al., *Comprehensive splice-site analysis using comparative genomics.* Nucleic Acids Res, 2006. **34**(14): p. 3955-67.
160. Abril, J.F., R. Castelo, and R. Guigo, *Comparison of splice sites in mammals and chicken.* Genome Res, 2005. **15**(1): p. 111-9.
161. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.
162. Buchner, N., et al., *Unhealthy diet and ultrafine carbon black particles induce senescence and disease associated phenotypic changes.* Exp Gerontol, 2013. **48**(1): p. 8-16.
163. Le Gouill, E., et al., *Endothelial nitric oxide synthase (eNOS) knockout mice have defective mitochondrial beta-oxidation.* Diabetes, 2007. **56**(11): p. 2690-6.
164. Corona, J.C. and M.R. Duchen, *PPARgamma as a therapeutic target to rescue mitochondrial function in neurological disease.* Free Radic Biol Med, 2016. **100**: p. 153-163.
165. Tanguy, Y., et al., *The PACAP-regulated gene selenoprotein T is highly induced in nervous, endocrine, and metabolic tissues during ontogenetic and regenerative processes.* Endocrinology, 2011. **152**(11): p. 4322-35.
166. Gehring, N.H., E. Wahle, and U. Fischer, *Deciphering the mRNP Code: RNA-Bound Determinants of Post-Transcriptional Gene Regulation.* Trends Biochem Sci, 2017. **42**(5): p. 369-382.
167. Schwartz, S., B.K. Felber, and G.N. Pavlakis, *Distinct RNA sequences in the gag region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of Rev protein.* J Virol, 1992. **66**(1): p. 150-9.
168. Schwartz, S., et al., *Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression.* J Virol, 1992. **66**(12): p. 7176-82.

8. Figure/Table contributions

Publication I

Table 1: JP collected and explained the SRE prediction algorithms

Figure 3: During literature search, JP found and described the TRIM62/HMSD minigene reporter

Table 2: JP applied the reviewed tools of table 1 on the TRIM62/HMSD reporter

Table 3: JP compared strengths between the variable 5’ss of the TRIM62/HMSD reporter

Publication II

Figure 4: JP analyzed 5’ss usage of the massively parallel splicing assay (MPSA), after processing the raw sequencing data.

Table 1: JP processed and analyzed of the raw mRNA sequencing data from primary fibroblasts, generated the 5’ss/GT-site pairs and analyzed the SSHW differences between weaker or stronger GT-sites

Figure 5: JP processed and analyzed of the raw mRNA sequencing data from primary fibroblasts, generated the 5’ss/GT-site pairs and compared the distribution of strength and SSHW between GT-sites and annotated 5’ss (i-vi). Final figures generated by ST.

Figure 6. JP processed and analyzed of the raw mRNA sequencing data from primary fibroblasts and generated the 5’ss/GT-site pairs, that were used for Figure 6A. JP calculated the receiver operating curves of the three logistic regression models, depicted in Figure 6B.

Unpublished Manuscript I

JP generated all figures and associated data

Publication III

JP developed the algorithm and generated the only Figure 1

Publication IV

Figure 2: JP generated Figure 1 depicting NOS3 mRNA isoforms and domains

Table 1: JP analyzed differential splicing in mRNA sequencing data of primary endothelial cells and measured splice NOS3 splice site usage upon treatment with high concentrations of low-density lipoprotein (LDL)

Table 2: JP analyzed mRNA sequencing data of primary endothelial cells upon treatment with high concentrations of low-density lipoprotein (LDL), comparing expression of mitochondrial and nuclear encoded genes.

Table 3: JP analyzed mRNA sequencing data of primary endothelial cells upon treatment with high concentrations of low-density lipoprotein (LDL), measuring expression mitochondrial transcripts.

Figure 9: JP analyzed mRNA sequencing data of primary endothelial cells upon treatment with high concentrations of low-density lipoprotein (LDL), calculating amounts of mtRNA precursor transcripts by measuring reads overlapping regions, separated in later mtRNA processing.

Publication V

Figure 1: JP analyzed mRNA sequencing data of primary endothelial cells upon treatment with lipopolysaccharide (LPS), either upon transduction with a lentiviral expression vector for APEX1(1-20) or with an empty vector. JP did the differential gene expression (DGE) analysis and depicted the results in the Venn diagrams and heatmaps

Figure 2: Regulation of SELENOT and PXDN, observed during DGE analysis of JP was validated in this figure by semi-quantitative real-time PCR

Figure 4: LPS-induced regulation of ICAM levels, observed during DGE of JP was validated in Figure 4B

Publication VI

Figure 1: JP developed the ModCon algorithm and generated the figure explaining the principle of the genetic algorithm

Figure 2: JP generated the figure explaining the applied methods of recombination used during the genetic algorithm

Figure 3: JP applied ModCon on 1000 TSL1 5’ss surroundings and measured the achieved change in SSHW, using 4 different ModCon settings

Figure 4: JP compared achieved SSHW optimization between ModCon settings with or without spike-in sequences from the sliding window approach per generation

Figure 6: JP applied ModCon on a firefly luciferase open reading frame (Figure 6A/B)

Publication VII

Table 1: JP determined the genetic coordinates of same HIV-1 sequence elements

Figure 4: JP determined the genetic coordinates of same HIV-1 sequence elements and helped analyzing the HEXplorer score per nucleotide