GENOME-WIDE SEQUENCE ANALYSIS OF HUMAN SPLICE ACCEPTOR REGIONS FOR MOTIF DISCOVERY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÜLŞAH KARADUMAN BAHÇE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

DECEMBER 2020

GENOME-WIDE SEQUENCE ANALYSIS OF HUMAN SPLICE ACCEPTOR REGIONS FOR MOTIF DISCOVERY

Submitted by Gülşah KARADUMAN BAHÇE in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in the Department of Health Informatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics, METU** _____

Examining Committee Members:

Assist. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Assoc. Prof. Dr. Tunca Doğan
Computer Engineering Dept., Hacettepe University _____

Assoc. Prof. Dr. Bala Gür Dedeoğlu
Biotechnology Institute, Ankara University _____

Assoc. Prof. Dr. Cem İyigün
Industrial Engineering Dept., METU _____

Date: **23.12.2020**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Gülşah Karaduman Bahçe

Signature : _____

ABSTRACT

GENOME-WIDE SEQUENCE ANALYSIS OF HUMAN SPLICE ACCEPTOR REGIONS FOR MOTIF DISCOVERY

KARADUMAN BAHÇE, Gülşah
Ph.D., Department of Health Informatics
Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

December 2020, 105 pages

For eukaryotic cells, alternative splicing of genes is a vital mechanism that drives protein diversity. Splicing signals on the genomic sequence controls the regulatory factors that orchestrate the alternative splicing. 3' and 5' splice sites and common branchpoint sequences are the primary splicing signals, and changes in these signals can be disease-causing. 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 the splice acceptor region. We analyzed 400 nucleotides long sequences (300 nucleotides upstream and 100 nucleotides downstream of 3') to identify motifs with potential functional roles. 207,583 sequences are retrieved from Ensembl Biomart and analyzed with MEME ChIP, resulting in 517 significant splice acceptor region motifs. We identified 457 known motifs and 60 novel motifs. Among the known motifs, 227 mapped to non-human mammalian genomes. Furthermore, proteins binding to the known motifs are mainly annotated for homeoboxes, homeodomains, DNA binding regions, and transcription regulation functions. 17 of the novel motifs comply with RBP binding motifs, and 10 of the novel motifs are computationally identified and supported with experimental evidence from branchpoint studies. Moreover, the acceptor region splice altering or disease-causing variants with experimental evidence are detected to co-locate with novel motifs and known motifs of homo sapiens and other mammalians. Here, we present these novel acceptor region motifs identified for the first time as splice acceptor motif candidates. Furthermore, we provide a set of 76 known mus musculus motifs that are novel to the human genome and highly co-locate with splice-altering SNPs. Experimental validation of the biological roles of novel motifs of this study with further functional studies will increase our understanding of the splicing mechanisms.

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

ÖZ

MOTİF KEŞFİ İÇİN İNSAN UÇBİRLEŞTİRME AKSEPTÖR BÖLGE SEKANSLARININ GENOM ÇAPINDA ANALİZİ

KARADUMAN BAHÇE, Gülşah
Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Aralık 2020, 105 sayfa

Ökaryotik hücreler için alternatif uçbirleştirme protein çeşitliliğini sağlayan önemli bir mekanizmadır. Genom sekansındaki uçbirleştirme sinyalleri, alternatif uçbirleştirmeyi yöneten faktörleri kontrol etmektedir. 3' ve 5' birleştirme uçları ve dallanma noktası sekansları birincil uçbirleştirme sinyalleridir ve bu sinyallerdeki değişiklikler hastalıklara neden olabilmektedir. Buna rağmen, bu sinyaller çevresindeki sekansların geniş ölçekli analizleri yapılmamıştır. Bu çalışma uçbirleştirme akseptör bölgelerinin genom çapında analizlerine odaklanmıştır. Fonksiyonel öneme sahip motifleri belirleyebilmek için 3' birleştirme ucunun 300 nükleotid yukarı bölgesi ile 100 nükleotid aşağı bölgesi arasında kalan sekanslar analiz edilmiştir. Bu bölgelerde bulunan 207,583 sekansın Ensembl Biomart'tan toplanarak MEME-ChIP ile analiz edilmesi sonucunda 517 anlamlı motif bulunmuştur. Bu motiflerin 457'si bilinen motiflerken 60'ı bu çalışmada keşfedilen yeni motiflerdir. Bilinen 457 motiften 227'si çeşitli memeli genomlarında gözlemlenmiştir ancak insan genomu için yenidir. Bilinen motiflere bağlanan proteinler çoğunlukla homeo-kutusu, homeo-alanı, DNA bağlanma bölgesi ve transkripsiyon düzenleme fonksiyonları ile ilişkilendirilmiştir. 17 yeni motif, RNA bağlayıcı motiflerle, 10 yeni motifse dallanma noktası motifleri ile örtüşmüştür. Ayrıca, uçbirleştirme değişiklikleri ya da hastalık ilişkisi doğrulanan varyantların, bulunan anlamlı motifler üzerine konumlandığı tespit edilmiştir. Bu çalışmada ilk kez belirlenen akseptör bölge motifleri, uçbirleştirme akseptör motif adayları olarak sunulmuştur. Ayrıca, insan genomu için yeni ve uçbirleştirme değişiklikleri ile ilişkili varyantlarla yüksek oranda bölgesel çakışmaya sahip 76 adet ev faresi motifi sunulmaktadır. Keşfedilen motiflerle yapılacak deneysel çalışmalar uçbirleştirme mekanizmalarının daha iyi anlaşılmasını sağlayacaktır.

Anahtar Kelimeler: Genom çapında motif analizi, Dallanma bölgesi motifleri, Uçbirleştirme akseptör bölge motifleri

To my endless love and my pretty chubby

ACKNOWLEDGMENTS

I express sincere appreciation to Assoc. Prof. Dr. Yeşim AYDIN SON for her perfect coaching and guidance for this study and doctorate duration. Her continuous guidance is invaluable to me in my theoretical education process and the writing of this thesis.

I am grateful to all thesis progress committee members Assoc. Prof. Dr. Yeşim AYDIN SON, Asst. Prof. Dr. Aybar Can ACAR, Assoc. Prof.Dr. Cem İYİGÜN; and examining committee members Assoc. Prof. Dr. Yeşim AYDIN SON, Asst. Prof. Dr. Aybar Can ACAR, Assoc. Prof.Dr. Cem İYİGÜN, Assoc. Prof. Dr. Tunca DOĞAN, and Assoc. Prof. Dr. Bala GÜR DEDEOĞLU for their participation and valuable comments.

I express sincere appreciation to all faculty members of The Graduate School of Informatics of Middle East Technical University, and all administrative staff personnel for their support throughout my doctorate studies. I learned a lot from both courses and scientific meetings.

I am very grateful to my husband, Gökhan BAHÇE, for his support throughout my Ph.D. studies. He is the person behind the scene, and this thesis would not happen without him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SNP	Single Nucleotide Polymorphism
ENA	European Nucleotide Archive
SRA	Sequence Read Archive
EMBL	The European Molecular Biology Laboratory
NIH	National Institutes of Health
NCBI	The National Center for Biotechnology Information
DDBJ	DNA Data Bank of Japan
NHGRI	National Human Genome Research Institute
UCSC	University of California Santa Cruz
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST Like Alignment Tool
VEP	Variant Effect Predictor
GO	Gene Ontology
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbSNP	The Single Nucleotide Polymorphism Database
dbGaP	The database of Genotypes and Phenotypes
GWAS	Genome-wide Association Study
PharmGKB	The Pharmacogenomics Knowledge Base
TCGA	The Cancer Genome Atlas
OMIM	Online Mendelian Inheritance in Man
TF	Transcription Factor
TSS	Transcription Start Site
RBP	RNA Binding Protein

CHAPTER 1

INTRODUCTION

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

In general, protein-coding genes are spliced, and most of them are alternatively spliced through a complex balance of the regulatory factors, which directs the splice-site selection. Splicing mechanism's misregulation results in the malfunction of the biological processes and diseases [2]. For example, spinal muscular atrophy is caused by a point mutation in an exonic regulatory element [3–6]. Tauopathies (neurodegenerative disorders characterized by the deposition of abnormal tau protein in the brain [7]), 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 [8–10]. The loss of a splicing factor causes frontotemporal lobar dementia, and aberrant alternative splicing events of several genes and variants influencing pre-mRNA splicing through exon exclusion are observed in patients with Parkinson's Disease [11, 12]. Furthermore, specific alterations in the expression of splicing factors are also observed in various cancers. Mutations of tumor suppressor genes affecting splice site selection are shown to cause cancer, as in the case of the BRCA1 gene for breast cancer [13–15].

The currently accepted molecular mechanism of splicing suggests that introns are excised away from the conserved sequences at 5' and 3' ends of the introns called splice sites. The excised intronic RNA sequence commonly begins with the dinucleotide GU at the 5' site and ends with AG at the 3' site. These are also known as donor and acceptor dinucleotides, respectively. These consensus sequences are essential for splicing, and changes in these sequences result in inhibition of splicing [16]. Another critical region for splicing resides at the branchpoint, which is located between 19 to 37 nucleotides upstream of the acceptor site [17] and contains the nucleotide (always adenine), which forms the intermediate intron lariat with the 5' splice site [18]. A recent genome-wide study on human splicing branchpoints has identified a set of 5 to 6 nucleotides long motifs around the branchpoint [17]. These

branchpoint 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), and GUnAn [17].

1.1. Motivation

Studies of splicing signals and motifs are restricted to the sequences around the 3' and 5' splice sites and branchpoints [17, 19, 20]. Genome-wide branchpoint studies [17, 19] concentrate mainly on intronic regions located up to 100 and 250 nucleotides upstream of 3' respectively for discovering novel branch site motifs by utilizing a single motif finder, namely MEME [21]. Other recent genome-wide splicing studies mostly concentrate on the mechanisms of the aberrant splicing events [22, 23] or analysis of specific gene families or motifs [24, 25]. Furthermore, there are studies on developing computational models for splice motif prediction and splice-altering variant discovery [26, 27]. Nevertheless, there is a lack of computational studies for genome-wide discovery and prioritization of the conserved motifs with a comprehensive approach using a combination of motif finders around the splice regions.

We have performed a genome-wide motif discovery analysis for the sequences around the splice acceptor sites with this motivation. We retrieved 400 nucleotides long DNA sequences around 207,583 3' splice sites (300 nucleotides upstream and 100 nucleotides downstream) and searched for the conserved patterns. Based on an extensive analysis through several motif finders (MEME [21], DREME [28], Centrimo [29], FIMO [30], and Tomtom [31]) and an experimental database (FootprintDB [32], ATtRACT [33]) search, we have identified various patterns, some previously identified as human DNA binding motifs and some others complying with RBP binding motifs. We have also identified several motifs that locate in the acceptor region for the first time in the human genome but previously observed on other mammalian genomes.

Furthermore, we have found novel motifs, most positioned up to 50 nucleotides around the acceptor site, and few co-localize on branchpoints indicating a potential role during splicing. We investigated their co-localization with the experimentally validated splice altering and/or disease-causing variants to validate the motifs' biological significance. Our results showed that the known motifs and the novel ones overlap with such validated variants of the acceptor regions. These findings strongly indicate that computationally significant motifs found here are biologically significant for mammalian genomes. We also proposed a prioritization scheme for the significant motifs by using co-located SNPs' splice altering effects. Besides known motifs, several of the novel motifs are prioritized at top ranks. Therefore, we believe that further analysis of the proposed motifs by experimental methods will further increase the understanding of splicing mechanisms.

1.2. Thesis Organization

This thesis consists of six chapters. In Chapter 1, we present the primary motivation of this study. We briefly describe the importance of pre-mRNA splicing for the organisms and biological signals and motifs, which play active roles in this process. Our proposed genome-wide motif analysis for splice acceptor regions concentrate on discovering significantly enriched novel motifs with no known splicing function.

Chapter 2 presents the theoretical background of the biological and bioinformatics concepts addressed in this thesis. First, an overview of splicing is provided. Afterwards, splicing motifs and genomic variations are explained. Finally, biological tools employed for this thesis are summarized.

In Chapter 3, motif analysis workflow and the methods used in each step are explained. This section provides various snapshots in order to describe the motif analysis process concisely.

In Chapter 4, results achieved through the motif analysis for splice acceptor regions are presented. Our results show that our analysis was able to detect several novel motifs besides the previously known ones. Furthermore, half of the known motifs belong to other mammalian genomes and therefore novel for the human genome.

In Chapter 5, discussions on this thesis are presented. The discussion on motifs' biological significance is presented. Furthermore, motif analyzers and their limitations are briefly described.

In Chapter 6, conclusive remarks of the presented genome-wide splice acceptor region motif analysis and possible future improvements are discussed.

CHAPTER 2

BACKGROUND

Bioinformatics is a computer science area with a primary focus on the collection, organization, and analysis of DNA and protein sequences. Various bioinformatics tools pave the way for genomic data analysis and increase our understanding of biological processes at the DNA and protein level. With this sense, this section briefly describes the biological background of the pre-mRNA splicing process and then describes bioinformatics tools employed for genome-wide motif analysis.

2.1. Central Dogma Of Molecular Biology

The Central Dogma, which Francis Crick first proposed in 1958, describes the flow of information from DNA to RNA and RNA to protein [34]. As shown in Figure 1, this dogma is a kind of framework defining the information transfer from nucleic acids to proteins. The first step in this framework is the process of copying from DNA to DNA which is called replication. This is followed by the process named transcription, in which a particular segment of DNA is copied to messenger RNA (mRNA). Finally, the resulting mRNA of the transcription process is translated into a protein molecule.

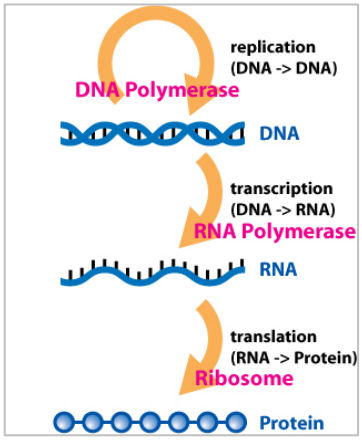


Figure 1: An overview of the (basic) central dogma of molecular biochemistry with all enzymes labeled (Credit: Creative Commons License)

2.2. Overview Of RNA Splicing

This section provides a brief overview of the general splicing mechanism and alternative splicing with its principles.

2.2.1. General Splicing Mechanism.

Protein molecules carry out all functions of life encoded through genes in a genome. Genes carry the codes of the responding protein through transcription, followed by translation [35]. Either during or immediately after transcription (DNA → RNA) in the nucleus of eukaryotic cells, genes are transcribed into pre-messenger RNAs (pre-mRNA) from which noncoding, intervening sequences, i.e., introns, are excised and the remaining coding sequences, i.e., exons, are joined for generating the mature mRNA [36]. This process, which is called splicing, allows different mRNAs and thereby different proteins to be produced from a single pre-mRNA by regulating the inclusion of exons in the mature transcript (alternative splicing).

Newly synthesized pre-mRNA in the eukaryotic nucleus is recognized by the splicing machinery, including small nuclear ribonucleoproteins, i.e., snRNPs. These snRNPs systematically interact with the pre-mRNA and trigger the basal machinery of splicing, as shown in Figure 2, which is directly adapted from the article "The splice of life: Alternative splicing and neurological disease" [1].

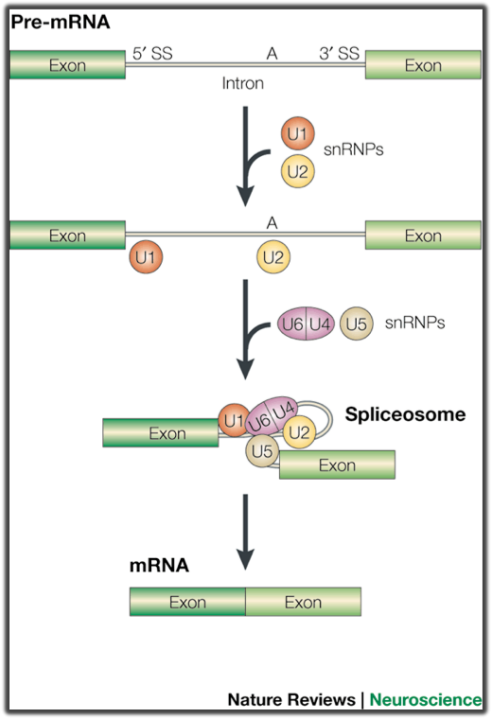


Figure 2: Splicing Machinery

Splicing is initiated by U1 snRNP recognizing and binding to the 5′ splice site. Afterward, U2 snRNP binds to the branchpoint, including an adenine nucleotide

involved in lariat formation. Next, U4/U6 and U5 are recruited to assemble the spliceosome, which enters its catalytically active conformation. After some transesterification reactions, spliceosomal components and the intron lariat detach from the ligated exons to form the mRNA [37].

2.2.2. Alternative Splicing.

The key element in eukaryotic gene expression, which increases its coding capacity, is alternative splicing. It is a regulated complex process; in which particular exons may be included in or excluded from the final produced messenger RNA [2].

There are five types of alternative splicing, which are summarized below:

- ***Exon Skipping/Inclusion (Cassette Exon):*** This is the most common type of alternative splicing where a particular exon is skipped or included in mRNAs under specific conditions or in particular tissues [38, 39] (Figure 3).

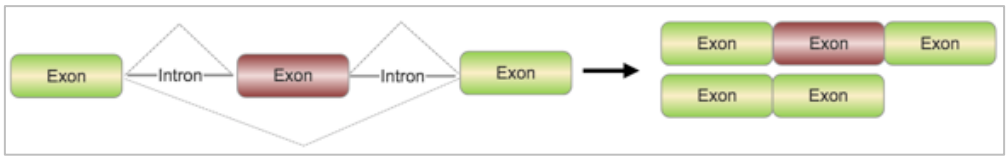


Figure 3: Exon Skipping/Inclusion (Figure is adapted from [40])

- ***Mutually Exclusive Exons:*** mRNAs are formed using only one of the two exons (Figure 4).

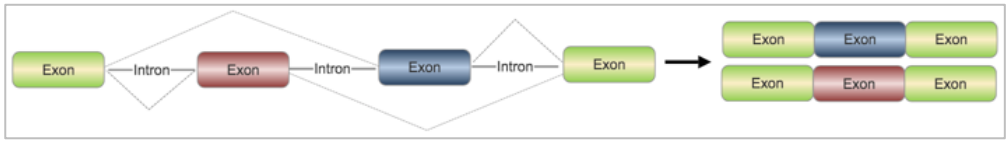


Figure 4: Mutually Exclusive Exons. Red and blue boxes indicate mutually exclusive alternative exons (Figure is adapted from [40]).

- ***Alternative 3' Splice Site:*** A frequently observed type of alternative splicing occurring at alternative 3' splice junction [39] (Figure 5).

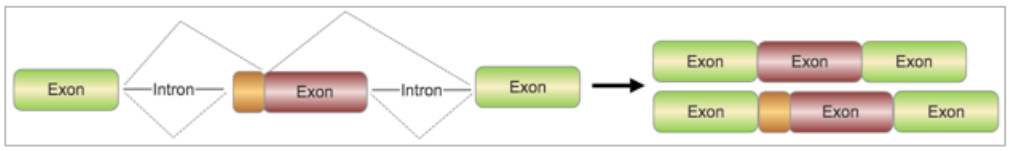


Figure 5: Alternative 3' Splice Site (Figure is adapted from [40]).

- ***Alternative 5' Splice Site:*** Another frequent type of alternative splicing where an alternative 5' splice junction is used [39] (Figure 6).

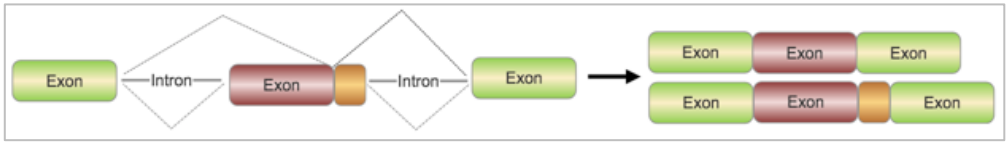


Figure 6: Alternative 5' Splice Site (Figure is adapted from [40]).

- ***Intron Retention:*** Rarest type of alternative splicing where a sequence is spliced as an intron or retained in the mRNA [39] (Figure 7).

Figure 7: Intron Retention (Figure is adapted from [40]).

Furthermore, there are several principles of alternative splicing as briefly described below:

- ***Alternative pre-mRNA splicing regulates the function of protein-coding genes:*** The most recent estimate by high-throughput RNA sequencing is that more than 90% of multi-exon genes of the human genome are alternatively spliced, which reveals the extent to which alternative splicing affects the regulatory and functional complexity of the genome [41]. Genomic signals regulate whether a part of a pre-mRNA will be removed as an intron or included in the mature mRNA as an alternative exon. The preprocess of mRNA is a crucial regulator of gene expression since numerous transcripts from a single protein-coding gene are generated. The transcripts generated through alternative splicing build the protein isoforms with various biological properties such as protein-protein interaction, subcellular localization, and catalytic activity [2].

- ***Changes in alternative splicing may be the cause or consequence of diseases:*** Since alternative splicing is a process affecting numerous genes, it is not surprising that the changes that occurred during this process produce harmful effects on organisms. The number of diseases reported to be associated with alternative splicing deficiencies increases each day [1, 2].

- ***Cis-regulatory elements and trans-acting factors regulate alternative exons:*** The ribonuclear proteins, namely trans-acting factors and cis-regulatory elements, carry splicing signals, enhancers, and silencers, and have an essential role in the regulation of exon recognition. The most essential cis splicing signals are 3' splice or acceptor site and 5' splice or donor site. Branch site and polypyrimidine tract, which lie upstream of the 3' splice site, are also important. The spliceosome recognizes these core cis splicing signals. In addition to the core signals, other auxiliary splicing elements can promote or inhibit splicing activity. The promoters are exonic splicing enhancers (ESEs) and intronic splicing enhancers (ISEs), whereas inhibitors are exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs). Trans-acting splicing regulators recognize these auxiliary elements to determine either an exon will be selected by the spliceosome or not.

For example, SR proteins promote exon selection by binding ESEs, while hnRNP proteins bind to ESSs and inhibit exon selection [2].

- ***Point mutations located outside the splice sites may change the usage of alternative exons:*** Numerous studies have shown that synonymous mutations in coding regions can influence splice site selection and usage of alternative exons. There is now strong evidence that single nucleotide polymorphisms (SNPs) and intronic mutations can also affect exon usage. However, because of the splicing mechanism's complexity, it is difficult to predict the effect of mutations unless they lead to diseases [2].

2.3. Splicing Motifs

Sequence motifs are short and repetitive DNA patterns that have biological functions. They generally indicate sequence-specific protein binding sites, such as transcription factor and nuclease binding sites. Furthermore, RNA motifs play essential roles in biological processes such as ribosome binding, mRNA processing (e.g., splicing, polyadenylation, etc.), and transcription termination [42].

Transcription factors (TFs) are proteins that have roles in the DNA to RNA transcription. TFs have DNA-binding domains that bind to specific sequences of DNA, namely promoters or enhancers. TFs binding to DNA promoter sequences at the transcription start site catalyze the formation of transcription initiation complex whereas, TFs binding to the regulatory sequences such as enhancers and silencers stimulate or repress the genes' transcription. Gene control is commonly handled by transcription regulation, and therefore, TFs action to allow unique expression of genes in different cell types [43]. Until the early 2010s, TFs were known to have indirect effects on splicing in terms of their influence on RNA Polymerase II elongation rates. However, according to recent studies, they also directly affect pre-mRNA splicing through pre-mRNA marking [44].

Prominent motifs in action for pre-mRNA splicing are the dinucleotides at the upstream (5' splice site) and downstream (3' splice), branchpoint motifs (also known as branchpoint B-boxes), and the polypyrimidine tract at the 3' end of the intron as shown in Figure 8:

- The splice donor site includes a GU at the 5' end of the intron.
- The splice acceptor site at the 3' end of the intron terminates the intron with an AG sequence.
- The AG's upstream region mainly composed of pyrimidines (C and U), called the polypyrimidine tract.
- Upstream from the polypyrimidine tract is the branchpoint, including an adenine nucleotide involved in lariat formation. Branchpoint adenine nucleotide is surrounded by other nucleotides, mostly as CUnAn, UUnAn, and GUnAn, called branchpoint B-boxes [17].

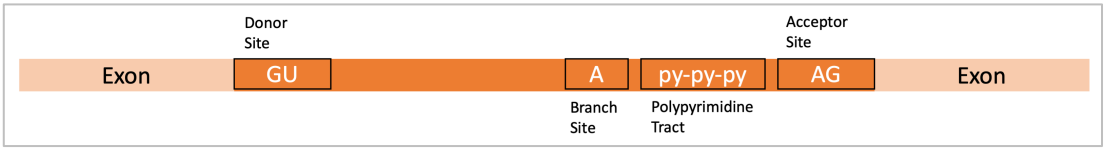


Figure 8: pre-mRNA Splicing Motifs

2.4. Single Nucleotide Polymorphisms (SNPs) and Other Genomic Variations

Genomic sequence variations are the genetic differences within or among populations. Considering the human genome, we can find a 99.9% similarity between individuals at the genomic level. The remaining 0.1% difference, i.e., genomic variation, is responsible for the differences. Such variations may vary in size and also in their impact on our bodies.

Types of variations can be grouped into three, namely, single base-pair substitutions, indels, and structural variations:

- Single base-pair substitutions are also known as single nucleotide polymorphisms (SNPs) and can occur as transition and transversion. Transition is the interchange of purines (Adenine/Guanine) or pyrimidines (Cytosine/Thymine), whereas transversion is the interchange of a purine with a pyrimidine. SNPs are the most common type of sequence variations accounting for 90% of variants [45].
- Indels are short polymorphisms corresponding to the addition or removal of a small number of bases in a DNA sequence. Indels are very common, although not as common as SNPs [46].
- Structural variations are the variations in the structure of an organism's genome. They are generally defined as changes in copy number (insertions, deletions, duplications), orientation (inversions), or chromosomal location (translocations) [47]. Insertions denote novel sequences compared to a reference genome, including mobile element insertions and DNA segments in a variable number of copies. There is a loss of genetic material in the deletions compared with the reference, and duplication of a sequence means that it is inserted contiguously in the genome. Inversions are defined as DNA segments in reverse orientation from the rest of the chromosome, and translocations are DNA segments that change their position without gain or loss of genetic material [47].

2.5. Effects Of SNPs On Splicing

SNP is the simplest form of DNA variation, which occurs throughout the genome approximately once every 100 to 300 bases. They can be grouped as synonymous and

nonsynonymous: there is no effect on the amino acid in the former group, whereas in the latter case, encoded amino acid is changed. SNPs may further influence promoter activity, mRNA conformation, and subcellular localization, and hence may cause various diseases [48].

According to a published report, 25% of the SNPs cause silent mutations, 25% lead to missense mutations, and the remaining 50% are found in non-coding regions [49]. Previously, Silent SNPs were believed to have no consequences on the gene function and phenotype, but recent findings have shown that this may not be the case, and such SNPs can affect in vivo protein folding and consequently function [50]. Nonsynonymous SNPs, which alter amino acids, may often produce diseases as expected.

Much interest in SNPs focuses on the ones located in coding regions since they are directly related to protein sequence change. However, many SNPs are located in genomic regions related to splicing events, such as canonical splice sites, exonic and intronic splice enhancers and silencers, and other DNA motifs. These kinds of SNPs may lead to the disruption of splicing machinery [51].

Large scale surveys on SNP-associated alternative splicing through the analysis of dbSNP and dbEST databases have shown that many alternative splicing events reported in human genes might be attributed to polymorphisms. One of the most exciting findings of splice-modifying SNPs is that the same SNP may affect the alternative splicing in one tissue but have no effect on the others, confirming the tissue-specific and tissue-independent nature of alternative splicing regulation.

SNPs causing changes in alternative splicing patterns may be deleterious and result in diseases. For example, common SNPs altering the alternative splicing of IRF5, OAS1, CTLA4, and CD45 genes were linked to several autoimmune diseases. Polymorphisms affecting the alternative splicing of the NPSR1 gene were found to implicate asthma. A SNP found in LDLR and causing exon skipping was strongly associated with pre-menopausal women's cholesterol levels, and there are many other examples [51].

The majority of disease-related splicing mutations are found to affect cis splicing signals. Nevertheless, such mutations can also act in trans-acting factors by disrupting RNA binding activity or splicing regulators' expression. While cis mutations affect a single transcript's splicing, transmutations' effects are more deleterious because such mutations can compromise the regulation of many splicing targets simultaneously [41].

2.6. Overview Of Genome Sequence Analysis Tools

In this section, we provide an overview of the tools used in this study. Initially, we collected genomic sequences stored in genomic sequence databases via genome browsers. Afterward, we utilized sequence motif finders to identify the motifs on the collected sequences and searched for the discovered motifs from motif databases

serving experimentally validated motifs. We also made some functional enrichment and repetitive sequence analysis for biological significance. Furthermore, we collected motif co-locating variants from variant databases and used splice effect predictors to find the splice altering effects of such variants. All these steps are explained in detail in the "Materials and Methods" section.

2.6.1. Genomic Sequence Databases.

The International Nucleotide Sequence Database Collaboration is an effort to collect DNA and RNA sequence databases. It primarily involves three databases European Nucleotide Archive, GenBank, and DNA Data Bank of Japan. These repositories contain sequence data from all organisms and synchronized daily with contributed sequences [52].

European Nucleotide Archive (ENA), which is composed of three primary databases, namely, Sequence Read Archive (SRA), Trace Archive, and EMBL Nucleotide Sequence Database (EMBL-Bank), is an archive consisting of DNA and RNA sequences [53]. SRA is the public repository of next-generation sequencing data, Trace Archive is the sequence reads catalog with quality information, and EMBL-Bank is the nucleotide sequences database of The European Bioinformatics Institute (EBI) [53].

GenBank is the publicly available genetic sequence database of the National Institutes of Health (NIH). It can be accessed through The National Center for Biotechnology Information (NCBI) Entrez retrieval system. As a very comprehensive database, it contains 260,000 species' nucleotide sequences submitted from individual laboratories and large-scale sequencing projects [54].

DNA Data Bank of Japan (DDBJ), which is developed by the Center for Information Biology and the National Institute of Genetics of Japan, is the sequence archive database of Asia. The database mainly consists of the original DNA data submitted by Japanese researchers. It also stores the sequence data submitted from China, Korea, Taiwan, and other Asian countries. Annotated sequences are publicly available [55].

2.6.2. Genome Browsers.

Genome browsers are the graphical interfaces displaying the genomic data information with genome coordinates. They also provide annotations and analyses of the genes, such as genes' frequencies and expression profiles [56]. Two genome browsers, namely UCSC Genome Browser [57] and Ensembl Genome Browser [58], are the most common ones.

The UCSC Genome Browser, hosted by the University of California, Santa Cruz, is an interactive online tool providing access to various species' genome sequences. Sequence data is presented with its extensive collection of aligned annotations and can be downloaded. This open-source browser is built on a MySQL database that provides rapid access and also easy querying mechanisms at many levels [57, 59]. Furthermore,

The UCSC Genome Browser provides various tools for the genome researchers' use such as Table Browser (for downloading data from the Genome Browser database) [60], BLAT (for rapid sequence alignment to the genome) [61], LiftOver (for converting genome coordinates between assemblies) [62], and UCSC DAS Server (for providing access to genome annotation data for all current assemblies featured in the Genome Browser) [63].

Ensembl Genome Browser provides a browser for vertebrate genomes with its main aim to support research in comparative genomics, evolution, sequence variation, and transcriptional regulation. Tools provided by Ensembl include BLAST/BLAT (for searching genomes for the provided DNA or protein sequences) [64], Biomart (for exporting custom datasets from Ensembl) [65], and the variant effect predictor (VEP, for analyzing the provided variants and predicting the functional consequences of known and unknown variants) [66].

2.6.3. Sequence Motif Finders.

Motif discovery is one of the primary steps in studying gene functions, and there are various algorithms in literature for designing motif finding tools. Recent studies classify such motif discovery algorithms into three main groups: enumerative algorithms, probabilistic algorithms, and nature-inspired algorithms [67].

Enumerative approaches mainly concentrate on counting and comparing nucleotide frequencies for all possible motifs based on a motif model description. They do an exhaustive search for the motifs which match the model description, and therefore this technique guarantees the global optimality. As it is exhaustively searching for the motif model, it requires a long time for processing and can be used for the discovery of shorter motifs [67].

In the probabilistic approach, a probability model called Position-specific Scoring Matrix (PSSM) is created. This matrix denotes the distribution of each nucleotide at each index position of the motif. The motif and non-motif sequences are distinguished from each other by using a specified set of search parameters [67].

There are also some motif finding algorithms inspired by nature. They simulate the behavior of bees, insects, and other kinds of animals for problem-solving. Widespread nature-inspired algorithms used for motif discovery are Genetic Algorithm [68] and Particle Swarm Optimization [69].

Some combinatorial approaches for motif discovery use a mixture of algorithms from different groups. In this study, we used such an approach by utilizing MEME-ChIP [70]. This tool initially executes two different motif finding algorithms, DREME [28] and MEME [21, 71], to identify novel sequence motifs. Afterward, it runs CentriMo [29] for detecting the enrichments in particular locations of the provided sequences. Then, FIMO searches the sequences for provided motifs to provide the locations of matches [30], and Tomtom compares the motifs against a database of known motifs [31]. We briefly describe all these tools below.

- ***MEME (Multiple EM for Motif Elicitation):*** This tool discovers the motifs in the provided DNA, RNA or protein sequences. It represents the motifs as position-specific probability matrices (PSPM) which shows the probability of bases, A,C,G, and T, at each position (e.g., Figure 9).

	A	C	G	T
1	0.600675	0.000000	0.399325	0.000000
2	0.000000	0.000000	1.000000	0.000000
3	1.000000	0.000000	0.000000	0.000000
4	1.000000	0.000000	0.000000	0.000000
5	1.000000	0.000000	0.000000	0.000000

Figure 9: Example PSSM and the corresponding motif logo.

MEME does not identify gapped motifs. Instead, it discovers such gapped patterns as two or more motifs. From a biological perspective, the algorithm discovers shared motifs in a set of provided sequences, whereas from the computational perspective, it finds a set of approximately matching, non-overlapping substrings found in the provided set of strings. This algorithm uses the Expectation-Maximization algorithm (EM). It searches for motifs in a four-step process. The first step is searching for words, which are likely to be good starting points for EM. In the second step, EM is run to convergence from each starting point identified in the first step. After convergence, there is a refinement step where MEME uses an objective function (e.g., classical enrichment which scores motifs based on the e-value, which is an estimate of the log-likelihood ratio of the motif to be generated randomly according to the background model) for refining the width and depth (i.e., number of sites) of the motifs. Finally, MEME performs a motif erasing by erasing the predicted sites for the last motif found to enable EM to find additional motifs. These four steps are repeated until a stopping criterion, such as the limit for the number of motifs to find, is reached [21, 71, 72].

- ***DREME (Discriminative Regular Expression Motif Elicitation):*** This tool discovers short regular expression motifs that are enriched in the provided set of sequences. This algorithm uses positive and control sequences. If the control sequences are not provided, DREME produces a set using the provided set of positive sequences. DREME's workflow can be explained with the following steps [73]:

    1. If control sequences are not provided, positive sequences are shuffled to create control sequences.
    2. All unique subsequences having lengths between 3 and 8 are found in the positive sequences.
    3. For each unique subsequence;
        a. The number of sequences that the subsequence observed is found for both positive and control sequences.

b. Fisher's exact test is used for computing the significance.
c. A subsequence is added to a sorted list of regular expression motifs (sorted by p-value).

4. The top 100 motifs of the sorted lists are repeatedly picked for generalization. One position is replaced with each possible ambiguity code, and the p-value is estimated with that code.
5. For each of the top 100 picked and generalized motifs;
    a. The number of sequences that the motif occurs in is found for both positive and control sequences.
    b. Fisher's exact test is used for computing the significance.
6. Motifs with lower p-values (less than the defined threshold) are picked and positive sequences are scanned for all matching sites for forming the resulting frequency matrices.

- ***CentriMo (Centrality Of Motifs):*** This tool identifies the motifs showing a significant preference for particular locations in the provided sequences. Given a set of motifs and a set of equal-length sequences, it produces a plot for reporting the positional distribution of each motif's best matches (Figure 10). With an assumption of union prior to best match positions, CentriMo uses a binomial test to compute the motifs' significance where the best match occurs on the provided set of sequences [74].

  CentriMo algorithm proceeds in the following three steps [74]:

  1. For each motif, each sequence is scanned, and the best site is recorded.
  2. A bin is assumed around the best sites, and each time the bin size is increased by 2 to keep it symmetric, and the binomial p-value is computed.
  3. Finally, the bin size with the lowest p-value is selected, and the site probability for the location of best sites' distribution is plotted.

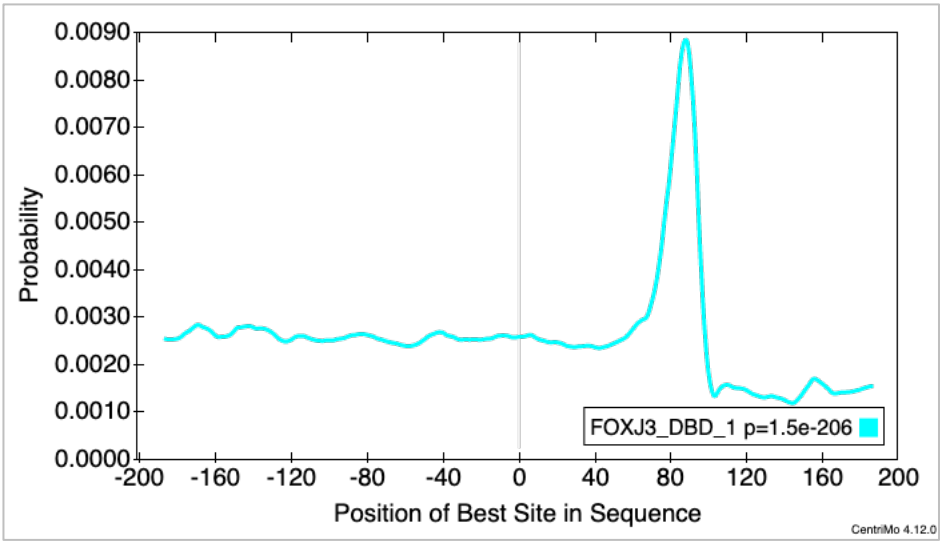


Figure 10: Sample CentriMo plot for motif FOXJ_3_DBD_1 on provided chromosome 1 splice acceptor region sequences.

- ***FIMO (Find Individual Motif Occurrences):*** This tool scans the provided set of sequences for individual matches of the provided motifs. The algorithm firstly converts the input motifs into log-odds position-specific scoring matrices (PSSM) and then scans each input sequence for each PSSM independently. Afterward, it reports the positions in each sequence that have a statistically significant match with a motif. The significance value, i.e., p-value, is a configurable parameter for the algorithm [75].

- ***Tomtom (Motif-Motif Similarity):*** This tool searches one or more query motifs against one or more databases of target motifs (e.g., JASPAR). Given a query motif, Tomtom searches the databases according to a statistical measure of motif-motif similarity. That measure is nothing but the definition of a motif similarity function, which takes all possible offsets and relative motif orientations into account [31, 76].

2.6.4. Motif Databases.

There are various motif databases, which are the comprehensive repositories for motifs identified through experimental studies. To use various databases through a single access point, we used FootprintDB [32] in this study. FootprintDB integrates 18 freely available DNA binding site libraries, and it also makes systematic annotations for the binding sites of the corresponding TFs [77]. Table 1 summarizes the databases served by FootprintDB.

Table 1: Motif databases included in FootprintDB

#	Database	Description
1	JASPAR	Open Access database of TFs, DNA binding sites, and DNA binding motifs across multiple species
2	CISBP	TFs and their DNA binding motifs
3	3D-footprint	DNA-binding protein structures database
4	HT-SELEX2	Database of human TFs SELEX motifs
5	UniPROBE	Data repository of universal protein binding microarray experiments
6	HOCOMOCO	Homo sapiens comprehensive model collection of hand-curated transcription factor binding site (TFBS) models constructed through the integration of binding sequences found via low and high-throughput methods
7	AthalianaCistrome	Genomic DNA motifs of in-vitro-expressed TFs for Arabidopsis thaliana
8	HumanTF	Human TFs' (identified via high-throughput SELEX and ChIP sequencing) sequence-specific binding preferences
9	HumanTF2	Sequence-specific binding preferences of cooperative-binding human TF pairs identified by CAP-SELEX
10	AthaMYB	Repository of the arabidopsis R2R3-MYB TF family's DNA-binding activities
11	FlyZincFinger	Drosophila Cys2-His2 zinc finger proteins' analyses repository
12	SMILE-seq	Repository of validated DNA-binding data obtained via SMiLEseq technology from a set of human, mouse and Drosophila TFs
13	ArabidopsisPBM	Repository of 63 plant TFs from 25 families
14	Athamap	Arabidopsis thaliana's genome-wide map of potential TFBSs
15	DBTBS	A collection of transcriptional regulation data of Bacillus subtilis with conservation information
16	RegulonDB	Transcriptional regulatory network data repository of Escherichia coli K12
17	DrosophilaTF	Drosophila melanogaster TFs' motif models
18	humanC2H2ZF-ChIP	DNA-binding data for C2H2-ZF domains

There are additional tools which are recently developed, such as oRNAment. It is an RNA binding motifs database. It is defined to be the first database detailing the transcriptome-wide distribution features of putative RBP target motifs across multiple species [78]. However, this tool catalogs the motif instances across the coding and non-coding transcriptomes but excludes intronic regions [78], and therefore we could not make use of this tool to analyze significant motifs identified in this study.

Another motif database is ATtRACT, which is a repository of RBP binding motifs. It provides 1583 experimentally validated RBP binding motifs and provides an online search feature for IUPAC coded sequences. However, this tool does not support searching for more than 12 bp long sequences, as the motifs in the database are at most 12 bps. Although it has a limitation, we used ATtRACT for searching experimentally validated RBP binding motifs since it is the most extensive repository to date.

2.6.5. Functional Enrichment Analysis.

Functional enrichment analysis is the process of identification of the over-represented genes or proteins that may have disease associations. During high-throughput experiments, researchers come up with a set of differentially expressed genes. Functional profiling of such genes paves the way for a better understanding of the actual biological events. This profiling can be handled by comparing the gene ontology (GO) terms with the differentially expressed genes of the experiment by means of a statistical test to detect the enrichment of the ontology terms for the input genes [79].

For this study's functional enrichment analysis purposes, we used The Database for Annotation, Visualization and Integrated Discovery (DAVID). This tool mainly provides annotation and gene-GO term enrichment analysis to identify the most relevant GO terms associated with a given list. DAVID provides a broad set of functional annotation tools which can [80, 81]:

- Isolate biological enrichments, especially the GO terms,
- Identify gene groups that have functional relations,
- Identify redundant group terms,
- Form visual maps of genes on BioCarta & KEGG pathways,
- Display genes-to-terms relations on 2D views,
- Search for functionally related genes not provided in the input gene list,
- Present cooperating proteins,
- Tie gene-disease associations,
- Highlight motifs and protein domains,
- Forward to related sources from literature,
- Transform gene identifiers between different types.

2.6.6. Repetitive Sequence Analysis.

Repetitive sequences are patterns of DNA or RNA that are found in multiple copies in the genome. In many organisms, genomic DNA is highly repetitive. For the human, repetitions occupy two-thirds of the genome [82]. Repetitive elements are mainly grouped into two. The first group is known as tandem repeats which denote the copies lying adjacent to each other, either directly or inverted. The second group of repeats is the interspersed repeats that are not adjacent, as in tandem repeats but dispersed throughout the genome in a nonadjacent manner [83].

This study used RepeatMasker program to search for interspersed repeats and low complexity DNA sequences on the human genomic sequence. RepeatMasker performs sequence comparisons with the cross match program, an efficient implementation of the Smith-Waterman-Gotoh local sequence alignment algorithm [84]. This tool needs the input sequence to be provided in FASTA format. It searches for the provided sequence in its library of repetitive elements from Repbase (a database of repetitive DNA elements) [85] and Dfam (a database of transposable elements and repetitive DNA families) [86].

2.6.7. Variant Databases.

The generation of vast amounts of genomic data led to various databases for storing and sharing that data. In this sense, variant databases are formed to store and share the identified genomic variants and also their effects.

We present a brief description of the variant databases that we used in this study below:

- ***The Single Nucleotide Polymorphism Database (dbSNP)*:** dbSNP is a freely available public archive for single nucleotide variations hosted by the National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI). This database contains SNPs, short indels, and short tandem repeats with their publication, population frequency, molecular consequence, and genomic mapping information [87].

- ***The Database of Genotypes and Phenotypes (dbGaP):*** dbGAP is a National Institutes of Health (NIH)-sponsored repository dedicated to association studies between phenotypes and genotypes. NCBI created this repository to provide a resource presenting individual-level genotype, phenotype, exposure, sequence data, and associations. Its publicly available data can be searched for study metadata, phenotype summary, documentation, and association analyses [88].

- ***The Cancer Genome Atlas (TCGA):*** TCGA is a repository for genetic mutations associated with cancer. This project aims to improve the understanding of the genetic basis of the disease, and it is managed by National Cancer Institute (NCI) and NHGRI. Funded by NIH, it is the first large-scale genomics initiative providing significant resources for bioinformatic studies [89, 90].

- ***GWAS Catalog:*** This is a curated collection of all published genome-wide association studies and their results. It was initially created by NHGRI but then became a collaboration between EBI and NHGRI [91].

- ***PharmGKB:*** This is a publicly available NIH funded repository that manages the aggregation, curation, integration, and dissemination of knowledge regarding the impact of human genetic variation on drug response [92].

- ***ClinVar:*** This is a public data archive that presents relationships between human variations and phenotypes with supporting evidence. It is maintained by NIH at NCBI [93].

- ***Online Mendelian Inheritance in Man (OMIM):*** OMIM is a regularly updated catalog that focuses specifically on the gene-phenotype relationships and presents genetic disorders and human gene traits [94].

2.6.8. Splice Effect Predictors.

Since splice-modifying variants have gained more attention recently, various tools have been developed to detect them. At the time of our analysis, we briefly investigated several popular tools of that time and selected SPANR [95] as the more appropriate tool for our analysis.

- ***MutPred Splice:*** Mutpred Splice is a computational model that uses human disease alleles gathered from the Human Gene Mutation Database (HGMD) to train a classifier, predicting the disruptive effects of coding region variants on pre-mRNA splicing. It takes the genomic coordinates, strand, and substitution for the variants in comma-delimited format or in Variant Call Format (VCF) and outputs a general score indicating the variant's potential to disrupt splicing [96]. MutPred Splice concentrates only on the variants at coding regions, i.e., exonic parts. It is not used for this study as a tool working for both intronic and exonic regions required.

- ***IntSplice:*** IntSplice is a model that predicts the splicing consequences of intronic single-nucleotide variations in the human genome. This tool's primary motivation is to identify the SNPs that affect intronic regulatory elements of splicing, and with this sense, it predicts splice affecting SNVs at intronic positions from −50 to −3 [97]. IntSplice cannot be utilized for this study as it does not provide predictions for the exonic regions.

- ***SPANR:*** SPANR is a deep learning-based technique that scores how strongly genetic variants affect RNA splicing. This tool can analyze synonymous, missense, and nonsense variants up to 100nt into exons, as well as intronic ones that are up to 300nt from splice junctions [95]. So, SPANR is an appropriate tool that can be utilized for this study, analyzing both exonic and intronic regions.

- ***SpliceAI:*** SpliceAI is one of the recent deep learning-based tools to identify splice variants. It annotates the genetic variants with their predictive splicing effect. It

mainly concentrates on the effect of variants in terms of acceptor/donor loss or gain. It makes predictions for the variants at all intronic and exonic locations on the genes [98]. SpliceAI may be an appropriate and extensive tool for our study but it was not yet published by the time of our analysis.

CHAPTER 3

MATERIALS AND METHODS

Our study's main workflow for the genome-wide motif discovery analysis on splice acceptor regions is outlined in Figure 11. First, we completed the genome-scale retrieval of acceptor sequences, accessed between March 2018 to June 2018, and then ran various motif finder algorithms to find patterns on these sequences. Next, we filtered the motifs with low significance during the pre-processing step and evaluated the selected motifs' biological relevance. Afterward, motifs found to be biologically significant are validated with experimental variant data from the literature. Finally, a motif prioritization scheme is developed by employing splice altering effects of co-located SNPs.

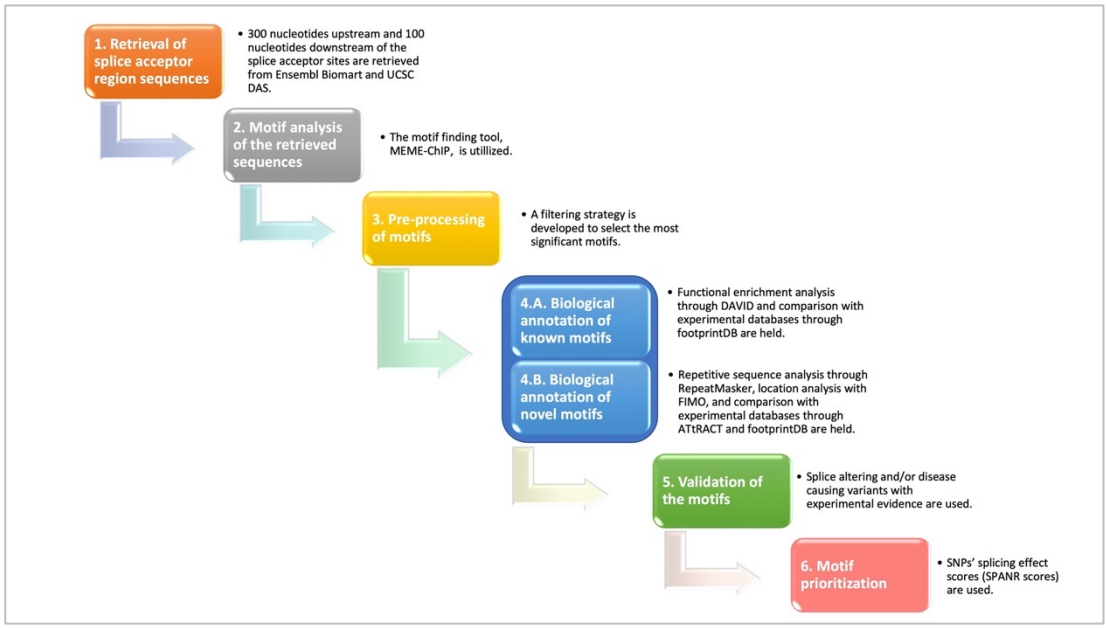


Figure 11: The workflow of the study (1. Retrieval of acceptor region sequences, 2. Motif analysis of the retrieved sequences, 3. Pre-processing of motifs, 4.A. Biological annotation of known motifs, 4.B. Biological annotation of novel motifs, 5. Validation of the motifs 6. Motif prioritization).

3.1. Acceptor Region Sequence Retrieval

This study focuses on the 400 nucleotides around the splice acceptor sites (300 nucleotides upstream and 100 nucleotides downstream of the acceptor site - Figure 12). The exact base pair locations of acceptor sites are retrieved through Ensembl Biomart [65]. 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 transcription start site (TSS) locations to eliminate 5'UTR sequences as our primary consideration is the splice acceptor sites. After gathering the exact acceptor positions by eliminating the exons with initiation codons through Ensembl Biomart, we have calculated the base pair locations of start and endpoints 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 (Appendix A - Figure 1 and Appendix A - Figure 2 present the acceptor site collection query for the chromosome 1 from Ensembl Biomart and the Perl script to retrieve corresponding sequences from UCSC DAS respectively).

Figure 12: 400 nucleotides long region around the splice acceptor site (i.e. 3' splice site).

3.2. Motif Analysis Of 400-Nucleotides Long Acceptor Regions

MEME-ChIP [70] is selected as the motif finding tool for this study. MEME-ChIP runs several motif analyzers, including MEME, DREME, Centrimo, FIMO, and Tomtom. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in the provided sequences using an expectation maximization algorithm extension [21]. DREME discovers short, ungapped motifs (recurring, fixed-length patterns) relatively enriched in the provided sequences. It works with a high speed as it restricts its search to regular expressions based on the alphabet of the provided sequences [28]. CentriMo identifies known or user-provided motifs that show a significant preference for particular locations in the provided sequences. It takes a set of motifs and a set of equal-length sequences and plots the positional distribution of each motif's best match [29]. FIMO searches the sequences for occurrences of provided motifs and provides the locations of matches. A p-value is calculated as a statistical threshold for motifs through a dynamic programming algorithm that converts log-odds scores into p-values, assuming a zero-order background model [30]. Tomtom compares the motifs against a database of known motifs (i.e., JASPAR CORE) [31, 99].

After sequence retrieval, negative strand sequences were reverse complemented by through an R code (Appendix A - Figure 3). Next, we have arranged all of the sequences in the 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), which is the most extensive set for known motifs of eukaryotic DNA. During the DREME and the Centrimo analysis, we have used the default parameters. However, we used a maximum number of 20 motifs for the MEME analysis. We set the motif length between 9 to 50 nucleotides to discover longer motifs not covered in the DREME analysis (Appendix A - Figure 4).

3.3. Pre-Processing Of Identified Motifs

The most significant motifs are selected through the filtering process based on inspecting the central enrichment information provided by Centrimo. Steps of the filtering workflow are as follows:

1. We have filtered the motifs with region width (the width of the most enriched region) less than three nucleotides since they mostly represent the dinucleotide signal, AG, of the acceptor sites.
2. We have filtered the motifs based on the e-values (i.e., the significance of motif enrichments). The cut-off value e^{-5} is used to filter the motifs with lower significance.
3. We have described a new threshold value, T-value, for prioritization of the identified motifs. This T-value is defined as "Region Match Ratio * Power of E-value." In this formula, "Region Match Ratio" indicates the ratio of the sequences whose best match to the motif falls into the Centrimo enriched region of the sequences provided to Centrimo run. The power of the E-value indicates the power of the motif enrichment significance (e.g., it is set as -8 if the E-value is e-8). With this sense, T-value is defined to select the motifs with higher significance but few sequence matches and also the motifs with lower significance but a high number of sequence matches, and set to -10 in this study.
4. We have filtered the insignificant motifs as classified by CentriMo and MEME or CentriMo and DREME.
5. We have removed the duplicated motifs. There were some duplicate motifs such as MA0634.1 and ALX3_full_1, which are represented by different identifiers as they exist in other motif databases (i.e., JASPAR 2018 [99] and HumanTF 1.0 [100] respectively).

3.4. Biological Annotation Of The Known Splice Region Motifs

Biological annotation of the known motifs is made through the following analyses:

- ***Functional Enrichment Analysis:*** The Database for Annotation, Visualization, and Integrated Discovery (DAVID) is a functional annotation tool, which mainly provides typical batch annotation and gene-GO term enrichment analysis to highlight the most relevant GO terms associated with a given list. We used DAVID v6.8 to evaluate the biological functions of transcription factors (TFs) that bind to the significant acceptor region motifs identified in this study [81, 101].

- ***Comparison With Experimental Databases:*** Initial comparison of the discovered motifs are made with Tomtom [31] using JASPAR CORE [99] as the single reference, but for further experimental investigation of the known motifs, FootprintDB [32] is utilized. 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) [32]. As a meta-database, it integrates several databases such as JASPAR [99], HumanTF [100], 3D-footprint [102], UniPROBE [103], 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 [32]. We have used FootprintDB's sequence search service to reveal the known motifs' biological significance by using the count matrix representation, a position frequency matrix (PFM) created by counting the nucleotide occurrences at each position.

3.5. Biological Annotation Of The Novel Splice Region Motifs

Biological annotation of the novel motifs is made through the following analyses:

- ***Analysis Of Repetitive Sequences:*** RepeatMasker, uses curated libraries of repeats, namely Dfam [86] and Repbase [85], to screen DNA sequences for interspersed repeats and low complexity DNA sequences [84]. 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 due to the high number of alternatives (especially for the long motifs), which was over the capacity of the online RepeatMasker. Furthermore, UCSC Blat [59, 61] was used for searching repetitions in case the whole motif is not recognized as a repetitive sequence by RepeatMasker, but it presents significant overlap with repetitive sequences.

- ***Comparison With RNA Binding Protein Associated Motifs:*** To detect any compliance of our novel motifs with RNA binding motifs, we utilized ATtRACT, which is a repository of 370 RNA binding proteins and 1583 experimentally validated RBP binding motifs [33]. As we identified motifs on genomic DNA sequences, we used their complementary RNA sequences for this search [104]. For example, 'ACATWY' is searched with its complementary RNA sequence 'UGUAWR'.

- ***Analysis Of Motif Locations:*** We have investigated the motif locations to detect any particular regions, which are predominantly preferred. First, the acceptor site sequences are separated into eight regions, as shown in Figure 13. Next, we have utilized the output of the FIMO program [30], which provides the exact locations of the motif occurrences in the analyzed sequences, to match the motif locations within the defined regions with $p < 0.01$ (Appendix B - Figure 1, Appendix B - Figure 2). Furthermore, occurrences of the motifs on the experimentally shown branchpoint locations, provided in the supplementary material of the genome-wide study on human splicing branchpoints [17], are analyzed ($p < 0.0001$). For handling this analysis, we loaded FIMO output into our local PostgreSQL database and ran some SQL queries (Figure 14).

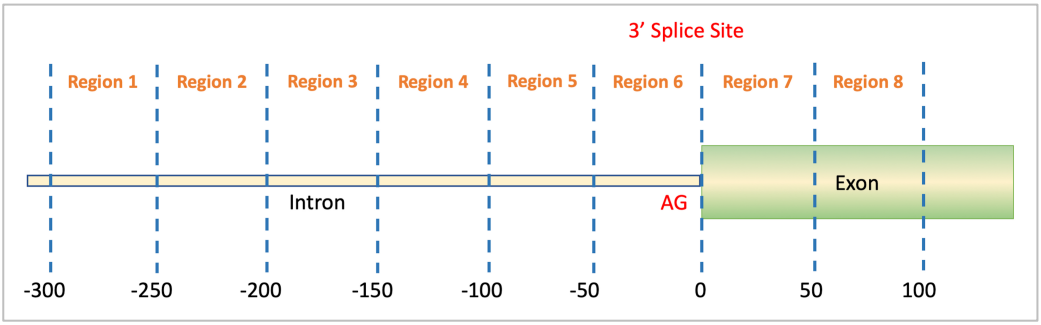


Figure 13: Eight regions are defined for the analyzed sequences. Region 1, Region 2, Region 3, Region 4, Region 5, and Region 6 are located in the intronic region at the upstream side of the acceptor dinucleotide, i.e., AG., Region 7 and Region 8 are located in the exonic region at the downstream side of the acceptor.

a.

	id	motif_id	start	real_start	stop	real_stop	strand	p-value
1	1	MA0528.1	356	52361541	376	52361561	+	0.000000000000595000
2	2	MA0528.1	317	10642286	337	10642306	-	0.000000000002870000
3	3	MA0528.1	9	201874860	29	201874880	+	0.000000000002870000
4	4	MA0528.1	329	201874860	349	201874880	+	0.000000000002870000
5	5	MA0528.1	335	1641766	355	1641786	+	0.000000000002870000
6	6	MA0528.1	335	1708898	355	1708918	+	0.000000000002870000
7	7	MA0528.1	359	52361538	379	52361558	+	0.000000000004020000
8	8	MA0528.1	332	1708901	352	1708921	+	0.000000000007890000
9	9	MA0528.1	332	1641769	352	1641789	+	0.000000000007890000
10	10	MA0528.1	233	20342364	253	20342384	-	0.000000000008290000

b.

	id	chr	chr_start	chr_end
1	670491	chr1	91661	91662
2	670492	chr1	169295	169296
3	670493	chr1	498477	498478
4	670494	chr1	769743	769744
5	670495	chr1	773130	773131
6	670496	chr1	781909	781910
7	670497	chr1	849444	849445
8	670498	chr1	852647	852648
9	670499	chr1	914759	914760
10	670500	chr1	918144	918145

c.

```
-- Retrieve motif occurrences on branch points
SELECT
  bpoints.id,
  motif_id,
  count(*)
FROM
  motif_locations ml,
  genome_wide_branch_point_locations bpoints
WHERE ml.p_value <= 0.0001 AND ml.real_start <= bpoints.chr_start
AND ml.real_stop >= bpoints.chr_end
GROUP BY bpoints.id, motif_id;
```

Figure 14: a. A snapshot from the "motif_locations" table, which stores FIMO output, b. A snapshot from the "genome_wide_branch_point_locations" table, which stores branchpoint locations published in [17], and SQL queries used to get motif occurrences on the branchpoints.

- ***Comparison With Experimental Databases:*** FootprintDB's [32] sequence search facility is utilized for novel motifs in the same manner as known motifs to reveal their biological significance.

3.6. Validation Of The Biological Significance Of The Motifs With Experimental Data

We followed two different approaches for validating the biological significance of the discovered motifs. As the motif sequences identified in this study reside in both intronic and exonic regions, we concentrated on intronic variants causing splicing alterations and intronic/exonic variants causing diseases.

3.6.1. Validation With Experimentally Proven Splice Altering Variants.

GTEx RNA-seq data provides a list of variants with their splicing effect based on gain or loss of acceptor and donor sites. This set is accessible from BaseSpace [105] as it is used for validation purposes of SpliceAI [98]. Here we have utilized this experimentally proven splice-altering variant set. After gathering the variants, we computed the number of co-located splice altering variants with the candidate splice acceptor motifs identified in this study. We should mention that we used the FIMO motif locations in our local PostgreSQL database for detecting the co-locations. Figure 15 presents one of such queries. The other queries differ only in the start and stop positions, which indicate the exact locations of the splice altering variants.

```
/* A query example for getting the motifs co-localized with the splice
altering variant located at chr1: 156134795*/

SELECT DISTINCT (motif_id)
FROM motif_locations
WHERE start <= 156134795 AND stop >= 156134795 AND chr = 'chr1';
```

Figure 15: SQL query for computing the co-location of motifs with splice altering variants. Exact location information of our significant motifs is stored in our local PostgreSQL database's motif_locations, table and variant locations are retrieved from the source of SpliceAI [98].

3.6.2. Validation With Experimentally Proven Disease-Causing Variants.

We used variant data from various databases: dbGaP [88], TCGA [89, 90], GWAS Catalog [91], PharmGKB [92], and ClinVar (only variants with OMIM entries are considered) [93, 94]. After gathering the disease-causing variants, we analyzed the ones which locate on the significant motifs identified in the splice acceptor regions.

- ***Gathering Variant Data From dbGaP:*** The Database of Genotypes and Phenotypes (dbGaP) is a National Institutes of Health (NIH) sponsored repository that archives and publishes information produced by genotype and phenotype interaction studies. Data published by dbGaP is organized as studies that contain all types of data discovered in genetic, clinical, or epidemiological research projects. Publicly available data is provided through an ftp address

(ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/), as shown in Figure 16 [88]. Entering one of the directories, we can see another directory structure, and the analyses directory in this structure provides the actual analysis results in a zipped format (Figure 17).

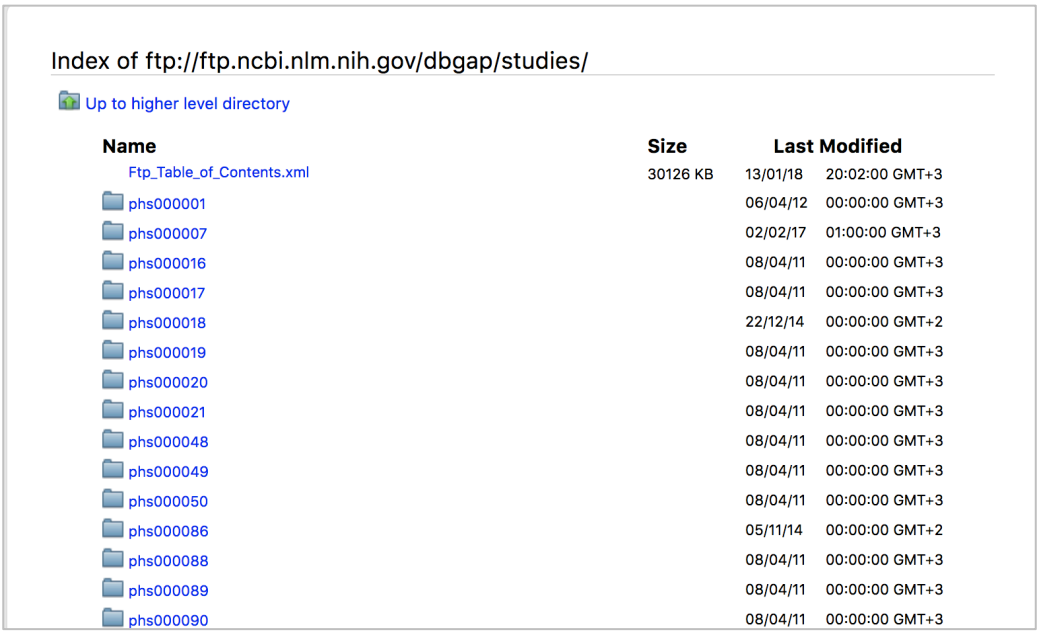


Figure 16: dbGaP public data repository published from ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/.

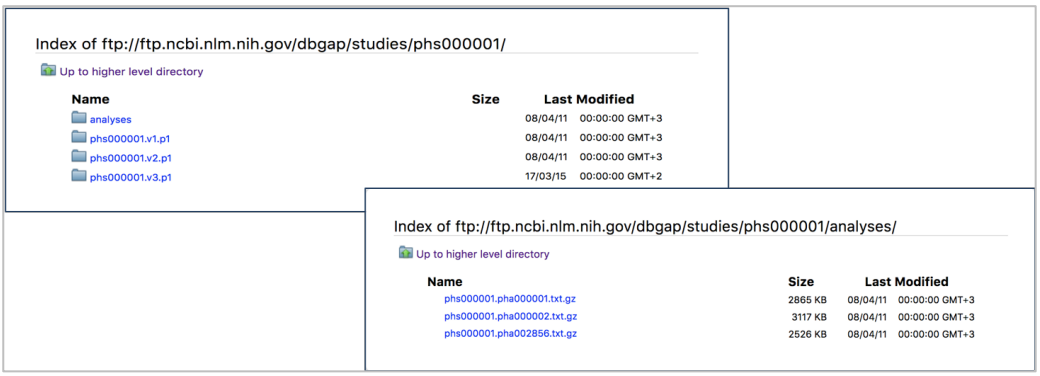


Figure 17: dbGaP public data repository's directory structure.

To download the data provided through dbGAP's ftp site, we have utilized GNU Wget, a free software package for retrieving files using HTTP, HTTPS, FTP, and FTPS that are the most widely-used Internet protocols.

```
wget -r -A txt.gz ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/ -P
/Users/gulsah/dbGAP/
```

After the download from ftp completed, we crawled through the downloaded files to extract the SNPs with higher significance (i.e., p-value <= 0.001 in our case). We created a database table in our local PostgreSQL database (Figure 18) and ran the Python code, which can be reached through the GitHub repository (https://github.com/gkaraduman/dbGAP), to parse the analysis files and insert the

data into the local database. This process resulted in 2,218,480 entries in our database.

```
create table dbgap

(
       rsid text not null,
       p_value numeric not null,
       study_id text not null,
       study_name text,
       study_description text,
       constraint dbgap_rsid_p_value_study_id_pk
              primary key (rsid, p_value, study_id)
);
```

Figure 18: dbGaP table creation SQL code snippet.

- ***Gathering Variant Data From TCGA:*** TCGA is the catalog of genetic mutations that are responsible for cancer [89, 90]. We downloaded the data through dbGaP and loaded it to our local database (Figure 19). We should mention that TCGA data is not publicly available through dbGaP, but we used authorized access to download the related data for this study.

```
create table tcga

(
       hugo_symbol text,
       entrez_gene_id text,
       center text,
       ncbi_build text,
       chromosome text,
       start_position text,
       end_position text,
       strand text,
       variant_classification text,
       variant_type text,
       reference_allele text,
       tumor_seq_allele1 text,
       tumor_seq_allele2 text,
       dbsnp_rs text,
       …
);

create index tcga_dbsnp_rs_idx on tcga (dbsnp_rs);

COPY tcga FROM
'/Users/gulsah/TCGA_dbGAP_Data/TCGA_2017/wgs.356.maf.txt' DELIMITER '
';
```

Figure 19: TCGA table creation and data loading SQL code.

- ***Gathering Variant Data From GWAS Catalog:*** GWAS Catalog is a curated collection of all published genome-wide association studies, produced by a collaboration between EMBL-EBI and NHGRI. A full copy of the database is

downloaded from its "Downloads" page (https://www.ebi.ac.uk/gwas/docs/file-downloads) and loaded to our local database (Figure 20).

```
create table gwas_catalog

(
	"DATE ADDED TO CATALOG" text,
	pubmedid text,
	"FIRST AUTHOR" text,
	date text,
	journal text,
	link text,
	study text,
	"DISEASE/TRAIT" text,
	...
);

copy gwas_catalog from '/Users/gulsah/GWAS Catalog/gwas_catalog_v1.0-
associations_e93_r2018-08-28.tsv' delimiter '  ';
```

Figure 20: GWAS Catalog table creation and data loading SQL code.

- ***Gathering Variant Data From PharmGKB:*** PharmGKB is a publicly available repository for presenting knowledge regarding the impact of human genetic variation on drug response [92, 106]. Publicly available clinical and variant annotations data of PharmGKB is downloaded through its "Downloads" page (https://www.pharmgkb.org/downloads) and loaded to our local database (Figure 21).

```
create table pharmgkb

(
	entity1_id text,
	entity1_name_rsid text,
	entity1_type text,
	entity2_id text,
	entity2_name_disease_name text,
	entity2_type text,
	evidence text,
	association_status text,
	pk text,
	pd text,
	pmids text
);

create index pharmgkb_entity1_name_rsid_index on pharmgkb
(entity1_name_rsid);

copy pharmgkb_old from
'/Users/gulsah/variant_disease_relationships.txt' DELIMITER '	';
```

Figure 21: PharmGKB table creation and data loading SQL code.

- ***Gathering Variant Data From ClinVar:*** ClinVar is a freely accessible, public archive for human variations and phenotypes, with supporting evidence [93], and

OMIM is an online catalog of human genes and genetic disorders [94]. Publicly available variant data of ClinVar is downloaded from its "Downloads/FTP" site (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/), and the variants with OMIM entries are considered for further analysis (Figure 22).

```
create table clinvar

(
       alleleid integer,
       type text,
       name text,
       geneid text,
       genesymbol text,
       hgnc_id text,
       clinicalsignificance text,
       clinsigsimple text,
       lastevaluated text,
       rsid text,
       dbvarid text,
       rcvaccession text,
       phenotypeids text,
       phenotypelist text,
       …
);

create index clinvar_rsid_index on clinvar (rsid);

create index clinvar_rsid_rs_inc_index on clinvar (rsid_rs_inc);

copy clinvar from '/Users/gulsah/clinvar.txt' DELIMITER '   ';
```

Figure 22: ClinVar table creation and data loading SQL code.

- ***Co-locating Disease Related Variants With Significant Motifs:*** Firstly, we downloaded dbSNP Build 150 data from UCSC (Download link: http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp150.txt.gz) and load that data into our local PostgreSQL database (Appendix C - Figure 1). Afterward, we ran some SQL code to produce a SNP subset that contains only the SNPs located 400 nucleotides around the splice acceptor sites (Appendix C - Figure 2). Our next step was finding the SNPs located on the significant motifs by running another piece of SQL code (Appendix C - Figure 3). For the exact location information of motifs, we used FIMO [30] output of MEME-ChIP [70] and set the p-value threshold to 10^{-4} for achieving significant results. Finally, we ran another set of SQLs to tag the motif co-located SNPs in terms of their disease relevance (Appendix C - Figure 4).

3.7. Motif Scoring With Splice Variant SNP Data

Our approach for motif scoring uses the splice altering potential of co-located SNPs. We present this approach in detail in the following subsections.

3.7.1. Collection Of SNPs On Splice Acceptor Region Sequences. We used dbSNP v150 data [87] gathered from the UCSC downloads page [107] and then identified the SNPs co-located with the significant motifs employing SQL queries (Figure 23).

```
INSERT INTO acceptor_site_snp_on_motif (snp_id, motif_id)
 (SELECT
   snps.rsid,
   'UP00128A_1'
  FROM motif_locations ml,
   acceptor_site_all_snps snps
  WHERE
   ml.start <= snps.chrstart AND
   ml.stop >= snps.chrend AND
   motif_id = 'UP00128A_1');
```

Figure 23: Sample SQL script for calculating SNPs co-location on significant motifs. "acceptor_site_snp_on_motif" table stores the SNP-motif pairs for co-locations. The query snippet shown in this figure is run in the same manner for all motifs.

3.7.2. SNPs' Splicing Effect Prediction With SPANR. SPANR is a computational technique that scores how strongly genetic variants affect RNA splicing. This tool can analyze synonymous, missense, and nonsense exonic variants, as well as intronic ones that are up to 300nt from splice junctions accurately (with up to 94% accuracy when clinical variants in spinal muscular atrophy and colorectal cancer genes are scored for validation purposes) [95]. Here, we gathered the SPANR scores through SPIDEX [108, 109], a pre-computed database of genome-wide SPANR scores of the splicing variants. After downloading SPIDEX data, we stored it into our local database and computed our SNPs' SPANR scores by running some SQL queries (Appendix D). We should mention that SPIDEX provides the scores according to the GRCh37 assembly. As the SNP locations were collected according to GRCh38 assembly in the previous steps, we did a locational mapping by using the LiftOver tool provided by UCSC [62] to make both data compatible.

3.7.3. Motif Prioritization. For each significant acceptor region motif, we have investigated the co-locating splice altering variants. We first computed the number of SNPs residing on that motif with |SPANR score| >= 5, where five is accepted to be splice disrupting in the corresponding study [95]. Afterward, we computed motifs' splicing effect score by considering the splice altering SNPs' frequency and how many times the corresponding motif is observed on the human genome by using the following formula:

$$\frac{\#\ SNPs\ With\ |SPANR\ Score|\ >\ =\ 5}{\#\ Motif\ Occurrences}$$

We are scoring the motifs by calculating the number of splice altering SNPs per motif occurrence. With this approach, we are providing higher scores to the motifs with high splice altering SNP density.

CHAPTER 4

RESULTS

This section briefly presents the results achieved through our genome-wide sequence analysis around splice acceptor sites. Results for known and novel motifs are provided in separate sections as we have applied further analysis for novel motifs.

4.1. Significant Motifs

MEME-ChIP analysis of 207,583 splice acceptor region sequences revealed 1491 significant Centrimo motifs with e-value ≤ 10 using the binomial test. After applying the filtering rules as defined in the pre-processing step, 517 motifs were selected for further investigation. Among these 517 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 27 were novel motifs found through DREME. All of these known and novel motifs are listed in Appendix E.

4.2. 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", did not have any footprintDB entry, although there was a link provided by MEME-ChIP. This finding may be due to some versioning inconsistencies between MEME-ChIP and footprintDB.

We list our common findings of known motifs below:

- ***According to previous experimental studies in FootprintDB, 51% of the known motifs were found in the human genome, while the remaining 49% found in other mammalian genomes.***

  We investigated the origin of the 455 motifs through footprintDB, and showed that 225 motifs were known motifs for non-human mammalian genomes. The

remaining 221 motifs were human genome motifs, and 9 of the motifs were observed in both Homo sapiens and non-human mammalian genomes (Figure 24).

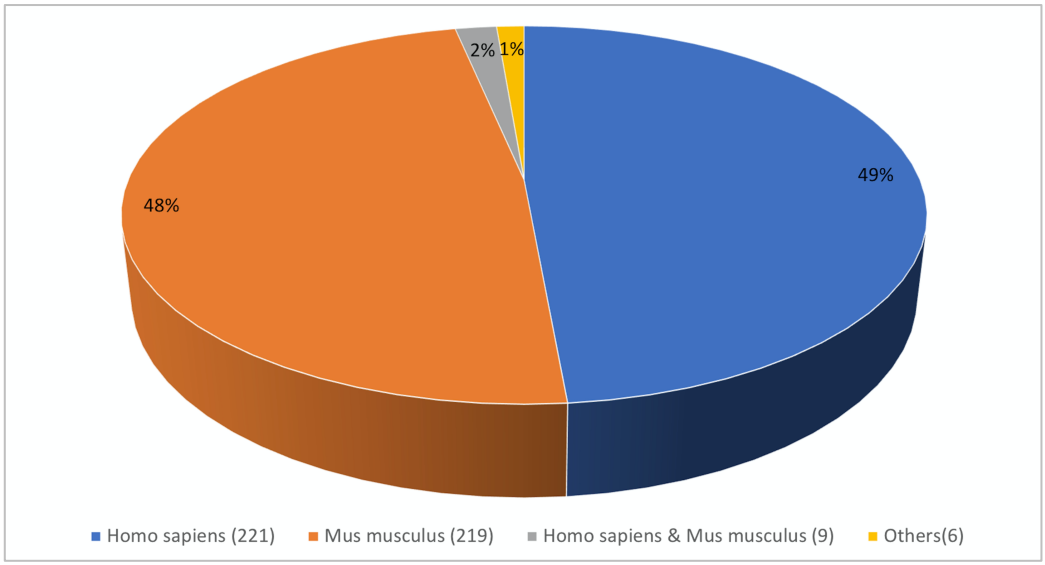


Figure 24: The distribution of known motifs according to footprintDB in different species. 49% (i.e., 221) of the motifs located in the human genome, 48% (i.e., 219) of the motifs located in the mouse genome, 2% (i.e., 9) of the motifs were seen on both Homo sapiens and Mus musculus, and 1% (i.e., 6) of the motifs were seen on some other rat types.

- ***Acceptor site motifs identified in this study were binding to TFs, 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 revealed by DAVID analysis. Analysis for Homo sapiens and Mus musculus performed separately. Six motifs found on several rat species 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 Appendix F). The most significant functional annotations with p-value < e^{-50} for 162 human and 164 mouse TFs are listed in Table 2 and Table 3. In both analyses, enrichment terms clustered in homeobox, homeodomain, DNA binding, and transcription regulation categories.

Table 2: Functional annotations of Homo sapiens TFs binding to known acceptor motifs identified in this study with p-value < e-50.

Category	Term	Count	P-Value
UP_SEQ_FEATURE	DNA-binding region:Homeobox	102	1.30E-
UP_KEYWORDS	Homeobox	107	9.30E-
INTERPRO	Homeodomain	107	2.30E-
INTERPRO	Homeobox, conserved site	98	4.30E-
UP_KEYWORDS	DNA-binding	160	1.90E-
INTERPRO	Homeodomain-like	107	6.00E-

Category	Term	Count	P-Value
SMART	HOX	107	1.50E-
GOTERM_MF_DIRECT	sequence-specific DNA binding	100	5.10E-
UP_KEYWORDS	Transcription	138	5.50E-
UP_KEYWORDS	Transcription regulation	137	5.30E-
UP_KEYWORDS	Nucleus	161	1.90E-94
GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	91	2.20E-69
INTERPRO	Homeodomain, metazoa	45	8.50E-68
GOTERM_CC_DIRECT	nucleus	152	4.60E-67
UP_KEYWORDS	Developmental protein	82	1.20E-64
GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	86	9.00E-62
GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	65	9.00E-55

Table 3: Functional annotations of Mus musculus TFs binding to known acceptor motifs identified in this study with p-value < e-50.

Category	Term	Count	P-Value
UP_KEYWORDS	DNA-binding	154	1.7E-172
GOTERM_MF_DIRECT	sequence-specific DNA binding	131	4.5E-164
INTERPRO	Homeobox, conserved site	96	7.2E-164
UP_KEYWORDS	Homeobox	100	1.2E-154
INTERPRO	Homeodomain	100	3.1E-152
UP_SEQ_FEATURE	DNA-binding region:Homeobox	90	1.7E-147
GOTERM_MF_DIRECT	DNA binding	152	2.0E-140
INTERPRO	Homeodomain-like	100	8.3E-140
GOTERM_BP_DIRECT	regulation of transcription, DNA-	152	6.7E-130
SMART	HOX	100	1.7E-124
UP_KEYWORDS	Transcription regulation	127	1.9E-109
UP_KEYWORDS	Transcription	128	2.5E-109
UP_KEYWORDS	Nucleus	157	1.2E-108
GOTERM_BP_DIRECT	transcription, DNA-templated	129	1.4E-98
GOTERM_CC_DIRECT	nucleus	158	4.6E-82
INTERPRO	Homeodomain, metazoa	50	5.0E-82
GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	95	3.5E-81
UP_KEYWORDS	Developmental protein	86	7.6E-74
GOTERM_BP_DIRECT	multicellular organism development	86	3.6E-64
GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	82	4.2E-60

Functional annotation analysis of TFs binding to known acceptor region motifs in the 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 the p-value of the enriched terms after Benjamini correction, four Homebox or homeodomain terms are ranked in the top 5. Functional annotation of the known acceptor region motif binding TFs in the mouse genome showed 20 significant enrichment terms 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 extracted for the known acceptor region motif binding TFs in the mouse genome were "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.

4.3. Analysis Results of Novel Motifs

Analyses of the novel motifs revealed the below list of results:

- ***Sequence lengths of novel motifs.***

  Sequence lengths of novel motifs. Genome-wide motif search on splice acceptor sequences through the MEME algorithm has produced 33 novel motifs, which are 9 to 50 nucleotides long. Searching motifs with the DREME algorithm has produced 27 motifs, which are 5 to 8 nucleotides long. The distribution of motifs, according to their lengths, is summarized in Figure 25.

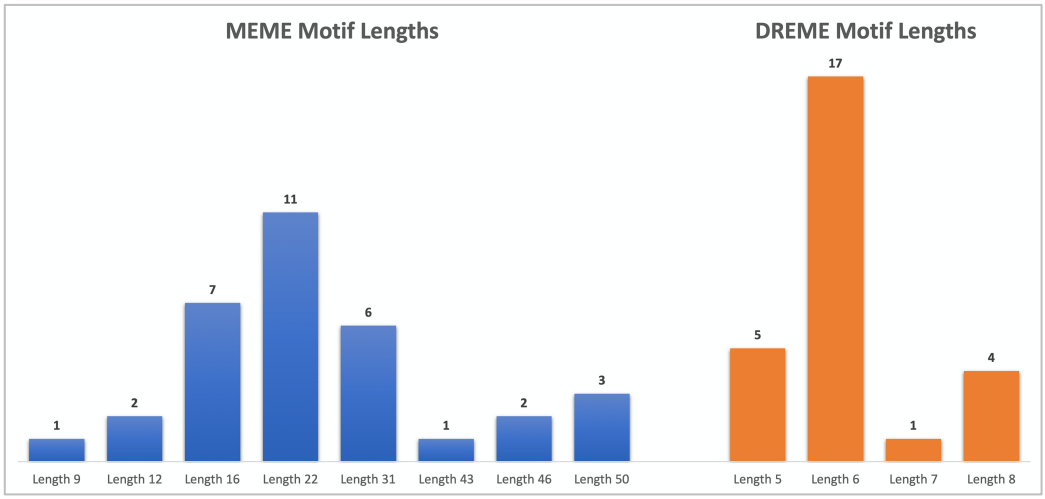


Figure 25: Distribution of the MEME and DREME motifs according to their length. Most of the MEME motifs are 16 to 31 bases long, whereas most of the DREME motifs are 6 bases long.

- ***17 novel motifs are annotated as RNA binding protein associated motifs.***

  Searching for 37 novel motifs from ATtRACT [33], we found that 12 of them did not match, but 25 motifs matched with RBP binding motifs of the ATtRACT repository. Out of 25, 17 of our motifs had at least one perfect match, and

interestingly 8 of them match with motifs of non-human organisms. Table 4 presents all novel motifs matching with the binding motifs. Quality score (Qscore) indicates the binding affinity between RNA binding proteins and binding sites and represents the probability of observing a given motif within the experiment [33]. With this sense, we count the novel motifs with Qscore = 1.00 as highly probable matches with RBP associated motifs.

Table 4. The list of novel motifs matching with RBP binding protein associated motifs in ATtRACT repository.

#	Motif Id	Complementary RNA Sequence	Matching RNA Binding Motifs	Gene Name	Organisms	Qscore
1	AAAAAMAA	UUUUUKUU	UUUUUUUU	MSS116	S. cerevisiae	**1.00**
				SXL	D. melanogaster	
2	AAAAHAAA	UUUUDUUU	UUUUUUUU	MSS116	S. cerevisiae	**1.00**
				SXL	D. melanogaster	
3	AAAAMAAA	UUUUKUUU	UUUUUUUU	MSS116	S. cerevisiae	**1.00**
				SXL	D. melanogaster	
4	AAARAAAA	UUUYUUUU	UUUUUUUU	MSS116	S. cerevisiae	**1.00**
				SXL	D. melanogaster	
5	AAATRH	UUUAYD	UUUAUA	HNRNPA1	Homo Sapiens	**1.00**
				HNRNPA2B1		
			UUUAUU	PPIE		0.015596
6	AAATRY	UUUAUA	UUUAUA	HNRNPA1	Homo Sapiens	**1.00**
				HNRNPA2B1		
			UUUAUU	PPIE		0.015596
7	AGAAA	UCUUU	UCUUU	PTBP1	Homo Sapiens	**1.00**
8	AGRAA	UCYUU	UCUUU	PTBP1	Homo Sapiens	**1.00**
9	DAAATR	HUUUAY	UUUUAU	CPEB1-A	Xenopus laevis	**1.00**
			AUUUAC	KHDRBS1	Homo Sapiens	0.437165
			UUUUAC			0.219670
			UUUUAU	PPIE		0.015596
			AUUUAU			
10	RAAATA	YUUUAU	UUUUAU	CPEB1-A	Xenopus laevis	**1.00**
			UUUUAU	PPIE	Homo Sapiens	0.015596
11	RAAATR	YUUUAY	UUUUAU	CPEB1-A	Xenopus laevis	**1.00**
			UUUUAC	KHDRBS1	Homo Sapiens	0.219670
			UUUUAU	PPIE		0.015596
12	RAAAYR	YUUURY	UUUUAU	CPEB1-A	Xenopus laevis	**1.00**
			UUUUAC	KHDRBS1	Homo Sapiens	0.219670
			UUUUAU	PPIE		0.015596
13	RGAAA	YCUUU	UCUUU	PTBP1	Homo Sapiens	**1.00**
			CCUUU			0.031230
14	RGAAR	YCUUY	UCUUU	PTBP1	Homo Sapiens	**1.00**
			UCUUC			0.031230
15	RGRAA	YCYUU	UCUUU	PTBP1	Homo Sapiens	**1.00**
			CCUUU			0.031230
16	RRGAAA	YYCUUU	CUCUUU	PTBP1	Homo Sapiens	**1.00**
			UUCUUU			
			CCCUUU	ZFP36		0.027732
			UCCUUU			
17	WCCCCR	WGGGGY	UGGGGU	HNRNPF	Homo Sapiens	**1.00**
				HNRNPH1		
				HNRNPH2		
18	TAWATR	AUWUAY	AUUUAC	KHDRBS1	Homo Sapiens	0.437165
			AUAUAU	PPIE		0.015596
			AUUUAU			
19	AYATWY	URUAWR	UAUAAA	PPIE	Homo Sapiens	0.015596
			UAUAUA			
			UAUAUA	RBMS3		0.353553
20	DTTTCY	HAAAGR	AAAAGA	PABPN1	Homo Sapiens	0.198788

#	Motif Id	Complementary RNA Sequence	Matching RNA Binding Motifs	Gene Name	Organisms	Qscore
21	CYTCCCW	GRAGGGW	GGAGGGA	B52	Drosophila	0.127280
			GGAGGGA	SRSF1	Homo Sapiens	0.087301
22	AGRAAR	UCYUUY	UCUUUC	ZFP36	Homo Sapiens	0.027732
23	RGRAAA	YCYUUU	CCCUUU	ZFP36	Homo Sapiens	0.027732
			UCCUUU			
			UCUUUU			
			CCUUUU			
24	RGRAAR	YCYUUY	CCCUUC	ZFP36	Homo Sapiens	0.027732
			CCCUUU			
			CCUUUC			
			CCUUUU			
			UCCUUC			
			UCCUUU			
			UCUUUC			
			UCUUUU			
25	TGGGRA	ACCCYU	ACCCUU	HNRNPK	Homo Sapiens	0.026640

- ***RepeatMasker analysis results.***

  For all the novel motifs, we have applied repetition analysis through RepeatMasker. 37 of the novel motifs shorter than 20nt long did not show any repetitive sequences. However, 22 of the novel motifs longer than 20nt had simple or low complexity repeats. The "GGRRACAGAGRCYCAGAGRGRG" motif was the only long motif without any repeats. We searched for this sequence's alternatives through UCSC Blat by replacing the IUPAC codes with possible nucleotides and observed that 20 nucleotides of this sequence are a SINE (Short Interspersed Nuclear Elements) repeat (Figure 26). We have excluded the motifs with repetitive sequences from our analysis, and continued the analysis of the 37 novel motifs with no repeats.

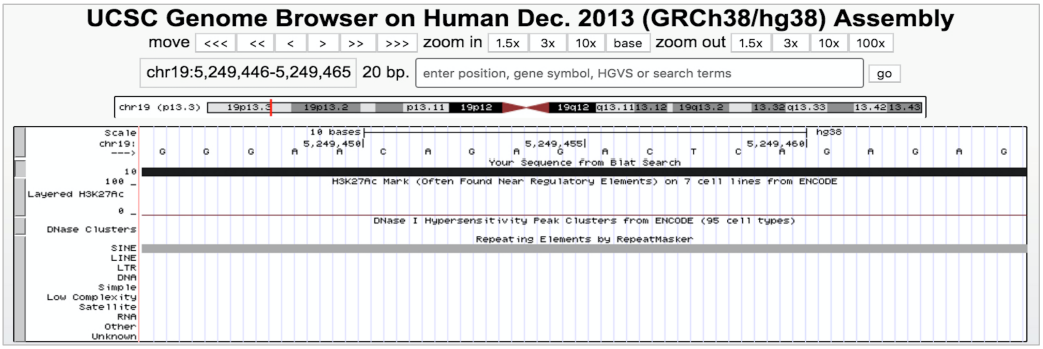


Figure 26: UCSC BLAT Search Results for an alternative of GGRRACAGAGRCYCAGAGRGRG (i.e. GGGAACAGAGACTCAGAGAGAG) has shown that the first 20nt long part of the sequence is a SINE repeat.

- ***Novel motifs with <20 nt length are most frequently located up to 50 nt upstream of the splice acceptor site.***

  We have investigated the position of 37 novel motifs (<20 nt long) in terms of the occurrence frequency in the eight identified regions of acceptor region sequences (shown in Figure 13 of Chapter 3 - Materials And Methods section). Based on the motif positions reported in the FIMO results;

- Region 6, i.e., the region of 50 nucleotides at the upstream side of the acceptor, was the first rank for 30 of the novel motifs,
- Region 5, i.e., the region from 50 to 100 nucleotides at the upstream side of the acceptor, was the first rank for the remaining 7 novel motifs,
- None of the novel motifs was observed in other regions in the first rank.

These results show that the novel motifs are dominantly located between 0 to 100 nucleotides upstream of the acceptor site. Table 5 provides the list of exact frequencies of all the 37 motifs in the defined regions.

- ***Ten novel motifs with <20nt length are located on branchpoints, and three of them conform to branchpoint b-boxes.***

We used the published list of experimentally validated branchpoint locations from the genome-wide human splicing branchpoints study [17]. A closer analysis of the occurrences of our 37 novel motifs on 143,963 branchpoints revealed that 10 of our novel motifs locate on a branchpoint, as shown in Table 6. Additionally, the exact branchpoint locations were detected through single mismatch errors in the cDNA [17]. So, we have independently investigated the novel motif occurrences for 59,325 branchpoints co-locating with a mismatch. This resulted in a nearly 5-fold decrease in 3 of the motifs namely AAAAAMAA, AAAAMAAA, AAARAAAA, but only 3-fold for the other seven motifs listed in Table 6. Another observation for these 10 motifs is that, 3 of them conform to branchpoint B-box elements: AAAAHAAA contains UUnAn, WAAAAAWAADAWVAAA contains CUnAn, UUnAn, and GUnAn, and WWAWRAAAAAAAWAAA contains CUnAn and UUnNn when ambiguity codes are replaced with possible nucleotides for both directions.

Table 5: Percent occurrence frequency of novel motifs with <20nt length in the identified 8 regions around the acceptor dinucleotide. Highest frequencies are marked as red. 31 motifs are most frequently observed in Region 6 and 7 motifs are most frequently observed in Region 5. Other regions are not preferred by any of the motifs in the first rank of occurrence.

Motif Id	**Reg1**	**Reg2**	**Reg3**	**Reg4**	**Reg5**	**Reg6**	**Reg7**	**Reg8**
AAAAAMAA	11.81	12.98	12.90	12.48	13.03	**25.89**	5.06	5.84
AAAAHAAA	11.81	13.00	12.95	12.60	13.14	**25.82**	4.95	5.72
AAAAMAAA	11.74	12.91	12.88	12.56	13.08	**25.95**	5.07	5.82
AAAARWAAAAAA	12.19	13.92	13.77	13.45	14.35	**24.39**	3.43	4.49
AAARAAAA	11.58	12.77	12.76	12.43	12.93	**25.96**	5.47	6.10
AAARCA	11.70	12.72	12.81	12.79	13.23	**18.33**	9.17	9.26
AAATRH	12.56	13.65	13.77	13.89	14.48	**17.23**	6.96	7.46
AAATRY	12.46	13.45	13.61	13.93	14.46	**15.48**	8.15	8.47
ACATWY	12.06	13.00	13.29	13.59	13.98	**14.08**	9.88	10.13
AGAAA	11.32	11.92	12.13	12.11	11.85	**14.72**	14.03	11.93
AGRAA	11.47	12.15	12.40	12.49	12.51	**15.93**	12.19	10.86
AGRAAR	11.13	12.04	12.25	12.33	12.39	**16.89**	11.97	10.99
AYATWY	12.49	13.42	13.61	14.04	**14.61**	13.28	9.34	9.20
CYCTCCYTNCCACMCT	10.96	12.95	12.75	13.11	15.52	**15.91**	9.00	9.79
CYTCCCW	12.09	12.69	12.51	12.62	**13.70**	13.21	11.28	11.91
DAAATR	12.64	13.69	13.81	13.89	14.51	**17.39**	6.76	7.32
DTTTCY	11.43	12.40	12.56	12.58	12.68	**18.30**	10.42	9.64
GGGCTGGGG	12.14	12.99	12.67	12.87	**15.10**	12.64	10.14	11.45
RAAATA	12.81	13.79	13.77	13.60	14.10	**20.01**	5.57	6.35
RAAATR	12.55	13.57	13.62	13.56	14.21	**19.29**	6.29	6.92
RAAAYR	12.32	13.29	13.38	13.23	13.74	**20.72**	6.35	6.96
RGAAA	11.27	11.86	12.09	12.12	11.98	**17.54**	12.26	10.88
RGAAR	10.80	11.46	11.69	11.74	11.55	**14.85**	14.72	13.19
RGGSAGGGGGRRGRRG	10.70	12.65	12.44	12.59	14.33	**19.03**	8.86	9.40
RGRAA	11.32	12.00	12.25	12.33	12.29	**16.11**	12.55	11.15
RGRAAA	11.15	12.06	12.24	12.34	12.40	**18.64**	11.05	10.11
RGRAAR	11.05	11.97	12.22	12.29	12.37	**17.13**	11.78	11.18
RGRRAARGRRAGAGAG	10.21	12.35	12.26	12.11	13.30	**23.19**	8.10	8.48
RRGAAA	11.21	12.06	12.13	12.15	12.25	**17.02**	12.06	11.12
STGGGGTGGGKG	11.32	12.77	12.50	12.80	**15.28**	12.82	10.78	11.74
TAWATR	12.84	13.83	13.98	14.42	15.04	**15.50**	6.76	7.64
TGGGRA	11.91	12.63	12.63	12.72	**13.51**	12.50	11.86	12.24
TTTTTTTTTTTTNAKW	11.83	14.19	14.03	13.80	14.79	**23.48**	3.51	4.37
WAAAAAWAADAWVAAA	12.02	14.41	14.34	14.19	15.27	**22.13**	3.38	4.26
WCCCCR	12.07	12.62	12.51	12.61	**13.56**	11.81	12.06	12.75
WWAWRAAAAAAAWAA	11.70	14.04	13.89	13.70	14.66	**23.77**	3.83	4.41
YCCYTCCCASCYCCYC	10.67	12.69	12.60	12.97	**15.21**	15.13	10.07	10.65

Table 6: Novel motifs' occurrences on branchpoints and genome-wide splice acceptor region sequences.

Motif Id	# Occurrence on Any Type of Branchpoint	# Occurrence on Match + Error Type Branchpoint	Genome-wide occurrence of the motif among 207,583 splice acceptor region sequences
AAAAAMAA	756	175	27.99%
AAAAHAAA	1074	326	30.80%
AAAAMAAA	813	182	30.94%
AAARAAAA	806	174	28.11%
CYTCCCW	663	206	34.38%
GGGCTGGGG	1640	636	42.83%
RGGSAGGGGGRRGRRG	6144	2607	32.79%
STGGGGTGGGKG	2023	743	33.97%
WAAAAAWAADAWVAAA	5315	2176	33.42%
WWAWRAAAAAAAWAAA	5741	2359	27.99%

- ***Most of the novel motifs with <20nt length found to be >80% similar with known homo sapiens motifs.***

  We used footprintDB's sequence search feature for the novel motifs with <20nt length, which showed that 33 of the inspected 37 motifs were at least 80% similar to known motifs of Homo sapiens (Figure 27). The four motifs that showed less similarity were "CYCTCCYTNCCACMCT", "RGRRAARGRRAGAGAG", "TTTTTTTTTTTTNAKW", and "YCCYTCCCASCYCCYC". Their similarity levels to previously known motifs of Homo sapiens were 67.75%, 69.56%, 78.19%, and 70.88%, respectively.

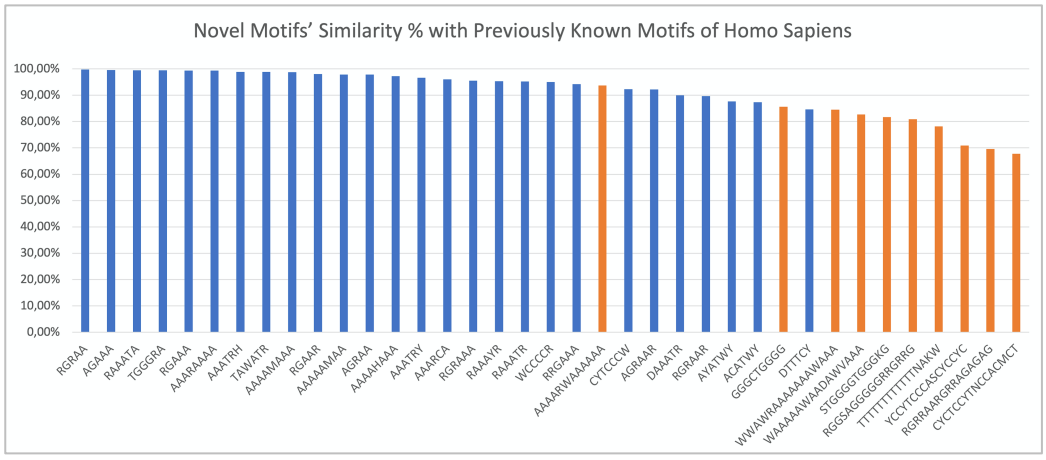


Figure 27: Novel motifs' (<20nt length) similarity percent with previously known Homo sapiens motifs. Motifs found by the DREME algorithm are shown in blue, and the MEME algorithm are shown in orange. DREME motifs have a higher similarity percent compared to MEME motifs.

4.4. Validation With Experimentally Validated Variants

The experimentally proven splice altering variants set that we use contains 3440 variants: 1612 variants causing splice acceptor alterations and 1828 variants causing splice donor alterations. We found that 634 of these variants overlap with our motifs. Furthermore, 412 (i.e., 65%) of these overlapping variants cause acceptor gain or acceptor loss. 313 of these acceptor variants overlap with only known motifs, while 22 of them overlap with only novel motifs. Also, 77 of the splice altering variants overlap with both known and novel motifs. Appendix G provides a list of these motif overlapping variants.

Experimentally proven disease associated variants of dbGaP [88], TCGA [89, 90], GWAS Catalog [91], PharmGKB [92], and ClinVar (only variants with OMIM entries are considered) [93, 94] were observed to locate on our motifs:

- 450 motifs overlap with at least 7 dbGaP variants,
- 438 motifs overlap with at least one TCGA variant,
- 423 motifs overlap with at least one GWAS Catalog variant,
- 206 motifs overlap with at least one pharmGKB variant,
- 393 motifs overlap with at least one ClinVar variant with OMIM entry.

With a closer look at the co-locations, we observed that 189 significant motifs of this study overlap with at least one variant from all these disease associated variant databases. Furthermore, 10 of them are the novel ones ("AAAAAMAA", "AAAAHAAA", "AAAAMAAA", "AAARAAAA", "CYTCCCW", "GGGCTGGGG", "RGGSAGGGGGRRGRRG", "STGGGGTGGGKG", "WAAAAAWAADAWVAAA", "WWAWRAAAAAAAWAAA"). Appendix H provides the list of these 189 motifs.

4.5. Prioritized Motifs By Splice Variant SNPs

By loading dbSNP Build 150 and collecting the SNPs located in the splice acceptor region, we produced a subset of SNPs, consisting of 6,196,185 entries. For each SNP in this subset, we collected the SPANR scores ranging from -92,5113 to 82,9138 and observed that 201,472 SNPs have |SPANR Score| > = 5, meaning that they are predicted to alter splicing (5 is the value used for the validation of SPANR). When we count the potentially splice altering SNPs residing on our motifs, we were left with 450 significant motifs (440 known and 10 novel) on which there exist such SNPs. The remaining 6 significant motifs were not co-locating with the splice altering SNPs. Computing the ratio of the number of splice altering SNPs on each motif and the number of the corresponding motif's occurrences on the human genome, we produced the splice affecting scores of our motifs. To clarify, our measure assigns higher scores to the motifs co-locating with higher number of splice altering SNPs.

450 significant motifs which co-locate either with a splice altering or disease-causing variant are scored as summarized in Figure 28. 113 high scoring motifs in the top

quartile are prioritized. 2 novel motifs, namely “RGGSAGGGGGRRGRRG” and “STGGGGTGGGKG”, are observed in the top quartile and therefore prioritized. Furthermore, 76 of the prioritized motifs are mus musculus motifs that are still novel for the human genome. The remaining 35 motifs are known homo sapiens motifs. Appendix I provides a list of these prioritized 113 motifs.

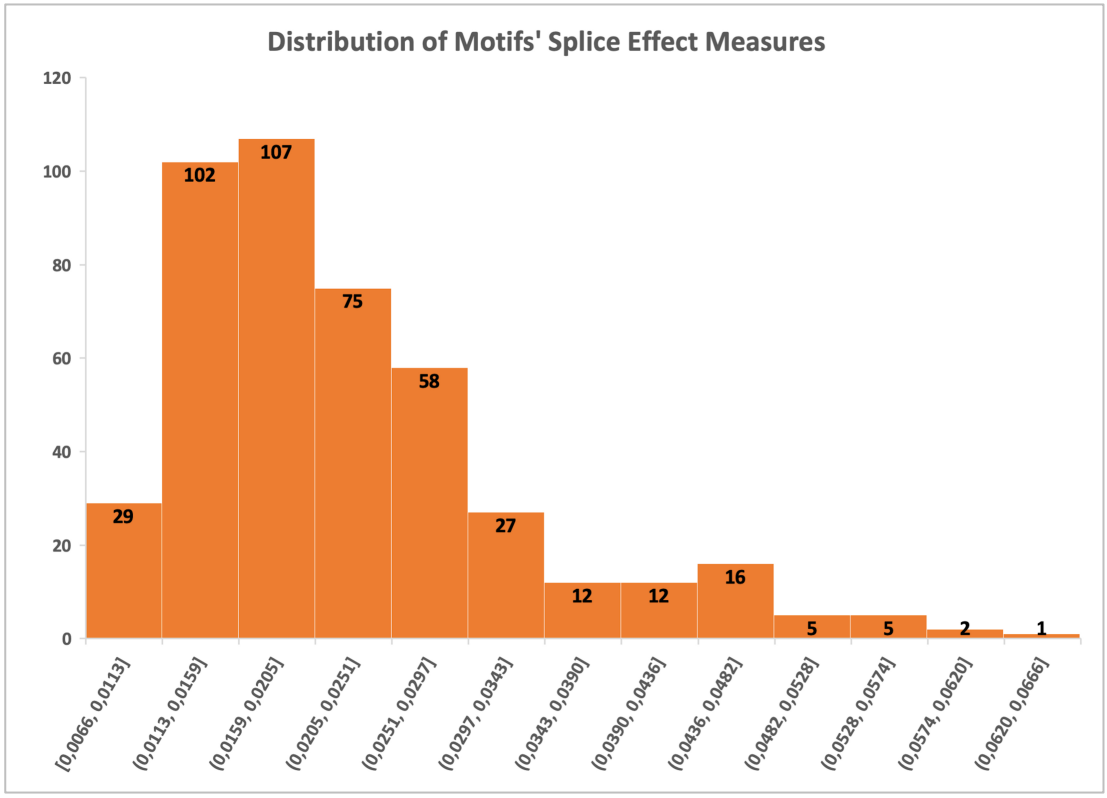


Figure 28: Distribution of motifs' splice effect measures. This measure is calculated by using splice effecting scores (i.e., SPANR scores) of co-located SNPs.

CHAPTER 5

DISCUSSION

In this study, we have performed a genome-wide computational motif analysis around the human acceptor region sequences. We have retrieved 400 nucleotides long sequences located 300 nucleotides upstream and 100 nucleotides downstream of the actual acceptor sites. We used a collection of tools included in MEME-ChIP to analyze the different aspects of motifs discovered by MEME and DREME. Their preference for particular locations is assessed with Centrimo, exact locations of significant motifs are collected through FIMO for more in-depth locational analysis, and novel and known motifs are differentiated with Tomtom. Tools included in the MEME-ChIP online were utilized for each chromosome independently due to the tool's computational limits (The sequence file input of MEME-ChIP can be at most 80MB, but our sequences are nearly 92MB in total). Completing this chromosome-wide analysis for all 24 chromosomes in the human genome, we observed several common motifs on different chromosomes. Therefore, we conducted our genome-wide analysis by aggregating the significant motifs of each chromosome-wide analysis. With this sense, we were able to include all the motifs found to be significant at the chromosome-wide level. This approach expanded the analyzed motifs set as it included the motifs that might be otherwise ignored due to the differences in the chromosomes' sizes. For instance, we analyzed 21041 sequences of chr1, whereas 458 sequences of chrY and a significant motif for chrY may be found less significant with a single step genome-wide analysis. The resulting aggregated set of motifs were filtered according to their length, significance, and consensus afterward. 517 motifs passed the filtering. Among the selected motifs, 457 were previously known as mammalian motifs, but 60 were novel motifs of the acceptor region.

The known motifs, computationally identified as significant motifs for acceptor sites in this study, were found to be experimentally proven DNA binding motifs located on human or other mammalian genomes. The TFs, binding to the known acceptor region motifs, are mainly annotated for homeobox and homeodomain, DNA binding region, or transcription regulation functions in both human and mouse genomes. Homeobox genes direct the formation of body structures during early embryonic development, and their alternative splicing is known to be under strict regulation [110]. The genomic organization and the complex arrangement of regulative elements in the homeobox gene structure are essential in their spatio-temporal regulation and function [111]. Also, conserved intronic regions containing homeobox sequences [112] or alternative

promoters [113] are known to have regulative roles. The homeobox sequences identified as acceptor region motifs in this study are likely to have a regulative role in the alternative splicing mechanism. Until the early 2010s, TFs were known to affect splicing indirectly by influencing RNA Polymerase II elongation rates. However, recent studies have provided evidence on the direct effects of TFs on pre-mRNA splicing through pre-mRNA imprinting [44], and aberrant TFs in cancer cells are found to alter the DNA binding activity [114]. These findings are also supporting that computationally significant TF binding motifs of this study are very likely to have regulatory roles in splicing. Therefore, we suggest that the computationally significant acceptor region motifs that are experimentally shown to be functional in non-human organisms can be prioritized to investigate their roles in the human genome further.

We were able to detect 60 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. A comprehensive analysis of the novel motifs' biological significance is done based on the length of the motifs.

Novel motifs with >20nt length were found to be simple or low complexity repetitive sequences. Shorter motifs marked as significant by DREME are all found to have sequence similarity between 84.60% 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 – [115]) 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 to 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 biologically meaningful patterns but mainly repetitive sequences.

ATtRACT is a database providing an experimentally validated set of 1583 RBP binding motifs. This repository does not provide a search mechanism with count matrices but offers a motif search interface that supports searching for sequences with IUPAC codes [46]. As ATtRACT is the most extensive experimentally validated publicly available repository, we used it to investigate our novel motifs. RBP binding motifs' lengths of ATtRACT range from 4 to 12 nucleotides. Even with this restriction, we found 17 novel motifs of this study are perfectly matching with motifs in the ATtRACT repository. Moreover, 8 of our novels match with RBP binding motifs that are novel to the human genome. This result is supporting our argument that computationally significant motifs of this study may also have biological roles.

For the novel motifs shorter than 20nt (i.e., 37 motifs), locational analysis of FIMO [30] outputs showed that 30 of the motifs most frequently occur in Region 6, 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 [116]) and the branchpoint (19-37 base pairs upstream [17]) maps, indicating a high probability for regulating acceptor site splicing. Like the donor site splice signals that facilitate the binding of U1 snRNP, these novel acceptor motifs might be facilitating the recognition of a site or a unique U2AF heterodimer [117]. Additionally, we have observed that 10 of the novel motifs locate on the experimentally shown branchpoints, and 3 contain branchpoint B-boxes. Interestingly, 3 of the novel motifs that co-localize with experimental branchpoints, namely "CYTCCCW", "GGGCTGGGG", "STGGGGTGGGKG", most frequently occur in Region 5 (50-100 nucleotides upstream of the acceptor site). This observation suggests that some branchpoints might be extending to the regions ≥ 50nt upstream [19]. For the remaining 4 novel motifs which most frequently occur in Region 5, we did not find a branchpoint match or conformity with B-boxes, except for the motif "AYATWY". Nevertheless, they might still be playing regulatory roles for splicing as intronic splicing regulation predominantly occurs within 100nt of the splice sites [118].

As motif overlapping variants proven to be related to biological changes are potentially disrupting the biological roles of the motifs [119], we searched for co-locating experimentally validated splice altering and/or disease associated variants to evaluate the biological significance of our motifs. In this analysis, we observed overlaps with known motifs and novel ones, which supports the biological significance of the novel motifs. We propose that the acceptor splice region motifs identified here, are candidate sequences for further experimental studies, as they can moderate alternative splicing.

As the final set of significant motifs co-locating with either splice altering or disease-causing variants consists of 450 entries, we concentrated on developing a prioritization measure for their further experimental investigation. Our final analysis step aimed to provide scores to the significant motifs in terms of their splice disruption competence. It was held in a genome-wide manner by co-locating the SNPs of dbSNP [87] and using the SNPs' splice disruption potential [95] to score the motifs. We propose that 113 motifs with higher scores (presented in Appendix I) have a higher potential to moderate splicing events and can be further prioritized in future experimental studies.

CHAPTER 6

CONCLUSION

6.1. Overview

The eukaryotic genome has exonic and intronic regions, and more than 90% of the genes are alternatively spliced to form the mRNA from the pre-mRNA. This process happens under the control of several regulatory factors, and disruptions of such a sensitive mechanism may result in operational malfunctions and diseases. GU dinucleotide at the donor 5' site and AG dinucleotide at the acceptor 3' site are the splicing machinery's main consensus sequences, which point to the introns' start and end positions. Another critical region is the branchpoint, an adenine nucleotide forming the intron lariat with a 5' splice site. Any alteration of these regulators may inhibit splicing.

Studies of splicing regulatory factors mainly concentrate on the sequences around acceptor and donor splice sites and branchpoints. However, there is a lack of genome-wide comprehensive studies of the genomic regions around the splice regulators. With this respect, our motivation is to perform a genome-wide motif discovery analysis for the sequences around the splice acceptor sites. So, we have identified novel biologically significant patterns that were not known to affect the splicing process.

6.2. Accomplishment

Performing an extensive analysis by combining the results of several motif finders, this study identified various significant motifs located around the splice acceptor site. Nearly half of the identified motifs were previously annotated as DNA binding motifs of the human genome. Nevertheless, we computationally identified various novel motifs in the splice acceptor region of the human genome. Some of these novels were previously proven to exist on other mammalian genomes, but several others were detected for the first time with this study. Furthermore, we have detected potential branch-site motifs supported with experimental evidence from DNA binding studies with the genome-wide computational analysis of the identified splice acceptor region motifs. Also, we observed many overlaps with splice altering or disease associated

variants. Therefore, we suggest that computationally significant acceptor region motifs identified in this study are potential candidates for functional analysis studies.

6.3. Future Studies

This study presents the results of a comprehensive motif analysis of human splice acceptor region sequences. It is evident that these results need wet-lab validation especially for the novel motifs, namely, "RGGSAGGGGGRRGRRG" and "STGGGGTGGGKG", which have the highest potential to be biologically functional for splicing regulation. Furthermore, previously detected motifs of other mammalian genomes that are computationally identified on the human genome should be prioritized for experimental functional studies.

We also propose that this thesis's comprehensive analysis method can be performed for the sequences around the splice donor site. A comprehensive analysis of SNPs has revealed that splice site disruptions occur more frequently in donor sites than in acceptor sites [120]. This may be an indicator of regulatory motifs around the splice donor sites waiting for discovery.

REFERENCES

[1] Dredge, B. K.; Polydorides, A. D.; Darnell, R. B. The splice of life: Alternative splicing and neurological disease. *Nat. Rev. Neurosci.*, **2001**. https://doi.org/10.1038/35049061.

[2] Tazi, J.; Bakkour, N.; Stamm, S. Alternative splicing and disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*. 2009, lpp 14–26. https://doi.org/10.1016/j.bbadis.2008.09.017.

[3] Cartegni, L.; Hastings, M. L.; Calarco, J. A.; de Stanchina, E.; Krainer, A. R. Determinants of Exon 7 Splicing in the Spinal Muscular Atrophy Genes, SMN1 and SMN2. *Am. J. Hum. Genet.*, **2006**. https://doi.org/10.1086/498853.

[4] Cartegni, L.; Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN. *Nat. Genet.*, **2002**. https://doi.org/10.1038/ng854.

[5] Kashima, T.; Rao, N.; David, C. J.; Manley, J. I. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Hum. Mol. Genet.*, **2007**. https://doi.org/10.1093/hmg/ddm276.

[6] Kashima, T.; Manley, J. L. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.*, **2003**. https://doi.org/10.1038/ng1207.

[7] Kovacs, G. G. Tauopathies. No *Handbook of Clinical Neurology*; 2018. https://doi.org/10.1016/B978-0-12-802395-2.00025-0.

[8] 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. https://doi.org/10.1016/j.bbadis.2004.08.010.

[9] Gallo, J. M.; Noble, W.; Martin, T. R. RNA and protein-dependent mechanisms in tauopathies: Consequences for therapeutic strategies. *Cellular and Molecular*

Life Sciences. 2007. https://doi.org/10.1007/s00018-007-6513-4.

[10] Hasegawa, M.; Smith, M. J.; Iijima, M.; Tabira, T.; Goedert, M. FTDP-17 mutations N279K and S305N in tau produce increased splicing of exon 10. *FEBS Lett.*, **1999**. https://doi.org/10.1016/S0014-5793(98)01696-2.

[11] Buratti, E. Multiple roles of TDP-43 in gene expression, splicing regulation, and human disease. *Front. Biosci.*, **2008**. https://doi.org/10.2741/2727.

[12] Gaweda-Walerych, K.; Mohagheghi, F.; Zekanowski, C.; Buratti, E. Parkinson's disease–related gene variants influence pre-mRNA splicing processes. *Neurobiol. Aging*, **2016**. https://doi.org/10.1016/j.neurobiolaging.2016.07.014.

[13] Liu, H. X.; Cartegni, L.; Zhang, M. Q.; Krainer, A. R. A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nat. Genet.*, **2001**. https://doi.org/10.1038/83762.

[14] Hoffman, J. D.; Hallam, S. E.; Venne, V. L.; Lyon, E.; Ward, K. Implications of a novel cryptic splice site in the BRCA1 gene. *Am. J. Med. Genet.*, **1998**. https://doi.org/10.1002/(SICI)1096-8628(19981102)80:2<140::AID-AJMG10>3.0.CO;2-L.

[15] Orban, T. I.; Olah, E. Emerging roles of BRCA1 alternative splicing. *Journal of Clinical Pathology - Molecular Pathology*. 2003. https://doi.org/10.1136/mp.56.4.191.

[16] Clancy, S. RNA splicing: introns, exons and spliceosome. *Nat. Educ.*, **2008**, *1* (2008), 3–6.

[17] Mercer, T. R.; Clark, M. B.; Andersen, S. B.; Brunck, M. E.; Haerty, W.; Crawford, J.; Taft, R. J.; Nielsen, L. K.; Dinger, M. E.; Mattick, J. S. Genome-wide discovery of human splicing branchpoints. *Genome Res.*, **2015**, *25* (2), 290–303. https://doi.org/10.1101/gr.182899.114.

[18] Fica, S. M.; Tuttle, N.; Novak, T.; Li, N.-S.; Lu, J.; Koodathingal, P.; Dai, Q.; Staley, J. P.; Piccirilli, J. A. RNA catalyses nuclear pre-mRNA splicing. *Nature*, **2013**, *503* (7475), 229–234. https://doi.org/10.1038/nature12734.

[19] Pineda, J. M. B.; Bradley, R. K. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev.*, **2018**. https://doi.org/10.1101/gad.312058.118.

[20] Baralle, F. E.; Singh, R. N.; Stamm, S. RNA structure and splicing regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2019. https://doi.org/10.1016/j.bbagrm.2019.194448.

[21] Bailey, T. L.; Elkan, C. Fitting a Mixture Model by Expectation Maximization

to Discover Motifs in Bipolymers. No *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*; 1994; lpp 28–36. https://doi.org/citeulike-article-id:878292.

[22] Feng, J.; Chen, K.; Dong, X.; Xu, X.; Jin, Y.; Zhang, X.; Chen, W.; Han, Y.; Shao, L.; Gao, Y.; u.c. Genome-wide identification of cancer-specific alternative splicing in circRNA. *Molecular Cancer*. 2019. https://doi.org/10.1186/s12943-019-0996-0.

[23] Yu, M.; Hong, W.; Ruan, S.; Guan, R.; Tu, L.; Huang, B.; Hou, B.; Jian, Z.; Ma, L.; Jin, H. Genome-wide profiling of prognostic alternative splicing pattern in pancreatic cancer. *Front. Oncol.*, **2019**. https://doi.org/10.3389/fonc.2019.00773.

[24] Perrone, B.; La Cognata, V.; Sprovieri, T.; Ungaro, C.; Conforti, F. L.; Andò, S.; Cavallaro, S. Alternative Splicing of ALS Genes: Misregulation and Potential Therapies. *Cellular and Molecular Neurobiology*. 2020. https://doi.org/10.1007/s10571-019-00717-0.

[25] Begg, B. E.; Jens, M.; Wang, P. Y.; Minor, C. M.; Burge, C. B. Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. *Nat. Struct. Mol. Biol.*, **2020**. https://doi.org/10.1038/s41594-020-0475-8.

[26] Monger, S.; Troup, M.; Ip, E.; Dunwoodie, S. L.; Giannoulatou, E. Spliceogen: an integrative, scalable tool for the discovery of splice-altering variants. *Bioinformatics*, **2019**. https://doi.org/10.1093/bioinformatics/btz263.

[27] Signal, B.; Gloss, B. S.; Dinger, M. E.; Mercer, T. R. Machine learning annotation of human branchpoints. *Bioinformatics*, **2018**. https://doi.org/10.1093/bioinformatics/btx688.

[28] Bailey, T. L. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **2011**, *27* (12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261.

[29] Bailey, T. L.; MacHanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **2012**, *40* (17). https://doi.org/10.1093/nar/gks433.

[30] Grant, C. E.; Bailey, T. L.; Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, **2011**, *27* (7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

[31] Gupta, S.; Stamatoyannopoulos, J. A.; Bailey, T. L.; Noble, W. S. Quantifying similarity between motifs. *Genome Biol.*, **2007**. https://doi.org/10.1186/gb-2007-8-2-r24.

[32] Sebastian, A.; Contreras-Moreira, B. FootprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*,

2014, *30* (2), 258–265. https://doi.org/10.1093/bioinformatics/btt663.

[33] Giudice, G.; Sánchez-Cabo, F.; Torroja, C.; Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database*, **2016**. https://doi.org/10.1093/database/baw035.

[34] CRICK, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.*, **1958**.

[35] Clancy, S.; Brown, W. Translation : DNA to mRNA to Protein. *Nat. Educ.*, **2008**.

[36] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, P. W. Molecular Biology of the Cell (Sixth Edition). No *Garland Science*; 2014.

[37] Dvinge, H. Regulation of alternative mRNA splicing: old players and new perspectives. *FEBS Letters*. 2018. https://doi.org/10.1002/1873-3468.13119.

[38] Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.*, **2003**. https://doi.org/10.1146/annurev.biochem.72.121801.161720.

[39] Sammeth, M.; Foissac, S.; Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, **2008**. https://doi.org/10.1371/journal.pcbi.1000147.

[40] Wang, Y.; Liu, J.; Huang, B. O.; Xu, Y.-M.; Li, J.; Huang, L.-F.; Lin, J.; Zhang, J.; Min, Q.-H.; Yang, W.-M.; u.c. Mechanism of alternative splicing and its regulation. *Biomed. reports*, **2015**. https://doi.org/10.3892/br.2014.407.

[41] Lu, Z. X.; Jiang, P.; Xing, Y. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdiscip. Rev. RNA*, **2012**. https://doi.org/10.1002/wrna.120.

[42] D'Haeseleer, P. What are DNA sequence motifs? *Nat. Biotechnol.*, **2006**. https://doi.org/10.1038/nbt0406-423.

[43] Philips, T.; Hoopes, L. Transcription Factors and Transcriptional Control in Eukaryotic Cells. *Nat. Educ.*, **2008**.

[44] Rambout, X.; Dequiedt, F.; Maquat, L. E. Beyond Transcription: Roles of Transcription Factors in Pre-mRNA Splicing. *Chemical Reviews*. 2018. https://doi.org/10.1021/acs.chemrev.7b00470.

[45] Collins, F. S.; Brooks, L. D.; Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **1998**. https://doi.org/10.1101/gr.8.12.1229.

[46] Mullaney, J. M.; Mills, R. E.; Stephen Pittard, W.; Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, **2010**. https://doi.org/10.1093/hmg/ddq400.

[47] Escaramís, G.; Docampo, E.; Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genomics*, **2015**. https://doi.org/10.1093/bfgp/elv014.

[48] Shastry, B. S. SNPs: impact on gene function and phenotype. *Methods in molecular biology (Clifton, N.J.)*. 2009. https://doi.org/10.1007/978-1-60327-411-1_1.

[49] Halushka, M. K.; Fan, J. B.; Bentley, K.; Hsie, L.; Shen, N.; Weder, A.; Cooper, R.; Lipshutz, R.; Chakravarti, A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **1999**. https://doi.org/10.1038/10297.

[50] Komar, A. A. SNPs, silent but not invisible. *Science*. 2007. https://doi.org/10.1126/science.1138239.

[51] Elsharawy, A.; Hundrieser, B.; Brosch, M.; Wittig, M.; Huse, K.; Platzer, M.; Becker, A.; Simon, M.; Rosenstiel, P.; Schreiber, S.; u.c. Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum. Mutat.*, **2009**. https://doi.org/10.1002/humu.20906.

[52] Karsch-Mizrachi, I.; Takagi, T.; Cochrane, G. The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **2018**. https://doi.org/10.1093/nar/gkx1097.

[53] Leinonen, R.; Akhtar, R.; Birney, E.; Bower, L.; Cerdeno-Tárraga, A.; Cheng, Y.; Cleland, I.; Faruque, N.; Goodgame, N.; Gibson, R.; u.c. The European nucleotide archive. *Nucleic Acids Res.*, **2011**. https://doi.org/10.1093/nar/gkq967.

[54] Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.*, **2013**. https://doi.org/10.1093/nar/gks1195.

[55] Mashima, J.; Kodama, Y.; Fujisawa, T.; Katayama, T.; Okuda, Y.; Kaminuma, E.; Ogasawara, O.; Okubo, K.; Nakamura, Y.; Takagi, T. DNA Data Bank of Japan. *Nucleic Acids Res.*, **2017**. https://doi.org/10.1093/nar/gkw1001.

[56] Wang, J.; Kong, L.; Gao, G.; Luo, J. A brief introduction to web-based genome browsers. *Brief. Bioinform.*, **2013**. https://doi.org/10.1093/bib/bbs029.

[57] Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, a. D. The Human Genome Browser at UCSC. *Genome Res.*, **2002**. https://doi.org/10.1101/gr.229102.

[58] Fernández, X. M.; Birney, E. Ensembl genome browser. No *Vogel and Motulsky's Human Genetics: Problems and Approaches (Fourth Edition)*; 2010. https://doi.org/10.1007/978-3-540-37654-5_45.

[59] Lee, C. M.; Barber, G. P.; Casper, J.; Clawson, H.; Diekhans, M.; Gonzalez, J. N.; Hinrichs, A. S.; Lee, B. T.; Nassar, L. R.; Powell, C. C.; u.c. UCSC Genome Browser enters 20th year. *Nucleic Acids Res.*, **2020**. https://doi.org/10.1093/nar/gkz1012.

[60] Karolchik, D.; Hinricks, A. S.; Furey, T. S.; Roskin, K. M.; Sugnet, C. W.; Haussler, D.; Kent, W. J. The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **2004**. https://doi.org/10.1093/nar/gkh103.

[61] Kent, W. J. BLAT - The BLAST-like alignment tool. *Genome Res.*, **2002**, *12* (4), 656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002.

[62] Lift Genome Annotations https://genome.ucsc.edu/cgi-bin/hgLiftOver (accessed jun 10, 2017).

[63] UCSC DAS Server. UCSC DAS Server http://genome.ucsc.edu:80/cgi-bin/das/hg38 (accessed jun 20, 2018).

[64] Madden, T. The BLAST sequence analysis tool. *BLAST Seq. Anal. Tool*, **2013**, 1–17.

[65] Kinsella, R. J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; u.c. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, **2011**, *2011*. https://doi.org/10.1093/database/bar030.

[66] McLaren, W.; Gil, L.; Hunt, S. E.; Riat, H. S.; Ritchie, G. R. S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.*, **2016**. https://doi.org/10.1186/s13059-016-0974-4.

[67] Hashim, F. A.; Mabrouk, M. S.; Atabany, W. A. L. Comparative Analysis of DNA Motif Discovery Algorithms: A Systemic Review. *Curr. Cancer Ther. Rev.*, **2018**. https://doi.org/10.2174/1573394714666180417161728.

[68] Evans, P. A.; Smith, A. D. Toward optimal motif enumeration. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, **2003**. https://doi.org/10.1007/978-3-540-45078-8_5.

[69] Yang, C. H.; Liu, Y. T.; Chuang, L. Y. DNA motif discovery based on ant colony optimization and expectation maximization. No *IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011*; 2011.

[70] Machanick, P.; Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, **2011**, *27* (12), 1696–1697. https://doi.org/10.1093/bioinformatics/btr189.

[71] Bailey, T. L.; Williams, N.; Misleh, C.; Li, W. W. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **2006**. https://doi.org/10.1093/nar/gkl198.

[72] MEME - MEME Suite http://meme-suite.org/doc/meme.html (accessed nov 1, 2020).

[73] DREME Tutorial - MEME Suite http://meme-suite.org/doc/dreme-tutorial.html?man_type=web (accessed nov 1, 2020).

[74] CentriMo Tutorial - MEME Suite http://meme-suite.org/doc/centrimo-tutorial.html?man_type=web (accessed nov 1, 2020).

[75] FIMO Tutorial - MEME Suite http://meme-suite.org/doc/fimo-tutorial.html?man_type=web (accessed nov 1, 2020).

[76] Tomtom - MEME Suite http://meme-suite.org/doc/tomtom.html?man_type=web (accessed nov 1, 2020).

[77] footprintDB - a database of transcription factors with annotated cis elements (DNA motifs and sites) and binding interfaces http://floresta.eead.csic.es/footprintdb/index.php?databases (accessed nov 1, 2020).

[78] oRNAment http://rnabiology.ircm.qc.ca/oRNAment/ (accessed nov 1, 2020).

[79] Gene set enrichment analysis - Wikipedia https://en.wikipedia.org/wiki/Gene_set_enrichment_analysis (accessed nov 1, 2020).

[80] DAVID Functional Annotation Bioinformatics Microarray Analysis https://david.ncifcrf.gov/ (accessed nov 1, 2020).

[81] Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **2009**. https://doi.org/10.1038/nprot.2008.211.

[82] de Koning, A. P. J.; Gu, W.; Castoe, T. A.; Batzer, M. A.; Pollock, D. D. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet.*, **2011**. https://doi.org/10.1371/journal.pgen.1002384.

[83] Interspersed repeat - Wikipedia https://en.wikipedia.org/wiki/Interspersed_repeat (accessed nov 1, 2020).

[84] Smit, A.; Hubley, R.; Green, P. RepeatMasker Open-4.0.6 2013-2015 . *http://www.repeatmasker.org*. 2017.

[85] 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**. https://doi.org/10.1159/000084979.

[86] Hubley, R.; Finn, R. D.; Clements, J.; Eddy, S. R.; Jones, T. A.; Bao, W.; Smit, A. F. A.; Wheeler, T. J. The Dfam database of repetitive DNA families. *Nucleic Acids Res.*, **2016**. https://doi.org/10.1093/nar/gkv1272.

[87] Smigielski, E. M.; Sirotkin, K.; Ward, M.; Sherry, S. T. dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Research*. 2000. https://doi.org/10.1093/nar/28.1.352.

[88] Tryka, K. A.; Hao, L.; Sturcke, A.; Jin, Y.; Wang, Z. Y.; Ziyabari, L.; Lee, M.; Popova, N.; Sharopova, N.; Kimura, M.; u.c. NCBI's database of genotypes and phenotypes: DbGaP. *Nucleic Acids Res.*, **2014**, *42* (D1). https://doi.org/10.1093/nar/gkt1211.

[89] Collins, F. S. The Cancer Genome Atlas (TCGA). *Online*. 2007.

[90] The Cancer Genome Atlas (TCGA) https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas (accessed nov 1, 2020).

[91] Buniello, A.; Macarthur, J. A. L.; Cerezo, M.; Harris, L. W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; u.c. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **2019**. https://doi.org/10.1093/nar/gky1120.

[92] Whirl-Carrillo, M.; McDonagh, E. M.; Hebert, J. M.; Gong, L.; Sangkuhl, K.; Thorn, C. F.; Altman, R. B.; Klein, T. E. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*. 2012. https://doi.org/10.1038/clpt.2012.96.

[93] Landrum, M. J.; Lee, J. M.; Benson, M.; Brown, G. R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; u.c. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **2018**. https://doi.org/10.1093/nar/gkx1153.

[94] Hamosh, A.; Scott, A. F.; Amberger, J. S.; Bocchini, C. A.; McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **2005**. https://doi.org/10.1093/nar/gki033.

[95] Xiong, H. Y.; Alipanahi, B.; Lee, L. J.; Bretschneider, H.; Merico, D.; Yuen, R. K. C.; Hua, Y.; Gueroussov, S.; Najafabadi, H. S.; Hughes, T. R.; u.c. The

human splicing code reveals new insights into the genetic determinants of disease. *Science (80-.).*, **2015**, *347* (6218), 1254806–1254806. https://doi.org/10.1126/science.1254806.

[96] Mort, M.; Sterne-Weiler, T.; Li, B.; Ball, E. V; Cooper, D. N.; Radivojac, P.; Sanford, J. R.; Mooney, S. D. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.*, **2014**, *15* (1), R19. https://doi.org/10.1186/gb-2014-15-1-r19.

[97] Shibata, A.; Okuno, T.; Rahman, M. A.; Azuma, Y.; Takeda, J.; Masuda, A.; Selcen, D.; Engel, A. G.; Ohno, K. IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome. *J. Hum. Genet.*, **2016**, *61* (7), 633–640. https://doi.org/10.1038/jhg.2016.23.

[98] 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.; u.c. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, **2019**. https://doi.org/10.1016/j.cell.2018.12.015.

[99] Mathelier, A.; Fornes, O.; Arenillas, D. J.; Chen, C.; Denay, G.; Lee, J.; Shi, W.; Shyr, C.; Tan, G.; Worsley-Hunt, R.; u.c. 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. https://doi.org/10.1093/nar/gkv1176.

[100] Jolma, A.; Yan, J.; Whitington, T.; Toivonen, J.; Nitta, K. R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; u.c. DNA-binding specificities of human transcription factors. *Cell*, **2013**, *152* (1–2), 327–339. https://doi.org/10.1016/j.cell.2012.12.009.

[101] Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **2009**. https://doi.org/10.1093/nar/gkn923.

[102] Contreras-Moreira, B. 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.*, **2010**, *38* (Database), D91–D97. https://doi.org/10.1093/nar/gkp781.

[103] Robasky, K.; Bulyk, M. L. 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). https://doi.org/10.1093/nar/gkq992.

[104] Search motifs https://attract.cnic.es/searchmotif (accessed dec 7, 2020).

[105] BaseSpace Sequence Hub. Analyses - BaseSpace Sequence Hub https://basespace.illumina.com/analyses/196845651/files/237314084?projectId=66029966 (accessed mai 29, 2020).

[106] Thorn, C. F.; Klein, T. E.; Altman, R. B. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol. Biol.*, **2013**, *1015*, 311–320. https://doi.org/10.1007/978-1-62703-435-7_20.

[107] UCSC Sequence and Annotation Downloads http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ (accessed sept 20, 2020).

[108] SPIDEX — Deep Genomics https://www.deepgenomics.com/spidex (accessed jun 10, 2017).

[109] SPIDEX. Download SPIDEX annotation database for use in ANNOVAR http://download.openbioinformatics.org/spidex_download_form.php (accessed sept 20, 2020).

[110] Stroeher, V. L.; Gaiser, J. C.; Garber, R. L. 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**. https://doi.org/10.1128/mcb.8.10.4143.

[111] Coulombe, Y.; Lemieux, M.; Moreau, J.; Aubin, J.; Joksimovic, M.; Bérubé-Simard, F. A.; Tabaries, S.; Boucherat, O.; Guillou, F.; Larochelle, C.; u.c. Multiple promoters and alternative splicing: Hoxa5 transcriptional complexity in the mouse embryo. *PLoS One*, **2010**. https://doi.org/10.1371/journal.pone.0010600.

[112] Keegan, L. P.; Haerry, T. E.; Crotty, D. A.; Packer, A. I.; Wolgemuth, D. J.; Gehring, W. J. A sequence conserved in vertebrate Hox gene introns functions as an enhancer regulated by posterior homeotic genes in Drosophila imaginal discs. *Mech. Dev.*, **1997**. https://doi.org/10.1016/S0925-4773(97)00038-5.

[113] Vacik, T.; Raska, I. Alternative intronic promoters in development and disease. *Protoplasma*. 2017. https://doi.org/10.1007/s00709-016-1071-y.

[114] Belluti, S.; Rigillo, G.; Imbriano, C. Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates. *Cells*, **2020**. https://doi.org/10.3390/cells9030760.

[115] Kulakovskiy, I. V.; Vorontsov, I. E.; Yevshin, I. S.; Soboleva, A. V.; Kasianov, A. S.; Ashoor, H.; Ba-Alawi, W.; Bajic, V. B.; Medvedeva, Y. A.; Kolpakov, F. A.; u.c. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **2016**. https://doi.org/10.1093/nar/gkv1249.

[116] Lodish H, Berk A, Zipursky SL, et al. *Molecular Cell Biology 5th Ed*; 2000. https://doi.org/10.1016/S1470-8175(01)00023-6.

[117] Thapar, R. Roles of prolyl isomerases in RNA-mediated gene expression.

Biomolecules. 2015. https://doi.org/10.3390/biom5020974.

[118] Wainberg, M.; Alipanahi, B.; Frey, B. Does conservation account for splicing patterns? *BMC Genomics*, **2016**. https://doi.org/10.1186/s12864-016-3121-4.

[119] Anna, A.; Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics*. 2018. https://doi.org/10.1007/s13353-018-0444-7.

[120] Kurmangaliyev, Y. Z.; Sutormin, R. A.; Naumenko, S. A.; Bazykin, G. A.; Gelfand, M. S. Functional implications of splicing polymorphisms in the human genome. *Hum. Mol. Genet.*, **2013**. https://doi.org/10.1093/hmg/ddt200.

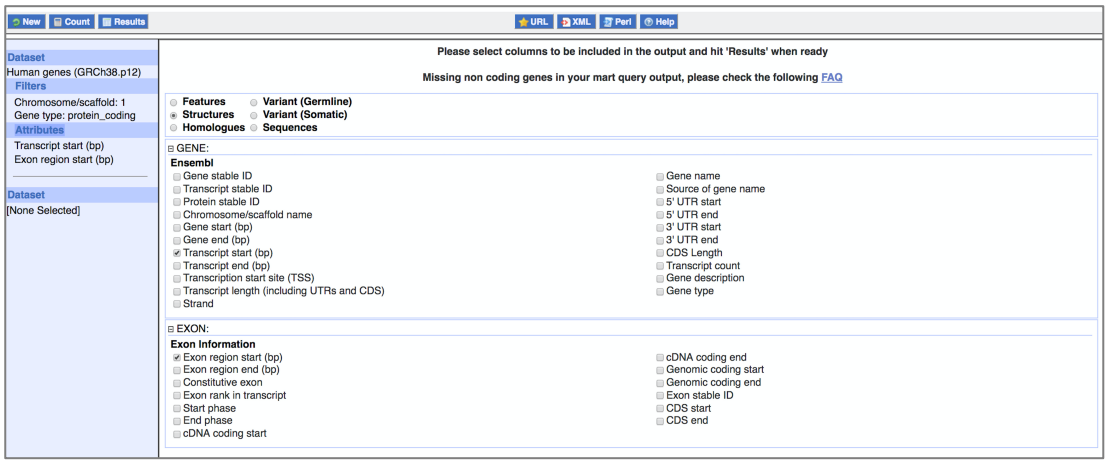
APPENDICES

APPENDIX A

TOOLS FOR MOTIF ANALYSIS

Ensembl Biomart query for acceptor splice-site collection

Exon start positions indicate acceptor sites in a genomic sequence except for the ones which contain the initiator codon. Therefore, we have collected transcript start and exon region start positions on protein coding-genes through Ensembl Biomart as shown in Appendix A - Figure 1. If the exon region start is the same as the transcript start, it is omitted since it is the exon with initiator codon and neighbor of 5' UTR but not the splice-site.



Appendix A - Figure 1: Acceptor site query through Ensembl Biomart for protein coding genes of chromosome 1. For all other 23 chromosomes, queries are formed in the same manner.

Perl script and R code utilized for downloading the sequences from UCSC DAS server

a

```perl
#!/usr/bin/env perl
use strict;
use warnings;
use LWP::Simple;
use XML::XPath;
use XML::XPath::XMLParser;

#Use DAS of UCSC to fetch specific sequence by its given chromosome position
my $URL_gene = $ARGV[0];

my $xml = get($URL_gene);

my $xp = XML::XPath->new(xml=>$xml);

my $nodeset = $xp->find('/DASDNA/SEQUENCE/DNA/text()'); # find all sequences
# remove white spaces
foreach my $node ($nodeset->get_nodelist) {
    my $seq = $node->getValue;
    $seq =~ s/\s//g;
    print $seq, "\n";
}
```

b

```r
# Read splice sites into a matrix
chr1_acceptor_sites = read.table("/Users/gulsah/chr1/chr1_acceptor_sites.txt");
chr1_acceptor_sites_m = as.matrix(chr1_acceptor_sites, ncol=1);

# For each site, call the perl script to download the related sequence from ucsc das and save output in a matrix
sequence_data_acceptor = matrix(, nrow=length(chr1_acceptor_sites_m), ncol=1);
for(i in 1:length(chr1_acceptor_sites_m)){
  chromosome_site = chr1_acceptor_sites_m[i,1];
  print(i)
  sequence_data_acceptor[i,1] = system(paste0("perl /Users/gulsah/sequencedownloder.pl
                                       http://genome.ucsc.edu/cgi-bin/das/hg38/dna?segment=", chromosome_site), intern = TRUE)
}

# Write the whole ouput in a csv file
chr1_acceptor_sequences_m = as.matrix(cbind(chr1_acceptor_sites_m, sequence_data_acceptor), ncol=2)
write.csv(chr1_acceptor_sequences_m, "/Users/gulsah/chr1/chr1_acceptor_site_sequences.csv")
```

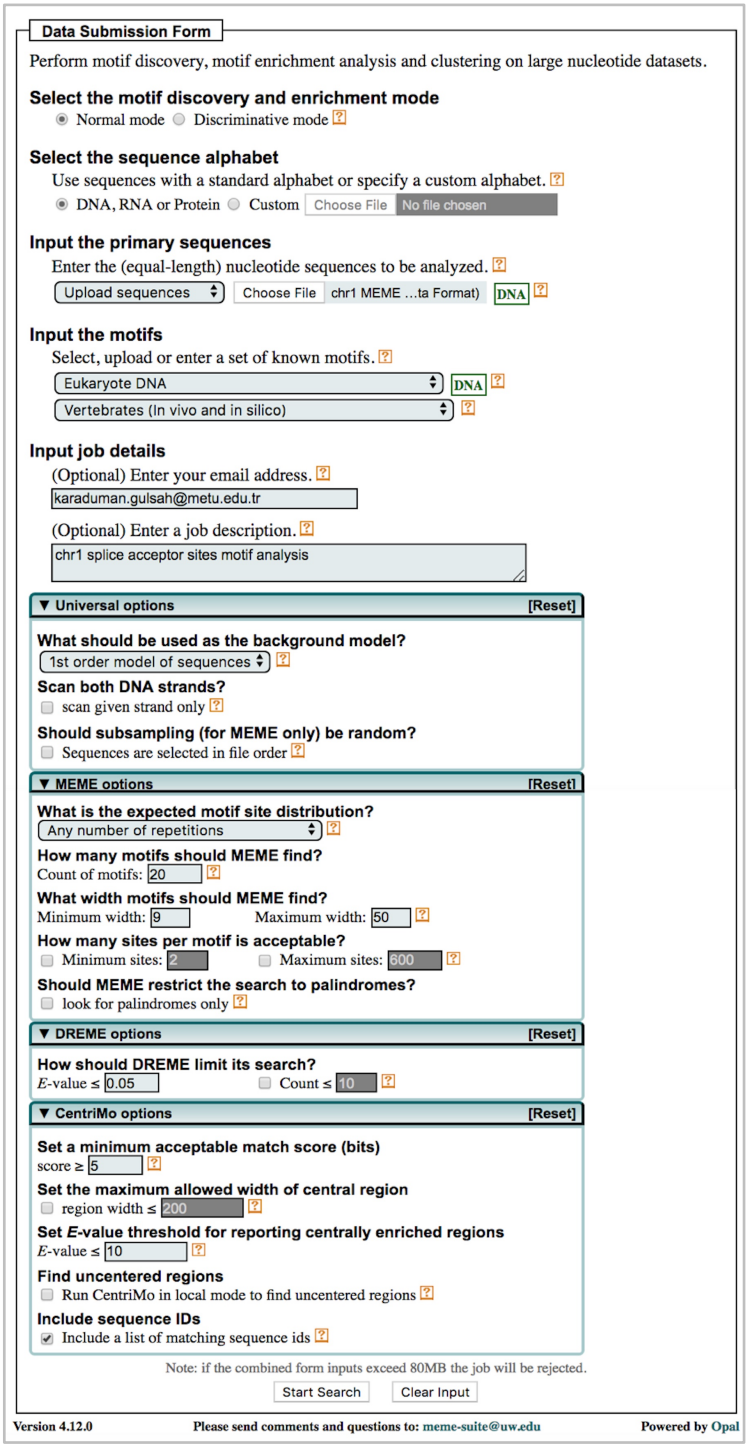
Appendix A - Figure 2: a: Perl script used for downloading sequences from UCSC DAS Server (directly taken from the forum entry reached through https://www.biostars.org/p/6156) b: R script for running Perl script and collecting its output in csv format for further inspection.

R code for DNA sequence reverse complementing

```r
revcomp = function(DNAstr) {
  step1 = chartr("acgt","tgca",DNAstr)
  step2 = unlist(strsplit(step1, split=""))
  step3 = rev(step2)
  step4 = paste(step3, collapse="")
  return(step4)
}

chr1_acceptor_site_seqs = read.table("chr1_acceptor_site_sequences.txt");
chr1_acceptor_sites_m = as.matrix(chr1_acceptor_site_seqs, ncol=4);

for(i in 1:length(chr1_acceptor_sites_m[,1])){
  if(chr1_acceptor_sites_m[i,3] == "-1") {
    chr1_acceptor_sites_m[i,4] = revcomp(chr1_acceptor_sites_m[i,4]);
  }
}

write.csv(chr1_acceptor_sites_m, "chr1_acceptor_site_sequences_rev_comp.csv");
```

Appendix A - Figure 3: DNA sequence reverse complementing R code snippet.

MEME ChIP run configuration



Appendix A - Figure 4: MEME ChIP run with default configuration except for MEME which is arranged to find 20 motifs with 9 to 50 nucleotides long.

APPENDIX B

SCRIPTS FOR LOCATIONAL MOTIF ANALYSIS

Motif locations and queries for getting motif occurrences on specified locations

```
-- Region 1
SELECT motif_id, count(*) FROM motif_locations WHERE start <= 50 -
(char_length(motif_id) / 2) GROUP BY motif_id;

-- Region 2
SELECT motif_id, count(*) FROM motif_locations  WHERE start > 50 -
(char_length(motif_id) / 2) AND start <= 100 - (char_length(motif_id) / 2)
GROUP BY motif_id;

-- Region 3
SELECT motif_id, count(*) FROM motif_locations WHERE start > 100 -
(char_length(motif_id) / 2) AND start <= 150 - (char_length(motif_id) / 2)
GROUP BY motif_id;

-- Region 4
SELECT motif_id, count(*) FROM motif_locations WHERE start > 150 -
(char_length(motif_id) / 2) AND start <= 200 - (char_length(motif_id) / 2)
GROUP BY motif_id;

-- Region 5
SELECT motif_id, count(*) FROM motif_locations WHERE start > 200 -
(char_length(motif_id) / 2) AND start <= 250 - (char_length(motif_id) / 2)
GROUP BY motif_id;

-- Region 6
SELECT motif_id, count(*) FROM motif_locations WHERE start > 250 -
(char_length(motif_id) / 2) AND start <= 300 - (char_length(motif_id) / 2)
GROUP BY motif_id;

--Region 7
SELECT motif_id, count(*) FROM motif_locations WHERE start > 300 -
(char_length(motif_id) / 2) AND start <= 350 - (char_length(motif_id) / 2)
GROUP BY motif_id;

--Region 8
SELECT motif_id, count(*) FROM motif_locations WHERE start > 350 -
(char_length(motif_id) / 2) AND start <= 400 GROUP BY motif_id;
```

Appendix B - Figure 1: SQL queries used for getting motif occurrences on the 8 regions around the splice acceptor sites.

	id	motif_id	start	real_start	stop	real_stop	strand	p-value
1	1	MA0528.1	356	52361541	376	52361561	+	0.000000000000595000
2	2	MA0528.1	317	10642286	337	10642306	-	0.000000000002870000
3	3	MA0528.1	9	201874860	29	201874880	+	0.000000000002870000
4	4	MA0528.1	329	201874860	349	201874880	+	0.000000000002870000
5	5	MA0528.1	335	1641766	355	1641786	+	0.000000000002870000
6	6	MA0528.1	335	1708898	355	1708918	+	0.000000000002870000
7	7	MA0528.1	359	52361538	379	52361558	+	0.000000000004020000
8	8	MA0528.1	332	1708901	352	1708921	+	0.000000000007890000
9	9	MA0528.1	332	1641769	352	1641789	+	0.000000000007890000
10	10	MA0528.1	233	20342364	253	20342384	-	0.000000000008290000

Appendix B - Figure 2: A snapshot from “motif_locations” table which stores FIMO output.

APPENDIX C

SCRIPTS FOR CO-LOCATING DISEASE RELATED VARIANTS WITH SIGNIFICANT MOTIFS

Storing dbSNP data in the local database

```
create table snp150_ucsc(

       bin numeric(5) not null,
       chrom varchar(31) not null,
       chromstart numeric(10) not null,
       chromend numeric(10) not null,
       name varchar(15) not null,
       score numeric(5) not null,
       strand varchar(1) not null,
       refncbi bytea not null,
       refucsc bytea not null,
       observed varchar(255) not null,
       moltype varchar not null,
       class varchar(1000) not null,
       valid varchar(1000) not null,
       avhet double precision not null,
       avhetse double precision not null,
       func varchar(1000) not null,
       loctype varchar(1000) not null,
       weight numeric(10) not null,
       exceptions varchar(1000) not null,
       submittercount numeric(5) not null,
       submitters bytea not null,
       allelefreqcount numeric(5) not null,
       alleles bytea not null,
       allelens bytea not null,
       allelefreqs bytea not null,
       bitfields varchar(1000) not null

);

create index snp150_ucsc_idx_1 on snp150_ucsc (chromstart);
create index snp150_ucsc_idx_2 on snp150_ucsc (chromend);
create index snp150_ucsc_idx_3 on snp150_ucsc (name);
create index snp150_ucsc_idx_4 on snp150_ucsc (chrom);

COPY snp150_ucsc FROM '/Users/gulsah/snps/snp150.txt' DELIMITER '  ';
```

Appendix C - Figure 1: Table creation code for storing dbSNP data into local PostgreSQL database. Data is loaded from the text file downloaded from UCSC downloads page (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp150.txt.gz).

Storing subset of SNPs located 400 nucleotides around the splice acceptor site

```
CREATE TABLE acceptor_site_locations(

  chromosome text,
  start_loc numeric,
  end_loc numeric

);

CREATE TABLE acceptor_site_all_snps(

  chromosome text,
  chrstart numeric,
  chrend numeric,
  rsid text,
  "alleleFreqCount" numeric,
  "allelesList" bytea,
  "alleleFreqsList" bytea,
  snp_func text,
  strand varchar(1)

);

INSERT INTO acceptor_site_all_snps
(chromosome, chrstart, chrend, rsid, "alleleFreqCount", "allelesList",
"alleleFreqsList", snp_func)
  (SELECT DISTINCT snp150_ucsc.chrom,
                   snp150_ucsc.chromStart,
                   snp150_ucsc.chromEnd,
                   snp150_ucsc.name,
                   snp150_ucsc.alleleFreqCount,
                   snp150_ucsc.alleles,
                   snp150_ucsc.alleleFreqs,
                   snp150_ucsc.func
   FROM snp150_ucsc
     INNER JOIN
     acceptor_site_locations
       ON (
       snp150_ucsc.chromStart
       >=
       acceptor_site_locations.start_loc
       AND
       snp150_ucsc.chromEnd
       <=
       acceptor_site_locations.end_loc

));
```

Appendix C - Figure 2: Table creation (acceptor_site_all_snps) for keeping the subset of SNPs located 400 nucleotides around the splice acceptor sites (acceptor_site_locations table is previously loaded with Ensembl Biomart query output).

Finding motif co-locating SNPs

```
CREATE TABLE acceptor_site_snp_on_motif(

       snp_id text not null,
       motif_id varchar(50),
       chr varchar(5) not null,
       id serial not null
              constraint acceptor_site_snp_on_motif_id_pk
                     primary key

);

CREATE TABLE motif_locations(

       id serial not null
              constraint motif_locations_pkey
                     primary key,
       motif_id varchar(16) not null,
       chr varchar(5) not null,
       start integer,
       stop integer

);

INSERT INTO acceptor_site_snp_on_motif (snp_id, motif_id, chr)
    SELECT
        distinct snps.rsid,
        ml.motif_id,
        chr
    FROM motif_locations ml INNER JOIN
        acceptor_site_all_snps snps
            ON (ml.chr = snps.chromosome AND
             ml.start <= snps.chrstart AND
             ml.stop >= snps.chrend);
```

Appendix C - Figure 3: SQL code snippets for finding SNPs located on the significant motifs. “acceptor_site_snp_on_motif” table is actually a list of motif and SNP pairs indicating that the SNP is located on the corresponding motif. Locations of the motifs retrieved from FIMO are used for finding the matching SNPs.

Tagging the SNPs for disease relevance

```
ALTER TABLE acceptor_site_snp_on_motif ADD has_dbgap_match BOOLEAN NULL;
UPDATE acceptor_site_snp_on_motif
SET has_dbgap_match = TRUE
WHERE
  exists(SELECT *
         FROM dbgap dg
         WHERE dg.rsid = acceptor_site_snp_on_motif.snp_id);

ALTER TABLE acceptor_site_snp_on_motif ADD has_gwas_match BOOLEAN NULL;
UPDATE acceptor_site_snp_on_motif
SET has_gwas_match = TRUE
WHERE
  exists(SELECT *
         FROM gwas_catalog gc
         WHERE gc.snp_id_current = acceptor_site_snp_on_motif.snp_id);

ALTER TABLE acceptor_site_snp_on_motif ADD has_tcga_match BOOLEAN NULL;
UPDATE acceptor_site_snp_on_motif
SET has_tcga_match = TRUE
WHERE
  exists(SELECT *
         FROM tcga tcga
         WHERE tcga.dbsnp_rs = acceptor_site_snp_on_motif.snp_id);



ALTER TABLE acceptor_site_snp_on_motif ADD has_pharmgkb_match BOOLEAN
NULL;
UPDATE acceptor_site_snp_on_motif
SET has_pharmgkb_match = TRUE
WHERE
  exists(SELECT *
         FROM pharmgkb phm
         WHERE phm.entity1_name_rsid = acceptor_site_snp_on_motif.snp_id);

ALTER TABLE acceptor_site_snp_on_motif ADD has_omim_match BOOLEAN NULL;
UPDATE acceptor_site_snp_on_motif
SET has_omim_match = TRUE
WHERE
  exists(SELECT *
         FROM clinvar o
         WHERE o.rsid = acceptor_site_snp_on_motif.snp_id);
```

Appendix C - Figure 4: SQLs for tagging motif co-located SNPs' disease relevance.

APPENDIX D

SCRIPTS FOR FINDING SPLICE ACCEPTOR REGION SNPS' SPLICING EFFECT

SPANR scores table creation

```
create table spidex(

       chromosome text not null,
       position numeric not null,
       ref_allele text not null,
       mut_allele text not null,
       dpsi_max_tissue double precision,
       dpsi_zscore double precision,
       gene text,
       strand text,
       transcript text,
       exon_number numeric,
       location text,
       cds_type text,
       ss_dist numeric,
       "commonSNP_rsid" text,
       constraint spidex_chromosome_position_ref_allele_mut_allele_pk
              primary key (chromosome, position, ref_allele, mut_allele)

);

create index spidex_position_idx on spidex (position);
```

Appendix D - Figure 1: Table creation for storing SPANR scores of splice region variants. This table is populated with SPIDEX data(http://download.openbioinformatics.org/spidex_download_form.php).

Arranging alleles

```
/* Creation of a table for arranging alleles and allele frequencies of the acceptor
SNPs*/
CREATE TABLE acceptor_site_all_snps_allele_list_arranged(
  id serial not null
                constraint acceptor_site_all_snps_allele_list_arranged_id_pk
                        primary key,
        chromosome text,
        chrstart numeric,
        chrend numeric,
        rsid text,
        "alleleFreqCount" numeric,
        "allelesList" bytea,
        "alleleFreqsList" bytea,
        snp_func text,
        strand varchar(1),
        allele1 text,
        allele2 text,
        allele3 text,
        allele4 text,
        ref_allele text,
        allelefreq1 text,
        allelefreq2 text,
        allelefreq3 text,
        allelefreq4 text,
        most_common_allele text,
        dpsi_most_common_allele double precision,
        chrstarthg19 numeric, -- this column is filled with liftover output
        chrendhg19 numeric -- this column is filled with liftover output
);

/* Indels are ignored, only single nuclotide changes are added to the table */
INSERT INTO acceptor_site_all_snps_allele_list_arranged(
  SELECT *
  FROM acceptor_site_all_snps
  WHERE ("allelesList" != '' AND length("allelesList") < 9
        AND split_part(encode("allelesList", 'escape'), ',', 1) != '-')
        AND length(split_part(encode("allelesList", 'escape'), ',', 1)) = 1
);

/* Alleles are arranged into columns */
UPDATE acceptor_site_all_snps_allele_list_arranged
SET allele1 = split_part(encode("allelesList", 'escape'), ',', 1);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allele2 = split_part(encode("allelesList", 'escape'), ',', 2);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allele3 = split_part(encode("allelesList", 'escape'), ',', 3);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allele4 = split_part(encode("allelesList", 'escape'), ',', 4);

/* Allele frequencies are arranged into columns */

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allelefreq1 = split_part(encode("alleleFreqsList", 'escape'), ',', 1);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allelefreq2 = split_part(encode("alleleFreqsList", 'escape'), ',', 2);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allelefreq3 = split_part(encode("alleleFreqsList", 'escape'), ',', 3);

UPDATE acceptor_site_all_snps_allele_list_arranged
SET allelefreq4 = split_part(encode("alleleFreqsList", 'escape'), ',', 4);
```

Appendix D - Figure 2: Queries for arranging alleles of acceptor region SNPs into separate columns.

Reference and most common allele detection

```
/* The allele with the highest frequency is set as the reference allele */
UPDATE acceptor_site_all_snps_allele_list_arranged
SET ref_allele = CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3, allelefreq4)
= allelefreq1
  THEN allele1
                ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3,
allelefreq4) = allelefreq2
                  THEN allele2
                     ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3,
allelefreq4) = allelefreq3
                       THEN allele3
                          ELSE CASE WHEN greatest(allelefreq1, allelefreq2,
allelefreq3, allelefreq4) = allelefreq4
                            THEN allele4 END END END END;

/* Setting the most frequent allele other than reference */
UPDATE acceptor_site_all_snps_allele_list_arranged
SET most_common_allele = CASE WHEN ref_allele = allele1
 THEN CASE WHEN greatest(allelefreq2, allelefreq3, allelefreq4) = allelefreq2
  THEN allele2
    ELSE CASE WHEN greatest(allelefreq2, allelefreq3, allelefreq4) = allelefreq3
     THEN allele3
      ELSE CASE WHEN greatest(allelefreq2, allelefreq3, allelefreq4) = allelefreq4
       THEN allele4
         END END END
            ELSE CASE WHEN ref_allele = allele2
             THEN CASE WHEN greatest(allelefreq1, allelefreq3, allelefreq4) =
allelefreq1
              THEN allele1
               ELSE CASE WHEN greatest(allelefreq1, allelefreq3, allelefreq4) =
allelefreq3
                THEN allele3
                  ELSE CASE WHEN greatest(allelefreq1, allelefreq3, allelefreq4) =
allelefreq4
                   THEN allele4
                    END END END
              ELSE CASE WHEN ref_allele = allele3
               THEN CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq4) =
allelefreq1
                THEN allele1
                  ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq4) =
allelefreq2
                   THEN allele2
                    ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq4) =
allelefreq4
                      THEN allele4
                        END END END
                 ELSE CASE WHEN ref_allele = allele4
                  THEN CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3) =
allelefreq1
                   THEN allele1
                    ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3) =
allelefreq2
                     THEN allele2
                       ELSE CASE WHEN greatest(allelefreq1, allelefreq2, allelefreq3) =
                              allelefreq3
                         THEN allele3
                          END END END
                   END
                  END
                END
              END;
```

Appendix D - Figure 3: Queries for detecting the reference and most common alleles of acceptor region SNPs. The allele with greatest allele frequency indicates the reference allele which refers to the nucleotide base on the reference assembly at the SNP's position. Then, the second greatest indicates the most common allele change.

Allele mapping to positive strand

```
/* Addition of two new columns for converting negative strand alleles to their positive
strand equivalents */
ALTER TABLE public.acceptor_site_all_snps_allele_list_arranged ADD
strand_arranged_ref_allele VARCHAR(1) NULL;
ALTER TABLE public.acceptor_site_all_snps_allele_list_arranged ADD
strand_arranged_most_common_allele VARCHAR(1) NULL;

/* Strand arranged allele setting for reference and most common alleles */
UPDATE acceptor_site_all_snps_allele_list_arranged
SET strand_arranged_ref_allele = CASE WHEN
  strand = '+'
  THEN ref_allele
                              ELSE CASE WHEN ref_allele = 'A'
                                THEN 'T'
                                   ELSE CASE WHEN ref_allele = 'C'
                                     THEN 'G'
                                        ELSE CASE WHEN ref_allele = 'G'
                                          THEN 'C'
                                             ELSE CASE WHEN ref_allele = 'T'
                                               THEN 'A' END END END END END;

UPDATE acceptor_site_all_snps_allele_list_arranged
SET strand_arranged_most_common_allele = CASE WHEN
  strand = '+'
  THEN most_common_allele
                              ELSE CASE WHEN most_common_allele = 'A'
                                THEN 'T'
                                   ELSE CASE WHEN most_common_allele = 'C'
                                     THEN 'G'
                                        ELSE CASE WHEN most_common_allele = 'G'
                                          THEN 'C'
                                             ELSE CASE WHEN most_common_allele = 'T'
                                               THEN 'A' END END END END END;
```

Appendix D - Figure 4: Queries for mapping alleles of acceptor region SNPs to positive strand.

Setting SNP scores

```
/* Arrangement of splicing effect score for the most common allele change */
UPDATE acceptor_site_all_snps_allele_list_arranged
SET dpsi_most_common_allele =
   (SELECT spidex.dpsi_max_tissue
    FROM spidex
        WHERE spidex.chromosome =
        acceptor_site_all_snps_allele_list_arranged.chromosome AND
        spidex.position =
        acceptor_site_all_snps_allele_list_arranged.chrEndHG19 AND
        spidex.ref_allele =
        acceptor_site_all_snps_allele_list_arranged.strand_arranged_ref_allele AND
        spidex.mut_allele =
        acceptor_site_all_snps_allele_list_arranged.strand_arranged_most_common_allele
);
```

Appendix D - Figure 5: Query for setting SPANR scores of acceptor region SNPs.

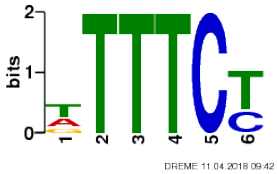
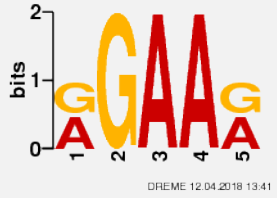
APPENDIX E

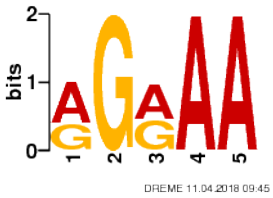
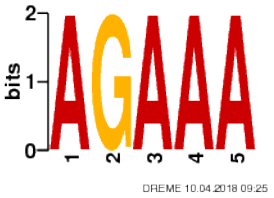
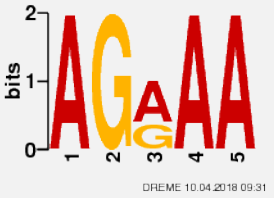
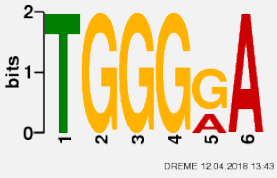
LIST OF SIGNIFICANT MOTIFS FOUND BY ACCEPTOR SITE DNA SEQUENCES

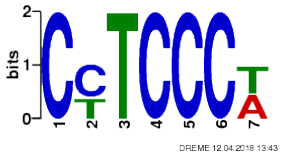
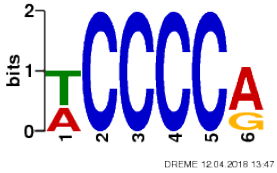
Appendix E - Table 1: List of previously known motifs.

1. Alx1_DBD_1	78. FOXO4_DBD_2	155. MA0050.2	232. NKX6-2_DBD	309. UP00024_1	386. UP00155_1
2. Alx1_DBD_2	79. FOXO4_DBD_3	156. MA0056.1	233. NKX6-2_full	310. UP00024_2	387. UP00157_1
3. ALX3_DBD	80. FOXO6_DBD_2	157. MA0057.1	234. NOTO_DBD	311. UP00025_1	388. UP00158_1
4. ALX3_full_1	81. FOXP3_DBD	158. MA0063.1	235. Nr2e1_DBD_1	312. UP00025_2	389. UP00164_1
5. ALX3_full_2	82. GBX1_DBD	159. MA0065.2	236. NR2E1_full_1	313. UP00028_2	390. UP00164_2
6. Alx4_DBD	83. Gbx1_DBD	160. MA0079.3	237. ONECUT3_DBD	314. UP00029_1	391. UP00166_1
7. ALX4_DBD	84. Gbx2_DBD	161. MA0084.1	238. OTX2_DBD_2	315. UP00030_1	392. UP00167_1
8. ARX_DBD	85. GBX2_DBD_1	162. MA0087.1	239. PAX4_DBD	316. UP00030_2	393. UP00168_1
9. Arx_DBD	86. GBX2_DBD_2	163. MA0108.2	240. PAX4_full	317. UP00032_1	394. UP00169_1
10. Barhl1_DBD_1	87. GSC2_DBD	164. MA0109.1	241. PAX7_DBD	318. UP00033_2	395. UP00170_1
11. Barhl1_DBD_2	88. GSX1_DBD	165. MA0125.1	242. PDX1_DBD_1	319. UP00034_1	396. UP00171_1
12. Barhl1_DBD_3	89. GSX2_DBD	166. MA0135.1	243. PDX1_DBD_2	320. UP00037_1	397. UP00172_1
13. BARHL2_DBD_1	90. HESX1_DBD_1	167. MA0142.1	244. PHOX2A_DBD	321. UP00039_1	398. UP00173_1
14. BARHL2_DBD_2	91. HMBOX1_DBD	168. MA0148.3	245. PHOX2B_DBD	322. UP00039_2	399. UP00174_1
15. BARHL2_DBD_3	92. HMX1_DBD	169. MA0151.1	246. PHOX2B_full	323. UP00041_1	400. UP00175_1
16. BARHL2_full_1	93. HMX2_DBD	170. MA0152.1	247. PITX1_DBD	324. UP00043_2	401. UP00179_1
17. BARHL2_full_2	94. HNF1A_full	171. MA0155.1	248. PITX1_full_1	325. UP00045_2	402. UP00180_1
18. BARHL2_full_3	95. HNF1B_full_1	172. MA0158.1	249. PITX3_DBD	326. UP00047_1	403. UP00182_1
19. BARX1_DBD_1	96. HNF1B_full_2	173. MA0442.2	250. POU1F1_DBD_1	327. UP00051_1	404. UP00187_1
20. BARX1_DBD_2	97. HOMEZ_DBD	174. MA0471.1	251. POU1F1_DBD_2	328. UP00054_2	405. UP00188_1
21. BSX_DBD	98. HOXA1_DBD	175. MA0480.1	252. POU2F1_DBD_1	329. UP00058_2	406. UP00189_1
22. CART1_DBD	99. HOXA10_DBD_1	176. MA0481.2	253. POU2F1_DBD_2	330. UP00059_1	407. UP00191_1
23. CDX1_DBD	100. HOXA10_DBD_2	177. MA0485.1	254. POU2F2_DBD_1	331. UP00061_1	408. UP00194_1
24. CDX2_DBD	101. Hoxa11_DBD_1	178. MA0493.1	255. POU2F2_DBD_2	332. UP00061_2	409. UP00196_1
25. CPEB1_full	102. Hoxa11_DBD_2	179. MA0497.1	256. Pou2f2_DBD_2	333. UP00062_1	410. UP00197_1
26. DLX1_DBD	103. HOXA13_DBD_1	180. MA0504.1	257. POU2F3_DBD_1	334. UP00063_2	411. UP00200_1
27. Dlx1_DBD	104. HOXA13_full_1	181. MA0507.1	258. POU2F3_DBD_2	335. UP00064_1	412. UP00200_2
28. Dlx2_DBD	105. HOXA13_full_2	182. MA0514.1	259. POU3F1_DBD_1	336. UP00069_1	413. UP00202_1
29. DLX2_DBD	106. Hoxa2_DBD	183. MA0515.1	260. POU3F1_DBD_2	337. UP00070_2	414. UP00206_1
30. DLX3_DBD	107. HOXB13_DBD_1	184. MA0516.1	261. POU3F2_DBD_1	338. UP00071_1	415. UP00207_1
31. DLX4_DBD	108. HOXB2_DBD	185. MA0517.1	262. POU3F2_DBD_2	339. UP00073_1	416. UP00208_2
32. DLX5_FL	109. HOXB3_DBD	186. MA0528.1	263. POU3F3_DBD_1	340. UP00073_2	417. UP00209_1
33. DLX6_DBD	110. HOXB5_DBD	187. MA0593.1	264. POU3F3_DBD_2	341. UP00074_1	418. UP00209_2
34. DRGX_DBD	111. HOXC10_DBD_1	188. MA0599.1	265. POU3F3_DBD_3	342. UP00074_2	419. UP00211_1
35. EGR4_DBD_1	112. Hoxc10_DBD_1	189. MA0601.1	266. POU3F4_DBD_1	343. UP00075_1	420. UP00212_1
36. EMX1_DBD_1	113. Hoxc10_DBD_2	190. MA0602.1	267. POU3F4_DBD_2	344. UP00077_1	421. UP00213_1
37. EMX1_DBD_2	114. HOXC10_DBD_2	191. MA0606.1	268. POU4F1_DBD	345. UP00077_2	422. UP00214_1
38. EMX2_DBD_1	115. HOXC10_DBD_3	192. MA0612.1	269. POU4F2_DBD	346. UP00078_1	423. UP00215_1
39. EMX2_DBD_2	116. HOXC11_DBD_1	193. MA0614.1	270. POU4F3_DBD	347. UP00079_2	424. UP00217_1
40. EN1_full_1	117. HOXC11_DBD_2	194. MA0618.1	271. POU5F1P1_DBD_1	348. UP00082_2	425. UP00218_1
41. EN1_full_2	118. HOXC11_full_2	195. MA0619.1	272. POU5F1P1_DBD_2	349. UP00086_1	426. UP00221_1
42. EN2_DBD	119. HOXC12_DBD_1	196. MA0621.1	273. POU6F2_DBD_1	350. UP00086_2	427. UP00224_1
43. En2_DBD	120. HOXC13_DBD_1	197. MA0625.1	274. PRDM1_full	351. UP00089_1	428. UP00225_1
44. ESX1_DBD	121. HOXD11_DBD_1	198. MA0627.1	275. PROP1_DBD	352. UP00090_2	429. UP00229_1
45. ESX1_full	122. HOXD11_DBD_2	199. MA0628.1	276. PROP1_full	353. UP00091_1	430. UP00234_1
46. EVX1_DBD	123. HOXD12_DBD_1	200. MA0630.1	277. PRRX1_DBD	354. UP00093_1	431. UP00236_1
47. EVX2_DBD	124. HOXD12_DBD_4	201. MA0648.1	278. PRRX1_full_1	355. UP00094_2	432. UP00238_1
48. FOXB1_DBD_3	125. Hoxd13_DBD_1	202. MA0722.1	279. PRRX1_full_2	356. UP00096_1	433. UP00240_1
49. FOXB1_full	126. HOXD13_DBD_1	203. MA0733.1	280. Prrx2_DBD	357. UP00096_2	434. UP00241_1
50. FOXC1_DBD_1	127. Hoxd13_DBD_2	204. MA0851.1	281. PRRX2_full	358. UP00097_2	435. UP00242_1
51. Foxc1_DBD_1	128. Hoxd3_DBD	205. MA0852.2	282. RAX_DBD	359. UP00099_2	436. UP00244_1
52. Foxc1_DBD_2	129. HOXD8_DBD	206. MA0853.1	283. RAXL1_DBD	360. UP00101_1	437. UP00248_1
53. FOXC1_DBD_2	130. Hoxd9_DBD_1	207. MA0897.1	284. SHOX_DBD	361. UP00101_2	438. UP00250_1
54. FOXC1_DBD_3	131. Hoxd9_DBD_2	208. MA0898.1	285. SHOX2_DBD	362. UP00105_1	439. UP00251_1
55. FOXC2_DBD_1	132. Hoxd9_DBD_3	209. MA0904.1	286. Shox2_DBD	363. UP00108_1	440. UP00252_1
56. FOXC2_DBD_2	133. IRF7_DBD_1	210. MA0910.1	287. SOX9_DBD	364. UP00113_1	441. UP00254_1
57. FOXC2_DBD_3	134. Irx3_DBD	211. MA1103.1	288. SP1_DBD	365. UP00114_1	442. UP00255_1
58. FOXD2_DBD_1	135. IRX5_DBD	212. MA1104.1	289. SP3_DBD	366. UP00115_1	443. UP00257_1
59. FOXD2_DBD_2	136. ISX_DBD_1	213. MA1107.1	290. SP8_DBD	367. UP00116_1	444. UP00260_1
60. FOXD3_DBD_1	137. ISX_DBD_2	214. MA1115.1	291. TEF_FL	368. UP00118_1	445. UP00262_1
61. FOXD3_DBD_2	138. ISX_full	215. MA1117.1	292. Uncx_DBD_1	369. UP00121_1	446. UP00263_1
62. Foxg1_DBD_3	139. KLF16_DBD	216. MA1125.1	293. UNCX_DBD_1	370. UP00124_1	447. UP00267_1
63. FOXI1_full_1	140. LBX2_DBD_2	217. MA1148.1	294. UNCX_DBD_2	371. UP00125_1	448. UP00390_1
64. FOXJ2_DBD_1	141. LHX2_DBD_1	218. MA1149.1	295. Uncx_DBD_2	372. UP00127_1	449. VAX1_DBD
65. FOXJ2_DBD_2	142. Lhx4_DBD	219. MA1152.1	296. UP00002_1	373. UP00128_1	450. VAX2_DBD
66. FOXJ2_DBD_3	143. LHX6_full_1	220. MEOX1_full	297. UP00004_1	374. UP00129_1	451. VENTX_DBD_1
67. FOXJ3_DBD_1	144. Lhx8_DBD_2	221. Meox2_DBD	298. UP00007_1	375. UP00130_1	452. Vsx1_DBD
68. Foxj3_DBD_2	145. LHX9_DBD_1	222. MEOX2_DBD_1	299. UP00007_2	376. UP00133_1	453. VSX1_DBD
69. Foxj3_DBD_3	146. LHX9_DBD_2	223. MIXL1_full	300. UP00012_1	377. UP00134_1	454. VSX2_DBD
70. FOXJ3_DBD_3	147. LMX1A_DBD	224. MNX1_DBD	301. UP00013_2	378. UP00140_1	455. Zfp740_DBD
71. Foxj3_DBD_4	148. LMX1B_DBD	225. MSX1_DBD_1	302. UP00014_1	379. UP00141_1	456. ZNF740_DBD
72. Foxk1_DBD_2	149. LMX1B_full	226. MSX1_DBD_2	303. UP00016_1	380. UP00142_1	457. ZNF740_full
73. FOXL1_full_1	150. MA0031.1	227. MSX2_DBD_2	304. UP00016_2	381. UP00144_1	
74. FOXL1_full_2	151. MA0036.3	228. Msx3_DBD_2	305. UP00021_1	382. UP00145_1	
75. FOXO1_DBD_1	152. MA0040.1	229. NKX6-1_DBD	306. UP00022_1	383. UP00149_1	
76. FOXO1_DBD_3	153. MA0041.1	230. Nkx6-1_DBD	307. UP00023_1	384. UP00150_1	
77. FOXO3_full_3	154. MA0047.2	231. NKX6-1_full	308. UP00023_2	385. UP00151_1	

Appendix E - Table 2: List of novel motifs detected by DREME algorithm.

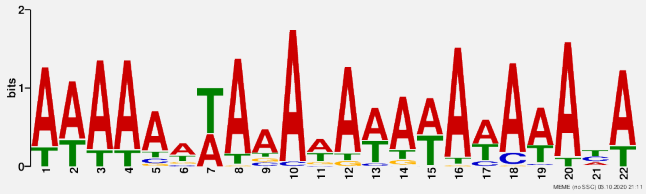
1. AAAAAMAA		**15. DTTTCY**	
2. AAAAHAAA		**16. RAAATA**	
3. AAAAMAAA		**17. RAAATR**	
4. AAARAAAA		**18. RAAAYR**	
5. AAARCA		**19. RGAAA**	
6. AAATRH		**20. RGAAR**	

7. AAATRY		**21. RGRAA**	
8. ACATWY		**22. RGRAAA**	
9. AGAAA		**23. RGRAAR**	
10. AGRAA		**24. RRGAAA**	
11. AGRAAR		**25. TAWATR**	
12. AYATWY		**26. TGGGRA**	

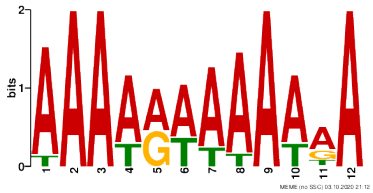
13. CYTCCCW		**27. WCCCCR**	
14. DAAATR			

Appendix E - Table 3: List of novel motifs detected by MEME algorithm.

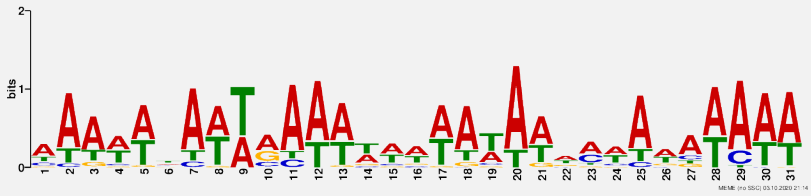
1. AAAAAWWAAAWAWAWAWAAAHA



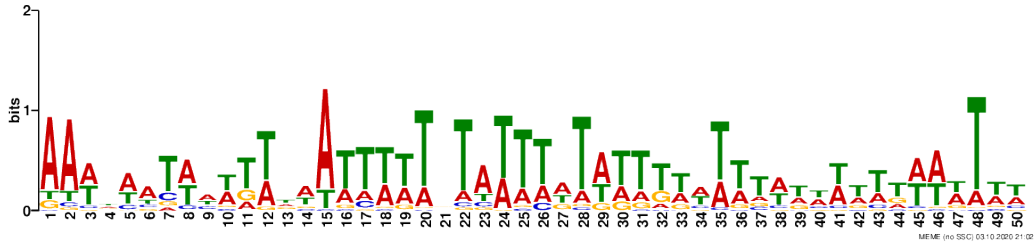
2. AAAARWAAAAAA



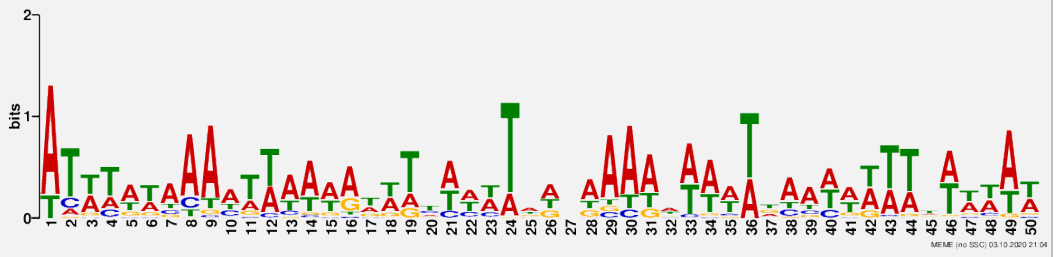
3. AAAWWHAWWRAAWWWWWAWAWWMWAWAWAWA



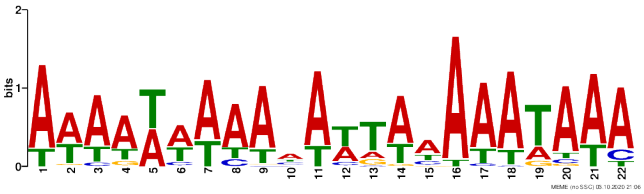
4. AAWDWATWHWTWDWATTWWTNTATTTDTWTTTWWTTTWWWWWTTWWWTWW



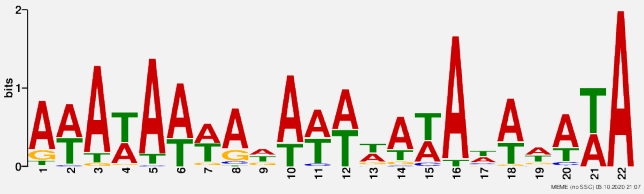
5. ATWTWWAAAATWAAWAWWTHWHWTDDNAAAAWWWWWDAAWWWWTDWWWAW



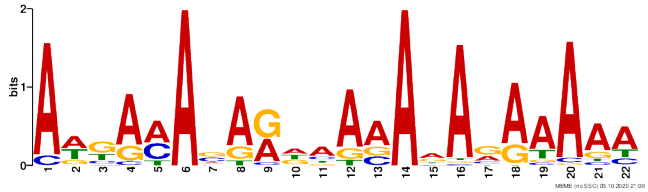
6. AWAAWWAAAHAWTAAAAATAAA



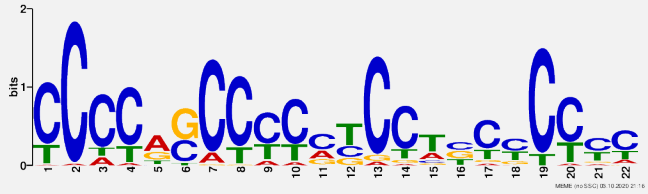
7. AWAWAWWADAWWWAWAWWWAWA



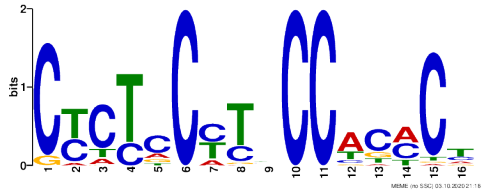
8. AWKAMASARDAAAARARAAAAA



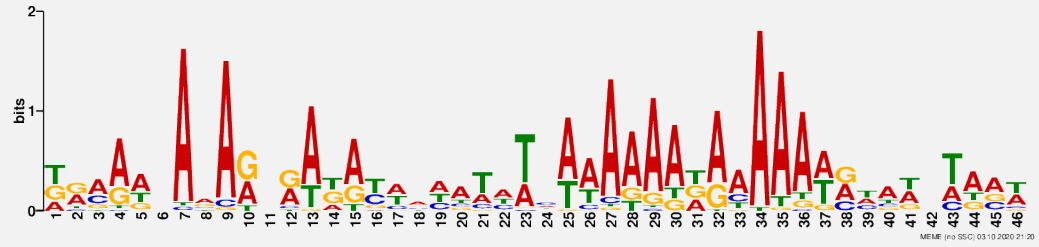
9. CCCCRSCCYCMYCCTBCYCCYC



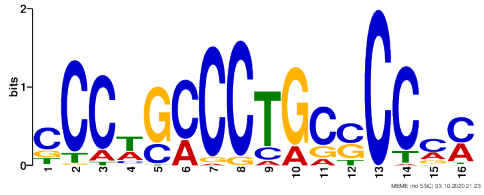
10. CYCTCCYTNCCACMCT



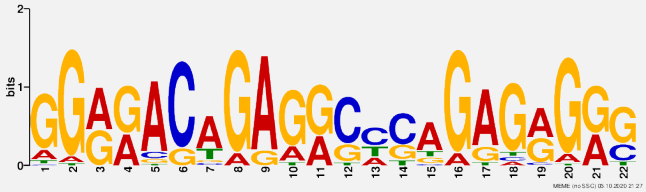
11. DRMAWNARARNRADAYWRHATHTHAAAAAADAAAAAWVHADNWAVW



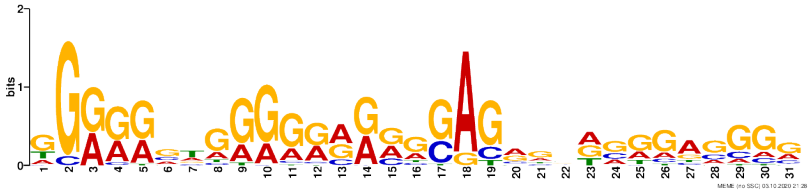
12. GGGCTGGGG



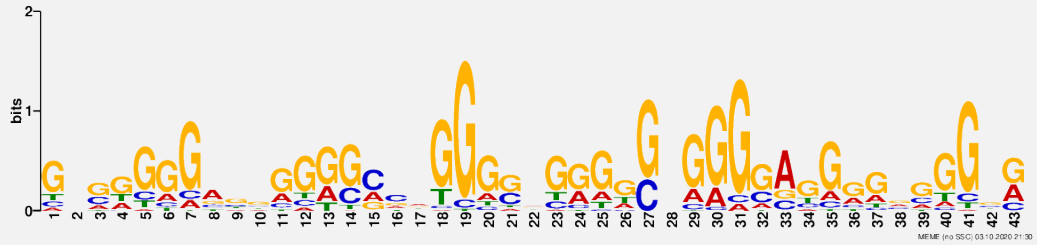
13. GGRRACAGAGRCYCAGAGRGRG



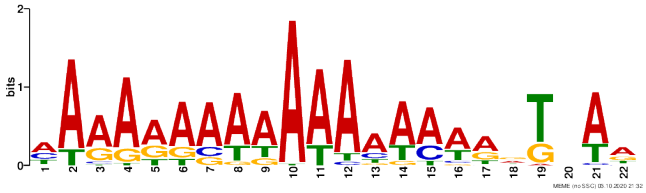
14. GGRRRSWGGGGRRRRGSAGRGBDSGGAGGGR



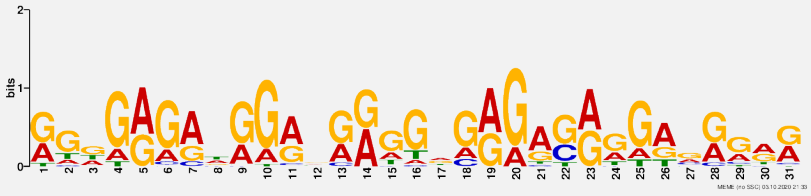
15. GNSGGGGASSGGGGCSDGGGSDGRGGGNGGGGAGGGGVGGGBR



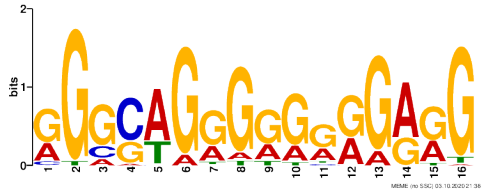
16. MAAARAAAAAAAAAARDTNAR



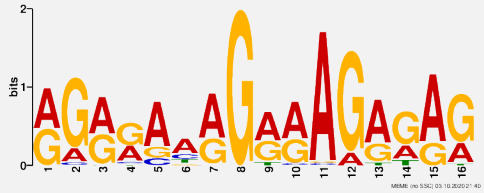
17. RGDGRGAWGGRDRRGGDGRGASRGGARRGRR



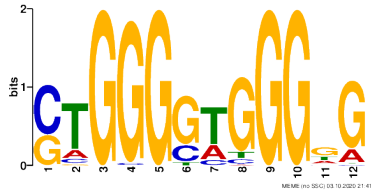
18. RGGSAGGGGGRRGRRG



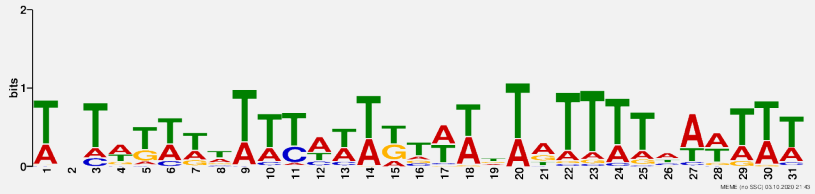
19. RGRRAARGRRAGAGAG



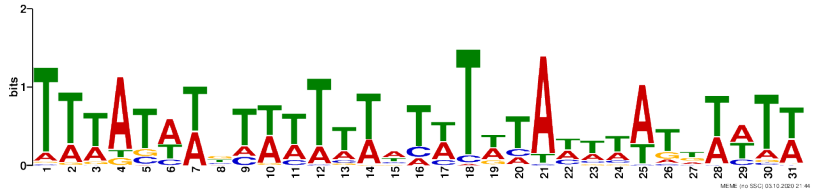
20. STGGGGTGGGKG



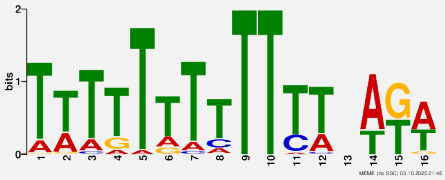
21. TNTWTWWWTTTWWWKTWWDWRTTTTWAWTWT



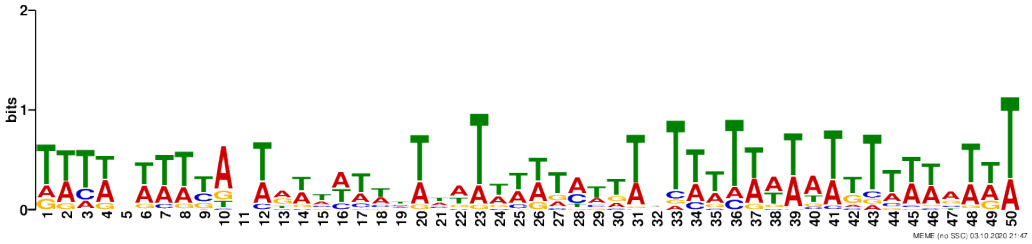
22. TTTATWWBTWWTTWWTWTWTAWWWATKWWTW



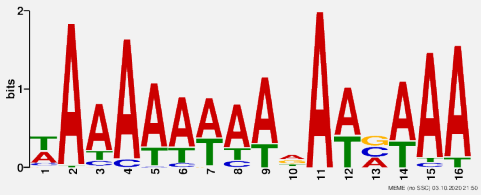
23. TTTTTTTTTTTTNAKW



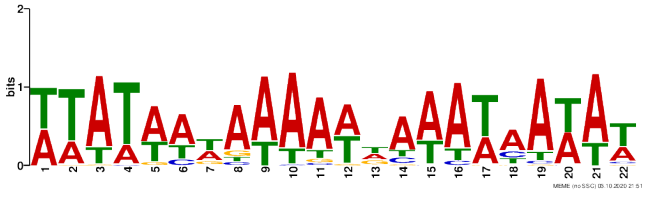
24. TWTWNWWTYANWRWWWTWHTHWTWWWTMTTTNTWWTWWWAWKTTWWAWWT



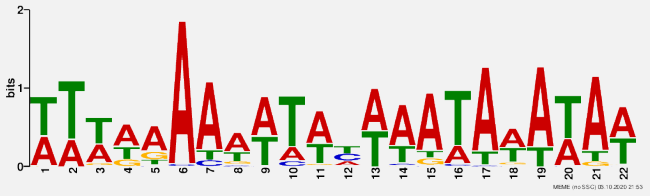
25. WAAAAAWAADAWVAAA



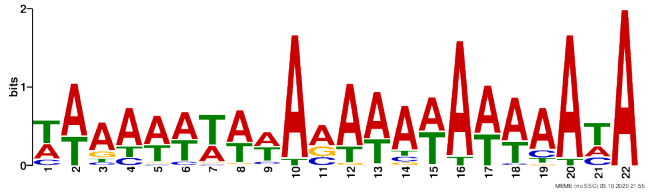
26. WTATWAWAAAAWDAAAWAAWAW



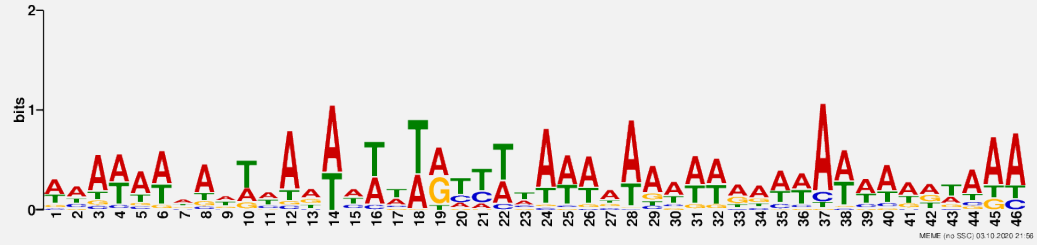
27. WTWWAAAWWTWHWAATAWAWAW



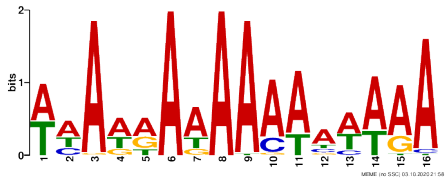
28. WWAAWWWWWARAWAWAWWAATA



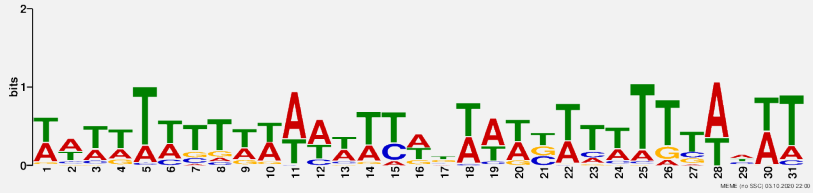
29. WWAWAWWADWHARWHWWWRTYWWAWWAAAWWWRAWWAWWAWDWAAA



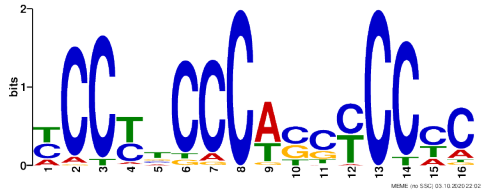
30. WWAWRAAAAAAAWAAA



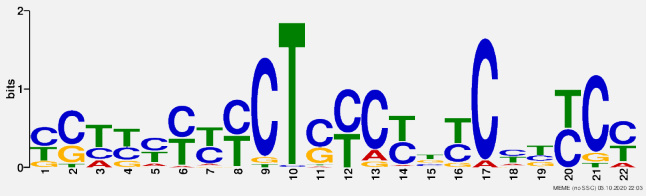
31. WWTWTWTTWWWWWTYWKWWWBWTWTTTWHWT



32. YCCYTCCCASCYCCYC



33. YSYYYYYYCTSYCYBYCYBYCY



APPENDIX F

GENE LISTS FOR HOMO SAPIENS AND MUS MUSCULUS TRANSCRIPTION FACTORS USED IN DAVID FUNCTIONAL ANNOTATION ANALYSIS

Appendix F - Table 1: Gene list for Homo sapiens TFs for which DAVID's functional annotation analysis held.

1. ALX3	28. FOXD2	55. HOXB2	82. SRY	109. PHOX2A	136. OTX2
2. ALX4	29. FOXD3	56. HOXB3	83. TBP	110. RAX2	137. PAX7
3. ARX	30. FOXI1	57. HOXB5	84. PDX1	111. RAX	138. PHOX2B
4. BARHL2	31. FOXJ2	58. HOXC10	85. FOXA1	112. UNCX	139. PITX1
5. BARX1	32. FOXJ3	59. HOXC11	86. NFATC2	113. VAX1	140. PITX3
6. BSX	33. FOXL1	60. HOXC12	87. INSM1	114. VENTX	141. POU1F1
7. ALX1	34. FOXO1	61. HOXC13	88. HOXA5	115. EGR4	142. POU2F1
8. CDX1	35. FOXO3	62. HOXD11	89. SOX10	116. ONECUT3	143. POU2F3
9. CDX2	36. FOXO4	63. HOXD12	90. E2F6	117. POU3F1	144. POU3F2
10. CPEB1	37. FOXO6	64. HOXD13	91. FOXP1	118. POU4F1	145. POU3F3
11. DLX1	38. FOXP3	65. HOXD8	92. MEF2C	119. POU4F3	146. POU3F4
12. DLX2	39. GBX1	66. IRF7	93. NR2C2	120. FOXK1	147. POU4F2
13. DLX3	40. GBX2	67. IRX5	94. POU2F2	121. FOXK2	148. POU5F1B
14. DLX4	41. GSC2	68. ISX	95. SP2	122. GATA6	149. POU6F2
15. DLX5	42. GSX1	69. KLF16	96. STAT1	123. KLF9	150. PRDM1
16. DLX6	43. GSX2	70. LBX2	97. ZNF263	124. POU5F1	151. PROP1
17. DRGX	44. HESX1	71. LHX2	98. FOXP2	125. RELB	152. PRRX1
18. EMX1	45. HMBOX1	72. LHX6	99. KLF5	126. ZNF384	153. PRRX2
19. EMX2	46. HMX1	73. LHX9	100. LBX1	127. PPARA	154. SHOX2
20. EN1	47. HMX2	74. LMX1A	101. NFATC3	128. RARA	155. SOX9
21. EN2	48. HNF1A	75. LMX1B	102. POU6F1	129. SOX15	156. SP3
22. ESX1	49. HNF1B	76. FOXD1	103. SHOX	130. MNX1	157. SP8
23. EVX1	50. HOMEZ	77. GATA2	104. GSC	131. MSX1	158. TEF
24. EVX2	51. HOXA1	78. IRF1	105. MEOX1	132. MSX2	159. VAX2
25. FOXB1	52. HOXA10	79. MZF1	106. MIXL1	133. NKX6-1	160. VSX1
26. FOXC1	53. HOXA13	80. PAX4	107. NKX6-2	134. NOTO	161. VSX2
27. FOXC2	54. HOXB13	81. SP1	108. MEOX2	135. NR2E1	162. ZNF740

Appendix F - Table 2: Gene list for Mus musculus TFs for which DAVID's functional annotation analysis held.

1. Alx1	29. Foxc1	57. Hoxb13	85. Klf1	113. Phox2b	141. Sox4
2. Alx3	30. Foxg1	58. Hoxb4	86. Klf7	114. Pitx2	142. Sox5
3. Alx4	31. Foxj1	59. Hoxb5	87. Lhx1	115. Pou1f1	143. Sox6
4. Arid3a	32. Foxj2	60. Hoxb7	88. Lhx2	116. Pou2f1	144. Sox7
5. Arid3b	33. Foxj3	61. Hoxb8	89. Lhx3	117. Pou2f2	145. Sox8
6. Arid5a	34. Foxk1	62. Hoxb9	90. Lhx4	118. Pou2f3	146. Sp4
7. Arx	35. Foxl1	63. Hoxc10	91. Lhx5	119. Pou3f1	147. Srf
8. Ascl2	36. Foxo1	64. Hoxc13	92. Lhx8	120. Pou3f2	148. Sry
9. Barhl1	37. Gabpa	65. Hoxc4	93. Lhx9	121. Pou3f3	149. Tcf1
10. Barhl2	38. Gata3	66. Hoxc5	94. Lmx1a	122. Pou3f4	150. Tcf3
11. Barx2	39. Gbx1	67. Hoxc6	95. Lmx1b	123. Pou4f3	151. Tcf7
12. Bbx	40. Gbx2	68. Hoxc8	96. Mafb	124. Prop1	152. Tcfap2e
13. Bcl6b	41. Gcm1	69. Hoxc9	97. Meox2	125. Prrx2	153. Tlx2
14. Cdx1	42. Glis2	70. Hoxd1	98. Msx1	126. Rhox6	154. Uncx
15. Cdx2	43. Gsh2	71. Hoxd10	99. Msx3	127. Rxra	155. Uncx4.1
16. Dbx1	44. Hmbox1	72. Hoxd13	100. Mtf1	128. Shox2	156. Vax1
17. Dbx2	45. Hmx2	73. Hoxd3	101. Nfatc2	129. Sox1	157. Vsx1
18. Dlx1	46. Hmx3	74. Hoxd8	102. Nkx2-5	130. Sox11	158. Zbtb7b
19. Dlx2	47. Homez	75. Hoxd9	103. Nkx6-1	131. Sox12	159. Zfp105
20. Dlx3	48. Hoxa10	76. Pdx1	104. Nkx6-3	132. Sox13	160. Zfp128
21. Dlx4	49. Hoxa11	77. Irf3	105. Nobox	133. Sox14	161. Zfp187
22. Egr1	50. Hoxa2	78. Irx2	106. Nr2e1	134. Sox15	162. Zfp281
23. Elf3	51. Hoxa3	79. Irx3	107. Obox5	135. Sox17	163. Zfp410
24. En1	52. Hoxa4	80. Irx4	108. Otx1	136. Sox18	164. Zfp740
25. En2	53. Hoxa5	81. Irx5	109. Otx2	137. Sox2	
26. Esrra	54. Hoxa6	82. Irx6	110. Pax6	138. Sox21	
27. Esx1	55. Hoxa7	83. Isgf3g	111. Pax7	139. Sox3	
28. Foxa2	56. Hoxa9	84. Isl2	112. Phox2a	140. Sox30	

APPENDIX G

LIST OF SPLICE ALTERING VARIANTS WHICH CO-LOCATE WITH MOTIFS

Appendix G - Table 1: Experimentally proven acceptor site splice altering variants set of SpliceAI that co-locate with the significant motifs found by this study. Locations of the variants and number of co-located motifs are presented.

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr1	1486104	0	2	2
chr1	9942715	0	5	5
chr1	11807123	0	1	1
chr1	17637891	0	3	3
chr1	20350764	1	0	1
chr1	24082691	2	2	4
chr1	26853438	1	2	3
chr1	27350799	1	0	1
chr1	28274674	0	1	1
chr1	32768592	5	5	10
chr1	46196862	0	1	1
chr1	54679839	0	2	2
chr1	63533438	6	7	13
chr1	66958835	0	29	29
chr1	92289110	0	21	21
chr1	108929693	0	2	2
chr1	109280915	1	0	1
chr1	109669593	2	7	9
chr1	110363926	0	2	2
chr1	153340595	0	2	2
chr1	153563547	0	10	10
chr1	154435956	2	9	11
chr1	158976709	0	1	1
chr1	160139911	0	2	2
chr1	160195458	0	6	6
chr1	197122305	0	2	2
chr1	224326470	1	2	3
chr1	225839874	1	0	1
chr1	228336230	1	0	1
chr1	228376010	0	1	1
chr1	230990117	0	2	2
chr1	235458763	0	1	1
chr1	246843930	0	2	2
chr1	246843930	0	2	2
chr2	8682633	0	1	1
chr2	27685018	0	8	8
chr2	27858572	0	5	5
chr2	27858572	0	5	5
chr2	46588129	0	2	2
chr2	53874403	0	9	9
chr2	60921756	1	0	1
chr2	61873109	0	1	1
chr2	84797535	0	2	2
chr2	85649016	0	1	1
chr2	95275841	0	12	12
chr2	96277240	0	4	4
chr2	112586193	1	4	5
chr2	127419885	0	4	4
chr2	127565278	1	0	1
chr2	142957586	1	2	3
chr2	165924678	0	2	2
chr2	166409937	0	3	3

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr2	169128828	0	2	2
chr2	170954647	0	7	7
chr2	178621751	1	3	4
chr2	197540297	0	2	2
chr2	206751028	0	16	16
chr2	216661830	0	2	2
chr2	218658861	0	4	4
chr2	232330571	0	2	2
chr2	237534564	0	1	1
chr2	237751178	0	1	1
chr2	241051736	0	1	1
chr2	241441319	0	1	1
chr3	37629188	1	1	2
chr3	44932942	0	1	1
chr3	46449034	0	1	1
chr3	46673388	0	2	2
chr3	46673388	0	2	2
chr3	47999158	0	1	1
chr3	51958277	0	6	6
chr3	123368063	0	2	2
chr3	123931507	0	2	2
chr3	129434139	0	4	4
chr3	131133858	0	1	1
chr3	132559063	0	1	1
chr3	183985471	3	3	6
chr3	184142275	1	0	1
chr3	185511738	0	4	4
chr4	681884	2	9	11
chr4	681884	2	9	11
chr4	2932155	0	6	6
chr4	38931415	0	17	17
chr4	47536718	0	2	2
chr4	47678064	0	2	2
chr4	56354827	0	1	1
chr4	75914364	0	1	1
chr4	83302273	0	1	1
chr4	100398505	0	3	3
chr4	109764664	0	1	1
chr4	119501236	0	5	5
chr4	155919826	0	5	5
chr4	185146579	0	54	54
chr4	185651827	0	1	1
chr5	748513	1	0	1
chr5	37142481	0	1	1
chr5	37142481	0	1	1
chr5	39122312	0	6	6
chr5	75581136	0	1	1
chr5	96730792	0	1	1
chr5	96762252	0	21	21
chr5	112205515	6	5	11
chr5	157734602	1	0	1
chr5	177095153	0	3	3
chr5	177605033	2	8	10
chr6	3129074	0	4	4
chr6	10724583	1	2	3
chr6	25770381	0	2	2
chr6	30891971	0	5	5
chr6	31411155	0	1	1
chr6	32062484	0	1	1
chr6	32062484	0	1	1
chr6	42232907	1	4	5
chr6	43675629	0	5	5
chr6	44304546	0	2	2
chr6	56603408	0	4	4
chr6	73500567	0	1	1
chr6	99376206	1	0	1
chr6	99376206	1	0	1
chr6	105281864	0	1	1
chr6	109449494	0	1	1
chr6	123278360	0	1	1
chr6	135388010	0	1	1
chr6	135428722	0	1	1
chr6	151436772	0	2	2
chr6	160060583	0	4	4

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr6	170584412	0	1	1
chr7	1476868	0	4	4
chr7	6588390	0	1	1
chr7	16862669	0	5	5
chr7	16862669	0	5	5
chr7	27794316	0	3	3
chr7	30459213	0	2	2
chr7	33015827	0	1	1
chr7	44115899	0	2	2
chr7	44139630	1	1	2
chr7	44573018	0	2	2
chr7	66994286	0	16	16
chr7	76618302	0	1	1
chr7	94639421	0	1	1
chr7	95367478	0	1	1
chr7	98233872	0	1	1
chr7	101212645	0	8	8
chr7	106254525	6	14	20
chr7	117666878	0	2	2
chr7	117725302	0	1	1
chr7	121267212	0	2	2
chr7	128845982	0	1	1
chr7	140027411	0	1	1
chr7	150340328	0	2	2
chr8	6936877	0	1	1
chr8	17950613	0	1	1
chr8	17964564	0	1	1
chr8	22617401	0	10	10
chr8	23289903	0	2	2
chr8	106745776	1	6	7
chr8	132923320	0	1	1
chr8	142318729	1	2	3
chr9	6255967	0	1	1
chr9	12698431	0	2	2
chr9	14740238	0	11	11
chr9	18889548	0	2	2
chr9	18889548	0	2	2
chr9	32984879	0	2	2
chr9	35834459	0	1	1
chr9	36058819	1	11	12
chr9	74750837	0	6	6
chr9	77221215	0	8	8
chr9	77227362	0	1	1
chr9	100300556	3	7	10
chr9	110370032	0	3	3
chr9	110389597	0	11	11
chr9	113194126	0	1	1
chr9	113284272	0	1	1
chr9	121332402	0	2	2
chr9	127894653	0	6	6
chr9	131518887	0	1	1
chr9	136791987	0	2	2
chr9	136941922	0	1	1
chr9	137049085	0	3	3
chr9	137049085	0	3	3
chr9	137088887	0	2	2
chr9	137349311	0	2	2
chr9	137349311	0	2	2
chr10	29533429	0	1	1
chr10	37856343	0	3	3
chr10	43394691	6	16	22
chr10	77819650	1	4	5
chr10	79394593	0	1	1
chr10	84214067	0	6	6
chr10	88743242	1	4	5
chr10	94338736	0	1	1
chr10	98385186	0	1	1
chr10	98461839	0	1	1
chr10	100219138	0	2	2
chr10	100498805	1	0	1
chr10	102404043	0	1	1
chr10	122600063	0	2	2
chr10	125807469	1	0	1
chr10	132228187	0	2	2

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr10	133288355	0	1	1
chr11	404541	0	1	1
chr11	551129	1	4	5
chr11	620543	0	1	1
chr11	680083	0	1	1
chr11	1840791	0	2	2
chr11	9416904	0	10	10
chr11	9440456	0	1	1
chr11	16748071	0	6	6
chr11	17442934	2	3	5
chr11	19166279	2	5	7
chr11	43889706	0	2	2
chr11	47161434	0	2	2
chr11	62831093	3	19	22
chr11	63915039	0	4	4
chr11	66335685	2	4	6
chr11	67406796	0	4	4
chr11	67586347	0	1	1
chr11	68043395	0	4	4
chr11	71435840	0	3	3
chr11	71444155	0	3	3
chr11	85711248	0	6	6
chr11	108294867	2	5	7
chr11	113364669	0	6	6
chr11	117192227	2	2	4
chr11	117458914	0	2	2
chr11	118532247	0	5	5
chr11	119091414	0	1	1
chr11	119172362	0	1	1
chr11	119188293	0	1	1
chr12	6236078	3	6	9
chr12	7102090	0	1	1
chr12	7108490	1	16	17
chr12	7108490	1	16	17
chr12	7482765	1	0	1
chr12	8857290	1	1	2
chr12	18399649	0	1	1
chr12	21864476	0	1	1
chr12	27671451	0	1	1
chr12	43777631	0	1	1
chr12	49042231	2	3	5
chr12	49657430	1	0	1
chr12	51913575	0	1	1
chr12	53077282	0	2	2
chr12	53504394	0	1	1
chr12	64420332	0	22	22
chr12	94303940	0	1	1
chr12	101662338	0	8	8
chr12	108648836	0	1	1
chr12	109590660	2	2	4
chr12	112987094	0	1	1
chr12	118238195	1	4	5
chr12	123743753	0	1	1
chr13	27273083	1	0	1
chr13	52133242	0	5	5
chr13	77604346	0	5	5
chr13	77617772	0	2	2
chr13	95612178	0	1	1
chr13	95612178	0	1	1
chr13	100515412	5	16	21
chr14	22771029	0	8	8
chr14	23352472	2	6	8
chr14	24211604	1	1	2
chr14	24237061	0	4	4
chr14	45224764	0	1	1
chr14	45224764	0	1	1
chr14	49771577	2	7	9
chr14	51249677	0	12	12
chr14	56603850	0	2	2
chr14	59486637	0	4	4
chr14	61766850	0	1	1
chr14	73264000	0	2	2
chr14	74295211	0	1	1
chr14	75819993	0	1	1

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr14	103727726	0	1	1
chr14	105492612	0	2	2
chr15	22914915	0	2	2
chr15	28255999	0	6	6
chr15	29720706	0	1	1
chr15	42083780	1	0	1
chr15	42083780	1	0	1
chr15	42149915	1	0	1
chr15	42665703	0	2	2
chr15	44570732	0	4	4
chr15	55356377	0	7	7
chr15	61991096	0	3	3
chr15	61991096	0	3	3
chr15	64406075	0	1	1
chr15	65016540	0	1	1
chr15	74343968	0	1	1
chr15	77028376	1	2	3
chr15	81279555	0	1	1
chr16	284920	0	4	4
chr16	1947426	0	15	15
chr16	1953100	3	19	22
chr16	3055800	0	1	1
chr16	3067569	0	1	1
chr16	3140684	1	1	2
chr16	10895659	0	1	1
chr16	12042765	0	1	1
chr16	28177374	0	2	2
chr16	30354014	0	4	4
chr16	46693077	0	3	3
chr16	46693077	0	3	3
chr16	57059377	1	3	4
chr16	67186838	0	4	4
chr16	67832564	0	20	20
chr16	67993225	1	1	2
chr16	69937543	0	1	1
chr16	71881406	0	6	6
chr16	88810148	2	2	4
chr16	88810149	2	2	4
chr16	88810149	2	2	4
chr16	88822675	0	3	3
chr16	89145247	0	3	3
chr16	89585409	0	20	20
chr16	89635923	0	1	1
chr16	89964122	4	36	40
chr17	1465915	1	7	8
chr17	1483966	0	4	4
chr17	3869642	4	14	18
chr17	5533976	0	1	1
chr17	7392990	0	8	8
chr17	28350926	0	5	5
chr17	28676052	1	1	2
chr17	30179094	0	1	1
chr17	32745304	1	6	7
chr17	39470877	0	4	4
chr17	39660138	0	1	1
chr17	43504645	0	1	1
chr17	44350525	0	2	2
chr17	44863010	0	2	2
chr17	47172025	0	2	2
chr17	48180156	0	1	1
chr17	50175349	1	2	3
chr17	57839918	0	1	1
chr17	63482375	1	2	3
chr17	63834375	0	1	1
chr17	69140759	0	4	4
chr17	69140759	0	4	4
chr17	74703126	2	30	32
chr17	74963082	2	0	2
chr17	78170545	1	2	3
chr17	80236931	0	1	1
chr18	9589026	0	1	1
chr18	13124502	0	2	2
chr18	23022125	1	14	15
chr18	23022125	1	14	15

Chromosome	Position	# Novel motifs	# Known motifs	Total # motifs
chr18	31022492	0	9	9
chr19	434848	0	2	2
chr19	2235901	0	8	8
chr19	3110153	0	1	1
chr19	3656638	0	1	1
chr19	4531763	0	3	3
chr19	9828315	2	50	52
chr19	12676141	0	6	6
chr19	12847110	1	1	2
chr19	15055389	1	2	3
chr19	15401066	0	6	6
chr19	17177809	1	0	1
chr19	17725759	0	1	1
chr19	36114715	0	12	12
chr19	38816526	2	4	6
chr19	38816526	2	4	6
chr19	42349511	1	10	11
chr19	45164257	0	6	6
chr19	45778618	0	14	14
chr19	46008164	0	3	3
chr19	46008164	0	3	3
chr19	46160414	0	4	4
chr19	49464161	2	6	8
chr19	49645292	0	1	1
chr19	49884493	0	2	2
chr19	50481882	0	30	30
chr19	55240408	2	7	9
chr20	3672848	0	3	3
chr20	3751293	2	2	4
chr20	17942399	0	2	2
chr20	33395130	0	3	3
chr20	36614393	0	2	2
chr20	38239704	0	1	1
chr20	38907784	0	10	10
chr20	47664201	0	13	13
chr20	62306549	0	3	3
chr20	63287344	0	3	3
chr21	25897659	0	2	2
chr21	33262540	4	8	12
chr21	37164063	0	1	1
chr21	41746399	0	3	3
chr21	43053626	0	5	5
chr21	44145578	0	4	4
chr21	44324487	0	2	2
chr21	44400294	0	2	2
chr21	44418341	0	4	4
chr21	45991998	0	2	2
chr21	46000753	0	2	2
chr21	46000753	0	2	2
chr21	46125387	0	2	2
chr21	46125392	0	2	2
chr22	23697164	0	3	3
chr22	23697164	0	3	3
chr22	35386515	0	2	2
chr22	37295746	0	3	3
chr22	37297557	0	2	2
chr22	37759151	0	2	2
chr22	37759151	0	2	2
chr22	38957330	0	4	4
chr22	41252962	0	7	7
chr22	41533027	2	2	4
chr22	41718057	0	2	2
chr22	44189079	1	2	3
chr22	50627096	1	3	4

APPENDIX H

LIST OF MOTIFS CO-LOCATED WITH DISEASE ASSOCIATED VARIANTS

Appendix H - Table 1: List of significant motifs overlapping with at least one disease associated variant from dbGaP, TCGA, GWAS Catalog, PharmGKB, or ClinVar.

1. AAAAAMAA	51. MA0481.2	101. UP00028_2	151. UP00164_1
2. AAAAHAAA	52. MA0493.1	102. UP00030_1	152. UP00169_1
3. AAAAMAAA	53. MA0497.1	103. UP00032_1	153. UP00171_1
4. AAARAAAA	54. MA0504.1	104. UP00033_2	154. UP00172_1
5. BARHL2_DBD_1	55. MA0514.1	105. UP00034_1	155. UP00175_1
6. BARHL2_DBD_3	56. MA0515.1	106. UP00037_1	156. UP00179_1
7. BARHL2_full_1	57. MA0516.1	107. UP00039_2	157. UP00182_1
8. BARHL2_full_3	58. MA0517.1	108. UP00041_1	158. UP00187_1
9. Barhl1_DBD_3	59. MA0528.1	109. UP00043_2	159. UP00188_1
10. CDX1_DBD	60. MA0599.1	110. UP00045_2	160. UP00189_1
11. CDX2_DBD	61. MA0606.1	111. UP00047_1	161. UP00191_1
12. CPEB1_full	62. MA0625.1	112. UP00051_1	162. UP00200_1
13. CYTCCCW	63. MA0648.1	113. UP00058_2	163. UP00200_2
14. EN1_full_2	64. MA0733.1	114. UP00059_1	164. UP00202_1
15. FOXJ2_DBD_1	65. MA0852.2	115. UP00061_2	165. UP00206_1
16. FOXJ2_DBD_3	66. MA0853.1	116. UP00062_1	166. UP00207_1
17. FOXO1_DBD_3	67. MA0898.1	117. UP00069_1	167. UP00208_2
18. FOXO3_full_3	68. MA1104.1	118. UP00070_2	168. UP00209_1
19. FOXO4_DBD_3	69. MA1107.1	119. UP00073_1	169. UP00209_2
20. Foxj3_DBD_2	70. MA1117.1	120. UP00073_2	170. UP00215_1
21. Foxj3_DBD_4	71. MA1125.1	121. UP00074_1	171. UP00221_1
22. GGGCTGGGG	72. MA1148.1	122. UP00074_2	172. UP00224_1
23. GSC2_DBD	73. MA1149.1	123. UP00075_1	173. UP00225_1
24. HOXA10_DBD_1	74. MA1152.1	124. UP00077_2	174. UP00234_1
25. HOXA13_full_1	75. Nr2e1_DBD_1	125. UP00078_1	175. UP00238_1
26. HOXC10_DBD_3	76. OTX2_DBD_2	126. UP00079_2	176. UP00241_1
27. HOXC11_DBD_1	77. POU1F1_DBD_1	127. UP00082_2	177. UP00242_1
28. HOXC11_DBD_2	78. POU2F3_DBD_2	128. UP00086_1	178. UP00244_1
29. HOXC11_full_2	79. POU3F1_DBD_2	129. UP00086_2	179. UP00248_1
30. HOXC12_DBD_1	80. POU3F3_DBD_2	130. UP00090_2	180. UP00251_1
31. HOXC13_DBD_1	81. PRDM1_full	131. UP00091_1	181. UP00252_1
32. HOXD11_DBD_1	82. Pou2f2_DBD_2	132. UP00093_1	182. UP00257_1
33. HOXD11_DBD_2	83. RGGSAGGGGGRRGRRG	133. UP00096_1	183. UP00262_1
34. Hoxa11_DBD_1	84. SOX9_DBD	134. UP00096_2	184. UP00263_1
35. Hoxc10_DBD_1	85. SP1_DBD	135. UP00097_2	185. WAAAAAWAADAWVAAA
36. KLF16_DBD	86. SP3_DBD	136. UP00099_2	186. WWAWRAAAAAAAWAAA
37. LHX9_DBD_2	87. SP8_DBD	137. UP00101_1	187. ZNF740_DBD
38. MA0041.1	88. STGGGGTGGGKG	138. UP00101_2	188. ZNF740_full
39. MA0050.2	89. UP00002_1	139. UP00105_1	189. Zfp740_DBD
40. MA0057.1	90. UP00007_1	140. UP00108_1	
41. MA0065.2	91. UP00007_2	141. UP00113_1	
42. MA0079.3	92. UP00012_1	142. UP00116_1	
43. MA0109.1	93. UP00013_2	143. UP00124_1	
44. MA0135.1	94. UP00014_1	144. UP00125_1	
45. MA0142.1	95. UP00016_2	145. UP00128_1	
46. MA0148.3	96. UP00021_1	146. UP00129_1	
47. MA0155.1	97. UP00022_1	147. UP00130_1	
48. MA0442.2	98. UP00023_1	148. UP00142_1	
49. MA0471.1	99. UP00024_1	149. UP00144_1	
50. MA0480.1	100. UP00025_2	150. UP00158_1	

APPENDIX I

LIST OF PRIORITIZED MOTIFS ACCORDING TO SPLICE ALTERING EFFECTS OF CO-LOCATED SNPS

Appendix I - Table 1: List of 113 prioritized motifs between Q3 and Q4 according to prioritization measure. This measure suggests higher scores for the motifs with higher co-localization with splice altering SNPs, computed as (#SNPs with |SPANR Score| >= 5) / (#Motif Occurrences).

Motif id	Motif type	Organism	Number of motif occurrences on splice acceptor regions	Number of SNPs with |SPANR Score| >= 5	Number of SNPs with |SPANR Score| < 5	Motif prioritization measure
1.UP00024_1	Known	Mus	23409	1560	47174	0.067
2.UP00032_1	Known	Mus	19277	1175	23888	0.061
3.MA0528.1	Known	Homo sapiens	302666	17529	504561	0.058
4.UP00007_2	Known	Mus	32069	1813	52138	0.057
5.UP00070_2	Known	Mus	16230	902	27737	0.056
6.UP00013_2	Known	Mus	68305	3778	115674	0.055
7.MA1149.1	Known	Homo sapiens	30192	1662	53757	0.055
8.MA1148.1	Known	Homo sapiens	24376	1318	33890	0.054
9.UP00030_1/ UP00069_1	Known	Mus musculus	35192	1752	35483	0.050
10.MA0065.2	Known	Mus	55413	2749	71024	0.050
11.UP00016_2	Known	Mus	9999	496	11427	0.050
12.UP00062_1	Known	Mus	35711	1739	36883	0.049
13.FOXO4_DBD_3	Known	Homo sapiens	26443	1274	27617	0.048
14.UP00077_1	Known	Mus	8921	427	9063	0.048
15.UP00033_2	Known	Mus	64919	3099	127525	0.048
16.MA0050.2	Known	Homo sapiens	104382	4924	117363	0.047
17.UP00047_1	Known	Mus	20086	942	34872	0.047
18.FOXO1_DBD_3	Known	Homo sapiens	35158	1633	36867	0.046
19.UP00079_2	Known	Mus	29351	1362	46725	0.046
20.UP00082_2	Known	Mus	36554	1694	53064	0.046
21.UP00007_1	Known	Mus	42211	1955	84926	0.046
22.UP00002_1	Known	Mus	56906	2581	127429	0.045
23.UP00043_2	Known	Mus	65193	2915	118966	0.045
24.IRF7_DBD_1	Known	Homo sapiens	27983	1243	20543	0.044
25.RGGSAGGGGGRR	Novel	Novel	185467	8151	241132	0.044
26.UP00086_1	Known	Mus	16734	735	14670	0.044
27.UP00101_2	Known	Mus	28616	1253	28214	0.044
28.UP00086_2	Known	Mus	26376	1150	29133	0.044
29.MA0504.1	Known	Homo sapiens	47933	2082	71567	0.043
30.UP00034_1	Known	Mus	22003	954	26591	0.043
31.MA0733.1	Known	Homo sapiens	31441	1353	63228	0.043
32.UP00039_2	Known	Mus	21939	922	20734	0.042
33.UP00022_1	Known	Mus	111335	4613	225018	0.041
34.MA0516.1	Known	Homo sapiens	124033	5127	193809	0.041
35.UP00058_2	Known	Mus	23035	951	21443	0.041
36.UP00099_2	Known	Mus	78468	3194	137467	0.041
37.UP00096_2	Known	Mus	33438	1352	47320	0.040
38.UP00028_2	Known	Mus	76823	3039	56806	0.040

Motif id	Motif type	Organism	Number of motif occurrences on splice acceptor regions	Number of SNPs with |SPANR Score| >= 5	Number of SNPs with |SPANR Score| < 5	Motif prioritization measure
39.PRDM1_full	Known	Homo sapiens	30804	1211	28058	0.039
40.UP00025_1	Known	Mus	28519	1118	25279	0.039
41.MA0155.1	Known	Homo sapiens	44585	1686	51569	0.038
42.UP00039_1	Known	Mus	27826	1052	23843	0.038
43.MA0517.1	Known	Homo sapiens	47912	1811	41215	0.038
44.UP00074_1	Known	Mus	14160	522	11623	0.037
45.MA0852.2	Known	Homo sapiens	15400	565	12015	0.037
46.UP00041_1	Known	Mus	24697	895	20458	0.036
47.UP00263_1	Known	Mus	11868	426	9623	0.036
48.UP00074_2	Known	Mus	19393	684	16244	0.035
49.ZNF740_DBD	Known	Homo sapiens	58544	2059	86691	0.035
50.UP00182_1	Known	Mus	11801	410	9122	0.035
51.MA0851.1	Known	Mus	25953	897	21605	0.035
52.UP00090_2	Known	Mus	63051	2172	50997	0.034
53.Zfp740_DBD	Known	Mus	51672	1751	73312	0.034
54.MA0142.1	Known	Mus	25917	878	19206	0.034
55.HOMEZ_DBD	Known	Homo sapiens	5403	183	4926	0.034
56.MA0480.1	Known	Mus	26536	892	15437	0.034
57.UP00114_1	Known	Mus	7185	241	8170	0.034
58.UP00061_2	Known	Mus	45524	1525	34720	0.033
59.UP00075_1	Known	Mus	16221	537	14914	0.033
60.UP00093_1	Known	Mus	57043	1886	104950	0.033
61.UP00054_2	Known	Mus	13364	440	9742	0.033
62.UP00091_1	Known	Mus	20020	658	16787	0.033
63.STGGGGTGGGKG	Novel	Novel	113354	3725	140375	0.033
64.UP00267_1	Known	Mus	8831	287	8109	0.032
65.UP00021_1	Known	Mus	202244	6529	353073	0.032
66.HOXA13_full_2	Known	Homo sapiens	10387	334	5360	0.032
67.FOXO3_full_3	Known	Homo sapiens	34367	1096	26052	0.032
68.MA0057.1	Known	Homo sapiens	37131	1174	35095	0.032
69.UP00390_1	Known	Mus	6749	213	5282	0.032
70.UP00077_2	Known	Mus	174360	5500	142906	0.032
71.MA0471.1	Known	Homo sapiens	49574	1526	46420	0.031
72.UP00073_1	Known	Mus	31714	971	25464	0.031
73.UP00051_1	Known	Mus	20431	624	17372	0.031
74.Foxj3_DBD_2	Known	Mus	33565	1021	25334	0.030
75.MA0481.2	Known	Homo sapiens	17134	520	10630	0.030
76.MA1117.1	Known	Homo sapiens	17749	537	15982	0.030
77.UP00023_1	Known	Mus	14682	440	14930	0.030
78.UP00217_1	Known	Mus	22866	683	16712	0.030
79.MA0108.2	Known	Homo sapiens	7990	238	6696	0.030
80.MA0898.1	Known	Mus	12186	362	10441	0.030
81.UP00197_1	Known	Mus	11229	332	9505	0.030
82.MA0148.3	Known	Homo sapiens	22736	671	17898	0.030
83.UP00173_1	Known	Mus	10904	321	8543	0.029
84.FOXC1_DBD_2	Known	Homo sapiens	15556	456	16270	0.029
85.UP00157_1	Known	Mus	13839	405	11902	0.029
86.UP00133_1	Known	Mus	15040	440	11983	0.029
87.UP00240_1	Known	Mus	15230	445	11797	0.029
88.UP00045_2	Known	Mus	22529	656	16474	0.029
89.ZNF740_full	Known	Homo sapiens	69449	2009	95389	0.029
90.POU1F1_DBD_1	Known	Homo sapiens	15410	445	12695	0.029
91.MA0593.1	Known	Homo sapiens	25855	743	13752	0.029
92.UP00155_1	Known	Mus	13722	394	11779	0.029
93.UP00025_2	Known	Mus	22479	645	19872	0.029
94.UP00064_1	Known	Mus	19460	557	16323	0.029
95.UP00101_1	Known	Mus	21446	613	15905	0.029
96.UP00208_2	Known	Mus	12358	352	10367	0.028

Motif id	Motif type	Organism	Number of motif occurrences on splice acceptor regions	Number of SNPs with |SPANR Score| >= 5	Number of SNPs with |SPANR Score| < 5	Motif prioritization measure
97.PAX4_DBD	Known	Homo sapiens	3863	110	1878	0.028
98.UP00097_2	Known	Mus	66259	1886	47906	0.028
99.FOXC2_DBD_1	Known	Homo sapiens	13878	394	14511	0.028
100.UP00012_1	Known	Mus	16185	459	15470	0.028
101.MA1104.1	Known	Homo sapiens	14294	404	10610	0.028
102.UP00229_1	Known	Mus	10134	286	8524	0.028
103.MA0079.3	Known	Homo sapiens	111725	3147	122798	0.028
104.UP00213_1	Known	Mus	18018	504	15074	0.028
105.MA1107.1	Known	Homo sapiens	87146	2431	98760	0.028
106.UP00236_1	Known	Mus	21867	609	19373	0.028
107.UP00089_1	Known	Mus	7470	208	5656	0.028
108.UP00073_2	Known	Mus	19900	553	16526	0.028
109.UP00214_1	Known	Mus	9938	276	7495	0.028
110.SP8_DBD	Known	Homo sapiens	61597	1708	87441	0.028
111.MA0497.1	Known	Homo sapiens	38717	1073	22445	0.028
112.MA0897.1	Known	Mus	11908	329	10274	0.028
113.UP00134_1	Known	Mus	18375	507	13404	0.028

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Karaduman Bahçe, Gülşah
Nationality: Turkish
Date and Place of Birth: 24.05.1983, Isparta
Marital Status: Married
E-mail: gkaraduman@gmail.com

EDUCATION

Degree	Institution	Graduation
M.Sc. Computer Engineering	METU	2010
B.Sc. Computer Engineering	METU	2006

WORK EXPERIENCE

Enrollment	Place	Years
Software Engineer	Glovo, Spain	2020 – Present
Senior R&D Engineer	SBG, Ankara	2017 – 2019
Senior Researcher	TUBITAK YTE	2012 – 2017
Researcher	TUBITAK ILTAREN	2008 – 2012
Software Engineer	Siemens E.C.	2006 – 2008

PUBLICATIONS

[1] D. Bozagaç, G. Karaduman, A. Kara, and M. N. Alpdemir. Sim-petek: A parallel simulation execution framework for grid environments. The Journal 91 of Defense Modeling and Simulation: Applications, Methodology, Technology, 9(4):303–319, 2012.

FOREIGN LANGUAGES

English: Fluent, Spanish: Basic