Genome-wide Sequence Analysis of Human Splice Acceptor Regions for Motif Discovery

Gülşah Karaduman Bahçe

Orta Dogu Teknik Universitesi

YEŞİM AYDIN SON (✉ yesim@metu.edu.tr)

Middle East Technical University <https://orcid.org/0000-0002-8118-4272>

Research article

Keywords: Genome-wide motif analysis, Branch site motifs, Splice acceptor region motifs

Posted Date: May 9th, 2019

DOI: <https://doi.org/10.21203/rs.2.9538/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Splicing mechanism in eukaryotic cells, drives the protein diversity. This important mechanism is orchestrated by regulatory factors controlled through various splicing signals. 3' and 5' splice sites, and common branch point sequences are known as initiating signals of splicing, and changes in these signals can be the underlying cause of diseases. Nevertheless, an extensive genome-wide analysis of the sequences around these signals is lacking. In this study, we focused on the genome-wide motif analysis of splice acceptor region and analyzed 400 nucleotides long sequences (300 nucleotides upstream and 100 nucleotides downstream of 3') to identify candidate motifs with functional importance. Results 207,583 sequences retrieved from Ensembl Biomart are analyzed with MEME ChIP resulting in 518 significant splice acceptor region motifs. Besides 457 known motifs, 227 of which are observed on various mammals other than Homo sapiens, 61 novel mammalian motifs are identified. Furthermore, proteins binding to the known motifs, are mainly annotated for homeobox, homeodomain, DNA binding region, and transcription regulation function. Additionally, 4 novel motifs, supported with experimental evidence from DNA binding studies, are identified as potential branch-site motifs.

Conclusions Genome-wide computational motif analysis around the human acceptor region sequences identified significant splice acceptor region motifs. The known mammalian motifs, among the computationally significant acceptor region motifs, had known biological roles with prior experimental support. The novel acceptor region motifs identified for the first time in this study are proposed as candidate motifs, and studying their biological roles in functional studies will further increase our understanding of the splicing mechanisms.

Background

In eukaryotes, genes are transcribed into pre-messenger RNAs (pre-mRNA) from which introns are excised and the remaining exons are joined to produce the mature mRNA through the splicing process. Alternative splicing allows production of variety of proteins from a single pre-mRNA by regulating the inclusion of different exon combinations in different mRNAs and thus in the mature transcripts [1]. In fact, gene splicing increases the genetic coding capacity to form structurally and functionally different protein isoforms.

Protein coding genes, are generally spliced and most of them are alternatively spliced by means of a fine balance within the regulatory factors, which controls the splice-site selection. Disturbance of regulation of splicing results in malfunction of the biological processes, which can lead to variety of diseases [2]. For example, spinal muscular atrophy, is caused by a point mutation in an exonic regulatory element [3–6]. Tauopathies (diseases of central nervous system) such as in Alzheimer's Disease, is known to be due to the changes in the ratio of the protein isoforms produced by alternative splicing [7–9]. Frontotemporal lobar dementia is caused by the loss of a splicing factor and aberrant alternative splicing events of several genes are observed in patients with Parkinson's Disease [10]. Furthermore, specific alterations in the expression of splicing factors are observed in various cancers. Mutations of tumor suppressor genes

affecting splice site selection are shown to cause cancer, as in the case of BRCA1 gene for breast cancer [11–13].

The currently accepted molecular mechanisms of splicing reveals that introns are excised away from the primary transcripts at conserved sequences called splice sites, which are found at the 5' and 3' ends of the introns. The excised intronic RNA sequence commonly begins with the dinucleotide GU at the 5' end and ends with AG at its 3' end, which are also known as donor and acceptor dinucleotides respectively. These consensus sequences are critical for splicing, and changes on these sequences result in inhibition of splicing [14]. Another important sequence for splicing resides at the branch point, which is located between 19 to 37 nucleotides upstream of the acceptor site [15] and contains the nucleotide (always adenine) which forms the intermediate intron lariat with 5' splice site [16]. A recent genome-wide study on human splicing branch points [15] has identified a set of 5 to 6 nucleotides long motifs around the branch point. These branch point sequences are called B-box elements as they are enriched for B-nucleotides (C, G, and U). B-box elements are the motifs in the form of UUNAN (U-motif), CUNAN (C-motif), G/AUNAN (G/A-motif), and others (B-boxes without a branch point adenine) [15].

Currently, studies of splicing signals and motifs have been restricted to the sequences around the 3' and 5' splice sites, and branch point. There is lack of computational studies for genome-wide discovery and prioritization of conserved motifs around the splice regions. In order to fill this gap for 3' splice sites of human genome, we have performed a genome-wide motif discovery analysis for the sequences around the splice acceptor sites. We have retrieved 400 nucleotides long sequences around the 3' splice sites (300 nucleotides upstream and 100 nucleotides downstream) for 207,583 sites and searched for the conserved patterns. Based on an extensive analysis through motif finders and an experimental database search afterwards (which is held from March 2018 to June 2018), we have identified various patterns, some of which are already known human DNA binding motifs. Additionally, we have identified several motifs that locates to the acceptor region for the first time in human genome, but previously observed on other mammalian genomes. Furthermore, we have found various novel motifs, most of which are enriched up to 50 nucleotides around the acceptor and few conforming to branch point B-boxes indicating that they may have a role during splicing. Our findings indicate that computationally significant motifs found in this study are biologically significant for mammalian genomes. Therefore, we believe that further analysis of the proposed motifs by experimental methods will further increase the understanding on splicing mechanisms.

Methods

Main workflow of our study for the construction of a genome-wide motif discovery analysis on splice acceptor sites is as outlined in Figure 1. First, genome scale retrieval of acceptor sequences is completed, then various motif finders are used to find patterns on these sequences. Afterwards, motifs with low significance are filtered during the pre-processing step. Finally, biological relevance of the motif sequences is evaluated.

Acceptor region sequence retrieval

In this study, we have focused on the 400 nucleotides around the splice acceptor sites (300 nucleotides upstream and 100 nucleotides downstream of the acceptor site). The exact base pair locations of acceptor sites are retrieved through Ensembl Biomart [17]. Exon start locations, which are the indicators of splice acceptor sites have been retrieved for all protein-coding genes based on GRCh38.p12 human assembly of Ensembl Biomart. By this query we have also gathered the transcript start site (TSS) locations in order to eliminate 5'UTR sequences as our main consideration is the splice acceptor sites (Supplementary A.1 presents the query for the chromosome 1 through Ensembl Biomart). After gathering the exact acceptor positions by eliminating the exons with initiation codons, we have calculated the base pair locations of start and end points for 400 nucleotides long sequences surrounding the acceptor sites. Next, UCSC DAS Server, located at <http://genome.ucsc.edu/cgi-bin/das>, is used to retrieve the sequences for the selected 400-nucleotide long acceptor regions (by the Perl script provided in Supplementary A.2).

Motif analysis of 400-nucleotide long acceptor regions

MEME-ChIP [18], is selected as the motif finding tool for this study. MEME-ChIP runs several motif analyzers including MEME, DREME, Centrimo,, FIMO, and Tomtom which are explained below:

- MEME, discovers novel, ungapped motifs (recurring, fixed-length patterns) in the provided sequences [19].
- DREME, discovers short, ungapped motifs (recurring, fixed-length patterns) that are relatively enriched in the provided sequences [20].
- CentriMo, identifies known or user-provided motifs that show a significant preference for particular locations in the provided sequences [21].
- FIMO, searches the sequences for occurrences of provided motifs and provides the locations of matches [22].
- Tomtom, compares the motifs against a database of known motifs [23].

After sequence retrieval, we have taken the reverse complement of the negative strand sequences. Next, we have arranged all of the sequences in FASTA format. We have run the MEME-ChIP with its default configuration for the matched known motifs, i.e. vertebrates (in vivo and in silico) of Eukaryote DNA, which is the largest set for known motifs of eukaryotic DNA. Furthermore, for DREME and Centrimo, we have used the default parameters suggested by MEME-ChIP. However, for MEME we have limited the number of motifs to 20 and the length of the motifs to from 9 to 50 nucleotides. The former limitation is held for making the algorithm run faster and finding the most significant motifs whereas the latter limitation helps for finding the novel motifs, which may be excluded by DREME, as it looks for the shorter motifs up to 8 nucleotides (Supplementary A.3 MEME-ChIP Analysis Parameters).

Pre-processing of identified motifs

A filtering strategy is developed to select the most significant motifs, by inspecting the central enrichment information provided by Centrimo. Steps of the filtering process is as follows:

1. We have filtered the motifs with region width less than 3 nucleotide (the width of the most enriched region), since they mostly represent the common dinucleotide signal, AG, of the acceptor sites.
2. We have filtered the motifs based on the E-values (the expected number motifs that would have at least one region as enriched for best matches to the motif as the reported region) The cut-off value e^{-5} is used to filter the motifs based on their significance.
3. We have described a new value for prioritization of the identified motifs, which equals to $\frac{1}{\log_{10}(E)}$ and set its threshold to -10 in order to select the motifs with higher significance but less number of sequence matches and also the motifs with lower significance but high number of sequence matches.
4. We have filtered the motifs, which were not found to be significant by all three motif finders, CentriMo, MEME and DREME.
5. We have removed the duplicate motifs. There were some duplicate motifs such as ALX3_full_1 and MA0634.1 which are represented by different identifiers as they exist in different motif databases (i.e. HumanTF 1.0 [24] and JASPAR 2018 [25] respectively).

Biological annotation of the known splice region motifs

Functional enrichment analysis. The Database for Annotation, Visualization and Integrated Discovery (DAVID) is a functional annotation tool, which mainly provided typical batch annotation and gene-GO term enrichment analysis to highlight the most relevant GO terms associated with a given list. Biological functions of transcription factors (TFs) that recognize the significant acceptor region motifs of this study are evaluated by DAVID v6.8 [26, 27].

Comparison with experimental databases. FootprintDB is a meta-database of TFs, which locates them to their experimentally determined DNA binding motifs and DNA binding sites (i.e. cis elements) [28]. As a meta-database it integrates several databases such as 3d-footprint [29], JASPAR [25], UniPROBE [30], HumanTF [24], and etc. which are experimental repositories for the binding preferences of TFs and proteins. FootprintDB's search engine requires a DNA consensus motif or site as an input to produce a list of DNA-binding proteins (mainly TFs), detected to bind with a similar DNA motif [28]. Here we have used FootprintDB's sequence search service to reveal the biological significance of the known motifs by using the count matrix representation, which is a basic position frequency matrix (PFM) created by counting the occurrences of each nucleotide at each position.

Biological annotation of the novel splice region motifs

Analysis of repetitive sequences. RepeatMasker, uses curated libraries of repeats namely Dfam [31] and Repbase [32] to screen DNA sequences for interspersed repeats and low complexity DNA sequences [33]. We have used RepeatMasker for the detection of the motifs on the repetitive sequences. For this search IUPAC codes in the sequences were replaced with the alternative nucleotides to represent up-to 3000

possible alternatives. The limitation was necessary as for many of the motifs, especially the long motifs, the number of alternatives were really high and over the capacity of the online RepeatMasker. Furthermore, UCSC Blat [34] was used for searching repetitions in case the whole motif sequence is not indicated to be repetitive by RepeatMasker but a huge portion of it may indicate a repeat.

Analysis of motif locations. In order to investigate the motif locations and detect whether there are any special regions, which are predominantly preferred, we have divided our acceptor site sequences into 8 regions as shown in Figure 2. Afterwards, we have utilized the output of FIMO program [22], which provides the exact locations of the motif occurrences in the analyzed sequences, in order to match the motif locations within the defined regions. The program run with a p-value of 0.01. Furthermore, Region 6 is further divided into 3 more regions in order to investigate the motifs that predominantly occur in the branch point region (i.e. between 19 and 37 nucleotides upstream of the acceptor site) as defined in [15].

Comparison with experimental databases. FootprintDB's sequence search facility is utilized for novel motifs in the same manner as known motifs to reveal their biological significance.

Results

518 motifs were found to be significant

MEME-ChIP analysis of 207,583 splice acceptor site sequences revealed 1491 significant Centrimo motifs with $e\text{-value} \leq 10$ using the binomial test. After applying the filtering rules as defined in the pre-processing step, 518 motifs were selected for further investigation. Among these 518 acceptor site motifs, 457 (i.e. nearly 88.22%) were mentioned by MEME-ChIP to be previously known, 33 were novel motifs found through MEME, and the remaining 28 were the novel motifs found through DREME. All of these known and novel motifs are listed in Supplementary B.

Analysis results of known motifs

FootprintDB annotation of 457 previously known motifs revealed 455 of the motifs as DNA binding motifs. The remaining two motifs, namely "HOXD12_DBD_4" and "PITX1_full_1", didn't have any footprintDB entry, although there was a link provided by MEME-ChIP. This may be due to some versioning inconsistencies between MEME-ChIP and footprintDB.

51% of the known motifs are experimentally observed in human genome and remaining 49% are experimentally observed in other mammalian genomes. When we investigate previously known 455 motifs through footprintDB in terms of the organisms that they were identified, we have observed that 51% of them have been experimentally shown to locate in human genome (221 of the motifs were identified in human genome and 9 of the motifs were identified in both Homo sapiens and Mus musculus). Remaining 49%, i.e. 225, of the known motifs previously identified on non-human mammalian genomes, mostly in Mus musculus genome, and they were novel to Homo sapiens (Figure 3).

Functional annotation of TFs binding to the known acceptor region motifs identified in this study are mainly classified as homeobox, homeodomain, DNA binding regions or transcription regulation activators. We investigated the functional annotations of the TFs which recognize the known acceptor region motifs by DAVID analysis. Analysis for Homo sapiens and Mus musculus were handled separately. 6 motifs found on several rat types have been discarded from DAVID analysis as the number of motifs was low (Gene lists for Homo sapiens and Mus musculus TFs are provided in Supplementary C). The most significant functional annotations with p-value > e-50 for 162 human and 164 mouse TFs are listed in Figure 4, and Figure 5, respectively. In both analysis, enrichment terms were clustered in homeobox, homeodomain, DNA binding and transcription regulation categories.

Functional annotation analysis of TFs binding to known acceptor region motifs in human genome by DAVID reported 17 terms in the first cluster of significant enrichment. When terms annotated for Homo sapiens motifs are sorted according to p-value of the enriched terms after Benjamini correction, four Homeobox or homeodomain terms ranked in top 5. When the functional annotation is repeated for the known acceptor region motif binding TFs in mouse genome 20 terms showed significant enrichment in the first cluster. All terms enriched in the home sapiens were also observed in the enrichment results for the mouse genome. Three new terms for the known acceptor site motif binding TFs in mouse genome was “regulation of transcription”, “transcription DNA-template”, and “multicellular organism development”. Like enrichment results for Homo sapiens TFs, homeodomain or homeobox terms were abundant in the top ranks.

Analysis results of novel motifs

Sequence lengths of novel motifs. Genome wide motif search on splice acceptor sequences through MEME algorithm has produced 33 novel motifs, which are 9 to 50 nucleotides long. Searching motifs with DREME algorithm has produced 28 motifs which are 5 to 8 nucleotides long. Distribution of motifs according to their lengths is summarized in Figure 6.

Repeat Masker analysis results. For all the novel motifs, we have applied repetition analysis through RepeatMasker which showed that 38 of the novel motifs shorter than 20nt long do not indicate any repeats. However, 22 of the novel motifs longer than 20nt were found by RepeatMasker as simple or low complexity repeats. The only remaining motif was “GGRRACAGAGRCYCAGAGRGRG”. Searching for the alternatives of this sequence through UCSC Blat by replacing the IUPAC codes with possible nucleotides we have observed that 20 nucleotides of this sequence is a SINE (Short Interspersed Nuclear Elements) repeat (Supplementary D). As we did not expect any further biological significance for the repetitive sequences, we continued our analysis of novel motifs with 38 motifs from this part on.

Novel motifs with <20 nt length are most frequently located up to 50 nt upstream of the splice acceptor site. Location analysis of 38 novel motifs (<20 nt long), in terms of the occurrence frequency in the 8 identified regions of acceptor region sequences (shown in Figure 2 of Methods section), through the output of FIMO program, produced the following results:

- Region 6, i.e. the region of 50 nucleotides at the upstream side of acceptor, was the first rank for 31 of the novel motifs,
- Region 5, i.e. the region from 50 to 100 nucleotides at the upstream side of acceptor, was the first rank for the remaining 7 novel motifs,
- Other regions were not preferred by any of the novel motifs in the first rank.

These results show that, frequency of novel motif occurrences is higher up to 100 nucleotides upstream of the acceptors. Supplementary E provides the list of exact frequencies of all these 38 motifs in the defined regions.

4 of the novel motifs with <20nt length are enriched between 19th and 37th nucleotides upstream of the splice acceptor site and conform to branch point B-boxes. According to an experimental study [15], branch-points have shown a restricted distribution upstream of the 3' splice site with 90% of branch-points occurring within 19 to 37 nucleotides upstream. This is the Region 6 (shown in Figure 2 of Methods section), which is the first rank for 31 novel motifs' occurrence frequencies. With a closer look at the positions of these novel motifs within Region 6, we have observed that 4 of the novel motifs namely, AAATRH, AAATRY, ACATWY, and TAWATR occur most frequently in Region 6_2. Furthermore, AAATRH, AAATRY, and TAWATR fully conform to the branch point B-box element UUNAN (i.e. AANTN for the DNA sequence) and ACATWY fully conform to another B-box element G/UNAN (i.e. C/TANTN for the DNA sequence) when the ambiguity codes are replaced with the possible nucleotides. The remaining 27 motifs mostly occur within 1 to 18 nucleotides upstream (i.e. Region 6_3). Supplementary E provides the list of exact frequencies of these 31 motifs in the sub regions of Region 6.

Most of the novel motifs with <20nt length are found to be >80% similar with previously known Homo sapiens motifs. Another analysis for the novel motifs with <20nt length, were searching them from footprintDB. For this purpose, footprintDB's sequence search feature was used and this analysis showed that 34 of the inspected 38 motifs were >80% similar with previously known motifs of Homo sapiens as shown in Figure 7. The 4 motifs that showed less similarity was namely "CYCTCCYTNCACMCT", "RGRRAARGRRAGAGAG", "TTTTTTTTTTTTTNAKW", and "YCCYTCCCASCYCCYC", and their similarity levels to previously known motifs of Homo sapiens were 67,75%, 69,56%, 78,19% and 70,88% respectively.

Discussion

In this study, we have performed a genome-wide computational motif analysis around the human acceptor region sequences. We have retrieved sequences of 400 nucleotides long regions, which are

located 300 nucleotides upstream and 100 nucleotides downstream of the actual acceptor sites. Motif finding algorithms included in the MEME-ChIP were utilized for each chromosome independently, due to the computational limitation of the tool (The sequence file input of MEME-ChIP can be at most 80MB, but our sequences are nearly 92MB in total). The resulting significant motifs were combined afterwards, and filtered according to their length, significance, and consensus. 518 motifs passed the filtering. Among the selected motifs 457 were previously known mammalian motifs, but 61 of them were novel motifs of the acceptor region.

Previously known motifs, identified as computationally significant motifs of acceptor site, were experimentally proven to locate on *Homo sapiens* and/or other mammalian genomes as DNA binding motifs. The proteins, binding to the known acceptor site motifs identified here, mainly annotated for homeobox and homeodomain, DNA binding region, and/or transcription regulation functions in both *Homo sapiens* and *Mus musculus* genome. Homeobox genes direct the formation of body structures during early embryonic development, and their alternative splicing is known to be under strict regulation [35]. The genomic organization and the complex arrangement of regulative elements in homeobox gene structure is essential in their spatio-temporal regulation and function [36]. Also conserved intronic regions, containing homeobox sequences [37] or alternative promoters [38] are known to have regulative roles. The homeobox sequences that are identified as acceptor region motifs in this study likely to have a regulative role in the alternative splicing mechanism. Also, the computationally significant acceptor motifs that are experimentally shown to be functional in non-human organisms can be prioritized for further investigation of their roles in human genome.

We were able to detect 61 novel motifs through two different algorithms, namely DREME and MEME. We have used both algorithms as DREME is designed for finding shorter motifs up to 8 nt long whereas MEME searches for the longer ones. We have arranged MEME to detect motifs up to 50 nt long. As functional annotation results of the known motifs showed their biological significance and supported the computational predictions, we investigated the relevance of the novel computationally significant motifs. Comprehensive analysis of the biological significance of the novel motifs are done based on the length of the motifs.

Novel motifs with >20nt length were found to be simple or low complexity repeats. Shorter motifs found to be significant by DREME are all found to have sequence similarity between 84,6% to 99,80%, to previously known *Homo sapiens* motifs without any perfect match. For example, the motif namely "RGRAA", showed 99,80% similarity to "yrsTTTCabTTyCyc" (IRF8_HUMAN.H10MO.DIM01272 model of *Homo sapiens* COMprehensive MODEL Collection (HOCOMOCO) v10 – [39]) when it is aligned as "——TTyCy-". This is not an exact match as the aligned motif is much longer, so MEME-ChIP categorizes it as a novel motif. Only 6 of the MEME motifs were found to be >80% similar with previously known motifs. Furthermore, such MEME motifs are at most 16 nt long. Our observation supports that as the novel motif length increases, the corresponding sequences are not actually biologically significant patterns but mainly repetitive sequences.

For the novel motifs shorter than 20nt (i.e. 38 motifs), locational analysis of FIMO [22] outputs showed that 31 of them most frequently occur in Region 6, which is the intronic region in the upstream part of the splice acceptor site. Furthermore, Region 6 overlaps with the biologically significant part of the intronic sequences, where polypyrimidine tract (located about 5-40 base pairs upstream [40]) and the branch-point (19-37 base pairs upstream - [15]) maps. 27 of the novel motifs located most frequently between the 1st and 18th nucleotides upstream of the acceptor site, indicating a high probability for regulating acceptor site splicing. Like the donor site splice signals that facilitates binding of U1 snRNP, these novel acceptor motifs can be proposed to be sequences that facilitates recognition of a site or a unique U2AF heterodimer [41]. Additionally, we have observed that 4 of the novel motifs (AAATRH, AAATRY, ACATWY, and TAWATR) most frequently occur between 19th and 37th nucleotides upstream of the acceptor site and fully conform to the branch-point B-boxes with an additional nucleotide. This observation based on the locations of the novel acceptor motifs identified by this genome-wide analysis of acceptor sequences, led us to suggest one nucleotide extension to the boundaries of the B-boxes regions. For the other 7 novel motifs which most frequently occur in Region 5, we have considered distal branch-points located ≥ 50 nt upstream [42], but did not find any conformity with B-boxes except for the motif AYATWY. Nevertheless, they might be playing regulatory roles for splicing as intronic splicing regulation predominantly occurs within 100nt of the splice sites [43].

CONCLUSIONS

To our knowledge, studies on splicing motifs have been restricted to the 3' and 5' consensus sequences, polypyrimidine tract, and the branch point B-boxes. In order to overcome this restriction for the 3' splice acceptor site and provide new routes for functional studies, we have provided the results of a genome-wide motif discovery around the splice acceptors up to 300 nucleotides upstream and 100 nucleotides downstream. This computational approach has revealed various novel motifs as well as the previously known ones (nearly half of which have been observed on other mammals but not on Homo sapiens yet). Furthermore, through genome-wide computational analysis of acceptor region motifs we have identified potential branch-site motifs that are also supported with experimental evidence from DNA binding studies. Therefore, we suggest that known motifs of other mammalian genomes and novel motifs conforming to B-boxes that are identified computationally significant as acceptor motifs in this study can be prioritized for further functional analysis.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data generated during this study are included in this published article and its supplementary materials. Acceptor site region sequences analyzed in this study are available in the GitHub repository (<https://github.com/gkaraduman/motifanalysis>).

Competing interests

The authors declare that they have no competing interest.

Funding

The authors declare that they do not have any funding for this study.

Authors' contributions

GKB collected and analyzed the sequence data. GKB and YAS commented on the results and wrote the paper. YAS coordinated the research. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Dredge BK, Polydorides AD, Darnell RB. The splice of life: alternative splicing and neurological disease. *Nat Rev Neurosci.* 2001;2:43–50. doi:10.1038/35049061.
2. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease.* 2009;1792:14–26.
3. Cartegni L, Hastings ML, Calarco JA, de Stanchina E, Krainer AR. Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2. *Am J Hum Genet.* 2006.
4. Cartegni L, Krainer AR. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN. *Nat Genet.* 2002.
5. Kashima T, Rao N, David CJ, Manley JI. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Hum Mol Genet.* 2007.
6. Kashima T, Manley JL. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet.* 2003.
7. Andreadis A. Tau gene alternative splicing: Expression patterns, regulation and modulation of function in normal brain and neurodegenerative diseases. *Biochimica et Biophysica Acta - Molecular Basis of Disease.* 2005.

8. Gallo JM, Noble W, Martin TR. RNA and protein-dependent mechanisms in tauopathies: Consequences for therapeutic strategies. *Cellular and Molecular Life Sciences*. 2007.
9. Hasegawa M, Smith MJ, Iijima M, Tabira T, Goedert M. FTDP-17 mutations N279K and S305N in tau produce increased splicing of exon 10. *FEBS Lett*. 1999.
10. Buratti E. Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease. *Front Biosci*. 2008.
11. Liu HX, Cartegni L, Zhang MQ, Krainer AR. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat Genet*. 2001.
12. Hoffman JD, Hallam SE, Venne VL, Lyon E, Ward K. Implications of a novel cryptic splice site in the BRCA1 gene. *Am J Med Genet*. 1998.
13. Orban TI, Olah E. Emerging roles of BRCA1 alternative splicing. *Journal of Clinical Pathology - Molecular Pathology*. 2003.
14. Clancy S. RNA splicing: introns, exons and spliceosome. *Nat Educ*. 2008;1:3–6.
[http://www.myteacherpages.com/webpages/sfenstemaker/files/RNA Splicing.doc](http://www.myteacherpages.com/webpages/sfenstemaker/files/RNA%20Splicing.doc).
15. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. *Genome Res*. 2015;25:290–303.
16. Fica SM, Tuttle N, Novak T, Li N-S, Lu J, Koodathingal P, et al. RNA catalyses nuclear pre-mRNA splicing. *Nature*. 2013;503:229–34. doi:10.1038/nature12734.
17. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*. 2011;2011.
18. Machanick P, Bailey TL. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*. 2011;27:1696–7.
19. Bailey TL, Elkan C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. 1994. p. 28–36. doi:citeulike-article-id:878292.
20. Bailey TL. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*. 2011;27:1653–9.
21. Bailey TL, MacHanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res*. 2012;40.
22. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*. 2011;27:1017–8.

23. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biol.* 2007.
24. Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. *Cell.* 2013;152:327–39.
25. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2015;44 November 2015:gkv1176. doi:10.1093/nar/gkv1176.
26. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009.
27. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009.
28. Sebastian A, Contreras-Moreira B. FootprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics.* 2014;30:258–65.
29. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.* 2010;38 Database:D91–7.
30. Robasky K, Bulyk ML. UniPROBE, update 2011: Expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 2011;39 SUPPL. 1.
31. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* 2016.
32. Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 2005.
33. Smit A, Hubley R, Green P. RepeatMasker Open-4.0.6 2013-2015 . <http://www.repeatmasker.org>. 2017.
34. Kent WJ. BLAT - The BLAST-like alignment tool. *Genome Res.* 2002;12:656–64.
35. Stroeher VL, Gaiser JC, Garber RL. Alternative RNA splicing that is spatially regulated: generation of transcripts from the Antennapedia gene of *Drosophila melanogaster* with different protein-coding regions. *Mol Cell Biol.* 2015.
36. Coulombe Y, Lemieux M, Moreau J, Aubin J, Joksimovic M, Bérubé-Simard FA, et al. Multiple promoters and alternative splicing: Hoxa5 transcriptional complexity in the mouse embryo. *PLoS One.* 2010.

37. Keegan LP, Haerry TE, Crotty DA, Packer AI, Wolgemuth DJ, Gehring WJ. A sequence conserved in vertebrate Hox gene introns functions as an enhancer regulated by posterior homeotic genes in *Drosophila* imaginal discs. *Mech Dev*. 1997.
38. Vacik T, Raska I. Alternative intronic promoters in development and disease. *Protoplasma*. 2017.
39. Kulakovskiy I V, Vorontsov IE, Yevshin IS, Soboleva A V, Kasianov AS, Ashoor H, et al. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res*. 2016.
40. Lodish H, Berk A, Zipursky SL et al. *Molecular Cell Biology* 5th Ed. 2000.
41. Thapar R. Roles of prolyl isomerases in RNA-mediated gene expression. *Biomolecules*. 2015.
42. Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev*. 2018.
43. Wainberg M, Alipanahi B, Frey B. Does conservation account for splicing patterns? *BMC Genomics*. 2016.

Figures

Figure 1

Workflow of the study (1. Retrieval of acceptor region sequences, 2. Motif analysis of the retrieved sequences 3. Pre-processing of motifs, 4.A. Biological annotation of known motifs, 4.B. Biological annotation of novel motifs).

Figure 2

8 regions defined for the analyzed sequences. Region 1, Region 2, Region 3, Region 4, Region 5, and Region 6 are located in the intronic region at the upstream side of the acceptor dinucleotide, i.e. AG. Region 7 and Region 8 are located in the exonic region at the downstream side of the acceptor. Region 6 is further divided into 3 regions. Region 6_1 is the far end from the acceptor dinucleotide AG. Region 6_2 is the region where branch-point locates with highest probability. Region 6_3 is the region nearest to the acceptor site dinucleotide AG.

Figure 3

Organism distribution of known motifs according to footprintDB. 49% (i.e. 221) of the motifs located in human genome, 48% (i.e. 219) of the motifs located to mouse genome, 2% (i.e. 9) of the motifs were seen on both Homo sapiens and Mus musculus, and 1% (i.e. 6) of the motifs were seen on some other rat types.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamin
☐	UP_SEQ_FEATURE	DNA-binding region:Homeobox	RT		102	63,0	1,3E-174	3,9E-172
☐	UP_KEYWORDS	Homeobox	RT		107	66,0	9,3E-169	8,9E-167
☐	INTERPRO	Homeodomain	RT		107	66,0	2,3E-165	2,3E-163
☐	INTERPRO	Homeobox, conserved site	RT		98	60,5	4,3E-161	2,2E-159
☐	UP_KEYWORDS	DNA-binding	RT		160	98,8	1,9E-158	9,3E-157
☐	INTERPRO	Homeodomain-like	RT		107	66,0	6,0E-150	2,1E-148
☐	SMART	HOX	RT		107	66,0	1,5E-139	4,0E-138
☐	GOTERM_MF_DIRECT	sequence-specific DNA binding	RT		100	61,7	5,1E-110	6,2E-108
☐	UP_KEYWORDS	Transcription regulation	RT		137	84,6	5,3E-103	1,7E-101
☐	UP_KEYWORDS	Transcription	RT		138	85,2	5,5E-103	1,3E-101
☐	UP_KEYWORDS	Nucleus	RT		161	99,4	1,9E-94	3,6E-93
☐	GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	RT		91	56,2	2,2E-69	1,3E-67
☐	INTERPRO	Homeodomain, metazoa	RT		45	27,8	8,5E-68	2,2E-66
☐	GOTERM_CC_DIRECT	nucleus	RT		152	93,8	4,6E-67	2,7E-65
☐	UP_KEYWORDS	Developmental protein	RT		82	50,6	1,2E-64	1,9E-63
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		86	53,1	9,0E-62	1,0E-58
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT		65	40,1	9,0E-55	5,1E-52

Figure 4

Functional annotations of Homo sapiens TFs binding to known acceptor motifs identified in this study.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	UP_KEYWORDS	DNA-binding	RT		154	97,5	1,7E-172	8,7E-171
☐	GOTERM_MF_DIRECT	sequence-specific DNA binding	RT		131	82,9	4,5E-164	4,3E-162
☐	INTERPRO	Homeobox, conserved site	RT		96	60,8	7,2E-164	7,1E-162
☐	UP_KEYWORDS	Homeobox	RT		100	63,3	1,2E-154	3,1E-153
☐	INTERPRO	Homeodomain	RT		100	63,3	3,1E-152	1,5E-150
☐	UP_SEQ_FEATURE	DNA-binding region:Homeobox	RT		90	57,0	1,7E-147	4,5E-145
☐	GOTERM_MF_DIRECT	DNA binding	RT		152	96,2	2,0E-140	9,3E-139
☐	INTERPRO	Homeodomain-like	RT		100	63,3	8,3E-140	2,7E-138
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		152	96,2	6,7E-130	7,2E-127
☐	SMART	HOX	RT		100	63,3	1,7E-124	3,9E-123
☐	UP_KEYWORDS	Transcription regulation	RT		127	80,4	1,9E-109	3,2E-108
☐	UP_KEYWORDS	Transcription	RT		128	81,0	2,5E-109	3,2E-108
☐	UP_KEYWORDS	Nucleus	RT		157	99,4	1,2E-108	1,3E-107
☐	GOTERM_BP_DIRECT	transcription, DNA-templated	RT		129	81,6	1,4E-98	7,5E-96
☐	GOTERM_CC_DIRECT	nucleus	RT		158	100,0	4,6E-82	2,0E-80
☐	INTERPRO	Homeodomain, metazoa	RT		50	31,6	5,0E-82	1,2E-80
☐	GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	RT		95	60,1	3,5E-81	1,1E-79
☐	UP_KEYWORDS	Developmental protein	RT		86	54,4	7,6E-74	6,5E-73
☐	GOTERM_BP_DIRECT	multicellular organism development	RT		86	54,4	3,6E-64	1,3E-61
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		82	51,9	4,2E-60	1,1E-57

Figure 5

Functional annotations of *Mus musculus* TFs binding to known acceptor motifs identified in this study.

Figure 6

Distribution of the MEME and DREME motifs according to their length. Most of the MEME motifs are 16 to 31 bases long whereas most of the DREME motifs are 6 bases long

Figure 7

Novel motifs' (<20nt length) similarity percent with previously known Homo sapiens motifs. Motifs found by DREME algorithm are shown in blue and MEME algorithm are shown in orange. DREME motifs have higher similarity percent compared to MEME identified motifs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.docx](#)