DOMINICAN UNIVERSITY of CALIFORNIA

Dominican Scholar

Graduate Master's Theses, Capstones, and Culminating Projects

Student Scholarship

5-2016

Evaluating the Performance of in Silico Predictive Models on Detecting Splice-altering Variants

Erica Cayton
Dominican University of California

https://doi.org/10.33015/dominican.edu/2016.bio.05

Survey: Let us know how this paper benefits you.

Recommended Citation
Cayton, Erica, "Evaluating the Performance of in Silico Predictive Models on Detecting Splice-altering Variants" (2016). *Graduate Master's Theses, Capstones, and Culminating Projects*. 226.
https://doi.org/10.33015/dominican.edu/2016.bio.05

This Master's Thesis is brought to you for free and open access by the Student Scholarship at Dominican Scholar. It has been accepted for inclusion in Graduate Master's Theses, Capstones, and Culminating Projects by an authorized administrator of Dominican Scholar. For more information, please contact michael.pujals@dominican.edu.

Evaluating the performance of *in silico* predictive models on detecting splice-altering variants

By
Erica C. Cayton

A culminating thesis submitted to the faculty of
Dominican University of California and BioMarin Pharmaceutical Inc.
in partial fulfillment of the requirements for the degree of
Master of Science in Biology

San Rafael, California
May 2015

This thesis, written under the direction of the candidate's thesis advisor and approved by the department chair, has been presented to and accepted by the Department of Natural Sciences and Mathematics in partial fulfillment of the requirements for the degree of Master of Science in Biology. The content and research methodologies presented in this work represent the work of the candidate alone.

Erica C. Cayton
Candidate

May 11, 2016
Date

Wyatt T. Clark, Ph.D
Graduate Research Advisor
Scientist I, Computational Biology, BioMarin Pharmaceutical Inc.

May 11, 2016
Date

Randall Hall, Ph.D
Second Reader
Professor of Chemistry, Dominican University of California

May 11, 2016
Date

Maggie Louie, Ph.D
MS Biology Program Director
Associate Professor of Biochemistry, Dominican University of California

May 12, 2016
Date

Copyright 2016, by Erica C. Cayton
All Rights Reserved

Contents

List of Tables

List of Figures

Abbreviations

3'ss 3' (Acceptor) splice site

5'ss 5' (Donor) splice site

AF allele frequency

ALS amyotrophic lateral sclerosis

ASSP Alternative Splice Site Predictor

AUC Area Under the (ROC) Curve

CDK6 cyclin-dependent kinase 6

cDNA coding DNA (deoxyribonucleic acid)

CF cystic fibrosis

CFTR cystic fibrosis transmembrane conductance regulator

EMDS Exonic Missense data set

ESE exonic splicing enhancer

ESR exonic splicing regulatory element

ESS exonic splicing silencer

ExAC Exome Aggregation Consortium

FN false negative

FP false positive

FTLD frontotemporal lobar degeneration

GWAS genome-wide association studies

HGMD Human Gene Mutation Database

hnRNP heterogenous nuclear ribonucleoprotein

HSF Human Splice Finder

ISE intronic splicing enhancer

ISS intronic splicing silencer

JRDS Junction Region data set

LOR log odds ratio

MaxEnt maximum entropy predictor of disruption to splice sites

MCAD medium-chain acyl-CoA dehydrogenase

MDD maximal dependence decomposition

MM Markov model

mRNA messenger RNA (ribonucleic acid)

NMD nonsense-mediated decay

NSAV non-splicing associated variant

PKU phenylketoneuria

PPT polypyrimidine tract

QTL quantitative trait loci

RNP ribonucleoprotein

ROC receiver operating characteristic

SAV splice-altering variant

SELEX systematic evolution of ligands by exponential enrichment

SLOS Smith-Lemli-Opitz syndrome

SMA spinal muscular atrophy

SNP single nucleotide polymorphism

snRNA small nuclear ribonucleic acid

snRNP small nuclear ribonucleoprotein

SNV single nucleotide variant

SPANR splicing-based analysis of variants

SR protien Serine-arginine protein

SRSF Serine-arginine protein splicing factor

TDP43 TAR DNA binding protein-43

TN true negative

TP true positive

TSL two sample logo

U2AF U2 associated factor

WMM weight matrix model

WT wild type

Abstract

As with any complex biological pathway, the splicing process has both advantages and obstacles with respect to the diversity and fidelity of protein production. The potential benefits of being able to produce multiple versions of a gene (isoforms) must be weighed against the additional complexity introduced by the noisy and mechanistically complicated process of splicing. Indeed, research has found that errors in splicing can be implicated in an increasing number of disorders.

Variants that cause disease may operate by disrupting splicing; however many of the variants are frequently annotated as disrupting function through a missense mutation, or via an unknown mechanism.

The objective of this study is to determine the ubiquity of splice-altering variants (SAVs) in the human genome with a focus on coding missense and silent synonymous polymorphisms that may impact splicing. As a first step, we evaluated the ability of *in silico* prediction tools to predict whether a given variant will disrupt splicing.

Top performing tools were then used to predict splicing disruption for two sets of variants in the genome; one data set contained variants located anywhere in an exon, and the second restricted variants by location with the focus specifically on those annotated as being involved in disease. The results demonstrate that for some of these prediction tools there is a bias in the results based on variant proximity to the exon-intron junction. Also, analysis of the data sets suggests that the variants listed as non-splice affecting in the database include a considerable number of false negatives. These results may be beneficial for updating the information in widely used databases to improve the usefulness of such resources.

The efforts summarized in this thesis will hopefully bring insights into the mechanisms by which splicing errors contribute to disease development and thus facilitate disease treatment improvements.

Acknowledgements

This project would not have been possible without the patience, guidance, and instruction from Dr. Wyatt Clark.

Thanks to Dr. Jonathan LeBowitz and Dr. Karen Yu for their input and leadership.

Thank you also to Kathryn Davidson, Gordon Vehar, and all my colleagues at BioMarin for their willingness to include me in their work and to encourage my education and scientific understanding in countless ways.

Additional thanks is extended to my husband, daughter, and all my family for their patience, support, and encouragement without which I would not have achieved this success.

1 Introduction

The initial product of transcription is the pre-mRNA which requires extensive modification prior to translation into a functional protein. Splicing is the process by which non-coding RNA segments (introns) are removed from a pre-mRNA transcript and the appropriate coding RNA segments (exons) are retained. The splicing process is important to the ability of a cell to produce mature transcripts, which are then translated. On average, 90% of a pre-mRNA transcript is removed as introns, leaving the remaining 10%, the exons, to be ligated and form the mature mRNA [Roy and Irimia, 2009]. While many researchers choose to focus on the beneficial aspect of splicing's ability to produce multiple isoforms of a single gene [Niu and Yang, 2011], it should not be overlooked that this mechanism's inherent complexity makes it susceptible to a myriad of sources of error, and thereby may enable the development of many diseases [Tazi et al., 2009]. The highly complex process of splicing often occurs cotranscriptionally and thus requires careful regulation to ensure the precision and accuracy of the resultant mRNA [Kornblihtt et al., 2004].

Research suggests that many human diseases are the result of aberrant splicing resulting from mutations that disrupt various components of the process [Cooper et al., 2009]. The connection between defects in the splicing process and disease occurrence has been demonstrated for various diseases such as Smith-Lemli-Opitz syndrome and Sandhoff disease in which disruptions in the splice site motifs result in aberrant splicing and pathology [Fitzky et al., 1998, Wakamatsu et al., 1992]. Furthermore, the severity of disease is likely proportional to how the splicing process affects the resulting protein function [Krawczak et al., 2007].

Splicing also interacts with the process of nonsense mediated decay (NMD) [Cartegni L, 2002]. As splicing mutations may result in frame shifts, or premature stop codons this then makes the resulting mRNA a potential target of NMD. NMD ultimately effects gene expression by degrading damaged mRNA thus preventing the translation

of potentially aberrant proteins [Wollerton et al., 2004]; it can result in complete loss of function even if the individual is heterozygous for the premature stop codon variant. Some bioinformatics tools have estimated that approximately 35% of human alternative splicing isoforms produce targets for NMD [Lewis et al., 2003].

Within the context of of splicing and disease, there are several areas that warrant further investigation. First, although most splice-altering variants (SAVs) have been documented as resulting in a loss of function from the destruction of splice sites, some may result in the gain of splice sites. Furthermore, since the discovery of splicing additional elements aside from the $5'$ and $3'$ motifs and the polypyrimidine tract have been discovered, such as exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) [Cartegni L and AR, 2003]. The inactivation or creation of any of these splicing regulatory elements as well as the creation of cryptic splice sites in ectopic locations can impact the resulting mature transcript and hypothetically be just as detrimental as mutations at the branch site, or $5'$ or $3'$ splice sites [Woolfe et al., 2010]. There is also evidence that mutations in the spliceosome complex, the cellular machinery that carries out splicing, may also play a significant role in the incidence of disease [Padgett, 2012]. This further illustrates the potential mistakes in splicing have to contribute to a variety of pathologies.

A variety of computer-based prediction tools have been developed for identifying variants in the genome and predicting the consequence on protein function of such variants [Mort et al., 2014, Xiong et al., 2015, Desmet et al., 2009]. Since confirming how a single variant disrupts function experimentally can be costly and time consuming, *in silico* prediction is advantageous with its potentially high level of precision in determining how a mutation may impact protein function.

Prediction tools utilize diverse modeling methods and data sets, and as a result vary in the features they predict. For instance, a tool such as MutPred Splice con-

siders only exonic variants that may act to disrupt pre-mRNA splicing [Mort et al., 2014]; while others such as Human Splice Finder (HSF) consider both intronic and exonic information to identify potential splice sites, and to predict possible branch sites [Desmet et al., 2009]. Tools such as these examine variants collected from databases like HGMD [Stenson et al., 2014] for their source data. Some of the information produced from these predictors may then be further investigated for input into reference portals that collect information on rare diseases that may be affected by such mutations. These reference portals provide a source for disease classification, drug inventories, and additional material for researchers and patients [Maiella et al., 2013] which may be enhanced by the information from the predictors.

However, there is a potential problem with the accuracy of these predictors in that a potential annotation bias exists in the reporting of information in them. For example, any mutations in the exonic sequence may be associated with disease but the reason for disease or protein malfunction may be incorrectly labeled. Often the disease is attributed to missense or possibly even synonymous mutations affecting the functionality of the protein on the amino acid level, rather than being recognized as the variant actually causing a disruption in the splicing of the mRNA and thereby changing protein function. More importantly, many synonymous variations are filtered out when the protein affected was being assessed for functionality and the investigation incorporated the amino acid substitution without observing how the variant may have impacted splicing. Variants may be mislabeled and as such not associated with pathology at all, yet they do in fact affect splicing and may thus be a contributing factor to the occurrence of disease.

The purpose of this study is to more accurately determine the contribution to disease of splice-altering variants (SAVs) located at and around the donor and acceptor splice sites. We utilize computer-based prediction tools and literature searches in an attempt to identify previously unknown, or incorrectly annotated SAVs with

gene-disease associations and identify SAVs which have been incorrectly annotated in various databases. Seven *in silico* prediction tools were evaluated for use in this context. The hope is that this investigation will provide information to improve upon the current models or develop an ensemble approach to improve the precision of prediction.

1.1 Splicing and Disease

The complexity of the splicing process leaves it vulnerable to error and these errors can be pathological. Multiple regulatory processes are necessary due to the intricacy of splicing mechanisms and these range from splice site recognition by the spliceosome, to alternative splice site selection [Keren et al., 2010]. It has been estimated that up to half of those point mutations which have been associated with causing genetic disease in humans do so as the result of defects in splicing [Cartegni L, 2002]. Forces such as changes in the environment of the cell can cause alterations to control elements and thus alter the proteins translated. Though some of this regulation is based upon the location and recognition of the splice sites themselves, this alone is not sufficient for management of the process [Tazi et al., 2009]. Additional regulatory protein complexes, RNPs (ribonucleoproteins), bind the pre-mRNA and aid in exon recognition [Tazi et al., 2009]. Mutations in any of the proteins or the site motifs may disrupt the regulatory processes and lead to splicing errors causing disease.

Depending upon the particular gene involved and its function, the presence of a variant may be more or less likely to cause aberrant splicing. Those genes whose function is not supported by an alternate pathway are particularly susceptible to mutations, as the system has no means for compensation if the function is interrupted. Examples of genes susceptible to aberrant splicing in the presence of suitable variants include *DHCR7*, responsible for Smith-Lemli-Opitz syndrome; *HEXB*, causing Sandhoff disease; *PAH*, which results in phenylketoneuria (PKU); and *CFTR*, that causes cystic fibrosis. The various mutations influencing these genes include disruptions to normal splice sites and/or the activation of cryptic splice sites, disruption of splicing enhancers, or the creation of splicing silencers. Any of these issues may lead to aberrant of splicing, dysfunctional or non-functional proteins, and disease pathology.

1.1.1 *DHCR7* Gene and Smith-Lemli-Opitz Syndrome

Smith-Lemli-Opitz syndrome (SLOS) is a disease which results from insufficient production of cholesterol and the accumulation of potentially toxic by-products of this process [National Library of Medicine (US), 2013d]. Cholesterol is a necessary component of the plasma membrane of every cell type hence this disease can have wide-spread consequences affecting multiple organs and systems, including cardiopulmonary, digestive, renal, and others [National Library of Medicine (US), 2013d].

SLOS results from mutations in the *DHCR7* gene (7-dehydrocholesterol reductase) which is involved in the manufacture of cholesterol in numerous cell types [National Library of Medicine (US), 2013a]. The most common mutation associated with SLOS is the IVS8-1G>C mutation which has an associated allele frequency (AF) of 4.2 x 10^{-3} in ExAC [The Broad Institute, 2015]. This mutation disrupts a $3'$ splice site and activates cryptic splice sites; the consequence is an alteration in the reading frame and introudction of a premature stop codon [Fitzky et al., 1998]. The abnormal mRNA produced is translated into an altered protein which lacks a C-terminal domain [Yu et al., 2000] which renders it non-functional.

1.1.2 *HEXB* Gene and Sandhoff Disease

Some mutations in the *HEXB* gene, which codes for the beta subunit of the hexosaminidase enzyme, do cause the lipid storage disorder Sandhoff disease. Beta-hexosaminidase A and B enzymes reside in lysosomes and function in the breakdown of sphingolipides, oligosaccharides, and gangliosides. Dysfunction of a subunit of this enzyme can inhibit these metabolic processes and cause substrate buildup resulting in destruction of neurons in the CNS [National Library of Medicine (US), 2013c].

Multiple splicing-associated mutations have been demonstrated to cause Sandhoff disease and many of them consist of either exonic or intronic mutations that interefere with the $5'$ or $3'$ splice sites [Furihata et al., 1999][Yoshizawa et al., 2002]. A C to T

mutation in exon 11 results in the incorporation of the amino acid leucine instead of proline at position 417 of the protein [Wakamatsu et al., 1992]; according to ExAC (Exome Aggregation Consortium) [The Broad Institute, 2015], this mutation has an allelic frequency of 6.6 x 10^{-4} and in spite of being a missense mutation has been demonstrated to cause the activation of a cryptic 3′ splice site.

In one particularly surprising instance this mutation presented in a Sandhoff patient in conjunction with an A to G substitution in exon 2 that resulted in the incorporation of a lysine to arginine substitution at position 121 of the protein. Here the investigators observed that there was an alteration of the splicing process yet the protein produced was not "biochemically defective" [Wakamatsu et al., 1992]; they thus determined that in this situation there existed "a novel mechanism for the cause of disease" [Wakamatsu et al., 1992].

The introduction of the IVS2-1G>A variant within intron 2 of the *HEXB* gene was reported to cause disruption of the 3′ splice site of intron 2; this resulted in the skipping of exon 3 in the translated beta subunit. However, due to the length of exon 3 (66 bases) no frameshift resulted from this change. As a result of the decreased length of the beta subunit, the effect of this SAV on the resulting protein was interference with protein folding and disruption to the secondary and tertiary structures, not damage to the protein active site or formation of a premature stop codon which is commonly anticipated [Yoshizawa et al., 2002].

1.1.3 *PAH* Gene and Phenylketoneuria

Another disease example is phenylketoneuria (PKU), one of the more common of these diseases with an incidence in the United States of 1 in 10,000 to 1 in 15,000 newborns [National Library of Medicine (US), 2013b]. PKU is the result of an inability to process the amino acid phenylalanine, present in all dietary protein, to tyrosine which is then further processed to generate homrones, neurotransmitters, and melanin

[National Library of Medicine (US), 2013b]. PKU has a range of severity of expression due to the levels of enzyme function present in the affected individual. In mild cases the disease can be controlled with dieteary restrictions while more severe cases may necessitate additional medication. If untreated hyperphenylalaninemia can lead to mental retardation seizures, and behavioral problems [National Library of Medicine (US), 2013b].

PKU results from a variety of different types of mutations in the *PAH* gene, some of which are associated with splicing. A common splicing variant is the c.30C>G mutation which occurs in exon 1 of *PAH* [Dobrowolski et al., 2010] with an allelic frequency of 5.6×10^{-4} [The Broad Institute, 2015]. This SAV results in the creation of an ESS located in the region of a weak $5'$ splice site and as such results in the skipping of exon 1. Of note, prior to recent investigations [Dobrowolski et al., 2010], this variant was categorized as neutral due to the fact that it is a synonymous mutation (p.G10G). It was not until further research was performed that it was recognized that pathology was induced by affecting the splicing process.

Another example of is the c.1144T>C [Heintz et al., 2012] SAV which is located within the motif of an ESE; its disruption results in the skipping of exon 11 due to alterations to the reading frame. Additionally, the abnormal mRNAs which stem from this variant are frequently the targets of NMD and thus degraded without producing a protein [Heintz et al., 2012].

1.1.4 *CFTR* Gene and Cystic Fibrosis

The extension of repeat elements within the genome can also be detrimental to the splicing process. Under normal circumstances these repeats help splicing factors to recognize their binding sites; however alterations to the number of repeats in the sequence can impede this recognition process [Tazi et al., 2009]. This has been seen in the context of the disease cystic fibrosis (CF), where an increase in the number

of UG repeats present causes abnormal splicing of the *CFTR* gene. It has been shown that approximately 13% of the mutations which cause CF by disruption of the chloride/bicarbonate ion channel encoded by *CFTR* are splicing mutations [Bell et al., 2015]. The SAV more commonly seen in CF is c.2657+5G>A located near the $5'$ splice site in intron 16 of the *CFTR* gene [Igreja et al., 2015].

1.1.5 *HEXA* gene and Tay-Sachs Disease

Tay-Sachs disease is a neurodegenerative disorder in which the lack of functional protein products from the *HEXA* gene causes toxic substances to accumulate in neurons, resulting in the disease pathology including loss of motor skills in infants, seizures, and cognitive impairment [National Library of Medicine (US), 2013e]. Of the various point mutations described in association with this disease, multiple examples have been identified at or near the exon-intron junctions frequently disrupting the splice site motifs [Myerowitz, 1988] and resulting in exon skipping and the production of multiple incorrect transcripts [Akli et al., 1990]. The most common SAV associated with Tay-Sachs according to ExAC is c.1073+1G>A with an allelic frequency of 2.2 x 10^{-4} [The Broad Institute, 2015]. This variant results in the disruption of a donor splice site. The abnormal transcripts produced are either degraded by the mechanism of NMD, or produce defective proteins [Levit et al., 2010]. In either case the end result is a lack of functional protein which permits the toxic accumulations in neurons and thus causes this disease pathology.

Clearly from the effects demonstrated in multiple disease processes, there are numerous ways in which splicing can be disrupted and thus lead to disease. This knowledge underscores the importance of identifying such aberrations in normal splicing and being able to identify those variants likely to result in pathology.

1.2 The Molecular Biology of Splicing

Splicing is a complex process by which segments of a pre-mRNA are cut and rejoined to form the instructions to build a protein. It must be meticulously carried out by the interplay of enzymes, mRNAs, proteins, and the pre-mRNA alternately coming together and dissociating in sequence. These component interactions are choreographed to ensure that the introns and select exons are removed, and appropriate exons are retained and ligated to form the final mRNA.

The process of splicing begins with transcription of pre-mRNA from the DNA template. This transcript contains a series of coding and non-coding regions; those elements which do not contain information needed for production of the desired protein must be removed, and the remaining segments then reattached to form a sequence which can be translated. The cellular machinery which performs the task of actually splicing the mRNA is the spliceosome. The spliceosome is an association of small nuclear ribonucleic proteins (snRNPs), U1, U2, U4, U5, and U6, which form a complex. Specific sequences at various positions on the pre-mRNA are recognized by particular spliceosomal components. Located at the junctions of the exons and introns are the $5'$ donor splice site and the $3'$ acceptor splice site (Figure 1), and within the sequence of the intron is the branch site. In the intron between the branch site and the $3'$ splice site is the polypyrimidine tract (PPT). The PPT does not have a conserved sequence, rather consists of a general enrichment of the region with pyrimidines, or the bases cytosine (C), thymine (T), and uracil (U). These four areas contain conserved motifs that are recognized by elements of the spliceosome and used for orientation of those components and the complex as a whole in the splicing process.

The first component of the spliceosome complex is the U1 snRNP which binds the GU residues of the $5'$ splice site of the pre-mRNA. Next the U2 associated factors recognize the AG residue of the $3'$ splice site and the PPT, and bind at that location [Padgett, 2012]. The U2 snRNP then recognizes the components of the branch point

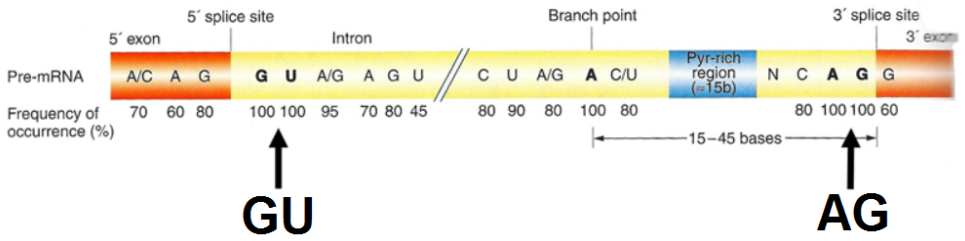

Figure 1: The nucleotides in bold are the most conserved portion of the sequence, however extended sections of the sequence are recognized in binding components of the spliceosome. The branch site of the intron is located within the body of the intron. The adenosine residue within the intron (shown in bold) is the site of attachment for formation of the intron lariat. The polypyrimidine tract consists of a sequence of nucleotides particularly saturated with pyrimidines (cytosine and uracil) and located between the branch site and the 3′ splice site. The 3′ acceptor splice site at the intron/exon boundary is a site of recognition with the nucleotides in bold being the most conserved in the sequence. [Szauter, 2015]

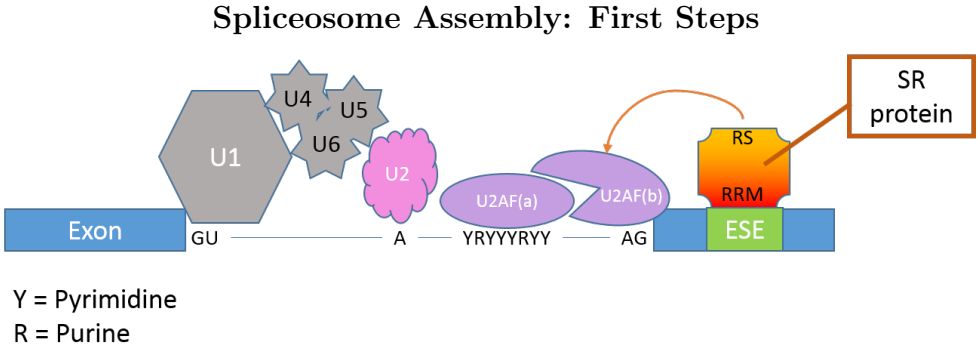

Figure 2: Assembly of the first components of the spliceosome on the pre-mRNA. The SR protein binds to the ESE within the exon and has a stabilizing effect on the two subunits of the U2 additional factor (U2AF(a) and (b)) bound to the PPT within the intron. This U2AF in turn promotes the binding of the U2 snRNP at the branch point. (Adapted with permission from [Cartegni L, 2002])

sequence around the main adenosine, it is stabilized by the effects of the U2AF, and binds with U1 to form the pre-spliceosome complex (Figure 2) [Matera and Wang, 2014]. The rearrangement of these elements allows three regions of the intron (5′ splice site, branch site, and 3′ splice site) to be brought close together.

Components U4, U6, and U5 form a complex independently and subsequently interact with the pre-spliceosome complex (U1, U2, and U2AF) already bound to the pre-mRNA (Figure 3). After further rearrangement of the elements, the U1 and U4

snRNPs dissociate. At this point the first cutting of the pre-mRNA takes place; a transesterification reaction cleaves the 5′ end of the intron. This free end is then bound to the adenine residue in the branch site, forming the lariat structure of the intron which will eventually be removed. Next a second transesterification reaction cleaves the 3′ end of the intron and the two exons are ligated. This forms the newly spliced mRNA which is now free and ready for translation. The intron lariat is released from the spliceosome and will be degraded, while the remaining components of the spliceosome dissociate to be available for the splicing of another transcript [Matera and Wang, 2014].

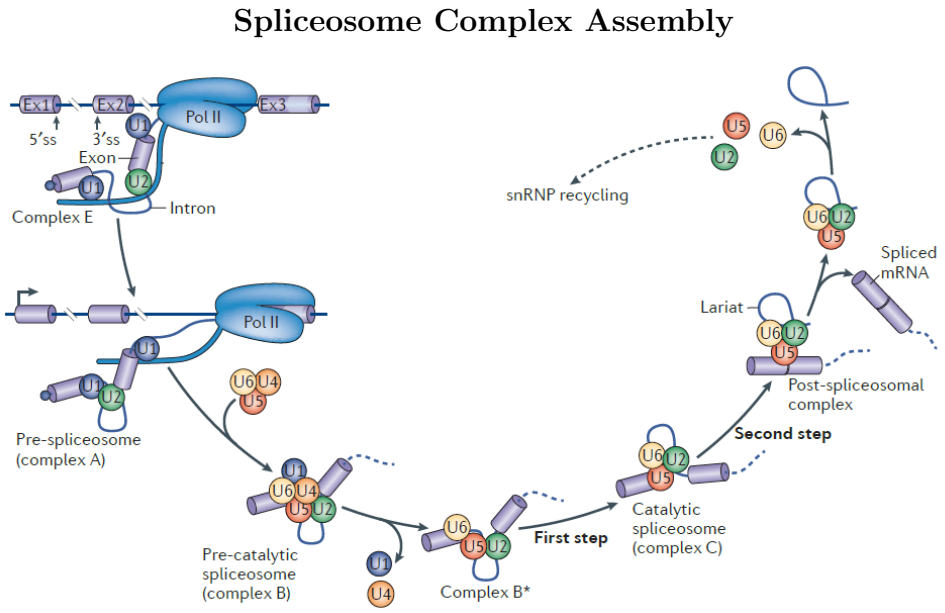

Figure 3: From [Matera and Wang, 2014]. The pre-spliceosome (U1 and U2) once assembled on the pre-mRNA, can interact with the U4, U6, U5 complex which formed separately. These two structures complex while bound to the pre-mRNA and subsequently release the U1 and U4 snRNPs. The remaining complex B (U2, U5, and U6) then performs the first transesterification reaction. The lariat structure is formed by the binding of the free end of the intron to the adenine residue of the branch point. Next the second transesterification reaction occurs to cleave the intron and the exons are spliced together. The remaining snRNPs separate from the intron lariat to be used in another splicing sequence.

1.3 Splicing in Evolutionary Context

Alternative splicing is an important and necessary component of human genetic regulation. It allows the coding sequences for multiple proteins to be contained within a single gene and produced by splicing in or out the appropriate exons. The human genome contains 24,000 genes which code for proteins, yet there are an estimated 100,000 protein isoforms that are actually synthesized in humans [Modrek and Lee, 2002]. However, the process of splicing in humans is far less fastidious than it is in other organisms.

Aberrant splicing and exon skipping has been associated with increased intron length; this may be because in the context of longer introns it is more likely that the intron sequence will include potential false splice sites. By contrast, exons associated with smaller introns are more frequently expressed in the spliced mRNA [Wu and Hurst, 2015]. As a result the donor and, particularly, acceptor sites flanking these extended introns tend to be weak; research has demonstrated, weak splice sites tend to be associated more frequently with the presence of ESEs [Wu and Hurst, 2015], since this enhances the ability of the spliceosome components to recognize said splice site and thus avoid errors. It follows logically that the presence of ESEs is more necessary in genes which contain more and/or longer introns. In this situation the $5'$ and $3'$ splice sites are further removed from each other and thus likely to be weaker due to the presence of potential decoy splice sites in the intron code. Hence it is beneficial to have more ESEs so that in conjunction with the support of SR proteins, the true splice sites can be identified and the components of the spliceosome may bind appropriately, enabling correct splicing of the mRNA.

The splicing process itself has evolved through various processes such as the accumulation of mutations in introns and exons resulting in both damage to constitutive splice sites and the creation of new ones [Keren et al., 2010]. When mutations arise in existing splice sites it can make them unrecognizable by the machinery of the

spliceosome or by SR proteins, and thus unusable. With the creation of new splice sites, these sites are then in competition with existing ones and their use can result in splicing aberrations. However, in circumstances where there is an effective negative selection pressure in place, this can push the new sites to exist only as "minor" splice sites, with the ancestral ones remaining the primary location for splicing; this would retain the function of the resulting protein [Keren et al., 2010].

The effects of evolution of the splicing process on the genome is not entirely negative. Aberrant splicing may result in the creation of a protein isoform which has slightly different protein-protein interactions from those of its source and which are beneficial to the organism [Roy and Irimia, 2009]. Splicing has evolved through additional processes such as changes to exon-intron structure, exon shuffling, and exonization [Keren et al., 2010]. Changes to the structure and relative sizes of exons and introns result in changes to the propensity that exons will or will not be retained in the product mRNA. Exon shuffling is a process by which new exons are incorporated into genes, or are duplicated within their original gene. It has been suggested that this process is advantageous to the organism since it may create a beneficial new protein isoform [Gilbert, 1978]; but it is much more likely that the altered protein will either have no function and consequently be degraded, or may be detrimental, interfering with the correctly functioning ancestral protein and potentially resulting in disease. An example of this is seen with the CFTR (cystic fibrosis transmembrane conductance regulator) membrane protein which causes cystic fibrosis [Lubamba et al., 2012] [Linsdell, 2015].

The process of exonization is the means by which regions of the genome become new exons. Frequently these regions are transposable elements, in humans the most common of which is the *Alu* element [Keren et al., 2010]. While in theory the addition of exons could confer an evolutionary advantage to the organism, in practice it seems much more likely that the addition of new exons would likely interfere with the

functioning of normal proteins and the cell, and thus result in disease [Kreahling and Graveley, 2004]. An example of this can be seen with respect to the *TRIM24* gene. In cases where *Alu* elements have become exons, the result was a lack of production of normal protein from this gene, which has been associated with leukemia [Amit et al., 2007].

Overall, though the process of genetic mutation and aberrant splicing has been associated theoretically with the evolutionary advancement of species, the process of splicing ultimately is less auspicious. It frequently is associated with the creation of errors in the resulting proteins, a phenomenon which often has the potential if not the actual consequence of causing disease. A noisy, and disruptive process, aberrant splicing may be more correctly viewed as doing more harm than good.

2 Review of Prediction Methods

The purpose of this project is to assess the performance of *in silico* predictors in the evaluation of variants in and around the splice sites and ESRs to determine the likelihood that their presence will result in aberrant splicing. In doing so, multiple tools were researched and screened, and of these seven *in silico* prediction tools were ultimately considered for evaluation.

SPANR (splicing-based analysis of variants) is the most recently developed of the predictors used in this project [Xiong et al., 2015]. This tool is unique in that it was not trained using disease annotations, functional genome annotations, or allele frequencies in the population. Even though disease phenotype associations were not utilized in training this predictor, according to its creators in practice it has proved able to detect variants relating to disease expression when combined with phenotype-matched genotype data [Xiong et al., 2015]. Furthermore, with respect to assessment of genome-wide association studies (GWAS) or quantitative trait loci data (QTL), though SPANR does not rely on allele frequency, it "can reliably detect rare and even spontaneous disease variants" [Xiong et al., 2015]. The combination of these techniques by SPANR serves to improve the specificity of GWAS and QTL in order to identify variants causing particular diseases [Xiong et al., 2015].

The dataset used in the development of the SPANR predictor consisted of 75 base pair single-end RNA-seq datasets that were collected from the Illumina Human BodyMap 2.0 project [Xiong et al., 2015]. The program was trained using cassette exons identified from RefSeq annotations, and screened to ensure that only high quality data from normal tissue was used. The data was further filtered to remove exons which had any overlap, were very similar to each other, or were excessively short (<10nt) or long (>6000nt) [Xiong et al., 2015]. SPANR looks for disruptions in splicing resulting from single nucleotide variants which occur throughout the exon and intron; it does not restrict its evaluation to the areas of the $5'$ and $3'$ splice sites, and

the branch point. While the scope of this tool's analysis is broad, it does not provide as much information in its output as some other prediction tools, yielding three values based on the value "psi" (ψ). Psi is defined as the percentage of transcripts which splice-in the exon containing the variant when said variant is present in the genome; thus the predictions returned relate to the propensity for exon-skipping when a particular variant is present.

Another prediction tool to identify single base substitutions that affect mRNA splicing is MutPred Splice [Mort et al., 2014]. MutPred Splice is a supervised learning method which used a training set of data from which its algorithm could build a model. New input is then applied to this model and assessed based on how similarly it compares. One of the goals of the MutPred Splice program is to be able to predict the severity of disease caused by a particular mutation by utilizing genotypic data and evaluating the resulting gain or loss of exonic splicing regulatory elements (ESRs) ensuing from a substitution [Mort et al., 2014]. In addition to evaluating ESRs, MutPred Splice considers the potential for splice site disruption, creation of cryptic splice sites, and exon skipping as a result of single nucleotide variants (SNVs). The results from all these considerations are combined into a single overall score. This predictor was trained using allelic information from instances of human disease to predict those exonic single nucleotide polymorphisms (SNPs) which result in aberrant splicing [Mort et al., 2014]. Unlike some other tools, MutPred Splice uses not only missense, but also synonymous and nonsense variant data in order to expand the scope of its predictions. However, its output is limited to variants located within the exon that may impact splicing [Mort et al., 2014].

Skippy is a web-based tool which scores exonic variants with respect to the likelihood that they will result in exon skipping during splicing, and looks to identify potential SAVs [Woolfe et al., 2010]. This tool focuses on the effects that alterations to ESRs may have on splicing, particularly with respect to exon skipping. It gauges

the likelihood that said variants will result in the creation, loss, or change in ESRs and thus may cause a splicing aberration. However, this predictor does not look at SAVs within the splice site sequences themselves or less than three base pairs from the $5'$ or $3'$ junction [Woolfe et al., 2010]. Analysis suggests that there may exist binding "hotspots" near but not within exon-intron junctions which are more important than other areas of the sequence for splicing regulation [Woolfe et al., 2010]. Though the implication is that these areas have a larger impact on the process of splicing, locations all throughout the exon still are important for splicing regulation [Woolfe et al., 2010].

Being web-based, this tool is easily and freely accessible; additionally it only requires the chromosome location and the identity of the variant alleles as input. The focus of Skippy on the effects of disruptions to ESRs is beneficial in that it does not limit the search for variants which may impact splicing to areas in the immediate vicinity of the splice sites. The authors argue that their predictor has demonstrated that SAVs occurring in regions removed from the exon-intron junctions in fact have a larger impact on splicing than mutations within the traditional splice site locations (i.e. the $5'$ and $3'$ ss, branch point, and PPT) [Woolfe et al., 2010]. Despite this however, Skippy may not give ample weight to the effects of features in those more traditional areas on the process of splicing. Furthermore, in its focus on the effects of ESRs may make its results less comprehensive than is needed for more general analysis purposes.

We also considered the predictor Human Splice Finder (HSF) [Desmet et al., 2009]. The HSF data set includes both intronic and exonic data from the Ensembl human genome database. This enables HSF to make predictions not only related to the effects of varians located at donor and acceptor splice sites, but also branch points, ESEs, and ESSs [Desmet et al., 2009]. In addition to evaluating variants in the aforementioned locations, this predictor was designed to predict potential branch

points, 5′ ss, and 3′ splice sites. Theoretically this might increase the precision in recignizing cryptic splice sites. HSF can also predict whether a variant will result in the gain or loss of functionality of these splice sites. Based upon the literature, HSF had good results in predicting mutations resulting in either exon skipping or cryptic splice-site activation or creation for variants at the 5′ and 3′ ss. Though there was some variation in precision depending upon the proximity of the variant to the splice sites, overall the predictors performance was efficient [Desmet et al., 2009]. Unfortunately the submission of variants to this predictor requires the separation of said variants by gene; only those variants located in the same gene can be submitted together. This unique requirement for data formatting was the dominant reason HSF was not incorporated in the predictor evaluation.

The Alternative Splice Site Predictor (ASSP) uses a neural network model, and a data set which looks at 70 base pairs in the exon and intron around skipped, cyptic, alternative, and constitutive splice sites [Wang and Marín, 2006]. In addition to evaluating the landscape of the 5′ or 3′ splice sites, this tool also examines the regions of the branch point, polypyrimidine tract, and regulatory elements in its assessment of the gain or loss of splice sites due to the effects of variants on splicing. Though thorough in its examination of the pre-mRNA and the components thereof, the creators recognize that a shortcoming of this program is the amount of false positives and false negatives in the identification of splice sites [Wang and Marín, 2006]. However, the authors did identify a phenomenon that those genes with cassette exons, i.e. skipped and cryptic exons, occur more frequently in GC-poor regions [Wang and Marín, 2006]. This discovery is demonstrative of the value of ASSP in identifying elements of the genetic code which could be used as markers in the search for variants affecting splicing, despite the shortcomings inherent in any tool. However a significant challenge with respect to the usefulness of this tool is the format in which data must be submitted. ASSP requires that the FASTA or raw sequence be

submitted including 70 base pairs before the first and after the last splice site. This was the primary factor in the decision not to include ASSP in the final comparison analysis.

ESE Finder [Smith et al., 2006] is a prediction tool that considers the effects of variants on not only the $5'$ and $3'$ splice sites and the branch sites, but also on the serine-arginine protein splicing factors (SRSFs) which affect the proclivity of the spliceosome to act on the pre-mRNA [Krainer Lab; Zhang Lab; Cold Spring Harbor Laboratory, 2007]. This method used SELEX (systematic evolution of ligands by exponential enrichment) to identify potential ESEs [Smith et al., 2006]. After multiple rounds of selection are performed using SELEX, the results were sequenced to determine the consensus motif for each of the five SRSFs this tool considers [Smith et al., 2006]. The sequences determined were then used to create a position-specific weight matrix for each SRSF [Cartegni L and AR, 2003]. These matrices allow the program to predict "the location of SR-protein-specific putative ESEs in exonic sequences" [Krainer Lab; Zhang Lab; Cold Spring Harbor Laboratory, 2007]. One of the benefits of this web-based prediciton tool is the fact that it considers the effects of variants not only on the splice sites and the branch site themselves, but also on these SR proteins which contribute to the regulation of the actions of the spliceosome itself. Originally the authors used sequences 20 nucleotides in length for the creation of libraries used in the SELEX process [Cartegni L and AR, 2003]. In the update to this tool in 2006, the length was changed to either seven or 14 nucleotides due to the fact that the sequence that is recognized by the SRSFs is actually a sequence of seven nucleotides [Smith et al., 2006]. As a result this reduces the likelihood of multiple ESEs being present on the same insert and thus improves exactitude of the predictions made.

Our final inclusion is the prediction tool MaxEnt, which uses the principal of maximum entropy to assess splice sites and evaluate the possibility that the presence

Predictors Considered for Assessment

Predictor	Variants Detected	Prediction Output
SPANR (Xiong et al. [2015])	SNV effects on splicing throughout the exon and intron	dPSI, dPSI percentile, PSI WT
MutPredSplice (Woolfe et al. [2010])	SNV effects on ESR gain/loss, splice site disruption, cryptic splice site formation, exon skipping	Overall splicing likelihood score
Skippy (Woolfe et al. [2010])	ESR gain, loss, or change	ESE/ESS gain/loss, LOR, RC, 5′ score, 3′ score
Human Splice Finder (HSF) (Desmet et al. [2009])	Location of branch points, 5′ and 3′ ss. Effect on splice site function, exon skipping, cryptic splice site creation	2 consensus value scores: HSF and Maximum entropy (5′ or 3′)
Alternative Splice Site Predictor (ASSP) (Wang and Marín [2006])	Effects of variants on branch point, PPT, ESRs, gain/loss of splice site	putative splice site type; confidence value 0-1
ESE Finder (Smith et al. [2006])	Effects on 5′ and 3′ ss, branch sites, and SRSFs	score for each of 5 different SRSFs
MaxEnt (Eng et al. [2004])	Evaluation of effect of SNVs on 5′ and 3′ ss	MaxEnt, MM, WMM

Table 1: *In silico* predictors of splice-affecting variants considered for assessment.

of a variant will disrupt normal splicing [Eng et al., 2004]. This predictor scores 5′ and 3′ splice sites in estimation of the efficiency of splicing occurring at each. MaxEnt focuses on the areas immediately around the 5′ and 3′ splice sites and does not consider the effects of variants located elsewhere in either the exon or intron [Eng et al., 2004]. For the 3′ splice site up to 23 base pairs can be input, while the maximum is nine base pairs for the 5′ splice site. Data entry for this predictor requires that both the positive and negative sequences be entered. Though the input of data for this predictor may be rather involved, MaxEnt performed well when its results were compared with cDNA data [Eng et al., 2004].

The *in silico* prediction tools we have selected for the assessment of SAVs represent a range of approaches to the evaluation of variants and their impact on splicing. The use of a diverse group of methods yet the same data sets as input provides greater insight into the results which emerge as it allows for comparisons between the different tools.

3 Data and Methods

3.1 Evaluated Predictors

We employed the prediction tools SPANR [Xiong et al., 2015], MutPred Splice [Mort et al., 2014], Skippy [Woolfe et al., 2010], and MaxEnt [Eng et al., 2004] to predict splice sites and evaluate the effects of the variants as well as those appearing in annotations in HGMD. We then quantified the performance of each tool by creating ROC (Receiver Operating Characteristic) curves and calculating the AUC (Area Under the Curve) for each tool. The tools were chosen to represent a variety of approaches to the prediction process regarding the impact of a particular variant on splicing. At the same time it was important that the same data sets could be used for all the predictors in order to facilitate legitimate comparisons of the results returned.

3.2 Evaluation Data Sets

To evaluate predictor performance we generated two data sets referred to as the Exonic Missense (EMDS) and Junction Region data sets (JRDS) which consist of disease causing variants selected from the 2014.3 version of the Human Gene Mutation Database (HGMD). In both data-sets positive, splice-affecting variants annotated as such in HGMD were selected. Any disease-causing variant not annotated as affecting splicing in HGMD was designated as a negative, putatively non-splice-affecting, variant.

In an attempt to ensure that performance was not influenced by variants belonging to the training set of any predictor listed in Table 1, we excluded all variants used in the training of the most recent predictor for which we could obtain such a list, MutPredSplice [Mort et al., 2014]. It must be noted that SPANR was a newer method we evaluated, but a training set of positive and negative data points was not available for it.

3.2.1 The Exonic Missense Data Set (EMDS)

The first data set generated consisted of 1000 exonic missense variants, 363 positive and 637 negative, located anywhere within the exon. This Exonic Missense data set enabled evaluation of the predictors with respect to how they appraise missense variants specifically. With this data set limited to exonic variants, four paramaters of three different *in silico* tools could be compared. It also facilitated an evaluation of the information catalogued in the variant databases, as most of the information therein primarily relates to exonic variants. The positive data points in the EMDS are missense mutations and are identified as being disease-causing in HGMD.

3.2.2 The Junction Region Data Set (JRDS)

The second data set generated consisted of both synonymous and missense variants collected within a specific range around the exon-intron junctions. These are exonic variants located within three exonic nucleotides of the $5'$ splice site or the $3'$ splice site. This is referred to as the Junction Region data set and includes 301 variants, 219 positive and 82 negative. The primary benefit of this data set is it restricts the locations of the variants considered and removes any which are far from the exon-intron junction. By so doing we can control for a confounding variable, the distance the variant lies from the junction. If those tools which rely on this variable to determine the likelihood of the variant affecting splicing vary in the results returned across the two data sets, that has implications regarding the predictor's accuracy. Additionally, the JRDS facilitates the inclusion of an additional predictor, MaxEnt, in the evaluation process.

3.3 Using ROC Curves to Quantify Predictor Performance

In order to use predictive methods to uncover novel SAV's it was necessary to first quantify the performance of each predictor to systematically identify predictors

which may or may not be useful. Each tool's performance was evaluated based on its Area Under the ROC curve (AUC) value calculated from the ROC (receiver operating characteristic) curve. The ROC curve is used to assess the performance of any test with two possible outcomes, given a continuous valued predictor. This is done by varying the decision threshold from the largest to the smallest value along the range of scores returned by the predictor. The decision threshold separates the scores into positives and negatives and based upon where it is placed, the proportion of positives and negatives will change. The confusion matrix in Table 2 illustrates the four possible outcomes when classifying a datapoint. A truly positive point may be labeled as positive (correctly) or negative (incorrectly). Likewise a truly negative point may be labeled by the tool as positive (incorrectly) or negative (correctly). These constitute the four possible designations: true positive (TP), false negative (FN), false positive (FP), and true negative (TN), respectively. A curve is produced from this information by plotting sensitivity, or the true positive rate (y axis), against the false positive rate (FPR), or 1-specificity (x axis) while using a sliding decision threshold; as shown in Figure 4A and Figure 5A.

Confusion Matrix

		True label	
		Positive	Negative
Predicted label	Positive	True positives	False positives (Type I)
	Negative	False negatives (Type II)	True negatives

Table 2: A sample confusion matrix showing the four potential outcomes when performing binary classification.

Sensitivity, also referred to as the true positive rate, is the probability that a test being performed will correctly identify a positive datapoint [Tape, 2015]; in this case, the proportion of actual splice affecting variants that the test correctly recognizes.

This is plotted on the y-axis of the ROC curve. Sensitivity is calculated as the number of true positives divided by the number of actual positives

$$sensitivity = \frac{|TP|}{|TP| + |FN|} \tag{1}$$

In the equation, vertical lines indicate the set cardinality operator (not absolute value). This refers to all the elements in a particular set of terms. For example, $|TP|$ refers to all the True Positive data points for a particular decision threshold.

Specificity is the probability that the classifier will accurately identify a negative datapoint [Tape, 2015]. Here it represents the proportion of non-splice affecting variants correctly reported by the test. Specificity is calculated as the number of true negatives divided by the number of actual negatives,

$$specificity = \frac{|TN|}{|TN| + |FP|}. \tag{2}$$

On an ROC curve sensitivity is plotted as a function of FPR. FPR is mathematically equivalent to 1-specificity and is defined as the fraction of negative datapoints incorrectly predicted to be positives:

$$FPR = \frac{|FP|}{|TN| + |FP|}. \tag{3}$$

A curve is generated using a sliding decision threshold; the value at which positive and negative prediction cutoffs are determined for a predictor. As the decision threshold is moved from one end of the x-axis to the other, a confusion matrix can ge generated for each point. By plotting each of these points the result is the ROC curve.

The ROC curve shows several things. First, there is always a trade-off between sensitivity and specificity: if there is an increase in sensitivity, there will be a decrease in specificity. If the decision threshold is shifted left, the threshold for identifying a

positive result is decreased. This results in a greater proportion of the datapoints in the overlapping section to be considered positives and consequently decreases the number of those overlap points which are considered negative. Second, the ROC curve also represents the probability that the predictor evaluated will score a randomly selected positive variant higher than a randomly selected negative variant [Tape, 2015]. Finally, the closer the curve follows the left hand border and the top border, the more accurate the test is. Conversely, if the curve more closely follows the 45 degree diagonal, it indicates that the results attained are just as likely to be generated at random. Additionally, should the curve appear under the diagonal, it likely that the predicted labels have been inverted. The AUC value calculated for a particular ROC curve indicates the accuracy of the predictor. The closer to 1.0 the more accurate the test; the closer to 0.5 the more likely the same results could be achieved at random. The AUC value is what is most commonly referenced regarding a test or predictors validity hence it is being employed as the measure for the predictors evaluated.

4 Results

In silico prediction tools have the potential to be valuable in examining large amounts of disease-associated variants to indentify SAVs. This process of reviewing thousands of data points by *in vitro* methods would take years to complete. However in order to have confidence in the results obtained, it is vital to first screen potential *in silico* methods in order to determine those most accurate and suitable for the project's circumstances.

To quantify and compare the performances of multiple prediction tools, we evaluated the performance of each predictor on the EMDS and JRDS. The results returned by the predictors were each plotted as an ROC curve on a single graph for each data set; this allowed for a visual comparison of how the tools performed in comparison with one another and enabled the area under the ROC curve (AUC) to be calculated with respect to each data set. This AUC value provided a numerical basis for comparison of the different tools.

4.1 Performance on Exonic Missense Data Set

The results in Figure 4 show the AUC scores calculated for each ROC curve and demonstrate appreciably good performance by two of the predictors, namely MutPred Splice [Mort et al., 2014] and Skippy's distance from the junction parameter [Woolfe et al., 2010]. As is seen in the ROC curve in Figure 4A, the lines for these two predictors demonstrate a much more convex shape. However the remaining curves follow a more linear path much closer to the 45 degree diagonal.

In looking at all four of the predictors, SPANR, MutPred Splice, Skippy, and MaxEnt, there were dramatic differences in the results returned by the various *in silico* tools which evaluated in the EMDS. The data indicates that MutPred Splice outperformed the other predictors with an AUC value of 0.87 (Figure 4B) followed by

Skippy's distance to the junction with 0.78. A distinguishing similarity between these two predictors that is not present for SPANR or Skippy LOR is the reliance on the variant location within the exon in assigning a score regarding the impact on splicing. Examination of all four predictor outcomes suggests that variant location with respect to the exon-intron junction appeared to be a confounding variable affecting the results returned by the predictors. To control for this variable, the Junction Region data set was evaluated.

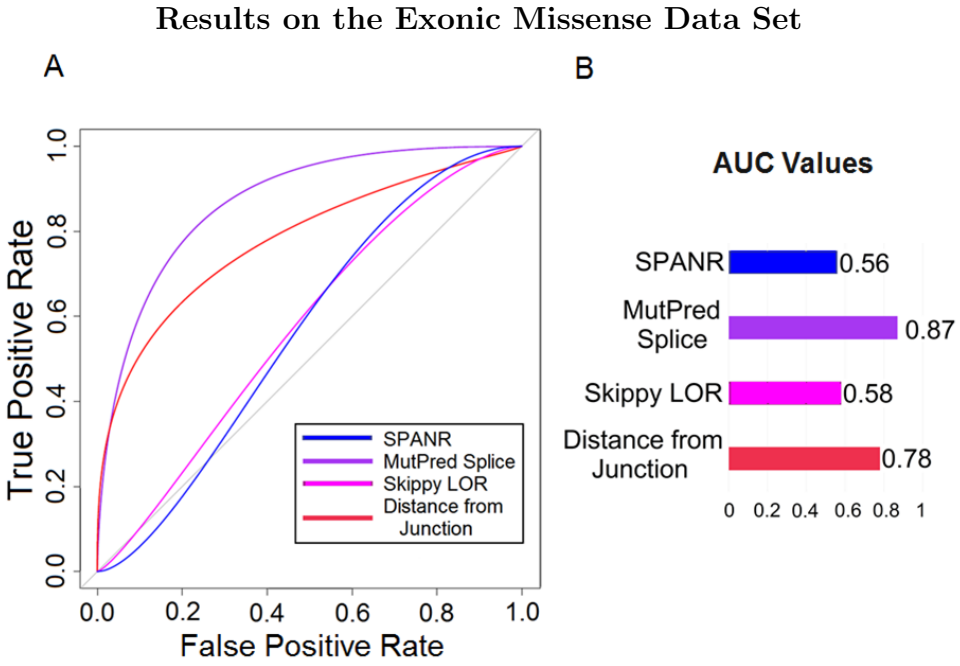

Figure 4: Figure A shows ROC curves for the Exonic Missense data set for the predictors SPANR, MutPred Splice, and Skippy. The Skippy LOR and Distance from Junction determinations were both from the Skippy predictor. Figure B shows a chart of AUC values from the ROC curve data using the Exonic Missense data set.

4.2 Performance on Junction Region Data Set

In the analysis of the Junction Region data set all of the predictors performed relatively similarly. As can be seen in Figure 5A, the ROC curves created from the predictors' results are all located closer to the 45 degree diagonal (shown in grey) than to the left-hand and top borders of the graph, and four of the five lines demonstrate a more linear shape rather than the parabolic curve that would be expected if the predictor performs well. Figure 5B quantifies the similar performances of all the predictors as demonstrated by their respective AUC scores. All are within eight one-hundredths of one another, and the range is substantially closer to a value of 0.5 than that of 1.0.

The variants included in the JRDS are all within three nucleotides of either the donor or acceptor splice sites, thus essentially removing any effect that variant location should have on the predictors' evaluation. An explanation of the results seen here for all the predictors might therefore lie in the characteristics of the dataset itself.

Results on the Junction Region Data Set

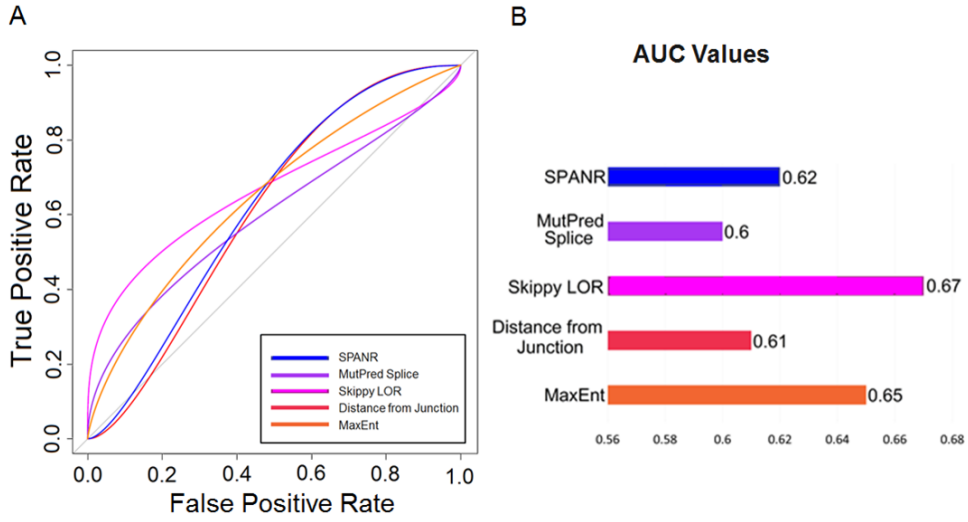

Figure 5: Figure A shows ROC curves for the Junction Region data set for the predictors SPANR, MutPred Splice, Skippy, and MaxEnt. The Skippy LOR and Distance from Junction determinations were both from the Skippy predictor. Figure B shows a chart of AUC values from the ROC curve data using the Junction Region data set.

5 Discussion

The process of splicing removes introns and particular exons from the pre-mRNA transcript in order to facilitate its translation into a protein. Due to the complexity of splicing, factors which result in disruption of the normal flow of this process have the potential to cause or contribute to the development of disease. This connection has been confirmed for a range of diseases from phenylketoneuria [Dobrowolski et al., 2010] to cystic fibrosis [Kuyumcu-Martinez et al., 2007]. Disruptions in splicing could be the result of mutations in the splice site or ESR motifs, or from damage to the components of the spliceosome itself that cause aberrant splicing of the pre-mRNA and thus disease.

The use of *in silico* methods to predict the effects from variants in the genetic code enables the evaluation of both their local impact and the likelihood that they will result in disease. By correctly identifying variants that contribute to gene-associated diseases, the possibility arises for developing new treatments. Many of the diseases associated with aberrant splicing cause serious disruption to the quality of life of the patients afflicted; thus, advances in treatment that might emerge from this information have the potential to be very beneficial.

5.1 Predictors are Biased by Distance to the Junction

As can be seen in Figure 4 when the performance of selected predictors on the EMDS was evaluated, the overall outcome from two of the predictors was distinctly better than the others. It was surmised from these results that with respect to the EMDS, predictors which take into consideration the distance the variant lies from the exon-intron junction perform better than those which do not. The prediction tool which performed the best based on its AUC score of 0.87 was MutPred Splice [Mort et al., 2014], which incorporates distance to the junction as a feature. The

Skippy distance to the junction parameter [Woolfe et al., 2010] though still inferior to MutPred Splice's combined approach, still out-performed the other two predictors.

Further explanation of the performance seen here is that this data set is enriched in positive variants and depleted in negative data points in the region of the exon-intron junction. If one looks at the distribution of positive (SAVs) and negative (NSAVs) points contained in the EMDS, 52% of those variants annotated as splice-affecting lie at the first position next to the exon-intron junction (Figure 6). The negative points, those annotated as causing disease by a mechanism not related to splicing, are overwhelmingly located far from the junction; 96% are six or more bases away. As MutPred Splice uses multiple features to make a determination with respect to how much a variant may impact splicing, this enrichment is expected to differently impact the apparent performance of the tool with respect to this data set.

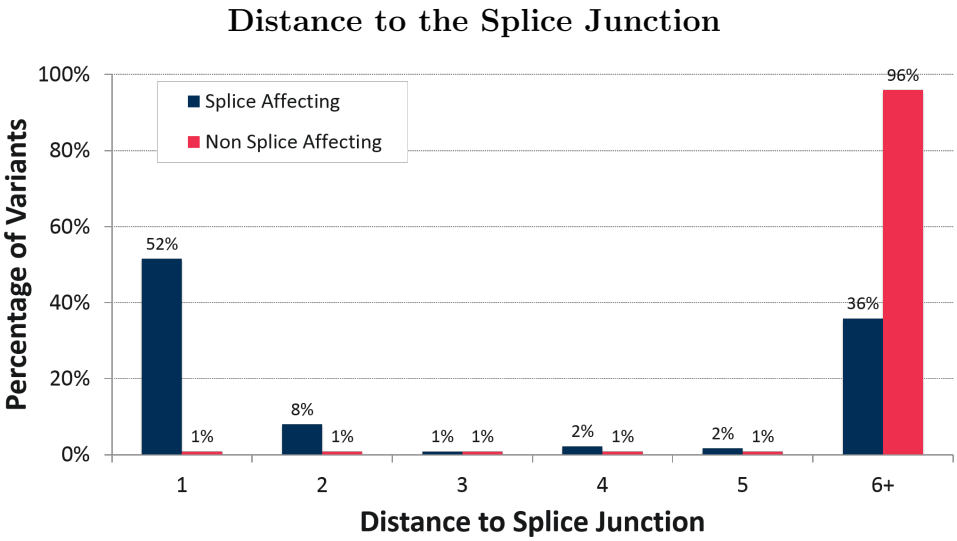

Figure 6: Distribution of the proportion of variants at each position one to six or more bases from the exon-intron junction.

However, when the distance to the junction is controlled as a confounding variable in the JRDS, the results returned are noticeably different. Under these circumstances, the tool that performs the best is MaxEnt. MaxEnt is a traditional motif-based method which only looks at sequences containing the splice site motifs, within three

exonic nucleotides of the border between the exon and the intron.

5.2 False Negatives at Donor Sites

As the AUC scores in Figure 5 demonstrate, all four predictors performed poorly on the JRDS; this is likely due to the characteristics of the data set itself. Figure 7 shows the proportions of the positive and negative variants in the region of the donor site. There are a large number of negative variants in the first position at the exon-intron border. Research has suggested that this location is highly conserved containing "key" nucleotides in the 5′ splice site motif, disruption of which is likely to cause aberrant splicing. Therefore this suggests that the dataset likely contains a large number of false negatives.

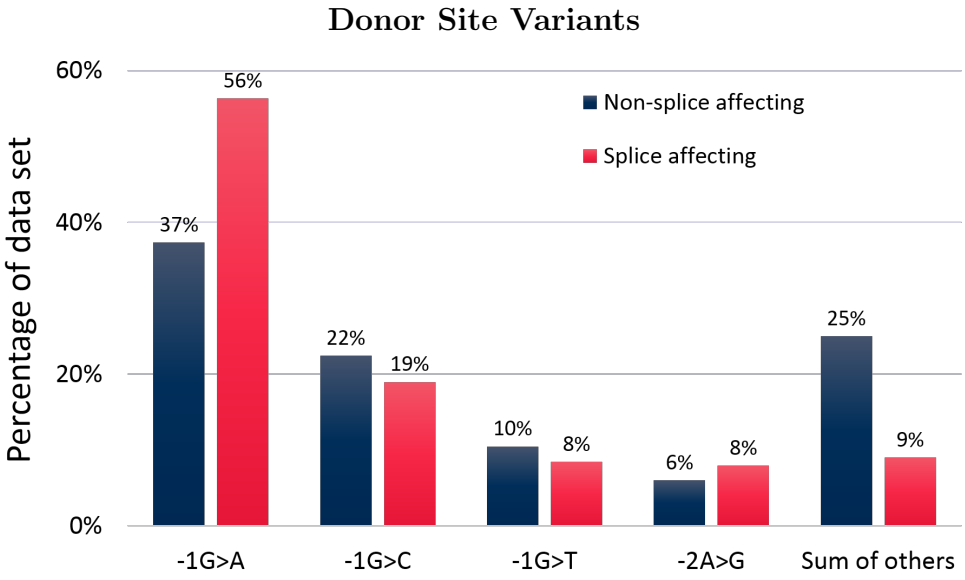

Figure 7: Distribution of positive and negative data points at the donor site for the JRDS. The vast majority of all the variants are located at the first position. The substantial number of putative non-splice-affecting variants suggests that the data set may contain a considerable number of false negatives. The last pair of bars is the sum of the additional variants not explicitly listed; the majority of these are variants which appear only once or twice.

Further explanation for the poor predictor performance on the JRDS is illustrated by the creation of two sample logos (TSLs). The TSLs were formed from the com-

parison of positive and negative, or mutant and wild-type nucleotide sequences at either the donor or acceptor sites. The web-based image generation program created a visual representation of which bases were enriched and depleted at each location within the splice site motif [Vacic et al., 2006]. Figure 8, Figure 9, and Figure 10 are two sample logos (TSLs) constructed from the JRDS variants. In examining the motifs in Figure 8 and Figure 9 one can see that mutations located at position three at the 5′ splice site cause both splicing associated and non-splicing associated diseases. This indicates that many of the NSAVs at this location may be misannotated in the database from which they were collected. These "key" nucleotides are those closest to the junction and appear to be the most highly conserved; thus any alterations are expected to have detrimental consequences [Lee, 2015].

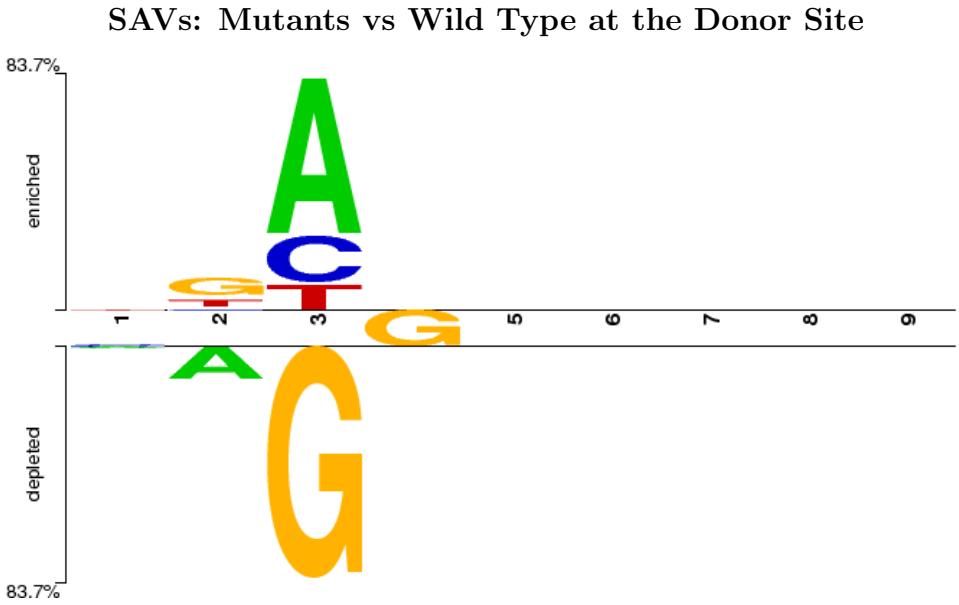

Figure 8: At position three adjacent to the junction the A, C, and T enrichments; and G depletion appear to cause splicing associated disease. These are the same enrichments and depletions seen in the NSAVs TSL.

The result of this enrichment of possible false negatives at the junction likely causes a disturbance in the overall results from the predictors and is the reason behind the surprisingly low AUC values.

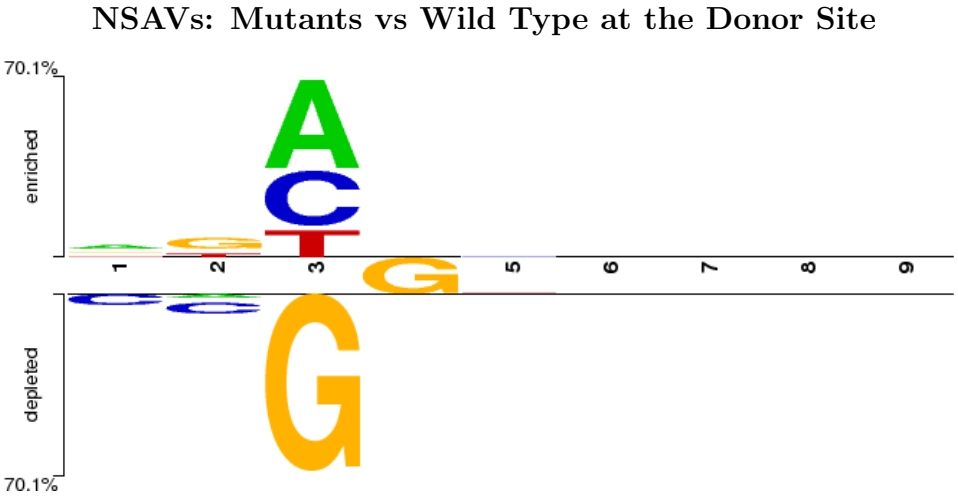

Figure 9: At position three adjacent to the junction the A, C, and T enrichments; and G depletion appear to cause non-splicing associated disease. These are the same enrichments and depletions seen in the SAVs TSL.

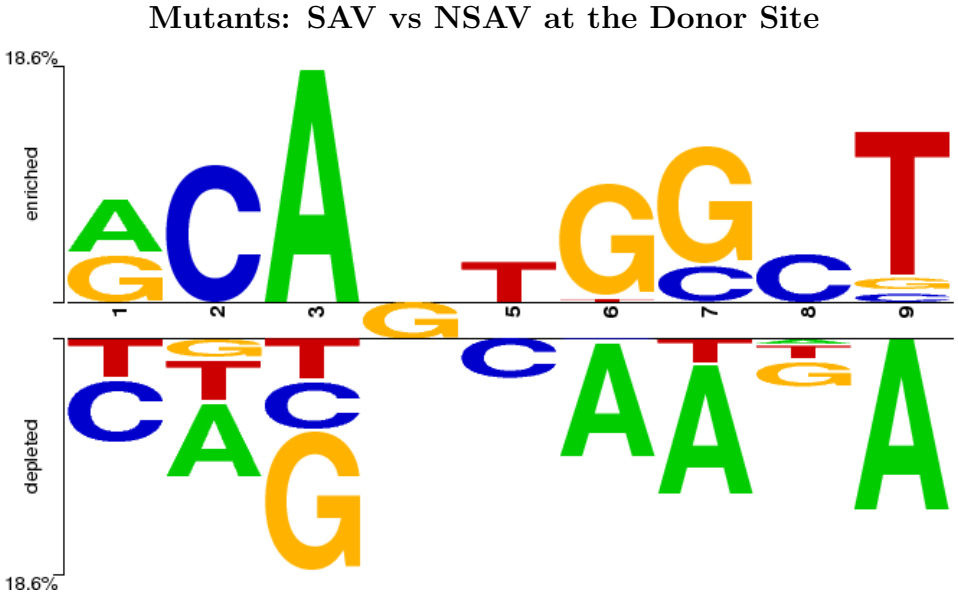

Figure 10: Based on the singular enhancement at positions two and three of the donor site, the TSL suggests that at least some of the SAVs and NSAVs which were evaluated in this TSL are misannotated.

5.3 Conclusions

The development of bioinformatics has been of great help in biology by facilitating the processing of immense amounts of data in much shorter periods of time than ever before. New tools are constantly being considered, created, and polished to further expediate and expand the ability of scientists to probe and discover; one type of such an *in silico* tool is the prediction tools developed to predict how genetic variants will impact the process of splicing and may thus cause disease.

In the evaluations of such predictors performed here, it has been revealed that multiple factors influence the efficacy of such tools. The type of analysis process employed by the predictor is a significant factor in determining how the tool will perform; and perhaps equally important is the type of data input. Depending on the evaluation method, features of the data being evaluated such as variant location, type of variant, etc. will influence the accuracy of the results provided.

We have determined that a key feature important for the value of any of the prediction tools is the quantity and ease of data processing. For these tools to be truly helpful, they must be high-throughput. Frequently the types of investigations which make use of these *in silico* methods are processing quantities of data with hundreds or thousands of individual points. Those tools which were not designed to process large amounts of data may still have utility for smaller inquiries, but are not likely to be employed for larger projects. Also, predictors which required the data to be input in formats that require significant amounts of additional research to find the information, or to construct the formatting are again of little use when it comes to selecting a tool for a large project.

From our consideration of numerous *in silico* prediction tools, and the systematic evaluation of four of these we have determined that bioinformatics prediction tools for the evaluation of splicing are useful to determine with respectable accuracy the likelihood that a given variant will affect splicing. Yet as some predictors have

demonstrated biases with respect to the distance the variant being evaluated lies from the exon-intron junction, this should be taken into consideration when selecting the appropriate tool for analysis in a particular project. It is also important to carefully evaluate the results obtained since data sets may contain characteristics, such as false positives or negatives, or enrichments of various types; and these elements are likely to influence the accuracy of the results returned.

There is clearly room for improvement in the construction of *in silico* predictors. As new prediction methods are being developed and refined, it would be prudent to revisit this comparison with the inclusion of new methods such as ΔtESRseq and ΔHZ_{EF}[Soukarieh et al., 2016].

Another future direction for predictor evaluation might be to consider their performance with respect to cancer-causing variants. The fact that these cell types behave substantially differently from those in other types of disease and from each other, may provide additional illumination with respect to how such predictors may be developed and improved. There is still a large potential for new innovations in the field of *in silico* predictors. And as these tools prove useful and even irreplaceable in new ways, scientists will undoubtedly continue to employ and advance them within various fields.

References

S. Akli, J. Chelly, C. Mezard, S. Gandy, A. Kahn, and L. Poenaru. A" g" to" a" mutation at position-1 of a 5'splice site in a late infantile form of tay-sachs disease. *Journal of Biological Chemistry*, 265(13):7324–7330, 1990.

M. Amit, N. Sela, H. Keren, Z. Melamed, I. Muler, N. Shomron, S. Izraeli, and G. Ast. Biased exonization of transposed elements in duplicated genes: a lesson from the TIF-IA gene. *BMC Molecular Biology*, 8(1):109, 2007.

S. C. Bell, K. De Boeck, and M. D. Amaral. New pharmacological approaches for cystic fibrosis: promises, progress, pitfalls. *Pharmacology & therapeutics*, 145:19–34, 2015.

K. A. Cartegni L, Chew SL. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, 3(4):285–98, 2002.

Z. Z. Z. M. Cartegni L, Wang J and K. AR. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, 31(13):3568–3571, 2003.

T. A. Cooper, L. Wan, and G. Dreyfuss. Rna and disease. *Cell*, 136(4):777–793, 2009.

F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud. Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, 37(9):e67–e67, 2009.

S. F. Dobrowolski, H. S. Andersen, T. K. Doktor, and B. S. Andresen. The phenylalanine hydroxylase c. 30c¿ g synonymous variation (p. g10g) creates a common exonic splicing silencer. *Molecular genetics and metabolism*, 100(4):316–323, 2010.

L. Eng, G. Coutinho, S. Nahas, G. Yeo, R. Tanouye, M. Babaei, T. Dörk, C. Burge, and R. A. Gatti. Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *Human Mutation*, 23(1):67–76, 2004.

B. U. Fitzky, M. Witsch-Baumgartner, M. Erdel, J. N. Lee, Y.-K. Paik, H. Glossmann, G. Utermann, and F. F. Moebius. Mutations in the δ7-sterol reductase gene in patients with the smith–lemli–opitz syndrome. *Proceedings of the National Academy of Sciences*, 95(14):8181–8186, 1998.

K. Furihata, A. Drousiotou, Y. Hara, G. Christopoulos, G. Stylianidou, V. Anastasiadou, I. Ueno, and P. Ioannou. Novel splice site mutation at ivs8 nt 5 of hexb responsible for a greek-cypriot case of sandhoff disease. *Human mutation*, 13(1): 38–43, 1999.

W. Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.

C. Heintz, S. F. Dobrowolski, H. S. Andersen, M. Demirkol, N. Blau, and B. S. Andresen. Splicing of phenylalanine hydroxylase (pah) exon 11 is vulnerable: molecular pathology of mutations in pah exon 11. *Molecular genetics and metabolism*, 106(4):403–411, 2012.

S. Igreja, L. A. Clarke, H. M. Botelho, L. Marques, and M. D. Amaral. Correction of a cystic fibrosis splicing mutation by antisense oligonucleotides. *Human mutation*, 37(2):209–15, 2015.

H. Keren, G. Lev-Maor, and G. Ast. Alternative splicing and evolution: diversi-

fication, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

A. R. Kornblihtt, M. de la Mata, J. P. Fededa, M. J. Munoz, and G. Nogues. Multiple links between transcription and splicing. *RNA*, 10(10):1489–1498, 2004.

Krainer Lab; Zhang Lab; Cold Spring Harbor Laboratory. ESE Finder 3.0: matrices & thresholds, 2007. URL http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=matrices. Reviewed: 2-27-2007; Accessed: 5-11-2015.

M. Krawczak, N. S. Thomas, B. Hundrieser, M. Mort, M. Wittig, J. Hampe, and D. N. Cooper. Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mrna splicing. *Human Mutation*, 28(2):150–158, 2007.

J. Kreahling and B. R. Graveley. The origins and implications of *Alu*ternative splicing. *Trends in Genetics*, 20(1):1–4, 2004.

N. M. Kuyumcu-Martinez, G.-S. Wang, and T. A. Cooper. Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. *Molecular Cell*, 28(1):68–78, 2007.

F. Lee. Molecular biology web book. 2015. URL http://www.web-books.com/MoBio/Free/Ch5A4.htm. Accessed: 2-24-2015.

A. Levit, D. Nutman, E. Osher, E. Kamhi, and R. Navon. Two novel exonic point mutations in hexa identified in a juvenile tay-sachs patient: role of alternative splicing and nonsense-mediated mrna decay. *Molecular genetics and metabolism*, 100(2):176–183, 2010.

B. P. Lewis, R. E. Green, and S. E. Brenner. Evidence for the widespread coupling of

alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences*, 100(1):189–192, 2003.

P. Linsdell. Interactions between permeant and blocking anions inside the CFTR chloride channel pore. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 2015.

B. Lubamba, B. Dhooghe, S. Noel, and T. Leal. Cystic fibrosis: insight into CFTR pathophysiology and pharmacotherapy. *Clinical Biochemistry*, 45(15):1132–1144, 2012.

S. Maiella, A. Rath, C. Angin, F. Mousson, and O. Kremp. Orphanet and its consortium: where to find expert-validated information on rare diseases. *Revue Neurologique*, 169:S3–8, 2013.

A. G. Matera and Z. Wang. A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, 15(2):108–121, 2014.

B. Modrek and C. Lee. A genomic view of alternative splicing. *Nature Genetics*, 30 (1):13–19, 2002.

M. Mort, T. Sterne-Weiler, B. Li, E. V. Ball, D. N. Cooper, P. Radivojac, J. R. Sanford, and S. D. Mooney. Mutpred splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology*, 15(1):R19, 2014.

R. Myerowitz. Splice junction mutation in some ashkenazi jews with tay-sachs disease: evidence against a single defect within this ethnic group. *Proceedings of the National Academy of Sciences*, 85(11):3955–3959, 1988.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethseda (MD): The Library; 2013 Sep 16. *DHCR7*, 2013a. URL https://ghr.nlm.nih.gov/gene/DHCR7. Reviewed: 07-2007; Accessed: 3-16-2016.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethseda (MD): The Library; 2013 Sep 16. Phenylketonuria, 2013b. URL https://ghr.nlm.nih.gov/condition/phenylketonuria. Reviewed: 02-2012; Accessed: 3-16-2016.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethseda (MD): The Library; 2013 Sep 16. Sandhoff disease, 2013c. URL https://ghr.nlm.nih.gov/condition/sandhoff-disease. Reviewed: 09-2008; Accessed: 3-16-2016.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethseda (MD): The Library; 2013 Sep 16. Smith-Lemli-Opitz syndrome, 2013d. URL https://ghr.nlm.nih.gov/condition/smith-lemli-opitz-syndrome. Reviewed: 07-2007; Accessed: 3-16-2016.

National Library of Medicine (US). Genetics Home Reference [Internet]. Bethseda (MD): The Library; 2013 Sep 16. Tay-Sachs disease, 2013e. URL https://ghr.nlm.nih.gov/condition/tay-sachs-disease. Reviewed: 10-2012; Accessed: 3-09-2016.

D.-K. Niu and Y.-F. Yang. Why eukaryotic cells use introns to enhance gene expression: Splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase i cutting activity. *Biol Direct*, 6:24, 2011.

R. A. Padgett. New connections between splicing and human disease. *Trends in Genetics*, 28(4):147–154, 2012.

S. W. Roy and M. Irimia. Splicing in the eukaryotic ancestor: form, function and dysfunction. *Trends in Ecology & Evolution*, 24(8):447–455, 2009.

P. J. Smith, C. Zhang, J. Wang, S. L. Chew, M. Q. Zhang, and A. R. Krainer. An increased specificity score matrix for the prediction of sf2/asf-specific exonic splicing enhancers. *Human molecular genetics*, 15(16):2490–2508, 2006.

O. Soukarieh, P. Gaildrat, M. Hamieh, A. Drouet, S. Baert-Desurmont, T. Frébourg, M. Tosi, and A. Martins. Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS genetics*, 12 (1):e1005756–e1005756, 2016.

P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. D. Phillips, and D. N. Cooper. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9, 2014.

P. Szauter. Donor splice site description, 2015. URL http://www.discoveryandinnovation.com/BIOL202/notes/lecture26.html. Accessed: 5-25-2015.

T. G. Tape. Interpreting diagnostic tests: Roc curves. 2015. URL http://gim.unmc.edu/dxtests/ROC1.htm. Accessed: 11-3-2015.

J. Tazi, N. Bakkour, and S. Stamm. Alternative splicing and disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1792(1):14–26, 2009.

The Broad Institute. Exome Aggregation Consortium (ExAC), 2015. URL http://exac.broadinstitute.org. Accessed: 5-25-2015.

V. Vacic, L. M. Iakoucheva, and P. Radivojac. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22(12):1536–1537, 2006.

N. Wakamatsu, H. Kobayashi, T. Miyatake, and S. Tsuji. A novel exon mutation in the human beta-hexosaminidase beta subunit gene affects 3'splice site selection. *Journal of Biological Chemistry*, 267(4):2406–2413, 1992.

M. Wang and A. Marín. Characterization and prediction of alternative splice sites. *Gene*, 366(2):219–227, 2006.

M. C. Wollerton, C. Gooding, E. J. Wagner, M. A. Garcia-Blanco, and C. W. Smith. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Molecular Cell*, 13(1):91–100, 2004.

A. Woolfe, J. C. Mullikin, and L. Elnitski. Genomic features defining exonic variants that modulate splicing. *Genome Biology*, 11(2):R20, 2010.

X. Wu and L. D. Hurst. Why selection might be stronger when populations are small: intron size and density predict within and between-species usage of exonic splice associated cis-motifs. *Molecular Biology and Evolution*, page msv069, 2015.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347 (6218):1254806, 2015.

T. Yoshizawa, Y. Kohno, S. Nissato, and S. Shoji. Compound heterozygosity with two novel mutations in the hexb gene produces adult sandhoff disease presenting as a motor neuron disease phenotype. *Journal of the neurological sciences*, 195(2): 129–138, 2002.

H. Yu, M.-H. Lee, L. Starck, E. R. Elias, M. Irons, G. Salen, S. B. Patel, and G. S. Tint. Spectrum of δ7-dehydrocholesterol reductase mutations in patients with the smith–lemli–opitz (rsh) syndrome. *Human molecular genetics*, 9(9):1385–1391, 2000.