Learning Alternative RNA Splicing with Applications to Mutagenesis

A thesis submitted for the degree of
Doctor of Philosophy of the University of Oxford



Edward Sanders

University College
University of Oxford

Wellcome Trust Doctoral Training Programme
in Genomic Medicine and Statistics

Weatherall Institute of Molecular Medicine

Michaelmas Term 2022

Abstract

Alternative RNA splicing is the process by which a single gene generates a large diversity of transcripts, but how the DNA sequence results in particular transcript abundances remains incompletely known. We present a two-stage machine learning algorithm that models this process by predicting the frequency of splice junction usage, with a particular focus on transcripts with low relative expression. We find that nonsense-mediated decay plays an important role in shaping transcript abundance, resulting in splice-altering mutations leaving no evidence in RNA-seq libraries. Using the model, we build a pipeline for annotating putative disease-causing mutations, and show state-of-art performance on HGMD mutations known to affect splicing. We also apply the model to common variation associated with disease phenotypes in GWAS, and show that variants annotated to affect splicing are significantly enriched, indicating that splicing plays a key mechanistic role in human disease aetiology arising from common genetic variation. These variants are found both near annotated splice-sites and in deep-intronic regions. Our model predicts that the deep-intronic variants often modify gene expression through activation of cryptic alternative splice sites followed by nonsense-mediated decay (AS-NMD), providing further evidence of AS-NMD as a common mechanism of gene regulation.

This thesis is dedicated to my wife Elisabetta. I am forever grateful to her for helping me during the challenging period of pandemic lockdown, which coincided with the birth of our beautiful boy Renato. During this PhD, my wife and son always managed to bring joy and love when work and life were otherwise challenging.

Acknowledgements

My deepest gratitude goes to my supervisors, Jim Hughes and Gerton Lunter. It has been an incredible experience to learn from their scientific wisdom and to work on projects with them. I am particularly grateful for the additional support they gave me when I was working remotely. It has been a fantastic experience scientifically and personally to work with these brilliant people.

I am also grateful to those involved in my programme, Genomic Medicine and Statistics, for all the training courses they provided, and to the Wellcome Trust for funding this programme.

I would also like to thank all the members of my labs who have been very helpful over the years, and in particular: Richard Brown, Chris Cole, Damien Downes, Ron Schwessinger, Lance Hentges, Jennifer Herrmann, and Emily Georgiades.

I am very grateful to Jenny Taylor, Edoardo Giacopuzzi and Alistair Pagnamenta for their collaboration. Being able to apply our work to real clinical cases was a most satisfactory experience. Similarly, I am grateful to Nick Proudfoot and Ruis Sérgio de Sousa Luís who provided us with additional transcriptomic datasets.

Finally, I am grateful to my parents, who unconditionally and patiently supported my education in music, mathematics and science, despite these being disciplines somewhat removed from their own field of work. My brother, Henry, who has supported me in all periods of my life. Special thanks to my son Renato, who refused to take my work seriously and who, during the pandemic period, demanded that I be involved in his playing with tractors and legos. Lastly, Elisabetta who has shown the utmost love and support.

Table of Contents

1 Introduction **8**

1.1 Acronyms 8

1.2 Background 9

1.3 Problem statement 12

1.4 Thesis outline 14

2 Alternative Splicing model **16**

2.1 Abstract 16

2.2 Introduction 16

2.3 Percent Spliced In 23

2.3.1 Defining the training dataset from read junction data 25

2.3.2 PSI plots for GTEx evidenced splice sites 27

2.3.3 PSI specific training loss 29

2.3.4 Evidence as a loss function – a path with a closed form formula 30

2.4 Modelling approach 32

2.4.1 Neural network architecture 33

2.5 Results 34

2.5.1 Comparison of empirical PSI and predicted PSI 36

2.5.2 Effect of transcription levels 40

2.6 Model generated hypotheses for splicing 46

2.6.1 Analysis of local splicing factors 47

2.6.2 Effect of acceptor loss 50

2.6.3 Effect of donor loss 53

2.6.4 Gain of Intron vs Loss of Exon 55

2.6.5 Recursive splicing 60

2.6.6 Estimating the prevalence of recursive splicing in the transcriptome 66

2.7 Discussion and conclusions 68

3 Predicting the effect of mutations **71**

3.1 The HGMD dataset 72

3.2 Applying Neural network to mutagenesis 73

3.3 Effect of combining independently models on mutagenesis 74

3.4 Applying downstream Machine learning to the Classifier 79

3.4.1 Improving classification using XGBoost 79

3.4.2 Building the trained networks 80

3.4.3 Cross validation parameter grid search 80
3.4.4 Importance scores for each parameter input 81
3.5 Results and Reapplication of pipeline to HGMD dataset 83
3.6 Annotations and their consistency 86
3.7 Collaboration with disease diagnostics platforms 95
3.8 Conclusion 98

4 Finding Nonsense amongst our Predictions 100
4.1 Motivation 102
4.2 Summary of mechanisms linked with NMD 103
4.3 Evidence amongst unannotated predicted alternative splice sites 105
4.3.1 Alternative acceptors 108
4.3.2 Alternative donors 111
4.3.3 Predicted inclusion of exon in annotated intronic region 117
4.3.4 Using model to estimate percentage of strong splice sites are AS-NMD sites 121
4.3.5 Intermediate conclusion 124
4.4 NMD model pipeline 124
4.5 Position of open reading frame 128
4.6 Application of model to NMD datasets 129
4.6.1 Worked gene examples 130
4.6.2 Gained NMD exon examples 132
4.6.2.1 SRSF11 132
4.6.2.2 DHX9 136
4.6.3 Alternative acceptors 138
4.6.3.1 ADAM15 138
4.6.4 Alternative donors 139
4.6.4.1 OBSCN 139
4.6.5 Gained intron 140
4.6.5.1 TRIP13 140
4.6.6 Alternative initial donor site 141
4.6.7 Estimates of NMD reducing expression 143
4.6.8 Application of NMD model to POINT-seq 150
4.6.9 Analysis of other experimental approaches to detect NMD associated splicing events. 157
4.6.10 Kovalev et al Exon Junction Complex analysis 158
4.7 Discussion 161

5 Application to common human variation in genetics 164
5.1 Read evidence supporting 164
5.2 Results of NMD suppression 166

5.2.1 Alternative acceptor site results 167
5.2.2 Alternative donor site results 170
5.2.3 Model predicts gain of exon 172
5.3 Conclusions 174

6 Applications in common disease fine mapping 176
6.1 Enrichment analysis introduction 178
6.2 Initial look at predicted splice altering GWAS SNPs 179
6.2.1 Balanced Genomic region enrichment analysis 187
6.3 Effect of combining multiple neural network models 196
6.4 Enrichment in or near annotated exons 203
6.5 Enrichment in deep intronic regions 205
6.6 LD-bin Enrichment analysis 208
6.7 Examples of predicted NMD splice altering GWAS catalogue polymorphisms. 210
6.7.1 SPATA1 210
6.7.2 UBP1 example 212
6.7.3 TCF19 gene 213
6.8 Discussion 214

7 Discussion 217
7.1 Why is the prediction of splicing and the effects of sequence variation important? 217
7.2 What was lacking in existing approaches? What motivated our work? 218
7.3 Technological advances 219
7.4 Biological of splicing 219
7.5 Future perspectives 222

8 References 225

1 Introduction

1.1 Acronyms

AS-NMD: alternative splicing and nonsense mediated decay

AUC: area under curve

AUCPR: area under the precision-recall curve

GWAS: genome wide association study

HGMD: Human genome mutational database

NGD: no go decay

NMD: nonsense mediated decay

NSD: non stop decay

nt: nucleotide

PSC: premature stop codon

PSI: percent spliced in

RS: recursive splicing

RT-PCR: Reverse transcription polymerase chain reaction

SNP: single nucleotide polymorphism

SNV: single nucleotide variant

SR: serine and arginine-rich

TPM: transcripts per million

TSS: transcription start site

UTR: untranslated region

1.2 Background

Splicing is a fundamental process involved in the processing of RNA molecules prior to translation of a gene from its nascent transcription from genomic DNA into a protein product. A gene can span a million base pairs or more of the genome; however, only a fraction of these base pairs actually code for the amino acids of the protein sequence and are known as *exons*, while the remaining bases are referred to as *introns*. *Splicing* is the process that removes intronic bases from a nascent transcript to form the mature mRNA transcript that is ready for translation into protein. The set of exons that can possibly contribute to the protein is termed the *gene definition*, and for a given transcription start site and polyadenylation region, it is therefore the process of splicing that defines which genomic regions are included in the mature transcript. Not all exons contribute to all mature mRNA transcripts in all cells and all situations; instead, some exons may be included only in some transcripts, and some exons can exist in multiple forms by moving the splice site at either end, in effect changing the length of that exon. This process is known as *alternative splicing* and is the major mechanism by which a single gene can result in many, typically 10 or so, different potential protein products.

Because of its profound impact on the final protein product, it is not surprising that splicing is implicated in human disease with various estimates suggesting a large proportion of human disease is associated with the loss or gain of splice sites [1]. A better understanding of splicing can therefore provide large diagnostic benefits and so there has been a strong focus on developing and evaluating *in silico* methods to better predict splicing [2]–[4] and the effect of sequence variation on it. The benefits of such approaches include better prioritisation of putative variants and the ability to perform *in silico* experiments. This provides the ability to formulate and test hypotheses before performing potentially more

expensive *in vivo* validation or experiments. Limited access to tissues, in particular during stages of human foetal or embryonic development, motivates an *in silico* approach. Obviously, the major logistical advantage of an *in silico* model is that a user could test millions of variants with only limited computational costs.

In addition to *alternative splicing* not all transcripts contribute to mature RNA due to RNA-surveillance, which degrades certain gene definition transcripts prior to translation. The three main RNA-surveillance pathways are Nonsense Mediated Decay (NMD), No-Go Decay (NGD) and Non-stop Decay (NSD). These are described by the presence of a premature stop-codon, the stalling of ribosomes, and the absence of a stop-codon respectively [5]. These surveillance pathways are typically described as quality control mechanisms however they also have a functional interaction with *alternative splicing* itself. In particular the coupling of *alternative splicing* and *nonsense mediated decay* (AS-NMD) can act as a means of gene regulation [6], particularly for genes involved in RNA processing. In this context the interaction of NMD with alternative splicing is not merely viewed as a mechanism for quality control but as a potential mechanism to increase/decrease gene-expression by increasing or decreasing the expression of AS-NMD transcripts. In particular AS-NMD has been seen to be used as a system of regulation for splicing factors themselves [7].

Despite substantial attempts to model the splicing process, attempts to model alternative splicing coupled to NMD are lacking. An important consideration in both modelling and validating these models is that AS-NMD isoforms have the potential to be missed by standard RNA-seq assays [8]. Consequently, the development of *in silico* methods based on these types of data tend to miss these transcripts in their training or evaluation processes.

It remains contested the extent to which AS-NMD is utilised purely as a quality control mechanism or as a conserved mechanism of gene-regulation [6], [9]. One hypothesis is that the vast majority of AS events which give rise to NMD are merely a consequence of "noisy splicing". From the perspective of developing and utilising *in silico* models both hypotheses motivate combining the two mechanisms into the model. Both the "noisy splicing" hypothesis and the conserved regulatory AS-NMD hypothesis predicate that *alternative splicing* includes a substantial proportion of splicing events which are not readily seen in standard RNA-seq libraries.

Therefore, any attempt to model and predict more fully the spectrum of *alternative splicing* from DNA sequence will be more prone to uncover these quality control mechanisms, even though they will be poorly validated in standard RNA-seq datasets. In intronic regions many pseudo-exons exist as "decoy exons" which evade annotation in standard gene-databases but exist as weak alternatively spliced-in exons to alter splicing [10]. This leaves the challenge in generating and deploying these models is to distinguish true but degraded splicing events from false positives in the model.

In addition to RNA-decay pathways, it is also important to better understand recursive splicing. It has been seen that recursive splicing is found in long introns [11]. Similar to coupled alternative splicing to RNA-decay, this process evades capture in RNA-seq libraries.

A consequence of developing an *in silico* approach which targets alternative splicing is to better understand its prevalence and the potential effect and functional consequences of sequence variation on it.

In summary, modelling the effects of splicing variation on disease biology is critically important to consider the effect on the overall functional output of the gene taking RNA surveillance into consideration. Changing the proportion of the output of a gene, which is

either degraded or contributes to the stable, translated pool, ultimately affects the amount of functional protein produced for the gene and so can drive human genetics. Therefore, both clinical mutations and normal variation associated with common diseases must be modelled, not only in terms of altered stable transcript structure but also on the effects on proportion of stable RNA and amounts of protein produced.

1.3 Problem statement

Our ambition is to model the process of *alternative splicing* as precisely as possible from local sequence. The ideal scenario would be to have a model which, taking the transcription start site, polyadenylation region along with the intermediate sequence, would predict the gene-definition in between. Such a model would be able to quantify the likely percentage of protein isoforms which utilise a particular splice site. In addition, such a model would be able to predict the presence of more hidden mechanisms, such as RNA-decay or recursive splicing, which are not typically visible in gene-definitions.

A key purpose of such a model is to enable the performance of *in-silico* experiments which enable the better prediction of the molecular phenotype. Such predictions can be utilised in various ways, from better understanding and determining disease variation to formulating hypotheses pertaining to the nature of splicing.

Existing models attempt to model *alternative splicing* categorically as present or absent. By designing a method of learning *alternative splicing* which utilises the full continuum of splice sites: from rarely used to constitutive, it provides the possibility of furthering the field. Furthermore, our analysis suggests that processes like nonsense mediated decay exaggerate

the appearance of bimodality of alternative splicing. Developing a model which can better account for these low-level splicing events has the potential to overcome the issue that this hidden signal poses to learning models. Moreover, we can add more subtle interpretations to how we believe a mutation or variant might be affecting gene-definition and its effect on the output of a given gene. In this thesis we will present clinical examples where our model correctly identified not only that splicing was affected but additionally the extent in which splice sites were affected.

Our model is deliberate in its attempt to model *alternative splicing* by design to modelling low frequency transcripts as well as more constitutive transcripts. This will direct the model to predict many transcripts which are otherwise unseen due to RNA-surveillance mechanisms. Therefore, in addition to modelling *alternative splicing* it is essential to link such predictions, as thoroughly as possible, to the main RNA-surveillance mechanism: NMD.
In this thesis we present a NMD model which combines with our trained neural network. We use these combined models to show that AS-NMD is abundant amongst *alternative splice* sites. We test these algorithms using both publicly available RNA-seq data as well as using a very well characterised in-house biobank of individuals of known genome sequence with matched transcriptomics. To overcome the challenge of validating such transient but functionally important splicing events using standard transcriptomics data, we test and determine the utility of data from more nascent forms of RNA-seq in addition to NMD inhibition experiments to overcome this challenge.
In summary, we wish to develop an *in silico* approach which not only predicts splicing in a more quantitative manner, but also incorporates the potential interactions with RNA

surveillance mechanisms via gene definition models, in order to predict the effect of sequence changes on the output of genes to decode human genetics.

1.4 Thesis outline

In chapter 2 we outline the neural network model designed to learn alternative splicing. We present validation results showcasing its ability to predict the different classes of alternative splicing. We apply the model to formulate hypotheses about aspects of splicing ranging from gene-expression levels, exon-definition, and recursive splicing.

In chapter 3 we apply the model to disease diagnostics which was motivated by a collaboration with Jenny Taylor's group. This includes validating the model's accuracy on known disease-causing splice-variants with comparison to *SpliceAI*. In addition, we present a machine learning approach which can be used to combine different diagnostic approaches to further increase accuracy.

In chapter 4 we show that a proportion of the model's novel predictions result from RNA-decay. We present a basic bioinformatic model which can partially segregate this. At the end of the chapter we show that an alternative protocol to generate splicing-datasets manages to better reveal nascent RNA-splicing. This has ramifications on our model's accuracy as discussed in chapter 2.

In chapter 5 we apply the model to our internal biobank of 50 individuals with RNA-seq data to showcase how the model can be used with common variation. In chapter 6 we perform enrichment analysis for the neural network model by comparing GWAS catalogue SNPs to

background 1000G SNPs. Combined with the work in chapter 5 this enables us to fine-map some GWAS catalogue examples found in unexpected genomic regions.

The thesis is then brought together with a discussion where we set out where the limitations and future directions of our work reside. Lastly, we discuss how we would take this work further were we to have the time and resources.

Code is available at https://github.com/luntergroup/ALTSPLICE which allows the ability to retrain any discussed model and rerun any discussed analysis.

2 Alternative Splicing model

2.1 Abstract

By building an AI model for alternative splicing we can achieve two goals. The more practical goal is that our model can be applied to help clinicians with diagnostics. This enables clinicians to design functional studies to validate the hypothesis that splicing is indeed the disrupted molecular mechanism.

The other goal is that the model can enable so-called in-silico experiments which allow one to formulate hypotheses about splicing and its disruption. To this end we have carefully designed our neural network so that it is unbiased when formulating these hypotheses.

For instance, we formulate a hypothesis on the importance of local sequence in determining *alternative splicing* based upon whether the gene is highly expressed or not. We formulate hypotheses relating to the extent of intron-definition and exon-definition in splice site selection. Finally, from the model, we formulate a hypothesis concerning how recursive splicing influences exon-skipping. To add confidence in this hypothesis we note that our model has a bias against learning this signal.

2.2 Introduction

Alternative splicing describes the regulatory process in which different isoforms may be selected for the same gene via the selection of differing splice sites. It is estimated that ~90% of genes undergo some form of alternative splicing [12]. Alternative splicing therefore

contributes significantly to protein diversity: a significant proportion of the proteome arises from the same genes expressing different isoforms.

In addition to this “canonical” diversity of the proteome of a species, genetic variation within a population that impacts the process of splicing also causes within-species phenotypic variation, and some of these variants may cause disease, or change disease risk. This motivates our desire to better understand the process of splicing, the effect of gene transcription, and to predict the protein diversity that results from the signals that regulate splicing.

When discussing alternative splicing, it is useful to outline the six distinct categories in which much of alternative splicing occurs (see figure 2.1) [13]:

1. Alternative splicing via alternative transcription start sites (TSS) or alternative terminal exon.
2. Alternative exon inclusion or exon skipping
3. Alternative acceptor (3’) site for an exon
4. Alternative donor (5’) site for an exon
5. Alternative intron retention
6. Alternatively spliced mutually exclusive exons

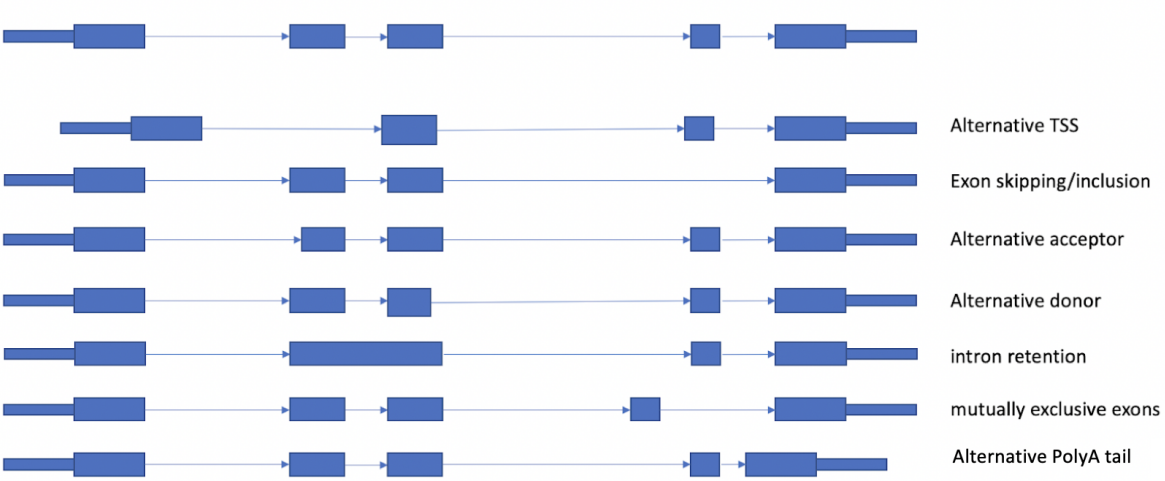


Figure 2.1. Typical variations arising from alternative splicing. Different forms of alternative splicing are described in comparison to the first transcript. Alternative TSS (1); Exon skipping/inclusion (2); alternative acceptor (3); alternative donor (4); intron retention (5),mutually exclusive exons (6) and alternative polyA tail (7)

These categories are generated by going through all the possibilities that could arise from disrupting a splice site. If we focus on an acceptor site that is formed as part of an exon, one possibility is that disrupting this acceptor site could lead to the exon being lost (2) by also disrupting the downstream donor. Alternatively, it could instead disrupt the upstream donor site leading to intron retention (5). Finally, the exon could be retained but modified from its previous form by inducing the use of a nearby alternative acceptor (3).

One way to better understand and predict this protein diversity is to build predictive models for splicing: given a sequence of DNA, where would splicing occur? Several authors have worked on this problem; we briefly discuss two state-of-art approaches.

SpliceAI [4] is a predictive model that predicts locations of splice donors and acceptor sites from primary sequence, using a neural network approach. It provides broadly accurate predictions and has been declared as state-of-the-art in making splice-site predictions from primary sequence [3]. The SpliceAI model is trained to predict the sites where splicing occurs in a transcript; its training data consists of a list of positions (within a gene body) and two columns of 0s and 1s, labelling whether or not the position acts as donor or acceptor site. *SpliceAI* was trained on canonical transcripts, supplemented with sites with rare read support, defined as being found in at least 5 individuals in the GTEx cohort [4]. Although it does not explicitly attempt to predict alternative splicing, by including these sites that are only infrequently included in observed transcripts the model does make predictions for alternatively transcribed sites. However, it does not make a distinction between constitutive splice sites, which are found in all transcripts, and those which are more selectively used.

MMSplice is an alternative model to *SpliceAI* which attempts to explicitly model alternative splicing at canonical sites [2]. The approach that *MMSplice* uses is to directly model the level of inclusion of known predefined exons. This allows users to predict the effect of variants near annotated splice sites [2]. This is powerful when investigating alterations to such splice sites but does not enable genome wide analyses or disease diagnostic analyses away from canonical sites. In addition, by referring to a canonical transcript, *MMSplice* needs to consider all possible alternative transcripts that may result from a mutation (listed in Fig. 2.1), complicating the model and analysis.

Both approaches have their strengths. *SpliceAI* can provide predictions away from canonical splice sites, and is not reliant on an existing gene model. *MMSplice* explicitly models

alternative splicing, which we believe is important as splicing is not an all-or-nothing affair and is a noisy process. This leads to a situation where most genes have several isoforms which involve splice sites that are rarely used. Analysis of GTEx data (figure 2.2) shows that most splice-sites (defined as being supported by >=5 reads) are low-frequency splice sites, but in the majority of transcripts no splicing occurs at these sites.

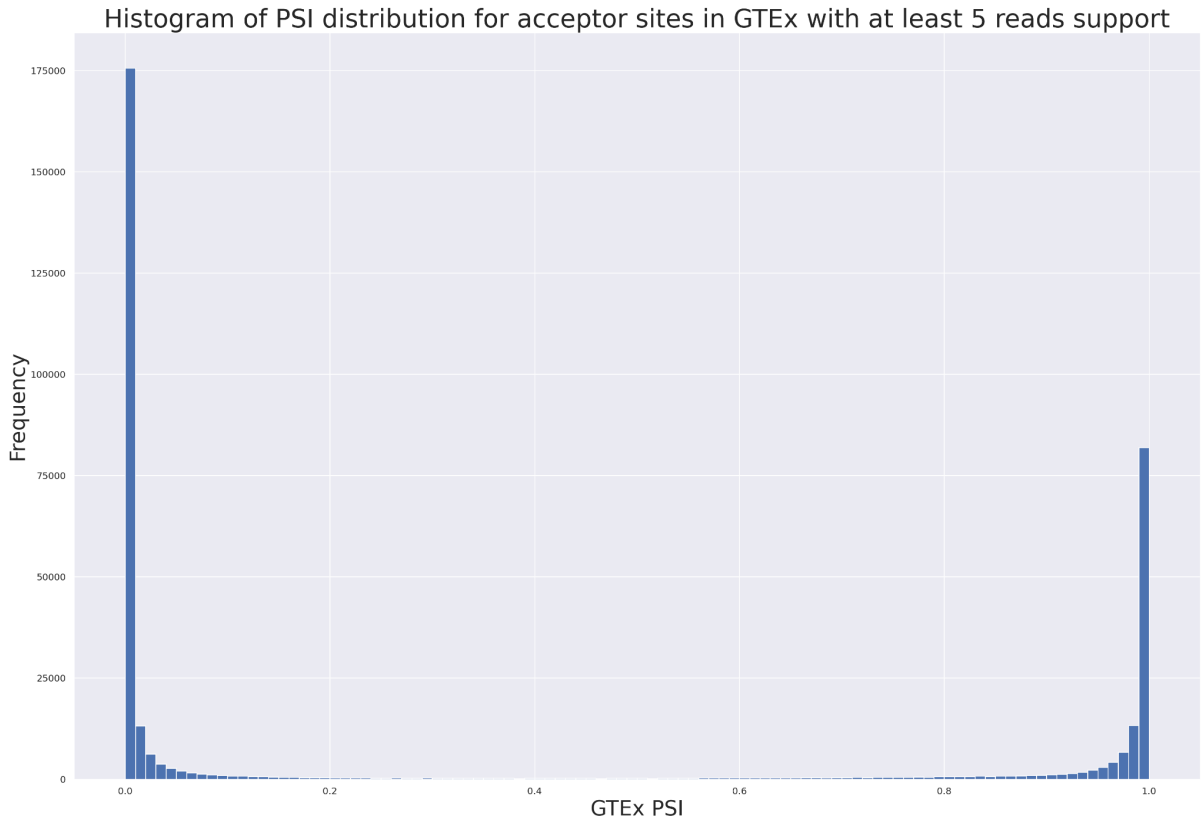


Figure 2.2: Distribution of PSI (percent spliced in) for acceptor sites found in GTEx with at least 5 reads support. The distribution is mostly dichotomous, with sites either being constitutively or very rarely spliced, while splicing at intermediate frequencies occurs less frequently. Percent spliced in, defined fully in section 2.3, measures splice site utilisation amongst all transcripts.

Our ambition is to maintain the gene-model-agnostic aspect of *SpliceAI* but follow *MMSplice* by explicitly modelling alternative splicing and predict splicing across the full spectrum from low-frequency to constitutive splice sites. At the time of writing this thesis we note that a direct ensemble method which combines *MMSplice* and *SpliceAI* is progressing towards this aim [14]. However, we still believe there is great utility in formulating a single model approach that aims to directly combine the strengths of the two approaches.

Modelling splicing as a stochastic process with an associated probability of splicing, rather than an all-or-nothing process, can be done by assuming that a true splice frequency *f* exists; the counts of transcripts with and without evidence for splicing are then assumed to follow a multinomial distribution. This is the approach of both *SpliceAI* and *MMSplice* (although for *SpliceAI* the training set consists of binary labels and the frequency is interpreted as a probability). Below we argue that this ignores an important source of uncertainty, and introduces a bias towards highly expressed genes.

It has been shown how histone marks and other chromatin factors influence splicing [15]. We do not want to limit the application of the model by requiring access to chromatin data, and therefore aim to model splicing based on local sequence features alone. In addition, the training data will contain some inaccuracies, some of a technical nature such as mapping biases, some biological (and later we discuss in detail one major biological source of inaccuracies, namely Nonsense Mediated Decay). Therefore, we do not expect to be able to predict the *true* splice frequency *f*; instead we could hope to predict a *distribution* of more or less likely splice frequencies (including 0 and 1) given local sequence features. If we ignore this aspect, the model would assign sites with similar sequence features with the same

splice frequency *f*, and the predicted frequency would then be biassed towards sites with high expression levels, as at such sites the high transcript counts give very precise estimates for *f*, whereas lowly expressed sites would contribute very little evidence and not contribute meaningfully to the estimate of *f*. This is the same context in which linear mixed models (or random effects models, or hierarchical models) are used to account for the fact that effect sizes may show group-level (here, site-level) variation.

To address this issue, as mentioned above, instead of aiming to predict a single true splice frequency *f*, we aim to predict a *distribution* of true splice frequencies *f*, and for a particular site the model assumes a single (unobserved) frequency drawn from this distribution. The actual frequency is unknown and also follows a distribution, because a limited number of transcripts per site is observed, and the precision with which the actual frequency can be estimated may vary widely depending on the transcript's expression level. However, given a predicted distribution and an observed distribution for *f*, the true frequency *f* can be integrated out, leaving a likelihood of the prediction given the observed transcripts. In this way, by limiting the precision of the predicted *f*, the contribution of even a very precisely observed splice frequency will remain limited, avoiding the bias towards highly expressed genes.

By properly dealing with the uncertainty and the dependence of evidence between sites, this approach allows us to include a rich variety of sites including those with very high and very low read support (i.e., expression levels), without biassing towards either category. By including more examples we hope that the model will be more apt to model the splicing

process, which we hope will increase our ability to identify disease-causing variation across a range of expression levels more reflective of the biology.

2.3 Percent Spliced In

Percent Spliced In (PSI) is the standard metric utilised to quantify alternative splicing [16]. For a given splice site it is given by the number of isoforms that support the splice site, divided by the number of isoforms that *could* support the splice site – that is, the isoforms whose immature transcript includes the splice site. In this manner every nucleotide in a gene-body can be assigned a PSI - a number between 0 and 1 - for both donors and acceptors. Sites within a gene-body with PSI of 1 are constitutive splice sites, while sites with PSI 0 are never used as splice sites; sites with PSI in between these extremes are involved in alternative splicing. This simple and observable metric (with a precision determined by expression level) provides a quantitative summary of a gene's constitutive and alternative splicing. An example is given in Fig. 2.3.

Summarising a gene's alternative transcripts using only PSI ignores some important information. In particular, splice site usage is not independent between sites, and by itself PSI does not tell which acceptor and donor pairs are combined to define exons.

Figure 2.4 shows the scenario where the acceptor-donor pair information is not recoverable from PSI alone. However, a simple analysis over the known exons in UCSC-V37 shows that such overlapping-exon scenarios are rare. Only 4.2% of internal exons overlap another exon which does not share either the acceptor or donor site.

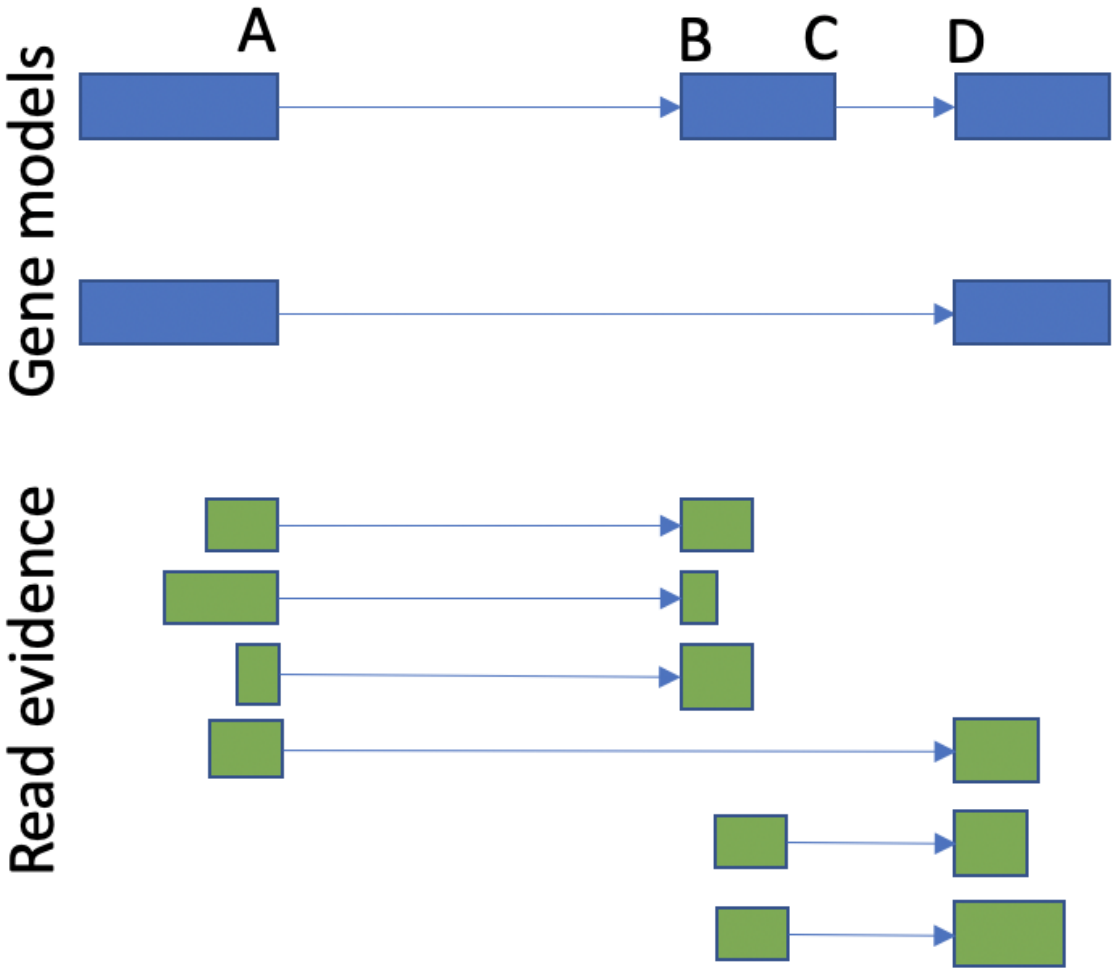


Figure 2.3: PSI used as a metric to summarise splice-site usage based upon RNA-seq read evidence. In this figure splice sites A, B, C and D have PSIs of 100%, 75%, 67%, and 100% respectively.

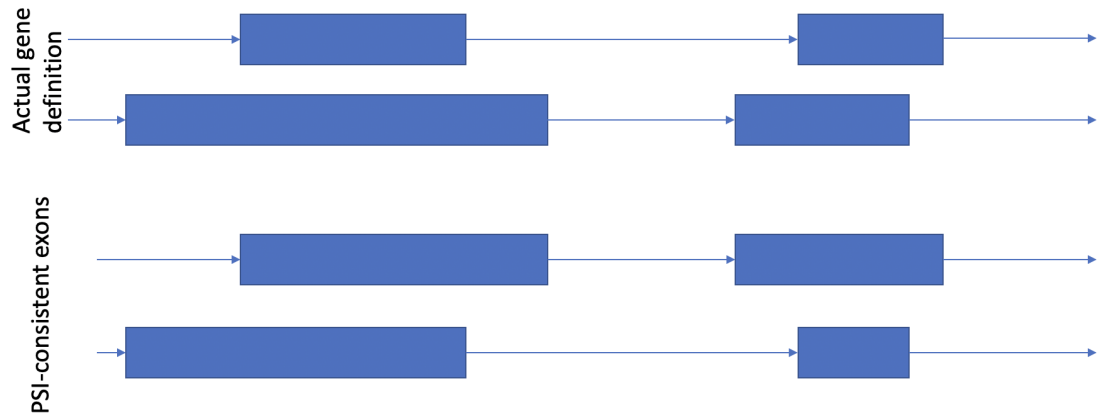


Figure 2.4: Figure showing edge case limitations of PSI for gene definition. Using PSI alone can make these two scenarios indistinguishable from each other. Overlapping exons which change both the acceptor and donor site lead to ambiguity when summarising gene-definition from PSI alone.

Another reason why we are not too concerned by the loss of information involved in focusing on PSI only, is that many downstream analyses would remain unchanged. Scenarios such as the prediction of a gain of an intronic exon or the switch between two competing alternative acceptors can be described using PSI alone.

In conclusion, although PSI ignores some aspects of the true complexity of splicing, it provides a useful summary and a simple and useful modelling target.

2.3.1 Defining the training dataset from read junction data

We built our training data using GTEx read-junction datasets, similar to the approach used by R. Brown in his thesis [17]. These read junction files were built by Bretschneider et al [18] using read-junctions obtained from GTEx by aligning FASTQ using the HISAT2 aligner [18],

[19]. Using these pre-processed read junction files avoided the need of reprocessing the raw and very large GTEx FASTQ files.

A read-junction is defined as an alignment gap arising from aligning RNA-seq data to the genome, which is taken as an indication of a splice event. Then on a chromosomal level our pipeline uses an interval tree data structure to load all the read-junctions. This data structure allows us to quickly identify all read junctions that overlap a certain interval, such as a gene body. Using these interval-trees we can compute for every nucleotide within a gene-body three pieces of information:

1. The number of reads supporting the position as a donor site (n_1)
2. The number of reads supporting the position as an acceptor site (n_2)
3. The number of reads overlapping the position (m)

n_1 and n_2 are trivially computed by adding all the contributions from different read-junctions which include that position. Every read-junction contributes to one donor and one acceptor. We initially compute m by summing all the contributions of read junctions that overlap the position. This results in an under-count in exon-regions, which have read coverage that is not reported in the read junction data. This is solved by smoothing the coverage values over the intron-exon boundaries, using the junction coverage data at exon boundaries as proxies for the coverage across the exon.

A read junction is associated to a transcript on one of the two strands of the genome, however, the raw reads are not stranded. Most sites have a single overlapping gene body, and the junctions are simply assigned to the gene's strand. However, for genes that overlap

and live on different strands, the strand may be undefined. To address this issue, we constructed a simple model which identifies the strand direction based on canonical junction motifs: they are assigned the strand for which the junction shows the canonical *GT-AG* motif.

From this data structure, for each nucleotide within a gene-body, we obtain the number of read-junctions which support the nucleotide as a donor and as an acceptor site, as well as the read-depth at each nucleotide. With this information the percent spliced-in estimates can be calculated.

2.3.2 PSI plots for GTEx evidenced splice sites

To verify that the pipeline works correctly, we plot the empirical PSI distribution on sites that show read-evidence in GTEx data, defined as having one individual with at least 2 reads supporting the site.

We find that the distributions are heavily biassed to PSIs close to 0 or 1. This hides some details such that there are still a large number of splice-sites between these extremes. Moreover, we find that many sites which have a PSI close to 0 have significant read support.

Figures 2.5 and 2.6 only include GTEx sites which have read evidence. Therefore, they do not include the biggest mode which are genomic sites with no evidence for splicing (PSI=0).

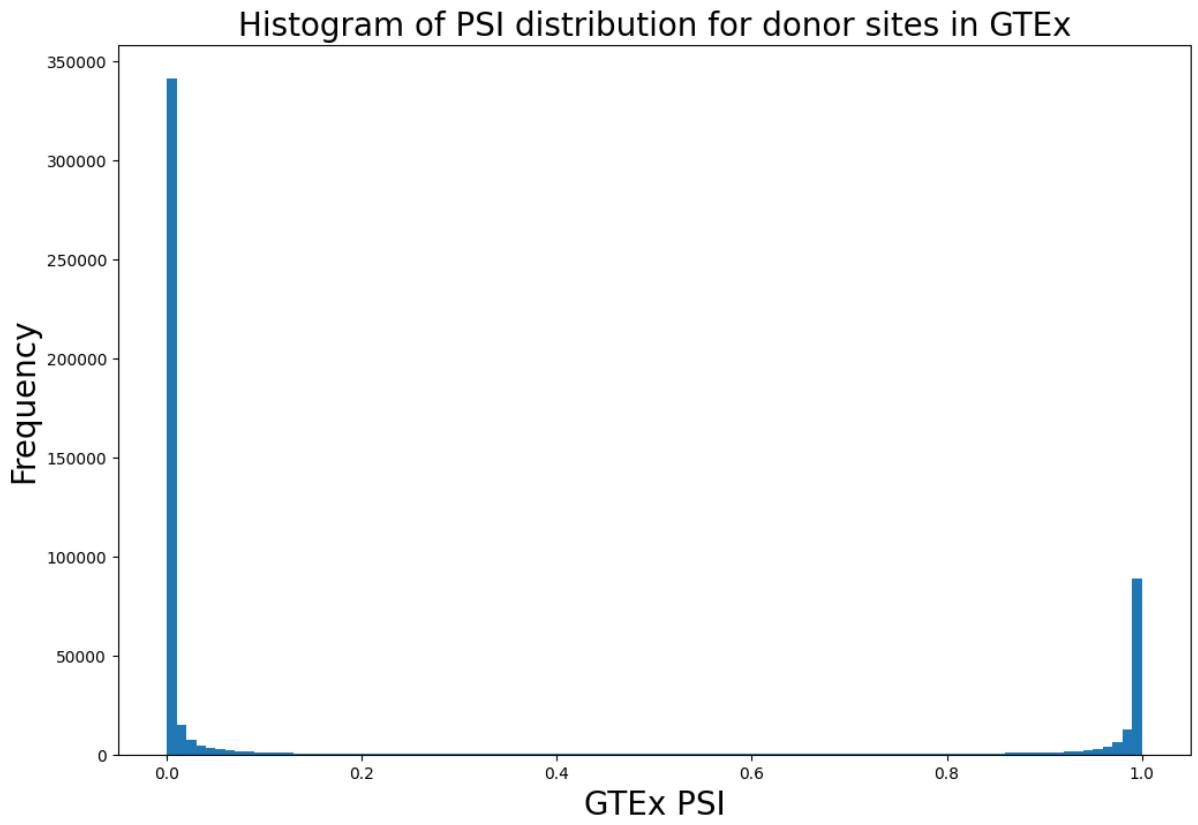


Figure 2.5: Empirical PSI for donor sites within GTEx, with sites with PSI exactly 0.0 removed. Distribution is defined by two modes near 0 and 1. Despite the dichotomous appearance, about a third of splice sites with a PSI > 0.05 are not constitutive (defined as PSI > 0.95).

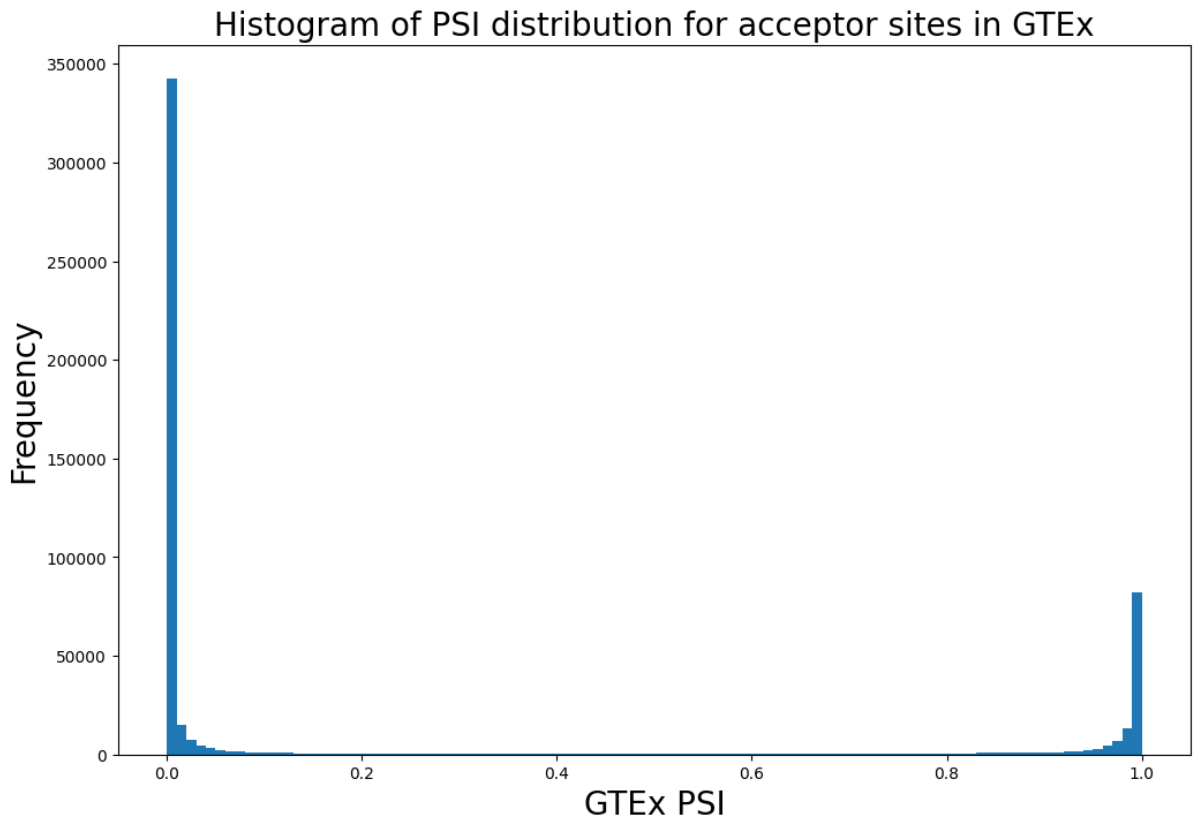


Figure 2.6:Empirical PSI for acceptor sites within GTEx, with sites with PSI exactly 0.0 removed. Distribution is defined by two modes near 0 and 1. Despite the dichotomous appearance, about a third of splice sites with a PSI > 0.05 are not constitutive (defined as PSI > 0.95).

2.3.3 PSI specific training loss

Typically, neural networks models for classification are trained using the cross-entropy loss, essentially the minus log likelihood for the binomial (or multinomial) distribution. This is a sensible choice of loss function when the label assignments are certain, and examples are independent of one another. As discussed above, in our application neither assumption holds particularly well. Although for a single transcript a site is either spliced or not, the process of Nonsense Mediated Decay means that we do not observe certain transcripts. In

addition, the probability (or frequency) of splicing varies between sites, due to unobserved features e.g. on the level of histone modifications, even if the local sequence features are similar, so that the independence assumption is also strongly violated.

Based on these considerations, we model the predicted splice frequency and its uncertainty as a Beta distribution; and we also model the read-data evidence given a true splice frequency - counts of reads with or without support for a splice site - as a Multinomial distribution. This allows us to mathematically express our uncertainty in both the observed and modelled true site-specific splice frequency, addressing both issues.

It remains to devise how these choices give rise to a particular loss function. Although not always emphasised in the machine learning literature, the loss function can usually be viewed as the minus log likelihood for a particular likelihood function; this shows that many neural network models are simply maximum likelihood models, with the difference that the prediction, linking expected observation given covariates, is done by a neural network rather than a more familiar regression line equation. The model therefore computes the (minus log) likelihood of observing the read data given the model and surrounding genomic sequence. A well trained model should report optimal likelihoods at each nucleotide for both donor sites and acceptor sites. Hence we attempt to maximise this likelihood, or equivalently, to minimise the loss.

2.3.4 Evidence as a loss function – a path with a closed form formula

For each base-pair model aims to predict f, the splicing-rate at the site for either donor or acceptor sites, and for a particular strand. As mentioned above, it is important to model the

uncertainty in *f*. We choose to model the actual splice rate *f* as a single draw from a Beta distribution,

$$f \sim Beta(\alpha_1 + \gamma_1,\ \alpha_2 + \gamma_2); \quad \alpha_{1,2} > 0;\ \gamma_{1,2} > 0$$

Here, $\alpha_{1,2}$ are fixed constants that determine the prior distribution of splice frequencies in the absence of additional information; $\gamma_{1,2}$ are the model predictions; together these determine both the mean predicted splice frequencies, $(\alpha_1 + \gamma_1)/(\alpha_1 + \gamma_1 + \alpha_2 + \gamma_2)$, and the certainty of the prediction. We do not observe f but instead observe N_1 reads that do not support splicing, and N_2 that do. The likelihood can be written as

$$p(N|\gamma) \;=\; \int p(N|f)\, p(f|\gamma) df \;=\; \int Mult(N|f)\, Beta(f|\alpha + \gamma) df$$

where we write $N = (N_1, N_2)$ and similarly for α and γ. This expresses that the likelihood involves the unknown splice frequency *f*, which however can be integrated out so that the final likelihood only depends on the prediction γ and the observation *N*. The integral can be calculated as follows:

$$\int_0^1 Mult(N|f) Beta(f|\alpha + \gamma) df$$

$$= \int_0^1 \frac{(N_1+N_2)!}{N_1!N_2!} f^{N_1} (1-f)^{N_2} \frac{1}{B(\alpha_1+\gamma_1, \alpha_2+\gamma_2)} f^{\alpha_1+\gamma_1-1} (1-f)^{\alpha_2+\gamma_2-1} df \qquad \text{(definitions)}$$

$$= \frac{(N_1+N_2)!}{N_1!N_2!} \frac{B(N+\alpha+\gamma)}{B(\alpha+\gamma)} \int_0^1 \frac{1}{B(N+\alpha+\gamma)} f^{N_1+\alpha_1+\gamma_1-1} (1-f)^{N_2+\alpha_2+\gamma_2-1} df \quad \text{(cancelling factors and…)}$$

$$= \frac{(N_1+N_2)!}{N_1!N_2!} \frac{B(N+\alpha+\gamma)}{B(\alpha+\gamma)} \int_0^1 Beta(f; N + \alpha + \gamma) df \qquad \text{(... vector notation; recognise as Beta)}$$

$$= \frac{(N_1+N_2)!}{N_1!N_2!} \frac{B(N+\alpha+\gamma)}{B(\alpha+\gamma)} \qquad \text{(integral of normalised Beta is 1)}$$

$\propto \frac{B(N+\alpha+\gamma)}{B(\alpha+\gamma)}$ (drop γ-independent factor)

Here, *B* is the Beta function, which is involved in the definition of the Beta distribution, and in the last line we left out factors that are constant in γ and therefore do not contribute meaningfully to the loss. Minus the logarithm of this quantity is a reasonable candidate for a loss function. The log Beta function can be easily computed numerically using the Log Gamma function, which is implemented in most numerical libraries.

2.4 Modelling approach

Initially the project started with an approach following *R. Brown's* DPhil thesis which models the acceptor site conditioned on the location of the previous donor site [17]. This was inspired by an intron/exon-definition mechanistic model, in which the splicing machinery, after having identified a donor site at the end of an exon, is thought to start "looking for" the next eligible acceptor site which may form the start of the subsequent exon. Having identified this acceptor site, the splicing machinery would then find a corresponding donor site, and so on. Specifically, the model implemented by *R. Brown* was tasked to predict a single acceptor site given the location of a particular donor site [17]. Our hypothesis was partly that sequence motifs around a donor site are relevant for identifying the subsequent acceptor site, and the model was given explicit access to these data.

This effort resulted in a model that worked reasonably well, but importantly, gave us confidence that in fact exon-definition, and not intron-definition, is in the main the key mechanism in selecting acceptor-donor pairs. Specifically, *R. Brown* found that, despite the

prediction target being only the location of the subsequent acceptor site, the model based its decision, in part, on the presence of motifs related to downstream *donor* sites [17]. The model apparently needed relevant motifs at both the acceptor site and downstream donor sites to decide that a particular site is a viable acceptor. This can be understood if the splicing machinery recognizes the acceptor and donor sites abutting an exon simultaneously as a pair, rather than individually as single sites, a process known as “exon definition”. Experimental evidence for this hypothesis was provided, for example by the work of Ule and Blencowe, 2019, which describes how SR proteins bind potential exons before other splicing factors are recruited to the acceptor site [20].

Having found that the motifs around the upstream donor site are not entirely relevant for the identification of subsequent acceptors (which would be the case for an “intron definition” model), but that (proximal) subsequent downstream donor sites *are* additionally relevant for the identification of acceptor sites (“exon definition”), we opted for a model that predicts both acceptor and donor sites simultaneously, but does not explicitly link them; because exons are limited in size, with the appropriate neural network architecture the model would be able to capture any interaction between these sites automatically. This approach is similar to that chosen by *SpliceAI*, and has been shown to give very good results. We therefore utilised *SpliceAI*’s model architecture as our initial version adapted to use with our novel training loss.

2.4.1 Neural network architecture

Following *SpliceAI* we implement their neural network architecture, but using our novel loss function. This enables us to separate differences between architectural choices and loss

function choices. The neural network architecture chosen by *SpliceAI* is known as a "residual convolutional neural network".

We additionally test our pipeline with a residual U-net architecture [21]. This architecture has been used extensively for medical image recognition [22]–[24]. It is characterised by an architecture which resembles a "U", first reducing (halving) the resolution of the input (image or sequence) at every step, followed by the upward leg of the U in which the resolution is doubled up to the initial resolution. At every step of the upward leg, information from the previous step is combined with that of the same-resolution layer of the downward leg. This allows the network to combine global (low-resolution) with local (high-resolution) information as appropriate, allowing the model to learn complex structural relationships while also providing high-resolution (base-pair-level) output.

We tried several variants of network architectures but we do not find that they significantly improve upon using SpliceAI's architecture combined with our loss function. The results we found for the u-net architecture were broadly equivalent to those we found with their original architecture.

2.5 Results

Understanding how well the model performs is a complex task. The model will minimise the loss function, but this in itself is only a proxy for the actual usefulness of a model. Instead we look directly at the predicted splice rates.

Figures 2.7-2.10 show box plots of empirical and predicted PSI for acceptors and donors, against their predicted or empirical PSI counterpart. For this analysis we use RNA-seq data from GTEx on several hold-out chromosomes, i.e. they are not used for training the model.

What is apparent from the plots is that the model broadly captures the observed splice rates. In particular, constitutively spliced sites (for the purpose of plotting defined as having PSI > 0.8) are broadly correctly identified, as are sites with PSI < 0.2. However, the model does show a tendency to overestimate the observed PSI as seen in GTEx RNA-seq data. This is most apparent from Figs. 2.7 and 2.8 that show empirical PSI as a function of predicted PSI - the distribution is heavily skewed towards low values, for all bins. Figures 2.9 and 2.10, which show predicted PSI as a function of empirical PSI, show an analogous but less strong skew towards high values.

The first observation, the skew towards low values seen in Figs. 2.7 and 2.8, could be explained if a substantial fraction of true splice sites were missing from the empirical data (and shown as PSI=0). We will come back to this observation when we discuss the impact of Nonsense Mediated Decay later.

2.5.1 Comparison of empirical PSI and predicted PSI

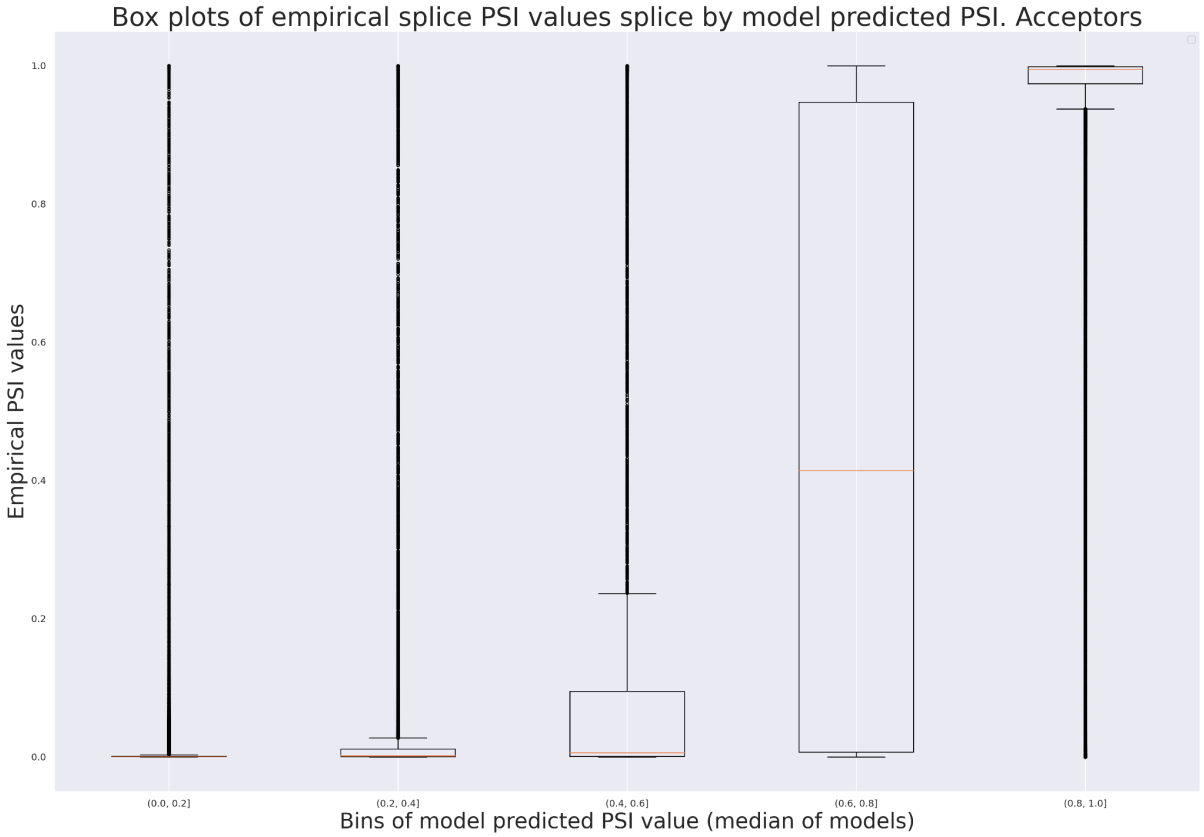


Figure 2.7. Empirical GTEx PSI vs. predicted PSI for acceptors. Each box-plot shows the distribution of empirical PSIs based upon their model predicted PSI. Sites with PSI=0 dominate the data set, but even for predicted PSI between 0.2 and 0.5 we see the model overestimating splicing rates. For sites predicted to be (near-)constitutive, PSI>0.8, we do see a high empirical PSI.

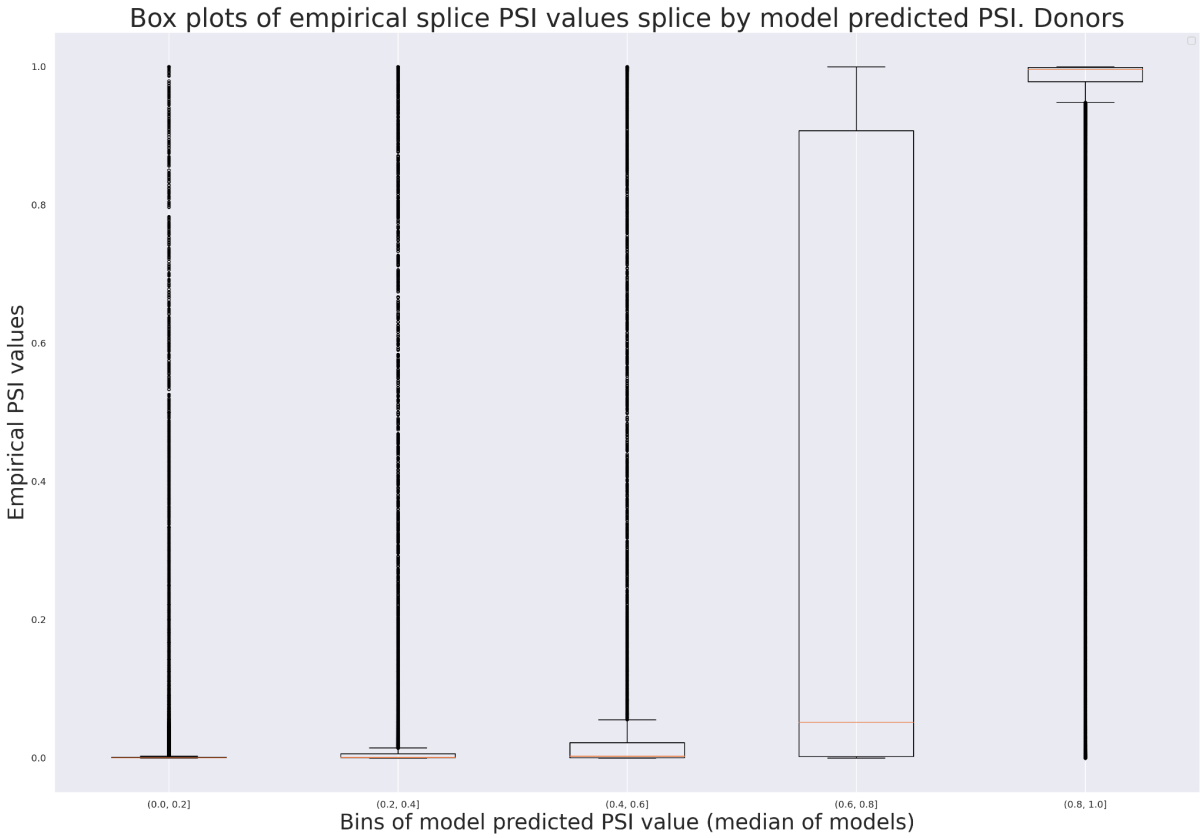


Figure 2.8.Empirical GTEx PSI vs. predicted PSI for donors. Each box-plot shows the distribution of empirical PSIs based upon their model predicted PSI. Sites with PSI=0 dominate the data set, but even for predicted PSI between 0.2 and 0.5 we see the model overestimating splicing rates. For sites predicted to be (near-)constitutive, PSI>0.8, we do see a high empirical PSI.

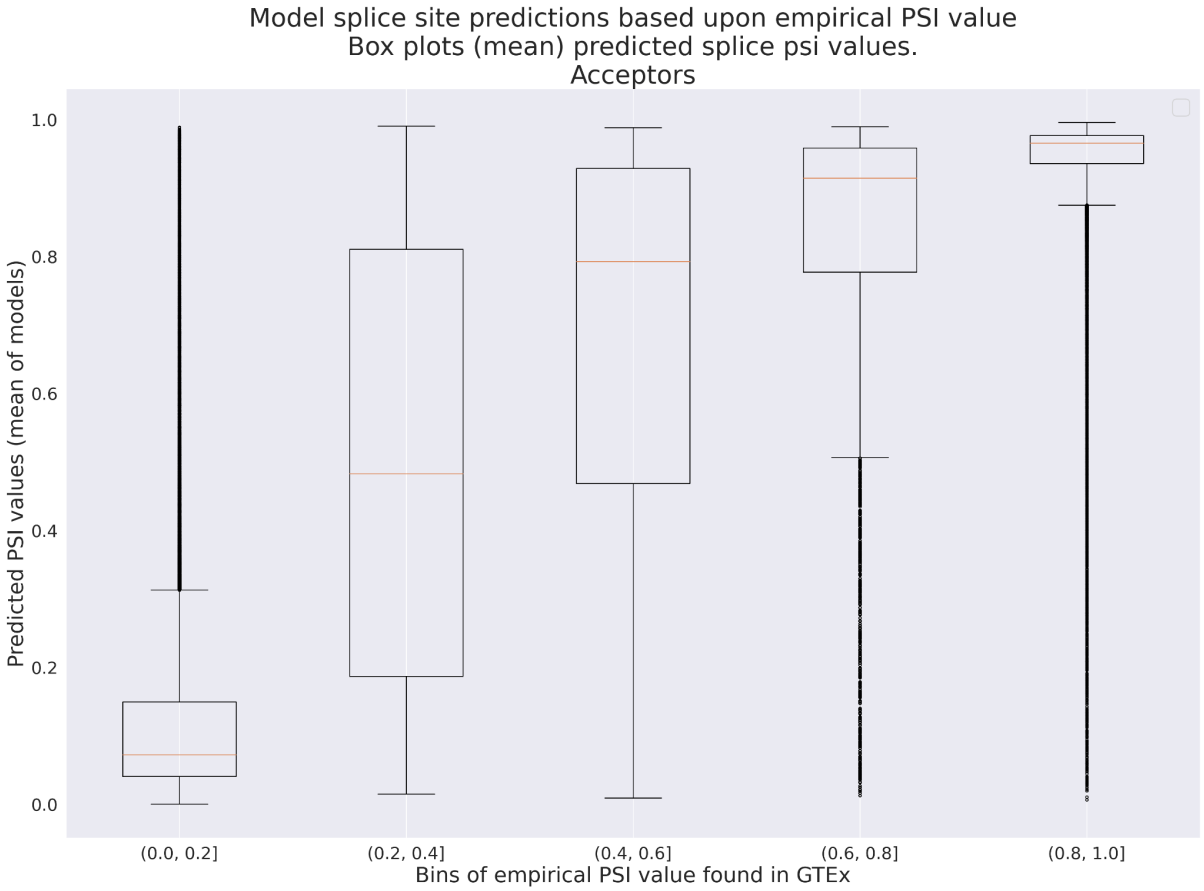


Figure 2.9: Empirical GTEx PSI vs. predicted PSI for acceptors. Each box-plot shows the distribution of predicted PSIs based upon their empirical PSIs. Empirically constitutive splice sites are predicted as such, rarely used splice sites predicted with low PSIs. As a class alternative splice sites with higher empirical PSIs are predicted with higher PSIs.

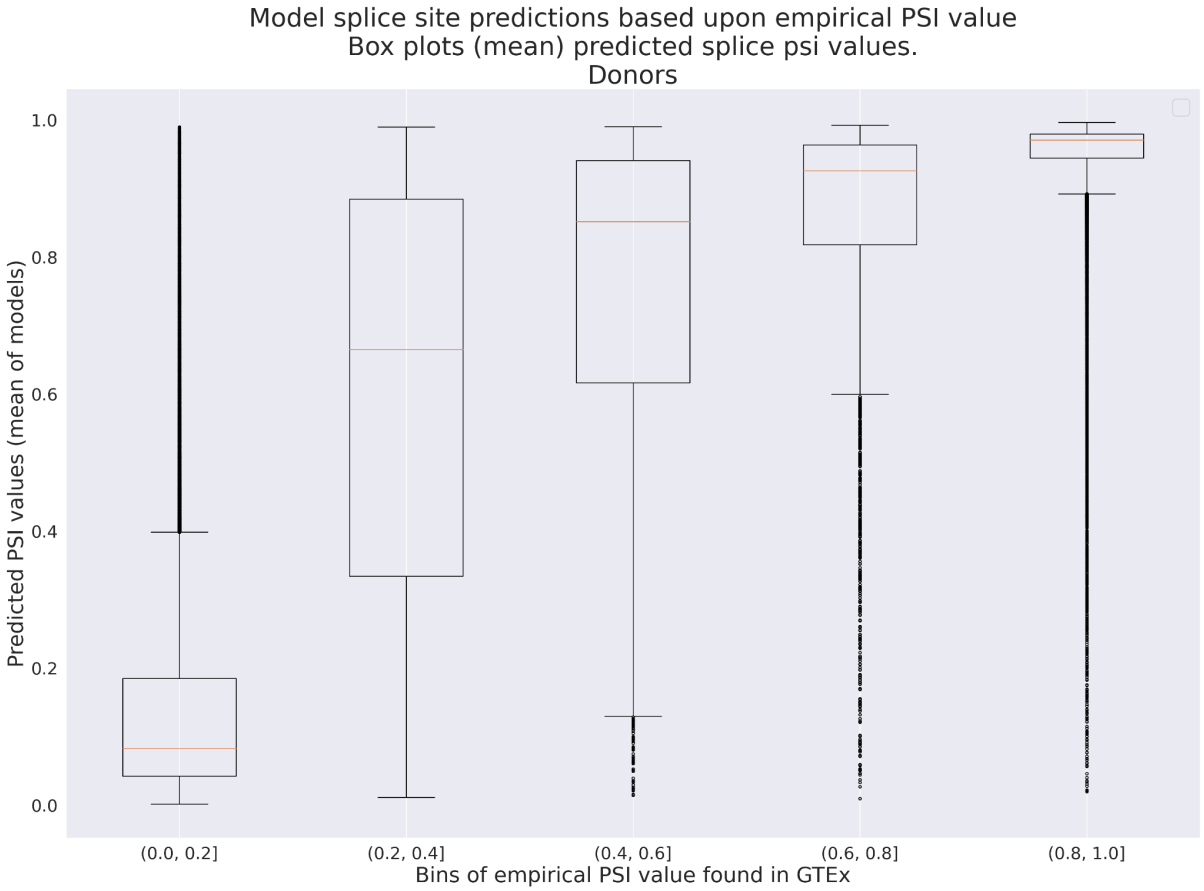


Figure 2.10: Empirical GTEx PSI vs. predicted PSI for donors. Each box-plot shows the distribution of predicted PSIs based upon their empirical PSIs. Empirically constitutive splice sites are predicted as such, rarely used splice sites predicted with low PSIs. As a class alternative splice sites with higher empirical PSIs are predicted with higher PSIs.

2.5.2 Effect of transcription levels

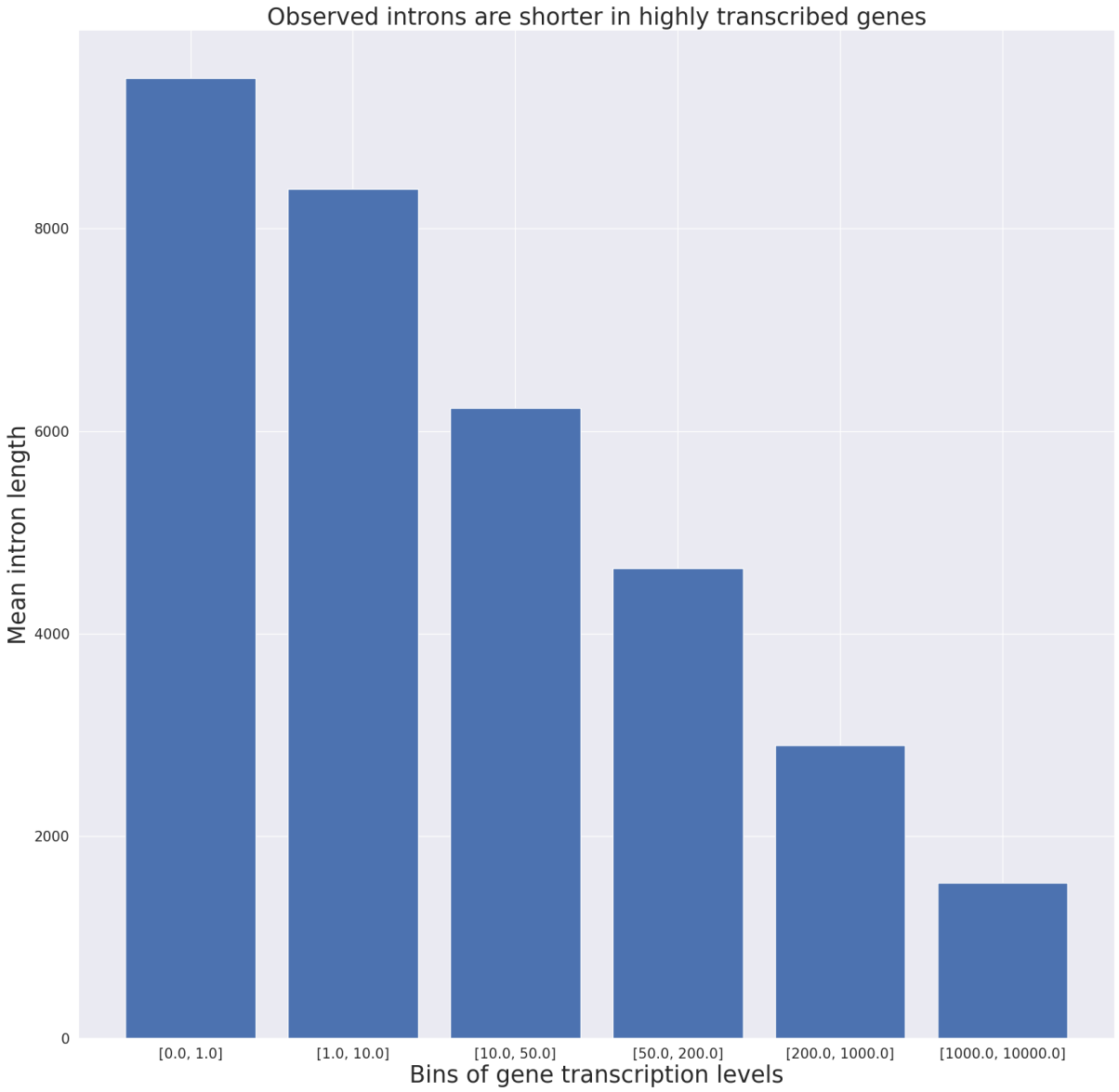


Figure 2.11: Introns obtained from UCSCv37 with gene transcript levels (TPM) obtained from GTEx. More highly expressed genes have, on average, shorter introns. One explanation would be the cost of having long introns in highly transcribed genes.

Gene expression levels vary across several orders of magnitude across genes and cell types, and due to external influences and cell stage. One observation that suggests that transcription level may influence other aspects of the gene structure is that the mean intron

length is strongly dependent on gene transcription levels (Fig. 2.11), with highly expressed genes having on average shorter introns (Fig. 2.11). Because of the huge range of gene expression across genes and other covariates, it is likely that splicing is also affected by gene expression. We therefore decided to investigate the model performance by gene expression level.

We annotate each gene with its total transcription levels obtained from GTEx [19]. This allows us to split our analyses into highly, medium and lowly expressed genes. We then perform our analyses for the previous sections splitting by gene expression levels. In particular we focus on acceptor sites, and on the analysis showing predicted PSI as a function of empirical PSI (Fig. 2.9), which shows a reasonably good overall agreement between predicted and empirical PSI.

We find that the model's accuracy indeed varies by the overall transcription level. For highly transcribed genes (>50 TPM), predictions are strongly skewed towards high PSI (Fig 2.12): sites with empirical PSI >0.2 are broadly predicted to be constitutive (median predicted PSI > 0.8); only sites in the low empirical PSI bin (<0.2) correctly show similarly low median predicted PSI. A similar but less pronounced skew is found for genes with medium expression levels (TPM between 1.0 and 10.0, Fig. 2.13), while for lowly expressed genes (TPM < 1.0, Fig. 2.14) empirical and predicted PSI are in good correspondence, although a residual over-estimation remains present, e.g. for empirical PSI in the range 0.4 to 0.8.

Genes with high transcription levels (>50 TPM) are a small subset of all genes - in our analysis this set comprises 820 genes. However, alternative splicing does occur among these

genes, and because of their high transcription levels, the correct regulation of splicing is presumably relatively important in this class of genes. However our model is apparently not well able to predict alternative splicing in this class, in contrast to the much larger class of lowly expressed genes where the model does a much better job. We will come back to this observation in the discussion.

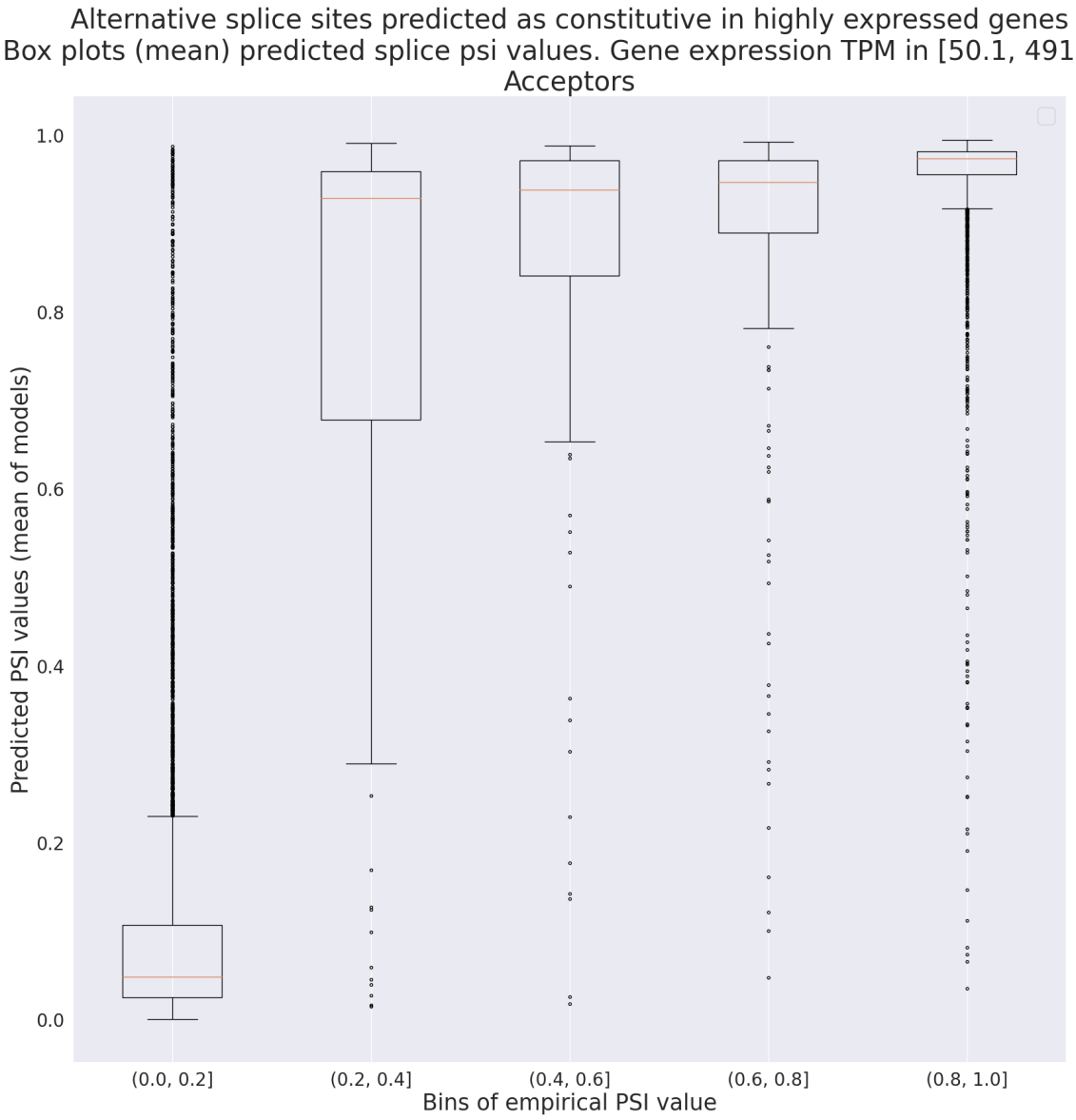


Figure 2.12: Box plots showing distribution of predicted PSIs based upon empirical PSIs obtained from GTEx. Analysis restricted to genes with high gene expression levels. For this set of splice sites only very weak splice sites (PSI in (0, 0.2]) are predicted as weak splice sites. As a set splice sites with PSI in (0.2, 0.4) are mostly predicted as almost constitutive.

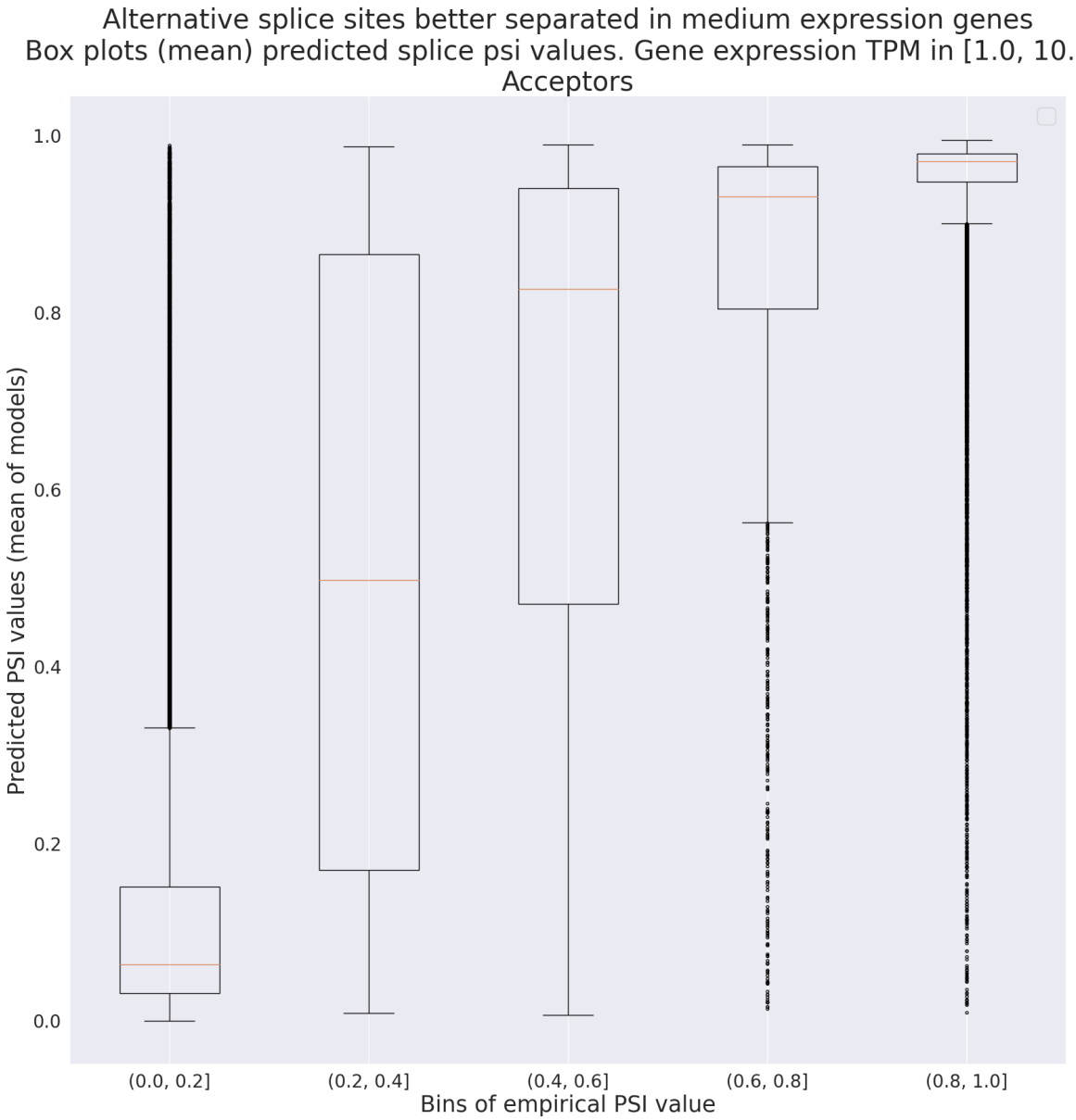


Figure 2.13: Box plots showing distribution of predicted PSIs based upon empirical PSIs obtained from GTEx. Analysis restricted to genes with medium gene expression levels. We observe that as a class weakly alternative splice sites (PSI in (0.2, 0.4]) are predicted as alternatively spliced.

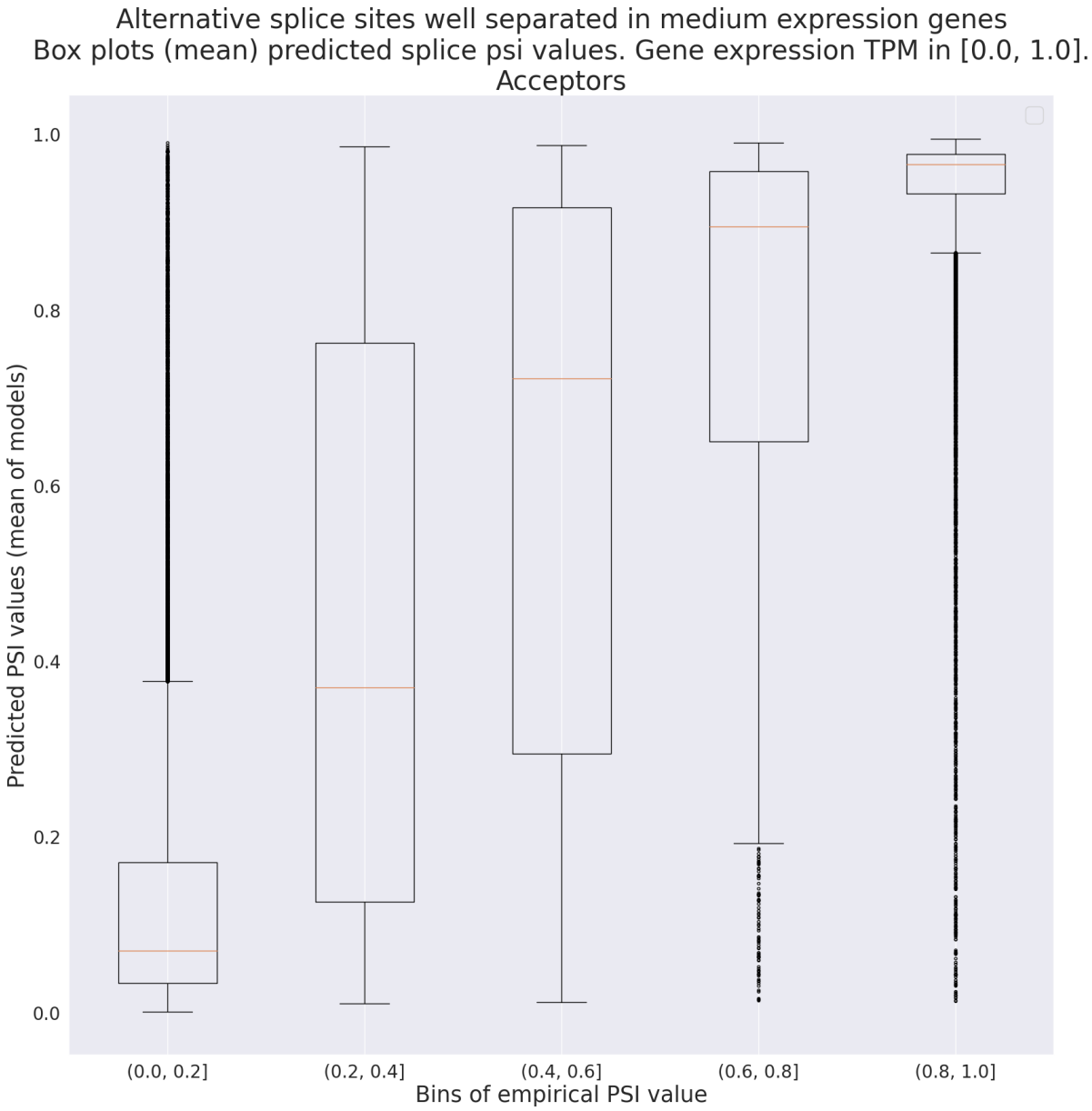


Figure 2.14: Box plots showing distribution of predicted PSIs based upon empirical PSIs obtained from GTEx. Analysis restricted to genes with low gene expression levels. We observe that the model is well separating the classes of splice sites..

2.6 Model generated hypotheses for splicing

In the last section of this chapter, we apply the model to try and better understand alternative splicing. The various forms of alternative splicing involve the spliceosome which for our purposes is best considered as a collection of trans-acting factors, including snRNAs, snRNPs, and proteins which must bind to the pre-mRNA to give rise to splicing [25]. By training our model to predict observed splicing events directly from sequence, we hope that the model has learnt the signals in the genome that give rise to splice events. We wish to understand how the model combines sequence features into a prediction of splicing. We achieve this by using *in silico* mutagenesis, that is, we mutate the reference genome and observe whether a mutation changes the model's predictions; if it does, we may hypothesise that the site plays a causal role in splicing.

We first investigate where the splice factors proximal to known (donor and acceptor) splice sites are located. We then investigate how predicted splice sites change if we knock out known acceptors and donors, leading to e.g. a change in exon length through use of an alternative acceptor or donor, exon skipping, or intron retention; this tells us something about how (the model predicts) acceptor and donor sites interact. Finally, we investigate the phenomenon of recursive splicing, which, although not observable in our training data, was apparently picked up by the model.

Broadly, we find that the model rarely only makes a local change in its prediction when a splice site is disrupted. Instead, it balances such a prediction with a corresponding prediction elsewhere within the gene body. For instance, changes that result in the removal of a predicted acceptor site are nearly always accompanied with a corresponding prediction of

an acceptor-gain, donor-loss or some combination of these. Likewise, changes that lead to a loss of a donor site prediction are nearly always accompanied with a corresponding prediction of acceptor-loss, donor-gain or some combination of these. We hypothesise that the fact that acceptor and donor predictions are not made independently of one another reflects the mechanism by which the spliceosome recognizes these sites. For instance, where the model has learnt that a downstream donor is key for the selection of an acceptor site, this provides support for an exon-definition model of splicing: the presence of a downstream donor is essential for a site to be able to act as an acceptor. Our findings in the remainder of this chapter appear to bear this out.

2.6.1 Analysis of local splicing factors

We first look at the placement of local splice factors around splice sites. In order to do this we take known constitutive splice sites. Around each splice site we spike in single nucleotide mutations for every position in a 200 base pair window ranging from 150 bps upstream to 50 bps downstream.

For each nucleotide position we run the model on each of the three alternative nucleotides and then provide summary information based on the model outcome differences; we then term “mutagenic activity”, which for a particular site (and across all three possible mutations) we defined as the maximum absolute difference of predicted splice probabilities across the window between reference and mutated sequences. This summary information tells us not only the effect on the specified acceptor/donor but it also provides additional information concerning nearby acceptors and donors. Precisely, we can determine whether

the loss of the acceptor/donor site leads to the creation of alternative acceptor/donor sites or the loss of the exon.

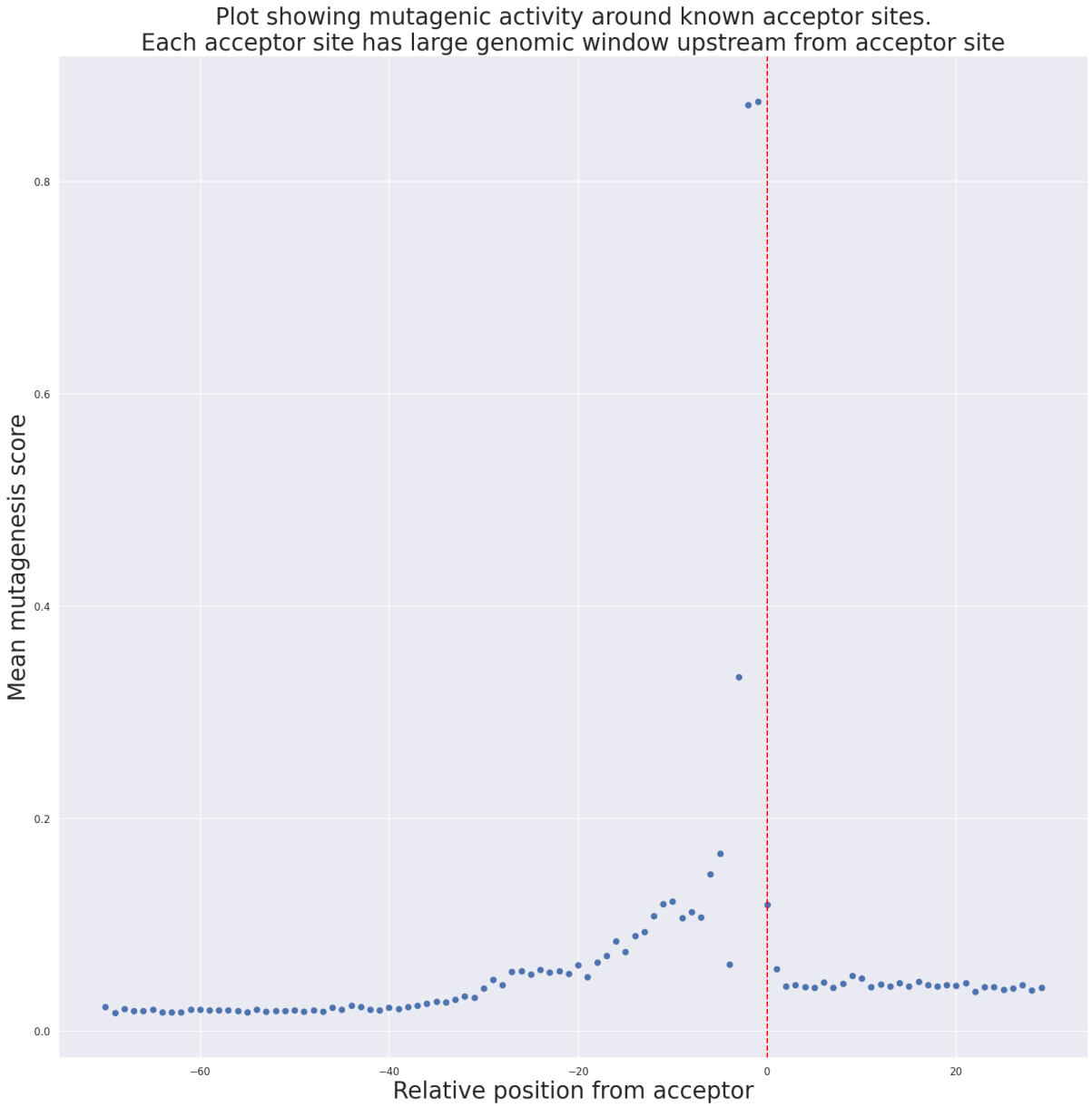


Figure 2.15: Model predicted effect of spiked in mutations based upon relative proximity to canonical acceptor sites. Very strong (average) predicted effect 2 base pairs preceding acceptor site (typically 'AG'). Limited effects found within exonic regions or when greater than 40 base pairs upstream.

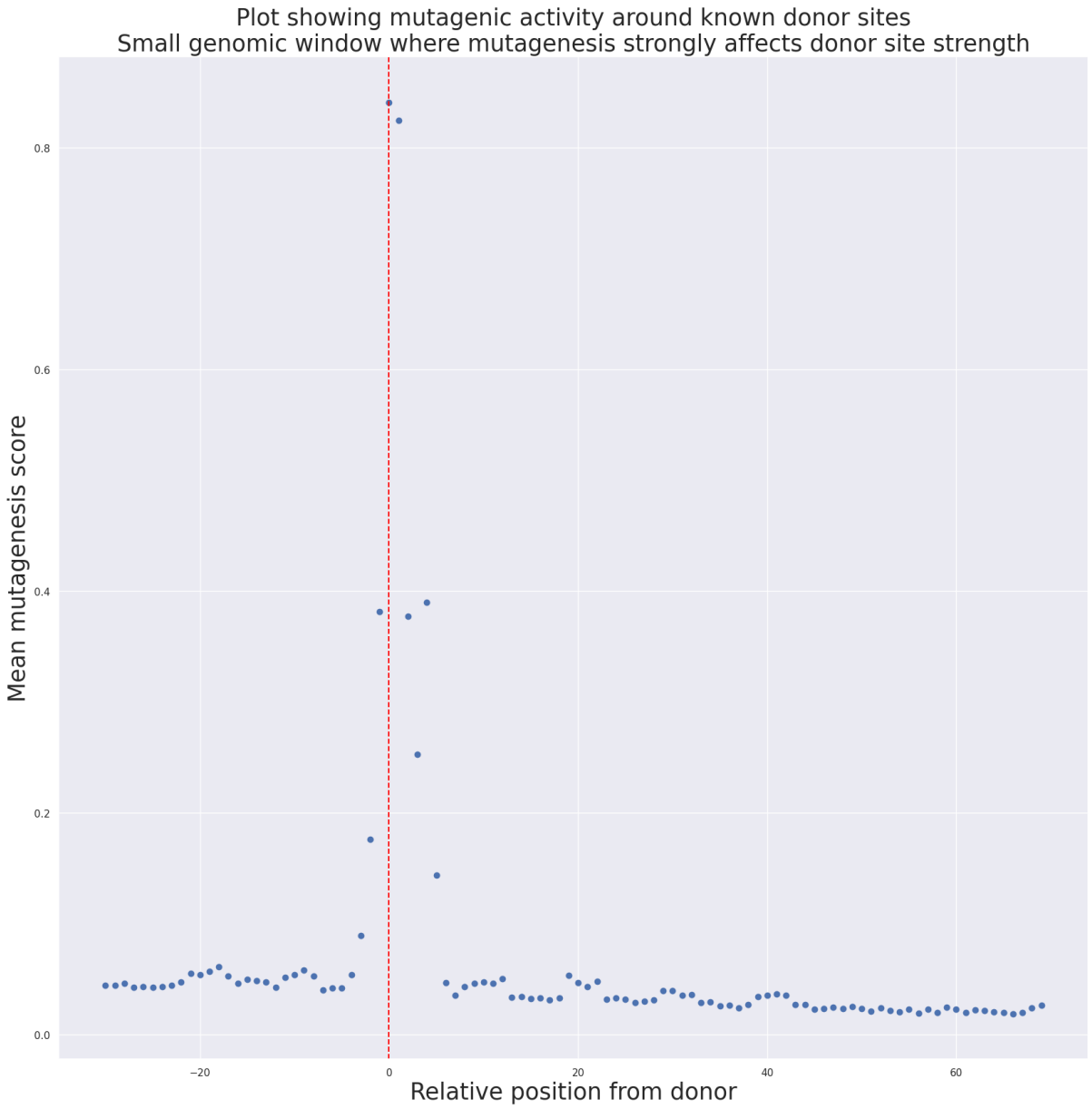


Figure 2.16: Model predicted effect of spiked-in mutations based upon relative proximity to canonical donor sites. Unlike for acceptor sites we observe strong average effects arising from the 2-3 nucleotides ending the exon. Do not observe high levels of intronic mutagenic activity beyond the first 6 nucleotides.

Figures 2.15 and 2.16 show that, according to our model, local splicing factors are restricted to around 10 nucleotides around the donor site, while loci up to 40 bp upstream of the acceptor site influence predictions of splicing strength. This is consistent with known features of canonical splice sites [26], in particular the larger region upstream of the

acceptor sites corresponds to the known placements of the polypyrimidine tract and the branch point locus. This does not imply that distal loci do not influence splicing but when averaged over multiple splice-sites only key local loci have a consistent effect.

For both acceptor and donor sites we find very limited mutagenic activity within the exonic regions. Despite acceptor definition utilising a substantial number of nucleotides within the intronic region, little mutagenic activity is observed within the exon. In chapter 3 we will observe that this seemingly disagrees with our analysis on disease causing variants. In this class many exonic mutations are observed to disrupt canonical splice sites, leading to exon-skipping. This is explained by the analysis in this section showing that deep exonic regions have on average low mutagenic activity. A plausible explanation is that exonic splicing regulatory sequence has no fixed distal relationship to the splice sites they regulate.

2.6.2 Effect of acceptor loss

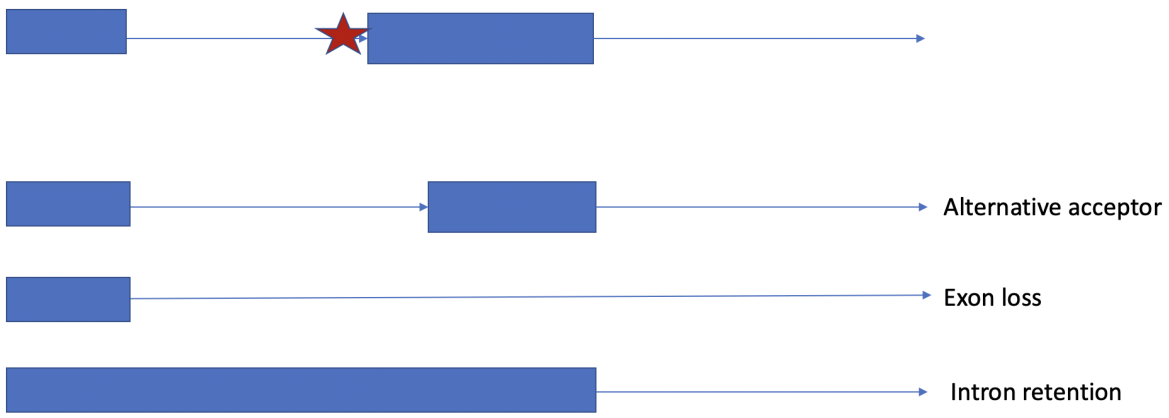


Figure 2.17: For a mutation targeting an acceptor site we find three main scenarios. From the model output both exon-loss and intron retention are explained by an additional loss of a donor site, either downstream or upstream of the lost acceptor site.

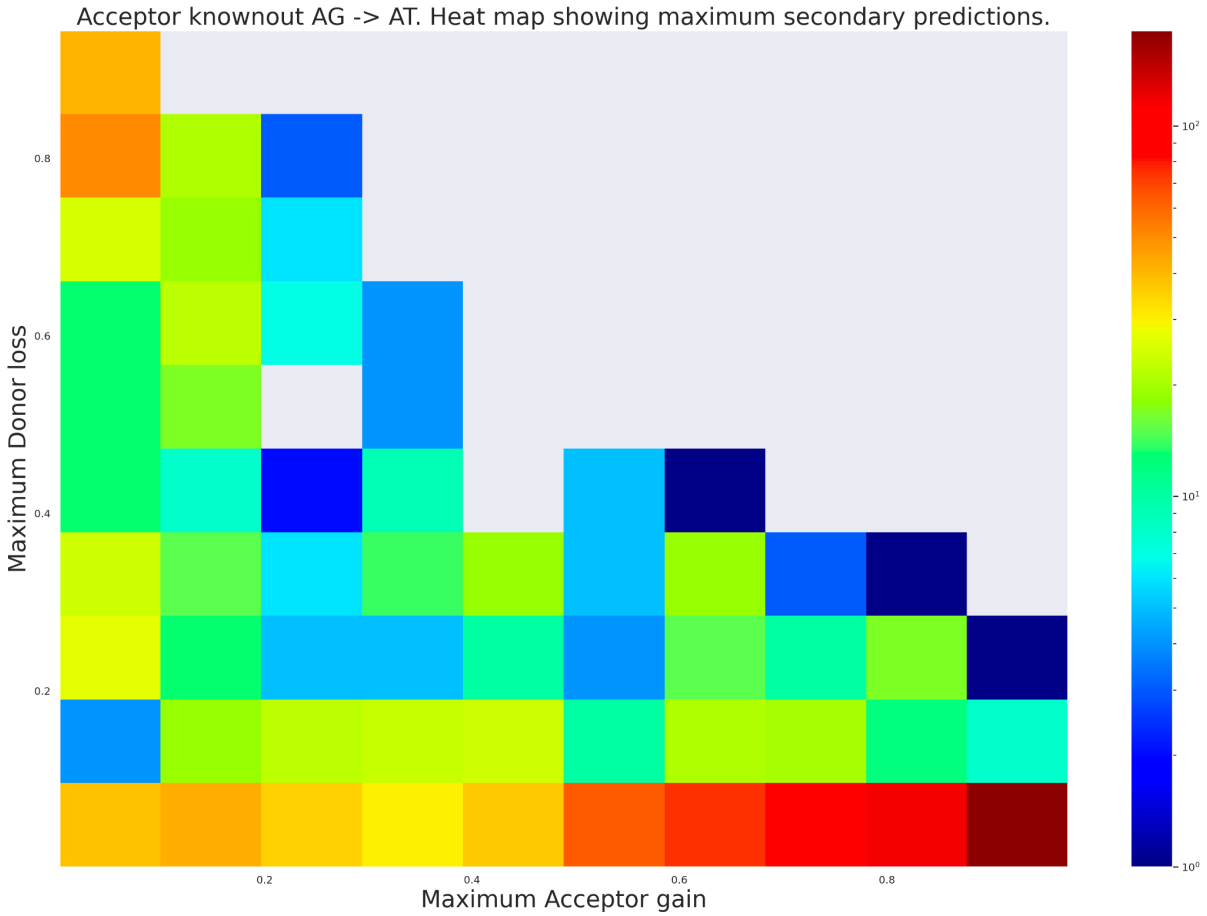


Figure 2.18. Effect of knocking out a known acceptor site. Shown is a 2D histogram of acceptor gains and donor losses as a result of knocking out a known acceptor site. We mutate a reference sequence to remove known acceptor sites, and record the resulting maximum increase of predicted acceptor activity across a window around the known acceptor (excluding the original acceptor site), and the maximum decrease of predicted donor activity within the same window. Concentration in the bottom-right corner shows that the prevailing transcript effect is to preserve the exon with a well-defined alternative acceptor.

We next analyse the impact of the loss of a strong acceptor site. We mutate the canonical acceptor motif ‘AG’ of known constitutive acceptor to ‘AT’, which should abolish any acceptor role for that site. We then look at how predictions for acceptor and donor sites, apart from the disappearance of the known constitutive acceptor site, change as a result of this mutation, in a window around the constitutive site. To achieve this, we record the largest gain of predicted acceptor site strength in the window, as well as the largest loss of predicted donor strength. The two dimensional histogram of the resulting acceptor gains and donor losses is shown in Fig. 2.18.

We find that our model predicts that acceptor knock-out broadly results in the gain of an alternative acceptor, leading to preservation of the exon although with an altered splice site (and possibly leading to a change in reading frame). In a minority of cases, the knock-out results in accompanying loss of a donor site, which could lead to either exon loss (when the lost donor is downstream) or intron retention (resulting from loss of an upstream donor). In 95% of cases either a strong gain of acceptor or strong loss of donor is predicted; in the remainder of cases the model predicts a mixture of the two.

The model preference for selecting an alternative acceptor site, does not, in itself, lead to any conclusions concerning intron-definition vs exon-definition. The null hypothesis of intron-definition is that splice sites are defined sequentially. Selecting an alternative acceptor site is consistent with this. However, this does suggest that putative acceptor sites operate in a competitive manner with the necessity of a downstream donor site for their selection. It is curious that an existing acceptor site can be weakly predicted until a nearby acceptor site is removed. The strength of the gained predictions suggests that the model

sees the alternative acceptor as a strong splice site in the absence of the original acceptor site. It also suggests that there is much more local information than the acceptor site motif in what makes a strong acceptor.

2.6.3 Effect of donor loss

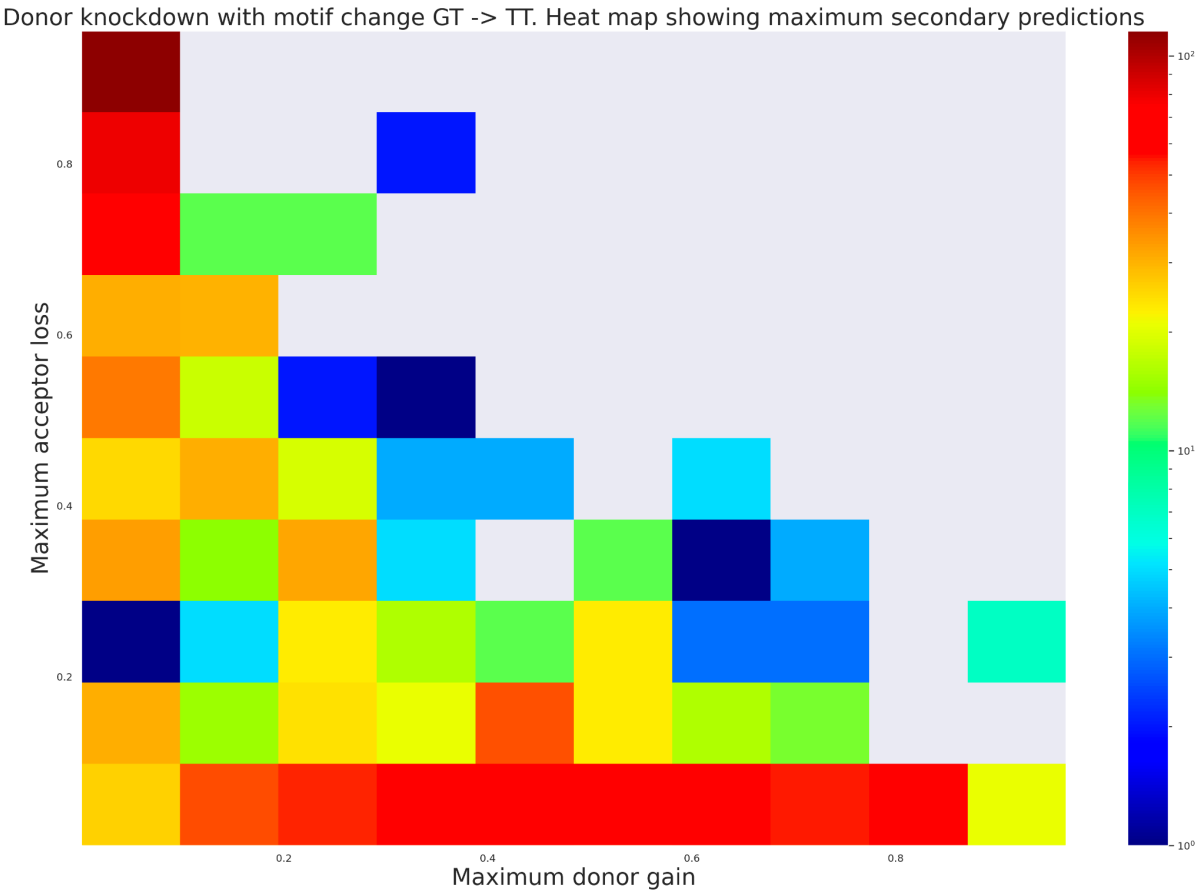


Figure 2.19. Effect of knocking out a known donor site. Shown is a 2D histogram of donor gains and acceptor losses as a result of knocking out a known acceptor site. We mutate a reference sequence to remove known acceptor sites, and record the resulting maximum increase of predicted donor activity across a window around the known donor (excluding the original donor site), and the maximum decrease of predicted acceptor activity within the same window. Concentration in the top-left corner shows that the prevailing transcript effect is to lose the exon. Spread of values in the bottom row shows that even when the exon is preserved the model does not have one well-defined alternative donor.

When performing the same *in silico* analysis for donor sites we observe a slightly different result. Again, in most cases the model predicts either mainly a loss of acceptor or a gain of an alternative donor, and rarely a combination of the two; but in contrast to acceptor knock-out, in this case either a strong loss of acceptor is predicted, or a gain of donor site of various strengths. This is seen in figure 2.19 by the relative concentration of additionally predicted acceptor losses vs alternative donor gains.

This provides clear model support for the exon-definition model whereby good definition of a subsequent donor site is key for acceptor-site definition. Knocking out the donor site leads the model to not identify the exon-associated acceptor site over finding an additional, less strong, donor site. Under a pure intron-definition model, one would expect the removal of a donor site to mostly lead to selection of a weaker alternative donor site, as the acceptor site strength would be left unchanged. In the null hypothesis of a pure intron-definition model, the choice of donor site would have almost no consequence on the strength of the upstream acceptor site.

In addition, we have a curious observation which we will discuss more in our rare disease variation work (chapter 3). When knocking out an acceptor leads to an alternative acceptor, there is quite a pronounced effect. The new acceptor is typically strongly predicted at one site. Conversely, when we witness the creation of alternative donors the model appears to split the probability between more competing options as observed by the gained donor size not being weighted towards 1. Essentially, the model finds it difficult to place its predictive weight on one clear alternative donor when it does not predict an exon-loss.

2.6.4 Gain of Intron vs Loss of Exon

It is observed that exons have a constrained length distribution, with over 80% of exons being shorter than 200 base pairs [27]. This has an impact on the prevalence of intron retention as opposed to exon loss, and we briefly consider this with its relationship to acceptor and donor loss.

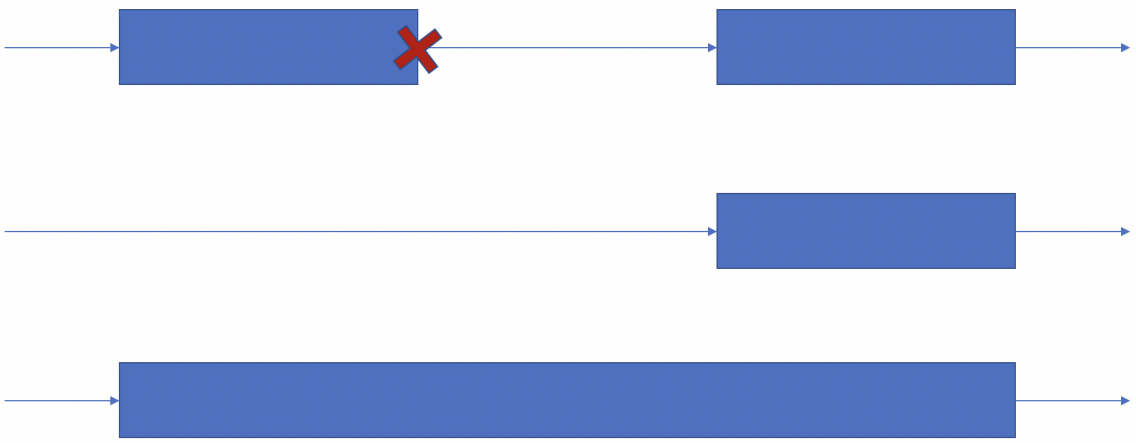


Figure 2.20: Balance of exon loss vs. intron retention as a result of donor knockout. We consider mutations that knock out a donor site (top), and that are predicted to also lead to loss of an acceptor site. Depending on whether this lost acceptor is upstream or downstream of the donor site, this results in exon loss (middle) or intron retention (bottom).

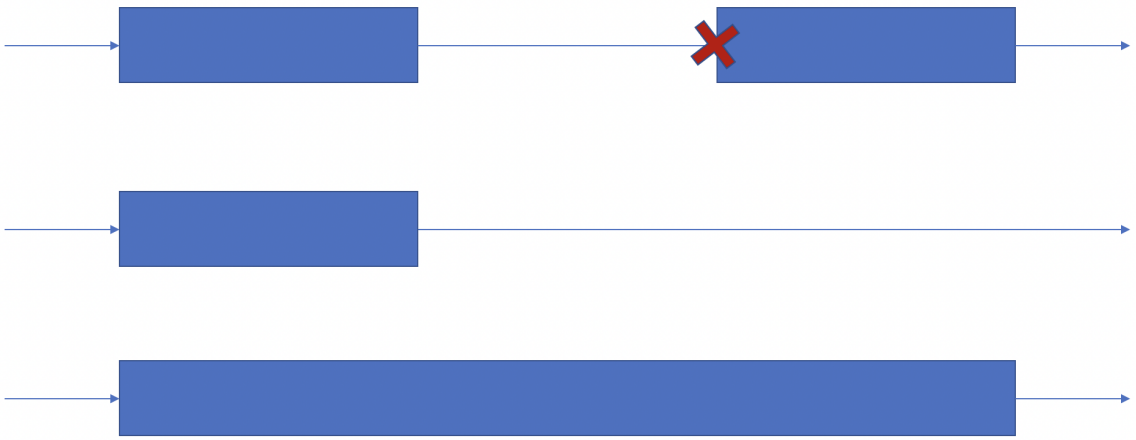


Figure 2.21: Balance of exon loss vs. intron retention as a result of donor knockout. We consider mutations that knock out an acceptor site (top), and that are predicted to also lead to loss of a donor site. Depending on whether this lost donor is upstream or downstream of the acceptor site, this results in intron retention (bottom) or exon loss (middle).

As discussed above, the model recognises that a weakening splice site leads to changes in splice-site strength at other splice sites. In particular, loss of a donor or acceptor often leads to the additional loss of an associated acceptor or donor site (see figures 2.18 and 2.19). If this associated splice site is adjacent to the same exon as the primary splice site, this leads to exon loss. However, if it is adjacent to the next or previous exon, it leads to intron retention. This leads us to ask which of the two scenarios, exon loss or intron retention, is predicted to be more prevalent.

To answer the question about the relative prevalence of exon-loss vs intron-gain, we consider artificial knockouts of donor and acceptor sites when the intron length is fully

contained within the model's window (16384 bp). The model predicts a substantial number of cases in which a loss of a splice-site would result in an alternative splice site of the same type (i.e., no intron retention or exon loss). These examples are inconclusive for this question. We therefore restrict our attention to examples of:

1. Donor losses leading to additional acceptor losses; and
2. Acceptor losses leading to additional donor losses.

Both of these situations may lead to either exon loss or intron retention, depending on the location of the additional splice site loss relative to the primary knock-out.

Results in figure 2.22 show that the model strongly predicts exon-skipping as the consequence of acceptor loss except when the intron is small. For the case of donor sites, figure 2.23, it predicts a strong preference for exon-skipping irrespective of intron length. In our simplistic understanding of "exon-definition" we would expect there to be symmetry in the likelihood of intron retention. It is not clear to us why the removal of a downstream acceptor site would lead to the selection of a further downstream donor site when defining the exon.

Removing the acceptor site (as in figure 2.21) can lead to the previous acceptor site pairing with the downstream donor site when the intron is short. This suggests that when an exon is relatively short, donor motifs can be present, without utilisation within the exonic sequence. Given that the model has learnt that this behaviour is possible, it implies that many canonical exons have relatively strong but not utilised exonic donor sites.

In summary it is of interest to us that the model picks out this asymmetry between acceptor and donor loss. However, we do not have a firm understanding of its meaning. It could be argued that there is a bias from the model to exon-loss over intron retention. However, this

does not answer the question why a stronger bias is observed for donor loss than acceptor loss.

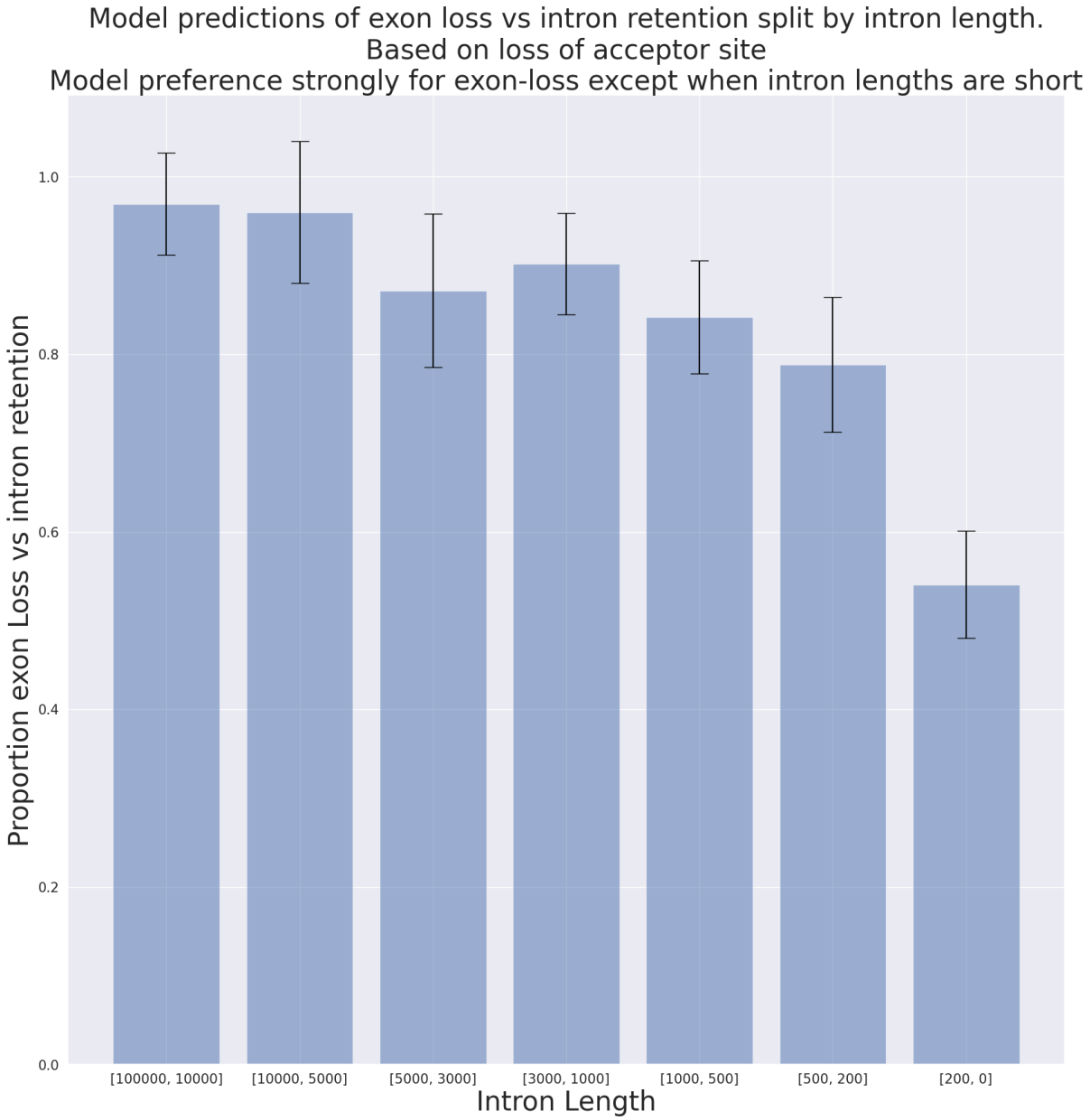


Figure 2.22. Analysis restricted to sites not leading to alternative acceptor sites. Split by intron length we compute the proportion in which the model predicts exon-loss (measured by downstream donor loss) vs intron-retention (measured by upstream donor loss). Except when intron lengths are short (< 200 bps) we observe that the model has strong preference for exon loss.

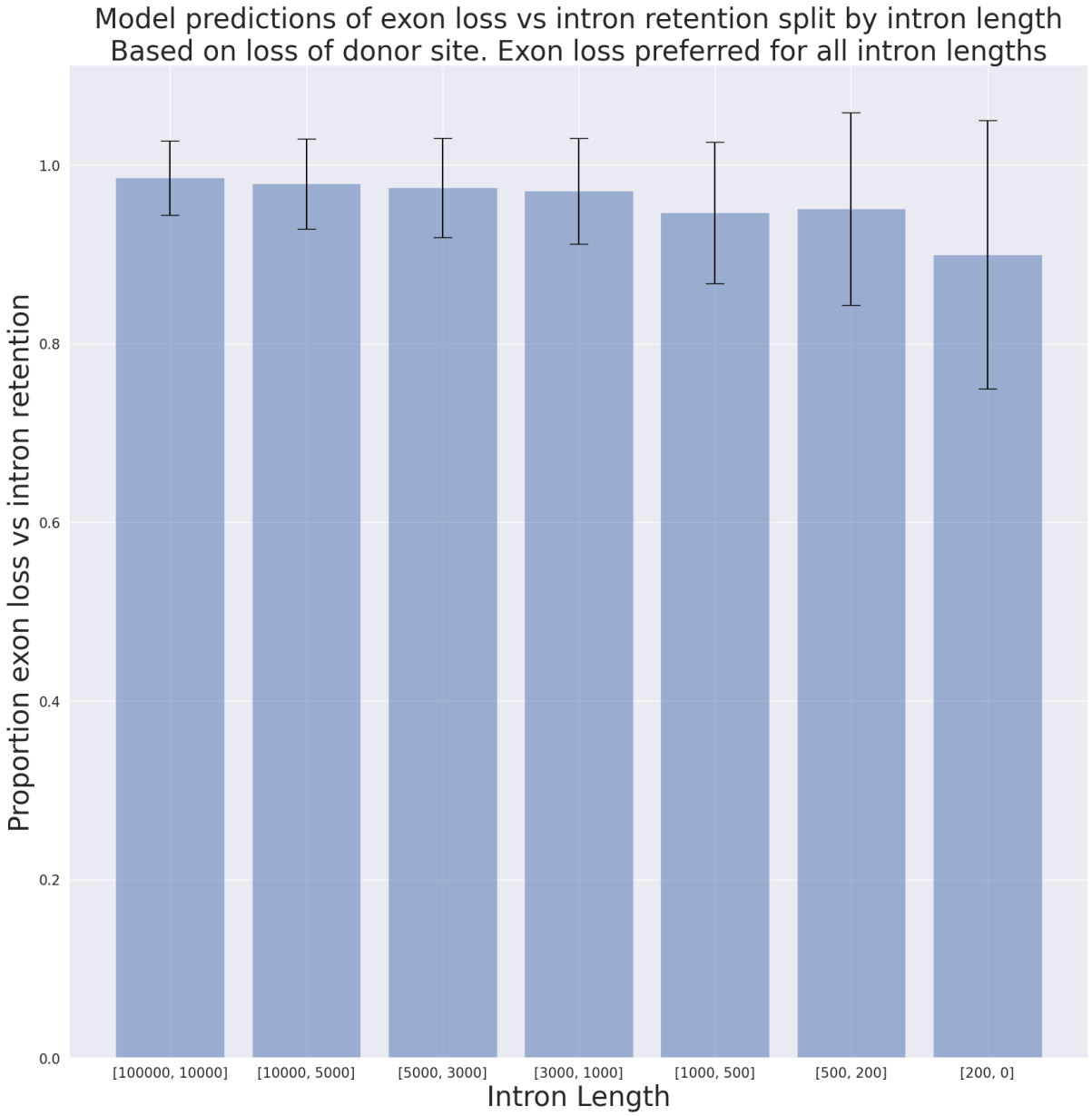


Figure 2.23. Analysis restricted to sites not leading to alternative donor sites. Split by intron length we compute the proportion in which the model predicts exon-loss (measured by upstream acceptor loss) vs intron-retention (measured by downstream acceptor loss). Even when intron lengths are short (< 200 bps) we observe that the model in the majority of cases predicts exon-loss.

2.6.5 Recursive splicing

Recursive splicing is a process by which splicing occurs but does not lead to an included exon. It has been shown that splicing can occur at these sites, particularly in long introns, where such sites may act as “stepping stones” to recursively splice out these long introns while not adding sequence to the mature mRNA [11]. Mechanistically, the process is thought to be similar to ordinary splicing, except that the “exon” defined by the acceptor and subsequent donor site has length 0.

As an initial analysis, we examine the model’s predictions using empirical data from Sibley et al [11] which details a list of 81 verified RS-sites, which are identified in long introns. We observe in figure 2.26 that the predicted donor strength at these sites is generally much higher than found in the genome at large, or when conditioning on acceptor sites of known exons. Although this is a small data set, it suggests that the model is partially learning these RS-sites despite no explicit RS sites being present in the training data.

The model for recursive splicing outlined in [11] is that the acceptor site is defined (as per exon-definition) by a downstream donor site at some distance from the acceptor, as usual. However, this downstream donor site is then not used, but instead a donor site directly adjacent to the acceptor site is used - resulting in a “zero-length alternative exon” configuration. This model suggests that recursive splicing can be a mechanism for exon skipping.

A quick look in the genome reveals a puzzling observation. If recursive splicing is used as a mechanism for exon skipping, we would expect that for alternative exons, the canonical

acceptor motif preceding the exon, *AG*, to be immediately followed by the canonical donor motif, *GT*, as the first two nucleotides of the exon. However, this *AGGT* pattern is actually very common at constitutive, and therefore unlikely recursive, splice sites. In fact, 18% of acceptor sites (defined by model PSI >= 0.8) have the additional *GT* motif following the acceptor site. Looking beyond the canonical sites, we find that of the 30 most frequent 5 bp nucleotide motifs following these acceptor sites (accounting for 12% of all sites), 22 are known donor site motifs (8.9% of all sites). The converse is also true: fully 56% of donor sites (defined by model PSI >= 0.8) have the additional presence of an *AG* motif before the donor site. In short, apparent recursive splice site motifs are more prevalent in gene models than might be expected from the known alternative splicing behaviour of these genes, and a simple bioinformatic analysis is apparently insufficient to identify true recursive splice sites.

Instead, we decided to rely on our model to try and investigate putative recursive splice (RS) sites. Apart from the 81 sites from Sibley et al, we did not have access to comprehensive data on experimentally identified RS sites, and our model was trained on GTEx RNA-seq data, in which RS sites leave no trace. However, to the extent that RS is simply ordinary splicing of a zero-length exon, the model's predictions learnt on exons with positive length might generalise to the RS case. This is what we aim to investigate.

We defined an acceptor site as a predicted RS site if the model also predicts a donor site at the same location. Focusing on known annotated acceptor sites, we find that there is a significant enrichment of predicted RS sites for acceptor sites that are known to be alternatively spliced. Figure 2.24 shows that exons at which our model predicts recursive splicing, by virtue of donor activity at the acceptor site, are more frequently skipped

compared to exons in general, and the frequency of exon skipping increases as the predicted donor activity is stronger.

Exon skipping could also result by recursively splicing at an exon's donor (3') end - that is, by selecting an alternative *acceptor* site at the same locus as an existing donor site. Similarly to the analysis above, we look at enrichment of exon skipping (ascertained in GTEx data) as a function of the strength of an acceptor site prediction at the exon's donor site (Fig 2.25). Again, we observe a clear enrichment of exon skipping that gets stronger as the predicted acceptor strength increases, although the enrichment is not as striking as it is in the case of RS at the exon's acceptor (5') end.

In summary, our model appears to have picked up a signal of RS, in two distinct contexts: in long introns (as ascertained by the analysis of sites from Sibley et al.) where RS is thought to aid splicing of long introns [11], and at known exons where RS at both the 5' and 3' ends appear to be a mechanism for exon skipping, a mechanism that has been shown before [28]. The fact that our model is able to pick up a signal of RS in both these contexts, despite having seen no explicit examples of RS, suggests that RS may simply be a special case of "ordinary" splicing where the intervening exon happens to have length 0. However, it is difficult to assess how well our model is able to predict RS, and it remains possible that RS is further regulated by additional features, which our model would not have picked up because of the lack of training data.

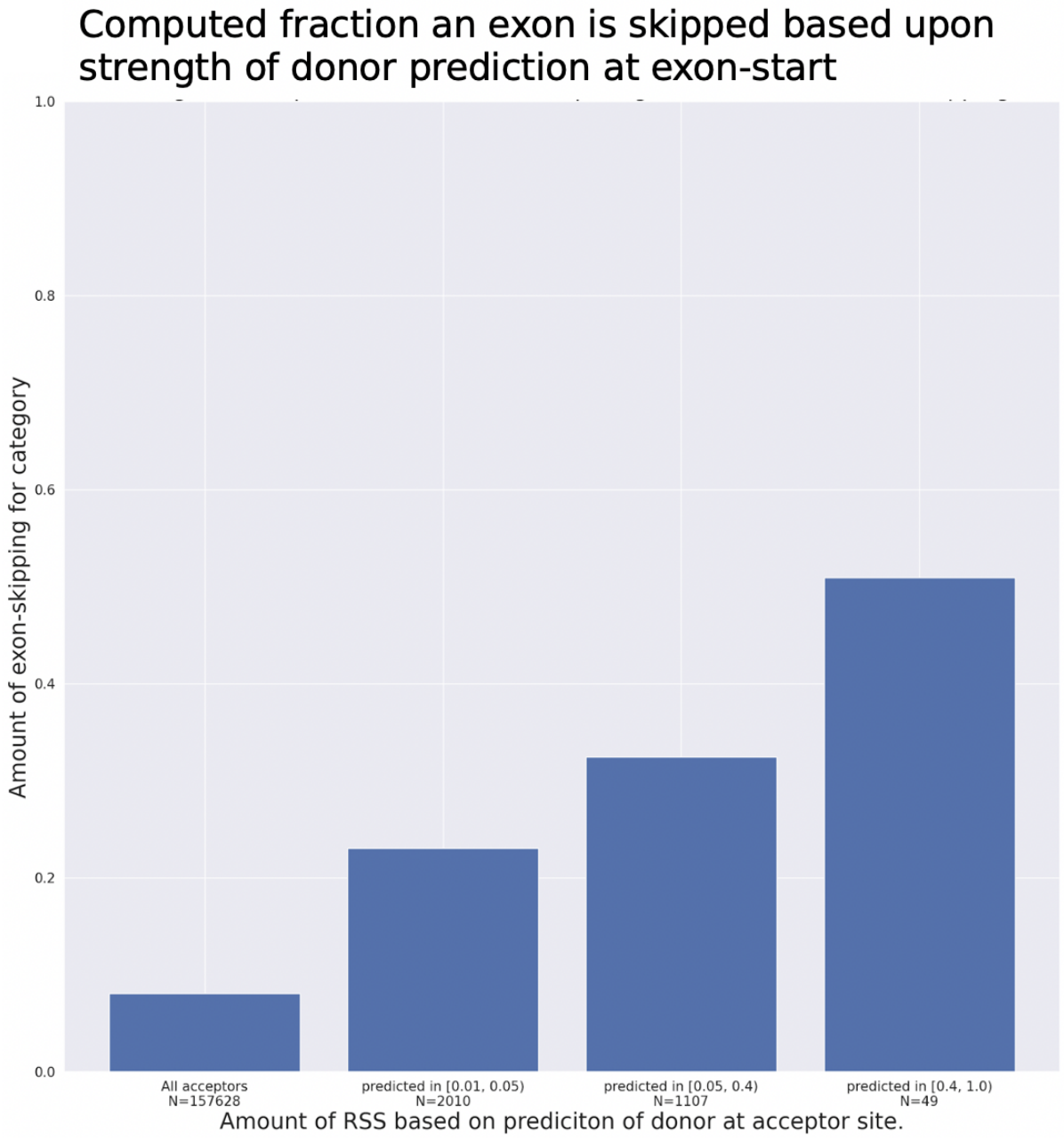


Figure 2.24: Recursive splicing at acceptor sites as a mechanism for exon skipping. Strengths of recursive splicing was defined as the predicted strength of a donor site, coinciding with the known acceptor at the exon start site. Bars show the proportion of exons exhibiting exon skipping, in bins of predicted strength of recursive splicing. Acceptor sites that show stronger predicted donors at the same site show more evidence of exon skipping. Analysis taken over known acceptor sites as defined by USCS V37 with the quantity of exon-skipping defined by GTEx.

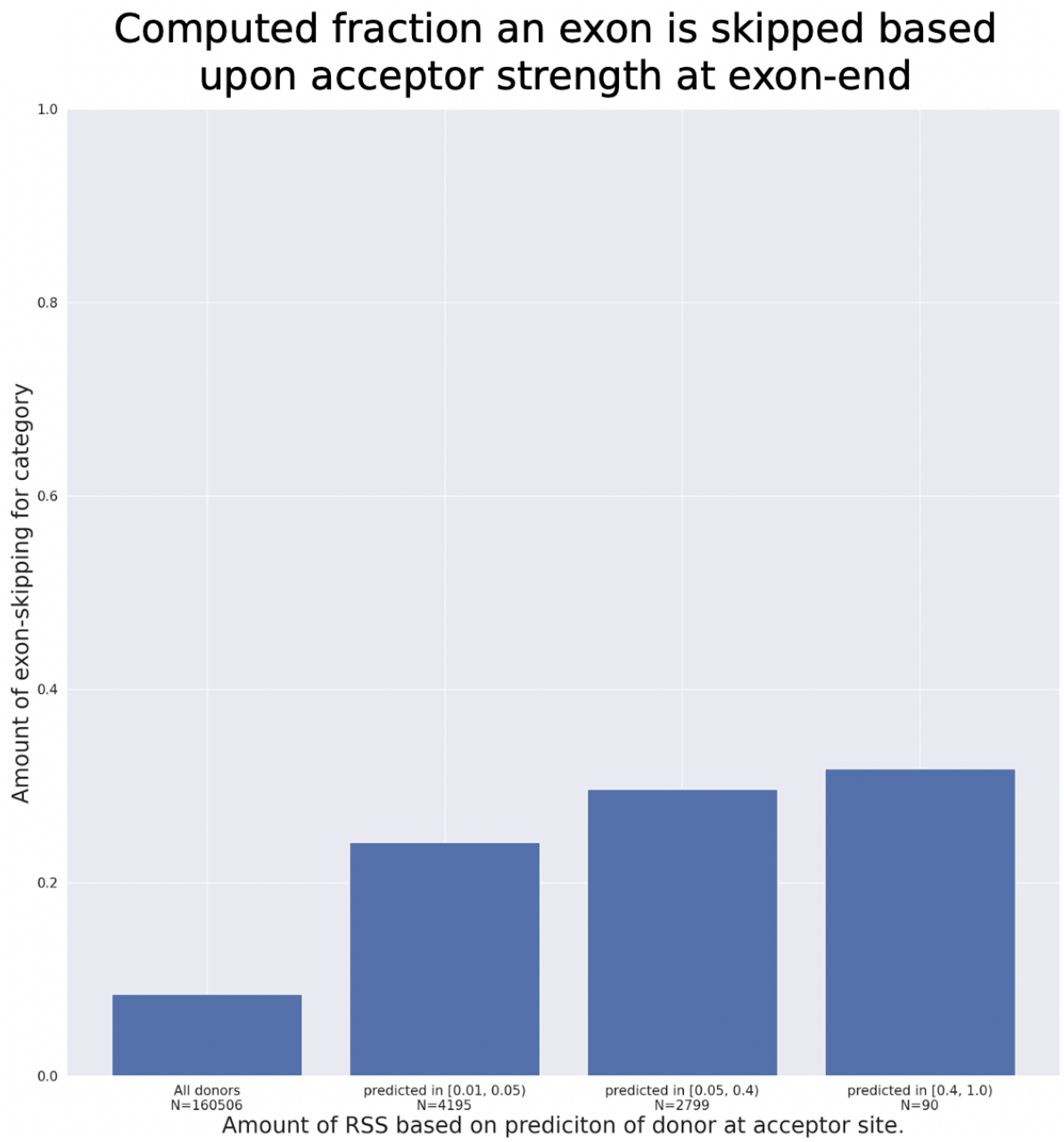


Figure 2.25: Recursive splicing at donor sites as a mechanism for exon skipping. Strengths of recursive splicing was defined as the predicted strength of an additional acceptor site, coinciding with the known donor. Bars show the proportion of exons exhibiting exon skipping, in bins of predicted strength of recursive splicing. Donor sites that show stronger predicted acceptors at the same site show more evidence of exon skipping. Analysis taken over known donor sites as defined by UCSC V37 with quantity of exon-skipping defined by GTEx.

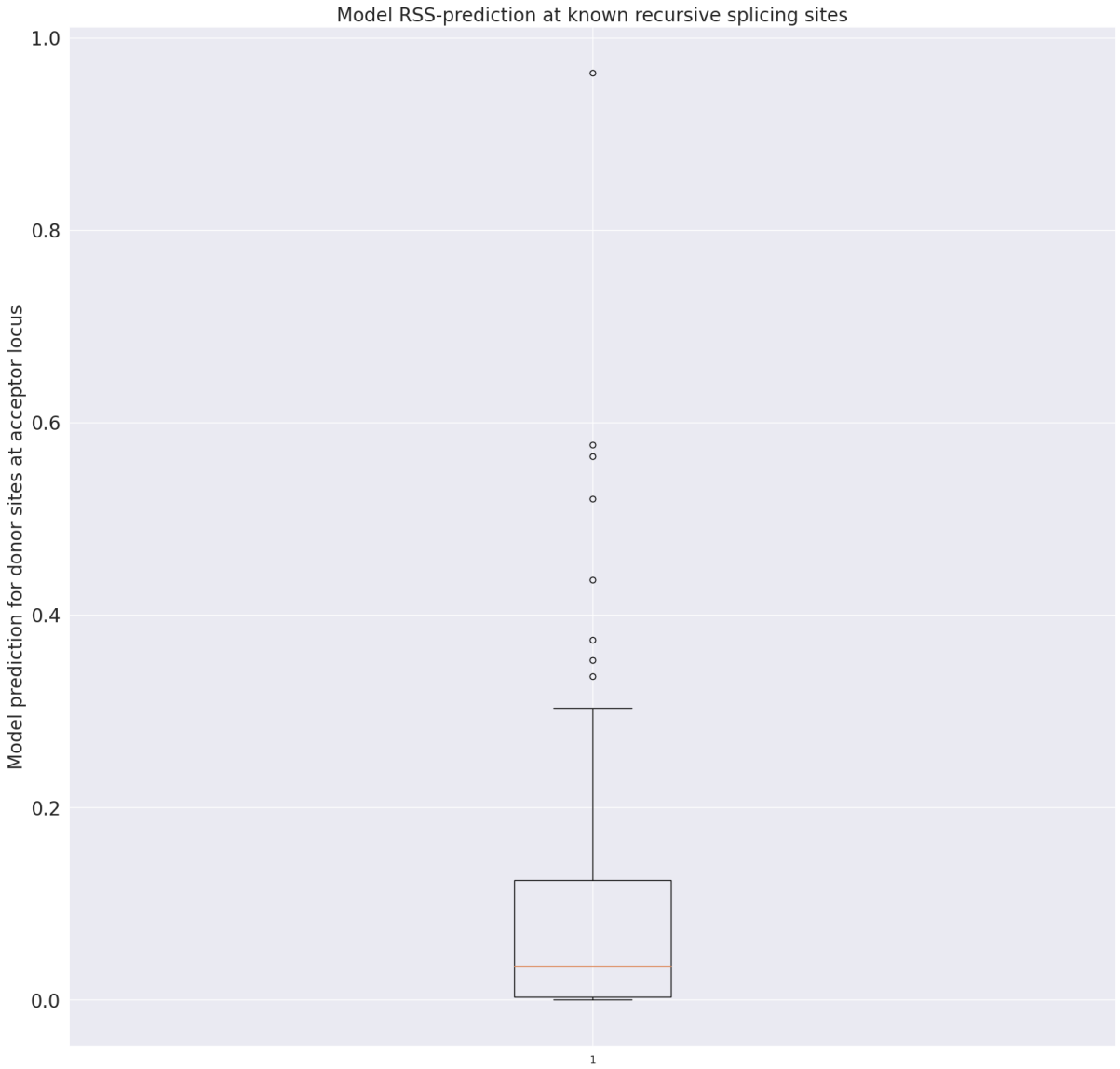


Figure 2.26. For validated RS-sites in Sibley et al we compute donor model scores for the recursive sites. We observe a median score for a donor site prediction at the recursive splice site location of 0.05 which suggests the model is learning the recursive signature despite no training.

2.6.6 Estimating the prevalence of recursive splicing in the transcriptome

In the absence of a good-sized data set of validated RS sites, we cannot train our model on RS sites, and neither can we estimate its accuracy at predicting them. Nevertheless, as we have seen, our model seems to have picked up a signal of RS. The question therefore arises, what is the prevalence of recursive splicing in the human transcriptome? A direct estimate based on model predictions is not possible. As observed, donor-like motifs arefound following acceptor motifs, but this occurs both on alternatively spliced and constitutively spliced exons, so a simple motif analysis is also inappropriate. Instead we devise an enrichment analysis to come up with a lower bound estimate.

We argue that if predictions of acceptors at sites where no GTEx evidence of splicing exists are in fact false positive predictions, we would then expect these to look similar to “normal” acceptors. In particular, they maybe followed by a GT donor-like motif as this appears to be common for acceptors; but we would not expect these to be followed by motifs associated with particularly strong donors. Instead, if this category are in fact true but recursively spliced acceptor sites, we *would* expect them to be followed by particularly strong donor motifs, in order to define an RS site. Therefore, the enrichment of motifs associated with strong donors is a signal that a fraction of strong predicted acceptors without GTEx evidence in fact undergo recursive splicing.

We collect a set of 196463 sites, defined as having a prediction for being an acceptor site (predicted PSI > 0.01), and followed by the *GT* motif. Among these, GTEx has evidence of constitutive splicing for 24755, and limited evidence of splicing for 170816; 892 appear to be alternatively spliced (0.2 <= PSI <= 0.8); these were ignored for this analysis.

Among the 196463, we find 10834 sites (5.5%) that are followed by a strong donor motif (*GTGAG* and *GTAAG*), and 52101 sites (26.5%) followed by an intermediate donor motif. Among the 24755 sites, these numbers are 515 (2%) and 6419 (25.9%). In all, we find that strong and intermediate donor motifs are enriched by 2.8% among sites with low evidence of splicing in GTEx, representing an expected 1498 sites. This indicates that at least 1498 sites, with strong acceptor sites, undergo recursive splicing in the transcriptome. Using this dataset, ignoring the motifs following the acceptor sites, there are 139837 sites which have evidence for constitutive splicing in GTEx. Therefore, this suggests ~1% as a lower bound for recursive splicing induced depletion of otherwise constitutive acceptor sites.

This estimate is likely to be an under-estimate, because (1) we only considered acceptor sites followed by *GT*; (2) we used simple 5-nucleotide motifs to stratify donor activity; and (3) we only consider the depletion of otherwise constitutive acceptor sites. It seems possible that recursive splicing is a reasonably important mechanism in defining the transcriptome.

	PSI <= 0.2	0.2 <= PSI <= 0.8	PSI >= 0.8
GT	115068 (86.2%)	639 (0.5%)	17821 (13.3%)
Weak donor motif	45452 (87.2%)	230 (0.4%)	6419 (12.3%)
Strong donor motif	10296 (95.0%)	23 (0.2%)	515 (4.8%)

Table 2.27: Model predicted acceptor sites are relatively depleted in GTEx datasets based upon presence of strong donor site following acceptor site. Strongest depletion for the two main donor sites. Using absolute counts of depleted sites it suggests 1% as a lower estimate for prevalence of recursive splicing amongst strong acceptor sites.

2.7 Discussion and conclusions

We aimed to build a model to predict alternative splicing directly from sequence, combining the features of *SpliceAI* and *MMSplice.* We presented a neural network model to predict the usage of transcribed sites as donor and acceptor splice sites, measured as Percent Spliced In (PSI), directly from sequence. We broadly followed the residual U-Net convolutional neural network architecture and the architecture of SpliceAI, but we employed a loss function that allows the model to predict a distribution of splice frequencies, which effectively deals both with uncertainties in the prediction and observations, and unobserved random effects due to epigenetic features such as histone modifications, and other features that may contribute to splicing. We trained the model on GTEx read junction data, and included the full spectrum of splice sites from low-frequency to constitutive.

We found that the model broadly captures the observed PSI frequencies, but has a tendency to over-predict PSI – in particular, in many sites where the model predicts PSI>0, our data shows no splicing at all. However, where alternative splicing was observed, the model usually predicts this, although not quite at the correct level. We also found that the model exhibits a particular bias for highly expressed genes: for those genes, where alternative splicing is observed, the model tends to predict constitutive splicing.

The first observation may be explained if our training data systematically excluded certain transcripts. In Chapter 4, we will see that Nonsense Mediated Decay, which degrades transcripts that have a stop codon before the terminal exon, causes many splice events to be unobservable in mRNA-seq data, and this will go some way towards explaining the upward bias.

The second observation, that alternatively spliced highly-transcribed genes are predicted to be constitutively spliced, is more difficult to explain. It may be that for these, presumably important, genes, alternative splicing is regulated differently, perhaps by a mechanism that is specific to a specific group of genes. This would be much harder for the model to pick up, resulting in the observed bias. A possible mechanism for regulation is by histone modification; it is known that splicing is partially regulated by histone modification signals [15], and perhaps this mechanism regulates splicing in these highly expressed and important genes.

We then interrogated the model to better understand how sequence features combine to result in a splicing prediction. We found that proximal loci contribute in expected ways (up to 40 base pairs upstream of acceptor sites, and 10 base pairs downstream of donor sites). Knockout of acceptor sites frequently lead to gain of alternative acceptors and inclusion of an alternative exon, and less frequently to loss of donors resulting in exon skipping or intron retention (figure 2.18). Similarly, knockout of donors frequently leads to gain of (sometimes weak) alternative donors resulting in alternative exons, and less frequently to acceptor loss resulting in exon skipping/intron retention (figure 2.19).

Finally, we looked at recursive splicing (RS), and found that our model has some power to pick up known RS sites, and that it finds a signal of RS in long introns, as well as at the splice sites of exons that exhibit exon skipping, consistent with a suggested role for RS in exon skipping. This raises the importance of RS, and broadens the mutational target for splicing. More generally, mutations that impact RS sites, either at exons involved in exon skipping, or

in long introns, may change transcripts or change the relative expression of isoforms, and thereby have an impact on phenotype. These mutations may easily be overlooked, and our model provides a way to assess the potential impact of such mutations.

It is surprising that the neural network predicts recursive splicing at all. The training data does not contain this signal, as RS is invisible in mature mRNA. Despite this, as we have seen in this chapter, the predictions of the model are consistent with both RS as a mechanism for exon-skipping, and intronic recursive splicing.

3 Predicting the effect of mutations

This chapter looks at applying the model to predict the downstream disease effect of rare mutations through their effect on splicing.

The main objective is the utilisation of the model to real-world examples of splice-altering mutations. A secondary aspect is to provide some way to objectively ascertain the quality of the model. As splicing has no perfect ground-truth dataset, it is unclear how to objectively assess how one model is superior to another. One could evaluate the ability to predict constitutive splice sites or the ability to predict all annotated splice sites in some database. Each of these approaches quickly meets the issue that the concept of "splice-site" is hard to define. Splicing is not an all-or-nothing process, and involves the interplay of multiple sites - at minimum, both the donor and acceptor sites.

Instead, in this chapter we evaluate how well models perform at predicting *known disease-causing mutations* which act by causing *splicing changes*. This is an objective end-point for assessing a splice prediction model. In addition, it evaluates the models on an important use case: diagnosing genetic disorders.

We use *SpliceAI*, both masked and raw scores, as our reference model to compare to. This tool is considered the leading *in silico* strategy for prioritising rare variants [3], [29].

At the time of writing we note that a recently published paper optimised the training procedure of SpliceAI, using a curated training set, resulting in improved prediction accuracy

[29]. This result came too late to be included in our comparison, however we use the same evaluation metric (area under the precision-recall curve, AUCPR) but applied to a different disease-control dataset.

3.1 The HGMD dataset

The Human Genome Mutation Database (HGMD) is a large database of known disease-causing mutations [30]. Every mutation in the dataset is included based on publications which link the mutation to a disease phenotype. *R.Brown* curated a subset of the HGMD dataset based upon being annotated as splice-altering, and we use this set of mutations to evaluate the models [17].

In addition, *R. Brown* curated a matching control set in his thesis to allow comparisons between different splicing algorithms [17]. The matching control set was taken from DbSNP [31] by selecting a set of SNPs at 1% allele frequency that has a similar positional relationship to exons as the HGMD variants [31].

One reason for utilising the analysis outlined by *R. Brown* is that it enables one to evaluate different approaches in an unbiased manner [17]. Metrics discussed in the previous chapter do not readily compare between different approaches, as the validation data is not necessarily complete due to RNA-decay or recursive splicing. In addition, some models explicitly target constitutive splicing in an attempt to best model splice events that occur in canonical transcripts. We are explicitly attempting to model alternative and cryptic splicing.

On the other hand, the set of known disease-causing mutations is strongly biassed towards known transcripts and canonical splice sites, because it is difficult to recognise a mutation as

affecting splicing if it does not affect a known and strong splice motif in a known transcript. We will come back to this aspect later in this chapter.

The other more direct motivation for the analysis presented in this chapter is to assess the ability to apply to disease diagnostics and SNV prioritisation. In this chapter we compare ourselves to *SpliceAI,* which has been shown to be the leading single algorithm approach in the prioritisation of pathogenic variants [3]. We compare to both their masked and raw score files.

3.2 Applying Neural network to mutagenesis

For a putative splicing mutation, we run the model on a window (16384 base pairs, chosen to span most introns, see figure 2.2) surrounding the mutation. We run the model with and without the mutation then combine the outputs to provide the annotated outputs.

The key annotations we provide include a score which is the maximum absolute difference between the splicing values. Additionally, we provide the locations and scores for the largest gain/loss of donor/acceptor sites along with their reference and alternative predictions.

In addition, for our collaborations on clinical datasets we further annotate whether we predict *nonsense mediated decay* for either the reference or alternative predictions. This helps with follow up validation, as knowing that the alternative transcript is likely to be subjected to NMD means that it might not be observed.

3.3 Effect of combining independently models on mutagenesis

Our comparison is relatively simple and is independent from thresholds. We compute precision-recall curves from the numerical scores calculated by a model, using the HGMD SNPs as our targets and the polymorphisms from dbSNP as our controls. The quality of a tool is quantified by computing the area-under these precision-recall curves. If a tool was perfect it would have an AUCPR of 1. Some tools might be better at precision or recall, but those distinctions are not reflected by the AUPCR value; we therefore also show the actual precision-recall curves.

SpliceAI computes both “raw” and “masked” scores, where the latter are calculated from the raw scores by removing a subset of predictions. Naturally, *SpliceAI*’s “masked” scores achieve a higher precision but a lower recall, particularly in intronic regions, compared to the “raw” scores. Although the “masked” scores are considered the final prediction, we assess both scores to ensure a fair comparison.

We note that the results of our model differ slightly each time the model is trained. Because of the stochastic nature of the neural network training procedure, this is not unexpected. In addition to choosing one of those results randomly, we also assess whether taking the mean of a few independently trained models (on the same training set) improves the results. This technique is also used by *SpliceAI*, which combines the results of 5 independently trained models [4].

We find that *SpliceAI* outperforms our model, when we use the output of a single training instance (figure 3.1). The effect of combining a few models using mean scores is substantial. Combining 2 models produces a comparable performance, while when we take the average of 6 predictions or more independently trained models our results outperform *SpliceAI*’s. A

caveat is that the differences are small which could signify a limit of the HGMD and control data.

The benefit of combining models can be observed from the figure 3.2. For this figure we bin the model predictions for one model output to compare the mean model scores for annotated splice-variants vs our controls. For instance, observing a score in range [0.1, 0.2] in one model, if it is annotated as splice-variant, we expect a splice-score of 0.21 from combining models compared to 0.11, if not annotated as a splice-variant.

A key observation is that the difference from combining models exceeds considerably the differences found by changing the model architecture. Moreover, these differences exceed considerably differences we found through different training strategies which obtained AUC-PRs in the range [0.92, 0.94].

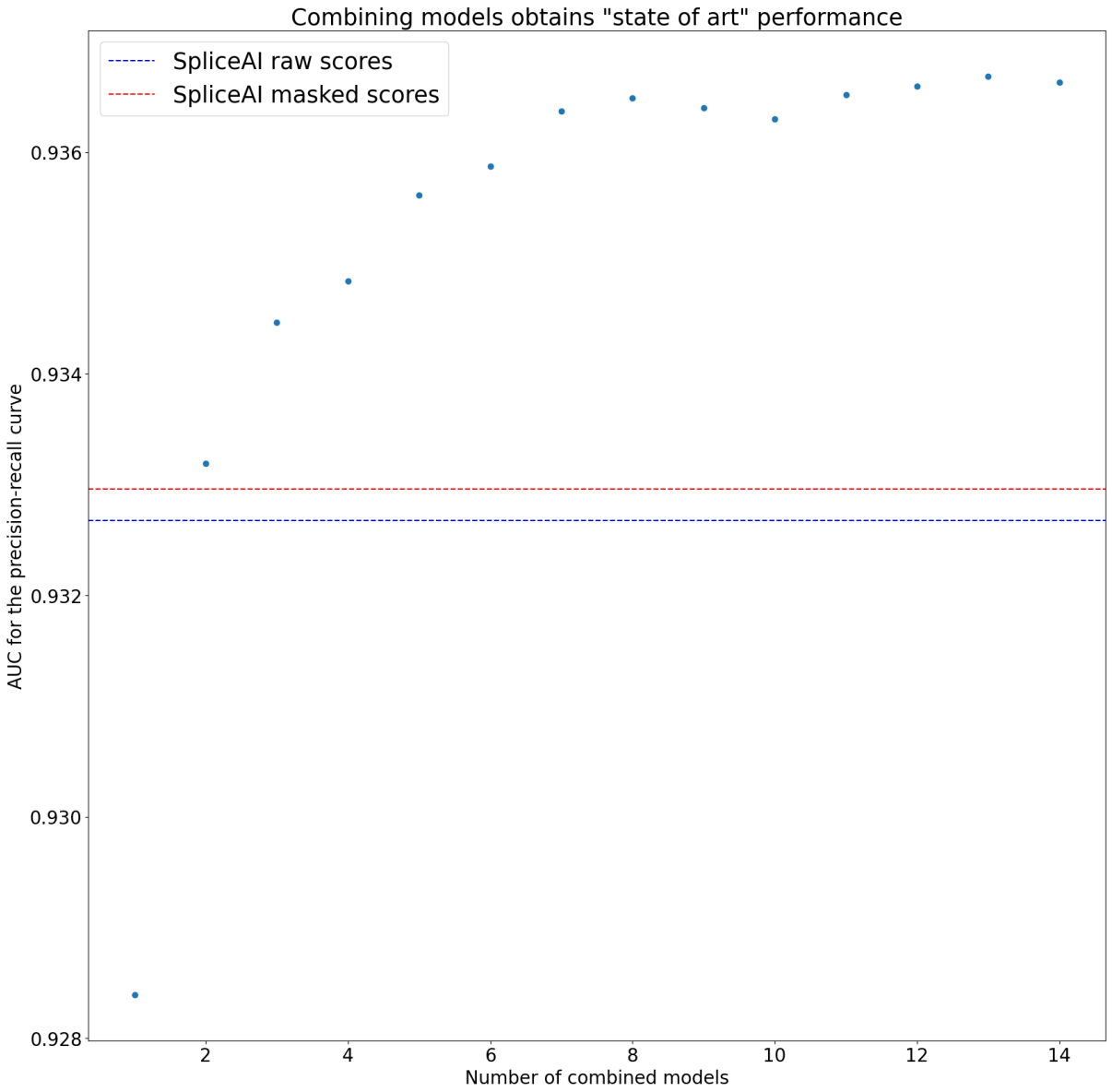


Figure 3.1: We compute the AUC for the precision-recall curve of dissecting HGMD splice variant sites. A high AUCPR shows that a good balance of recall and precision is achieved. With 2 or more models equivalent accuracy to *SpliceAI* is achieved.

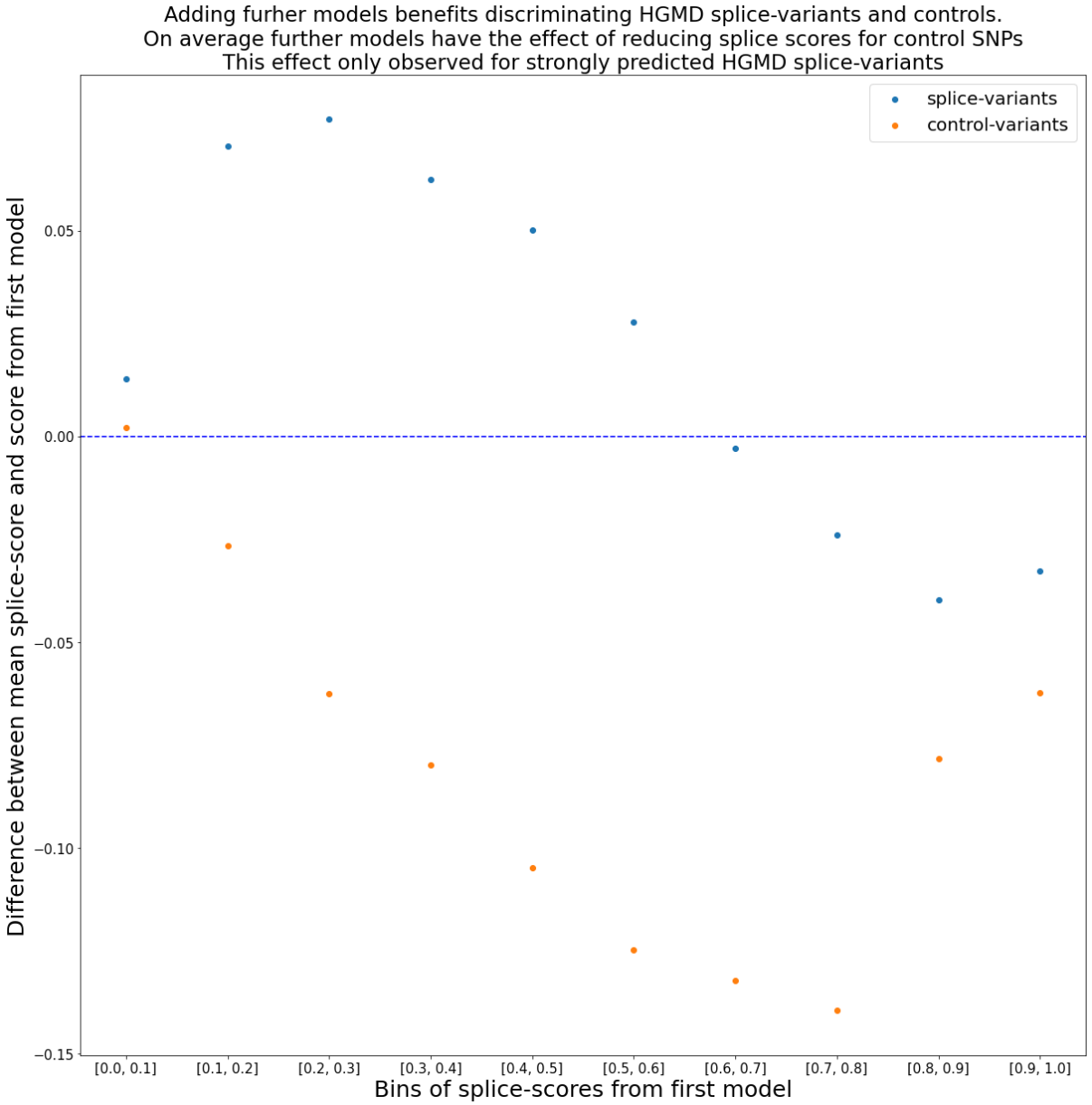


Figure 3.2: In this analysis we explore the change to model scores from using one model to combined model approach. HGMD splice-variants and control-variants are split by the model-scores obtained from one model. We observe that for low to medium scores HGMD splice-variants' scores increase from combining models. Scores which are very strong based upon one model tend to decrease. We observe that scores tend to decrease for control variants irrespective of the score size from one model.

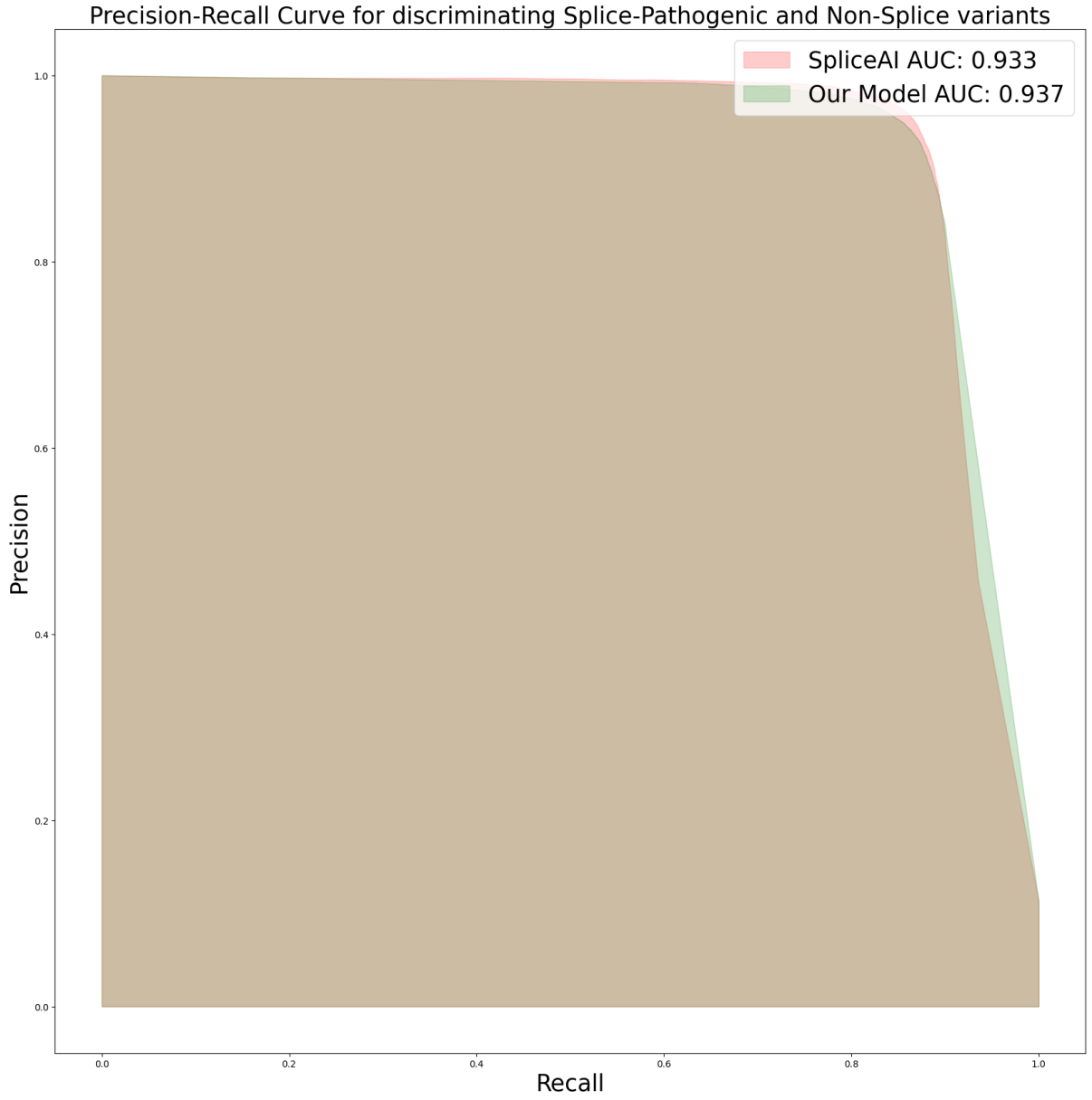


Figure 3.3: Precision-recall curve for modelling approaches at predicting HGMD splice-altering variants against a control set. For this splice-variant disease classification challenge we find that the models obtain broadly consistent results with *SpliceAI* obtaining better precision (lower false positive rates) when one is prepared to sacrifice recall. Our modelling approach obtains a slightly better recall when one is less concerned about precision.

3.4 Applying downstream Machine learning to the Classifier

3.4.1 Improving classification using XGBoost

XGBoost is a python library for classifying data (version 1.6.2). It is a form of Gradient Boosted decision tree which has been applied to clinical diagnostics [32]. For our purpose, XGBoost allows us to make use of many different covariates that could help to predict whether a mutation is likely to cause disease or not.

Previously, we used the maximum change as the criteria to discern whether a splice change is significant. One set of covariates, obtained from our model's output, arises from splitting this metric into 4 further metrics:

1. Size predicted acceptor loss
2. Size predicted acceptor gain
3. Size predicted donor loss
4. Size predicted donor gain

This allows the XGBoost classification algorithm to differentiate between the different splice disruption categories such as exon-skipping, exon-inclusion, alternative donor sites etc. As a combination of a donor and acceptor loss/gain could be an intron loss/gain or an exon loss/gain, we additionally include a boolean value for this prediction. Furthermore, some variants would give rise to splicing changes which do not neatly fit into 1 single category. The use of these covariates enables the XGBoost classifier to access and use this information. Moreover, it is possible that the classifier might provide more weight to a splice-score if associated with a donor gain than an acceptor loss. Donors and acceptors have subtly

different properties, as discussed in section 2.6, so treating scores in the same manner for both is not necessarily sensible.

The next relevant covariate is whether the affected splice sites are known or not. Instead of masking predicted donor/acceptor losses of unknown sites, we include a boolean value to distinguish these.

The final covariates we have chosen relate to the genomic loci of the variant, distinguishing between exonic and intronic variants. We also add gene tolerance scores to add weighting to variants based on the genes they affect.

Additionally, we can use the XGBoost classifier model as an independent means of determining which covariates are important when discerning splicing as it acts as an agnostic unbiased way to rank the covariates.

3.4.2 Building the trained networks

We utilise subsets of the HGMD dataset and subsets of the control SNPs to train the model. For each mutation we create an array of inputs (covariates) as described. The datasets are split three-ways to generate a test, train and validation dataset for the HGMD variants (targets) and control SNPs. This enables us to have confidence that the validation datasets are unaffected by parameter or hyperparameter overfitting. The results presented in this chapter are based upon the validation subset.

3.4.3 Cross validation parameter grid search

We utilise cross validation to build a model which is robust on our validation data. Our validation data is a subset of the HGMD and control SNPs. When training our model,

XGBoost utilises a grid search to optimise hyperparameters which define the architecture of the trees.

3.4.4 Importance scores for each parameter input

Once a model is trained, we can obtain the importance scores which rank covariates by how much they are utilised by the trained model. This allows us to interpret the model to see which covariates it uses to make the predictions.

Unsurprisingly, XGBoost identifies the predicted loss of splice activity at existing splice donors and acceptor loci as important variables for predicting disease impact. It also assigns importance to "oe_lof", which quantifies gene tolerance to mutations, which is also expected.

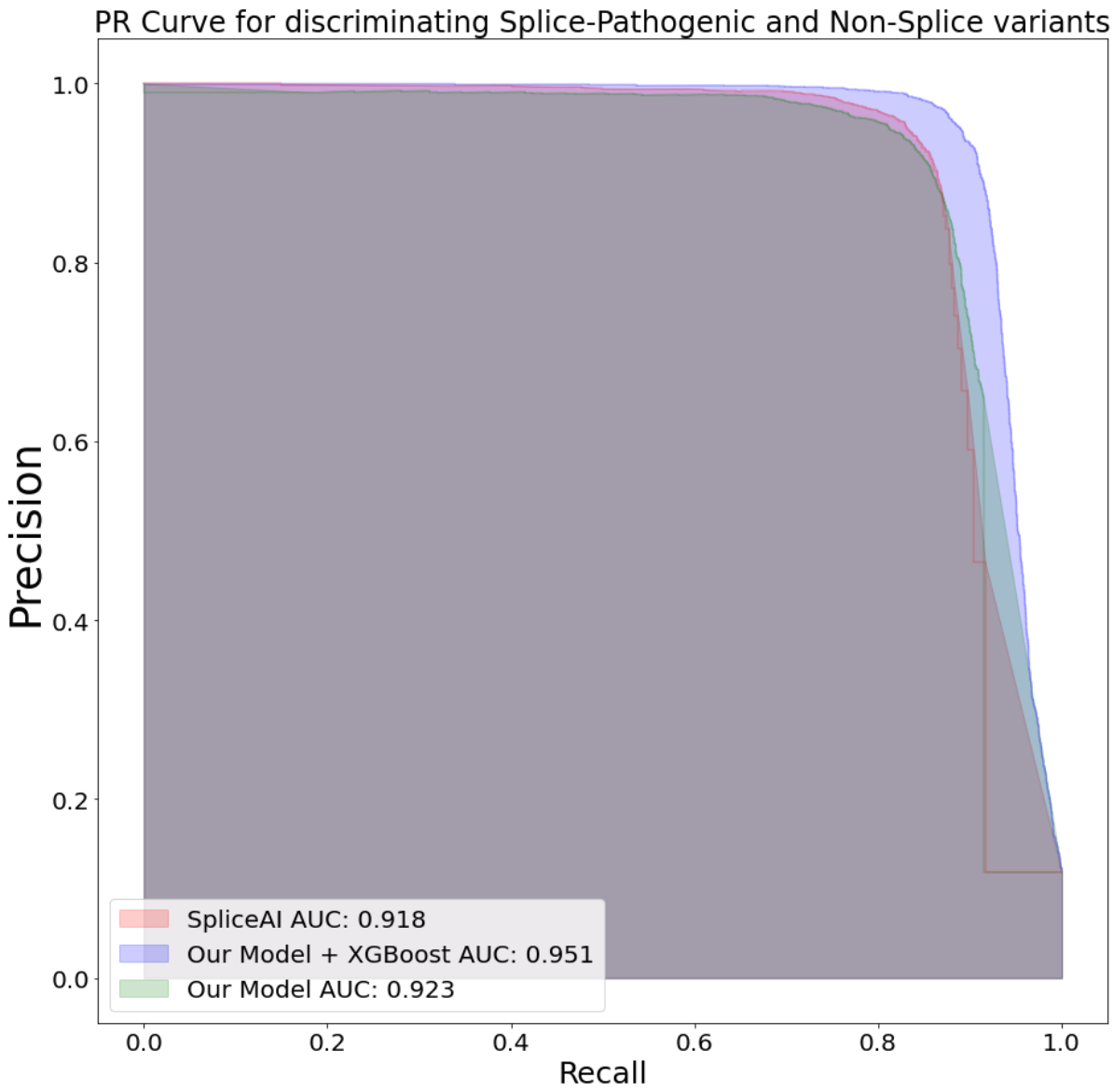


Figure 3.4: Precision-recall curve for modelling approaches at predicting HGMD splice-altering variants against a control set. Applying XGBoost significantly improves our model's AUCPR on a validation set. The improvements made using XGBoost dwarf the improvements made from our alternative neural network approach.

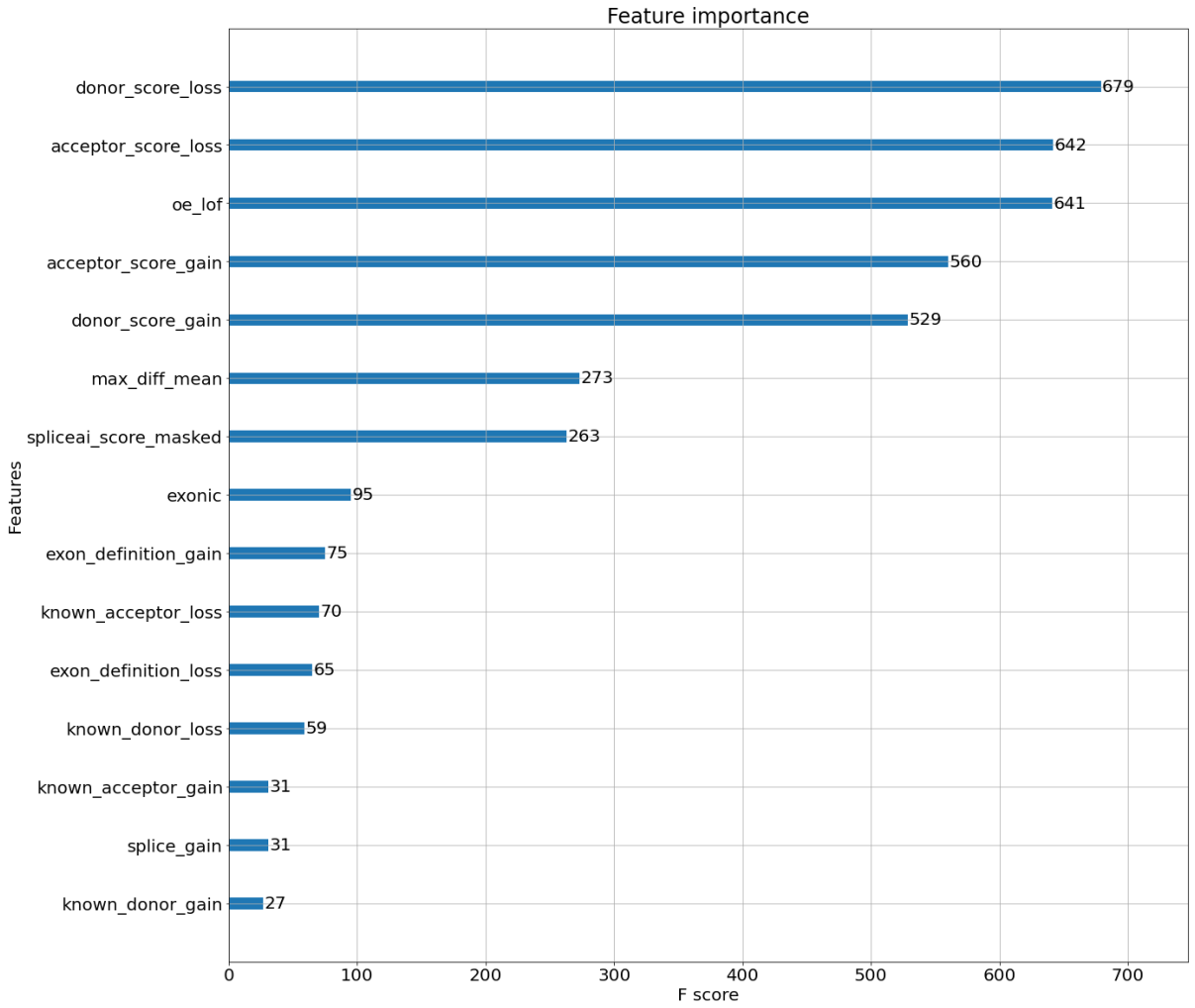


Figure 3.5: Extracted model 'feature importance' scores from trained XGboost model. The learning model appears to care about the kind of splice site disruption which occurs with accent placed on whether it is donor/acceptor gain/loss. Of interest the gene-tolerance scores (oe_lof) which appears high up in feature importance.

3.5 Results and Reapplication of pipeline to HGMD dataset

We see from figure 3.4 that this approach can build a combined model which achieves substantially better precision and recall than using *SpliceAI* or our neural network model alone. Part of this improvement comes from the addition of gene tolerance scores as seen by its feature importance in figure 3.5.

When we now apply the final model to known HMGD splice variants, we find that it correctly identifies 13547 out of 15047 (90.0%) as splice-altering, at a threshold of 0.5, compared to 285 of our control set out of 110249 (0.26%) control SNPs. We choose our threshold of 0.5 to minimise the number of control SNPs marked as splice-altering.

However, we can also apply the model to HGMD SNVs which are not annotated as splice-altering. Under the same criteria 4108 out of 97987 (4.5%) are recovered, which further underlines how splicing is a key, but potentially missed mechanism in disease databases.

Of the annotated splice-altering HGMD sites, 19.4% (2631 out of 13547) are intronic, whilst 81.5% (3322 out of 4108) of the recovered sites, which are not annotated as splice altering in HGMD, are exonic. Of the already annotated exonic HGMD splice altering SNVs they have a median distance to a known splice site of 1. This means they are typically affecting the first base pair within the exon. Conversely, amongst the exonic HGMD SNVs not already identified as splice-altering they have a median distance of 15 base pairs.

In numbers, 632 of the annotated splice-altering exonic HGMD variants are 5 or more base pairs from a known splice-site whilst 2073 of the recovered unannotated HGMD variants are. These (relatively) deep-exonic variants can either be predicted to gain a cryptic splice-site, leading to an alternative splice site, or lead to the loss of the exon. In the latter category this could arise from disrupting regulatory sequences such as exonic splicing enhancers.

Of the annotated 632 splice-altering deep-exonic HGMD variants, 470 of them are predicted to create a cryptic splice-site compared to 1329 out of the 2073 recovered unannotated sites. Of the remaining sites, we predict that only 81 (of the 162) annotated splice-altering deep-exonic HGMD sites lead to the loss of the exon. This compares to 674 of the remaining 744 deep-exonic unannotated sites which our model predicts to distally lead to the loss of the exon.

This suggests that HGMDmisses the splice-altering annotation when splice-altering exonic-variants are not near known splice-sites. In particular, this is most pronounced when the likely mechanism is that the variant is affecting splicing regulatory sequence which is not readily identified by donor or acceptor motifs.

We do find similar quantities of recovered deep-intronic HGMD variants. However, we cannot rule out selection bias in the HGMD itself. This is also a criticism of our approach of reapplying machine learning to classify these variants. As HGMD lacks deep-intronic splice-altering variants (311 out 15047 are 100 base pairs from a known exon), this means that the classifier has limited to no information to train upon.

In addition, as with the later GWAS chapter, we do not rule out that these exonic HGMD variants, identified by our model, as not having additional mechanisms. For this class of variants our model predicts an average reduction in splicing of 42% compared to a predicted 82% predicted reduction when at exon-intron boundaries. This suggests that the model believes that the exons would still be alternatively spliced. This is consistent with the observation that many are likely to alter splice regulatory sequence instead of splice-motifs.

3.6 Annotations and their consistency

In our collaborations with clinical genetics colleagues it has been frequently asked what is the predicted consequence to the full transcript. Our prediction scores focus on the effect at one splice-site, but typically a splice-change alters 2 or more splice sites. For instance, the skipping of an exon involves the loss of an acceptor and donor. Another example is a simple alternative donor site, leading to a longer exon, and involves both the loss and the gain of a donor site. This is similar to the work found in section 2.6 done in a more artificial setting.

This motivates us to ask the following question of the model:

1. Contingent on the disruption of a (primary) acceptor/donor site, are secondary effects predicted, and what are the most common effects?
2. How consistent are the secondary effects due to an acceptor/donor site being disrupted?

In figure 3.7 we see that the model produces easily interpretable transcript predictions. The predicted loss of an acceptor site is predicted to result in an alternative acceptor site or the loss of a donor site. There are only a few instances where the model is unsure about which of the two scenarios it prefers, or where a mixture of the two is predicted. This means a clinician or geneticist using the model would obtain a clearly interpretable transcript prediction to experimentally follow up.

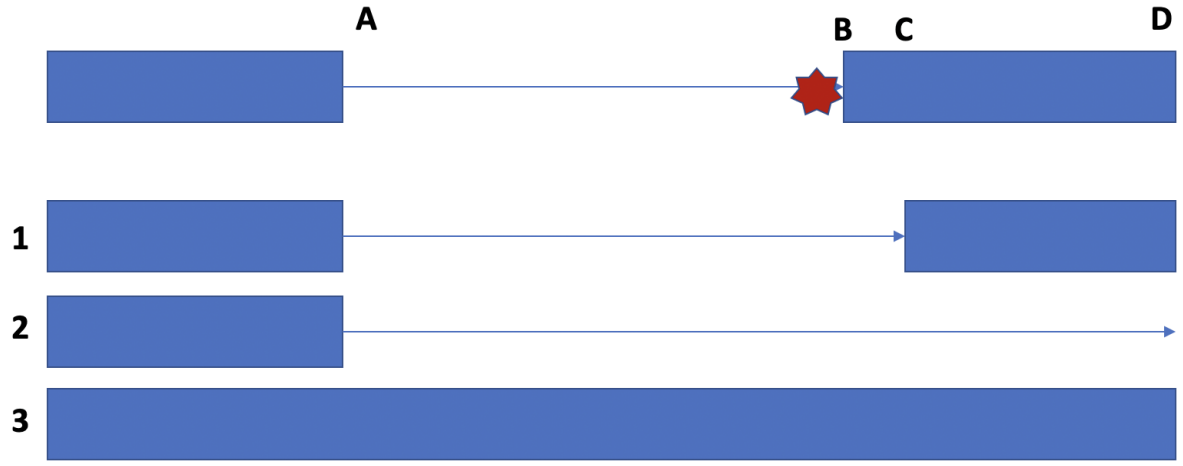


Figure 3.6: When the model predicts loss of acceptor site at B it additionally reports secondary effects which lead to a fuller transcript interpretation. Scenario 1 occurs from a predicted gain of acceptor at C. Scenarios 2 and 3 occur from a predicted loss of donor site at A or D.

In figure 3.8 the transcript level interpretations are less clear cut. This means that when a donor site is predicted to be lost the overall transcript interpretation is less evident. A large proportion are not predicted to also cause an acceptor loss, but they do not predict a strong donor gain either. What explains this discrepancy?

By analysing different model runs we find that there is less consistency for the gained donor sites. The predicted gained donor sites are only 70% consistent between different model runs compared to over 94% consistency for acceptor sites. This remains true even when we adjust for predicted splice-differences. Precisely, even when adjusted for only strong splice site gains we find that the gained acceptor site location is more consistent than the donor site location. This suggests that the model truly believes, on average, that a selection of alternative donors are viably gained, whilst typically only one alternative acceptor is viably

gained. In this context this would mean that a clinician or geneticist using the model would lack a clear single predicted alternative transcript.

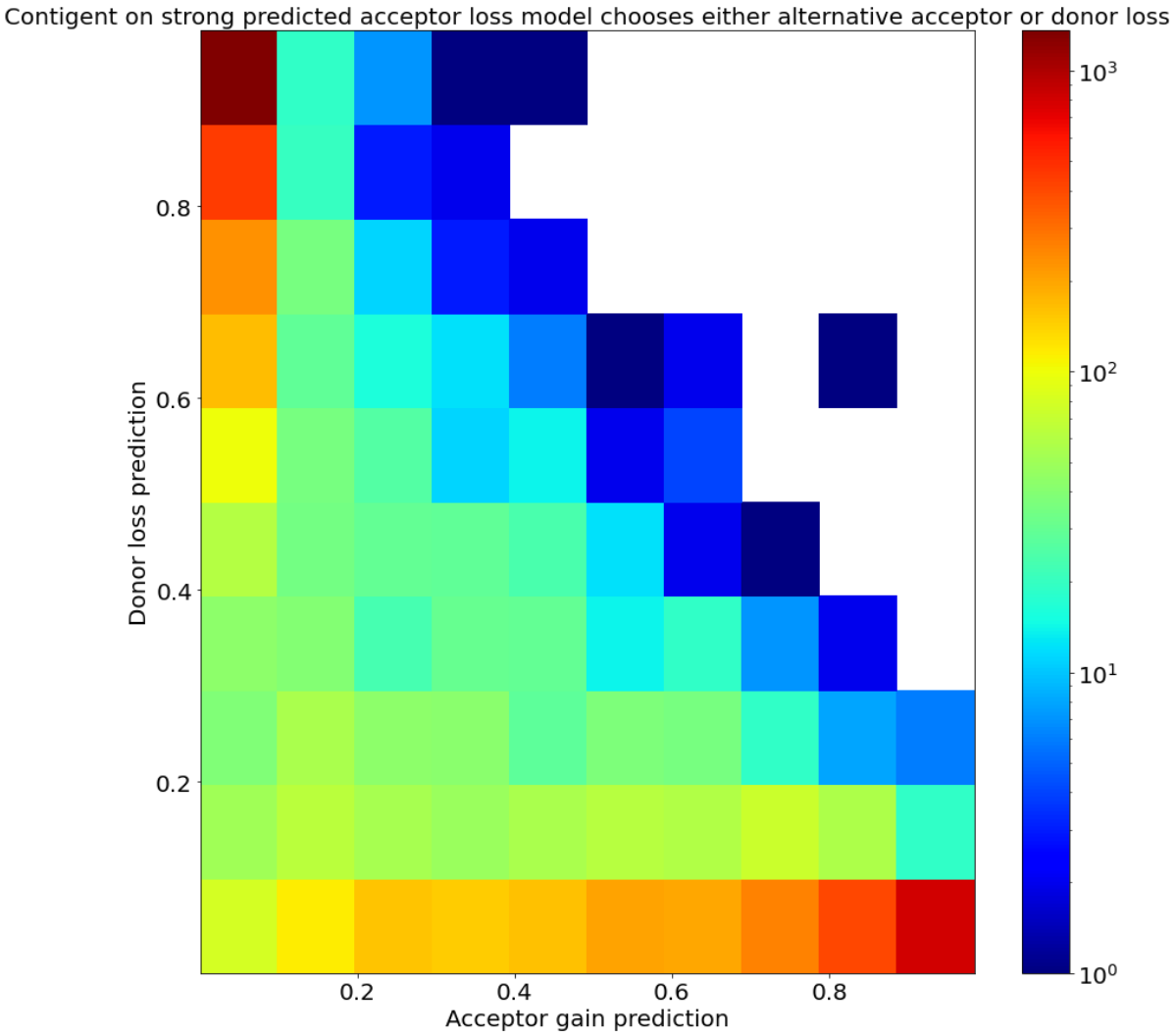


Figure 3.7: For the class of strongly predicted acceptor losses measure whether the model predicts an exon-loss or intron retention (measured by maximum donor loss) or an alternative acceptor (measured by maximum acceptor gain). For this class it is notable that the model splits between the two categories in almost a binary manner. The model either predicts a strong acceptor gain or a strong donor loss but rarely a mixture of the two.

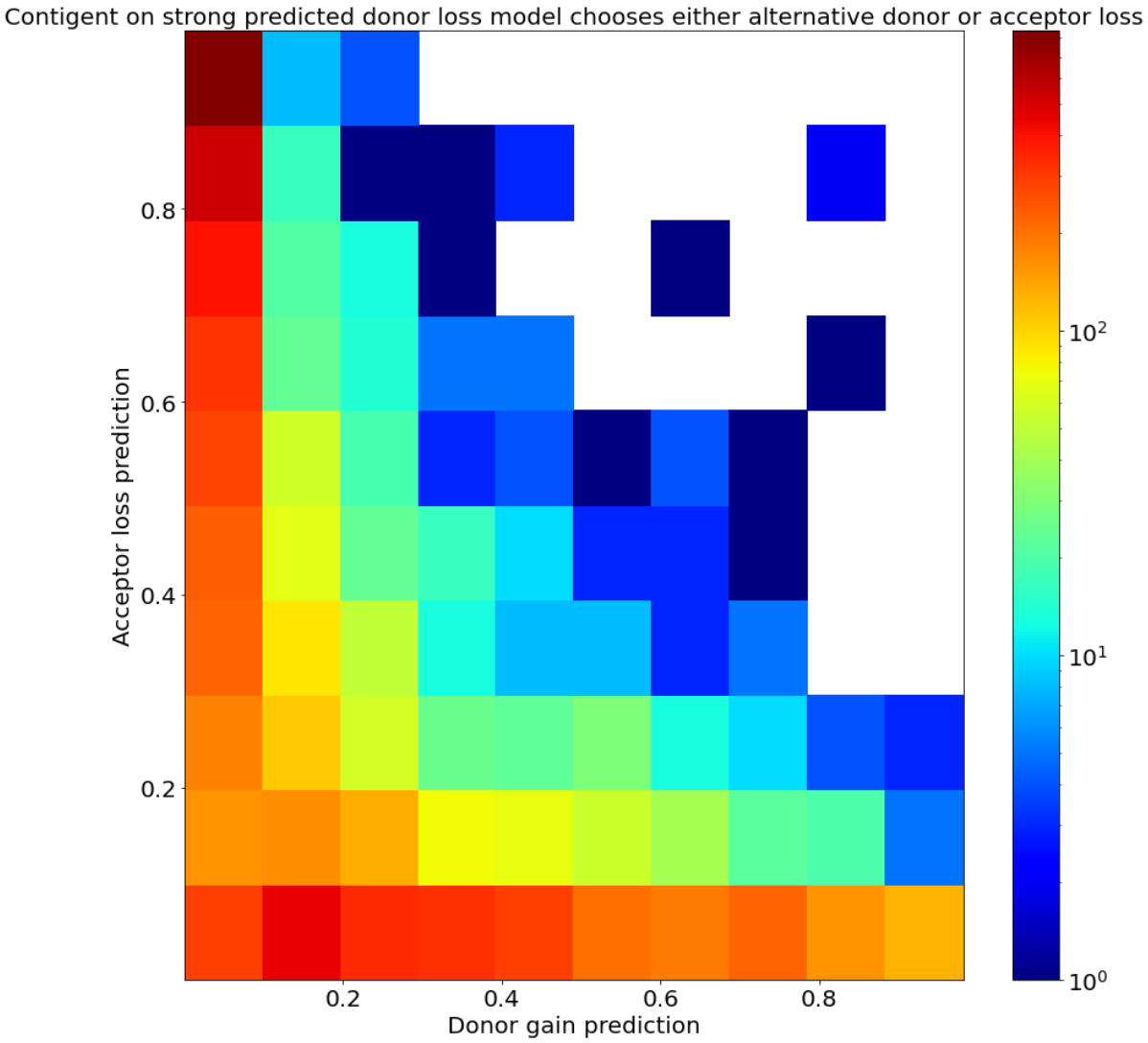


Figure 3.8: For the class of strongly predicted donor losses measure whether the model predicts an exon-loss or intron retention (measured by additional acceptor loss prediction) or an alternative donor (measured by donor gain prediction). For this class it is notable that the predicted donor gains (measured on an individual donor site) are typically weaker than we saw in figure 3.7.

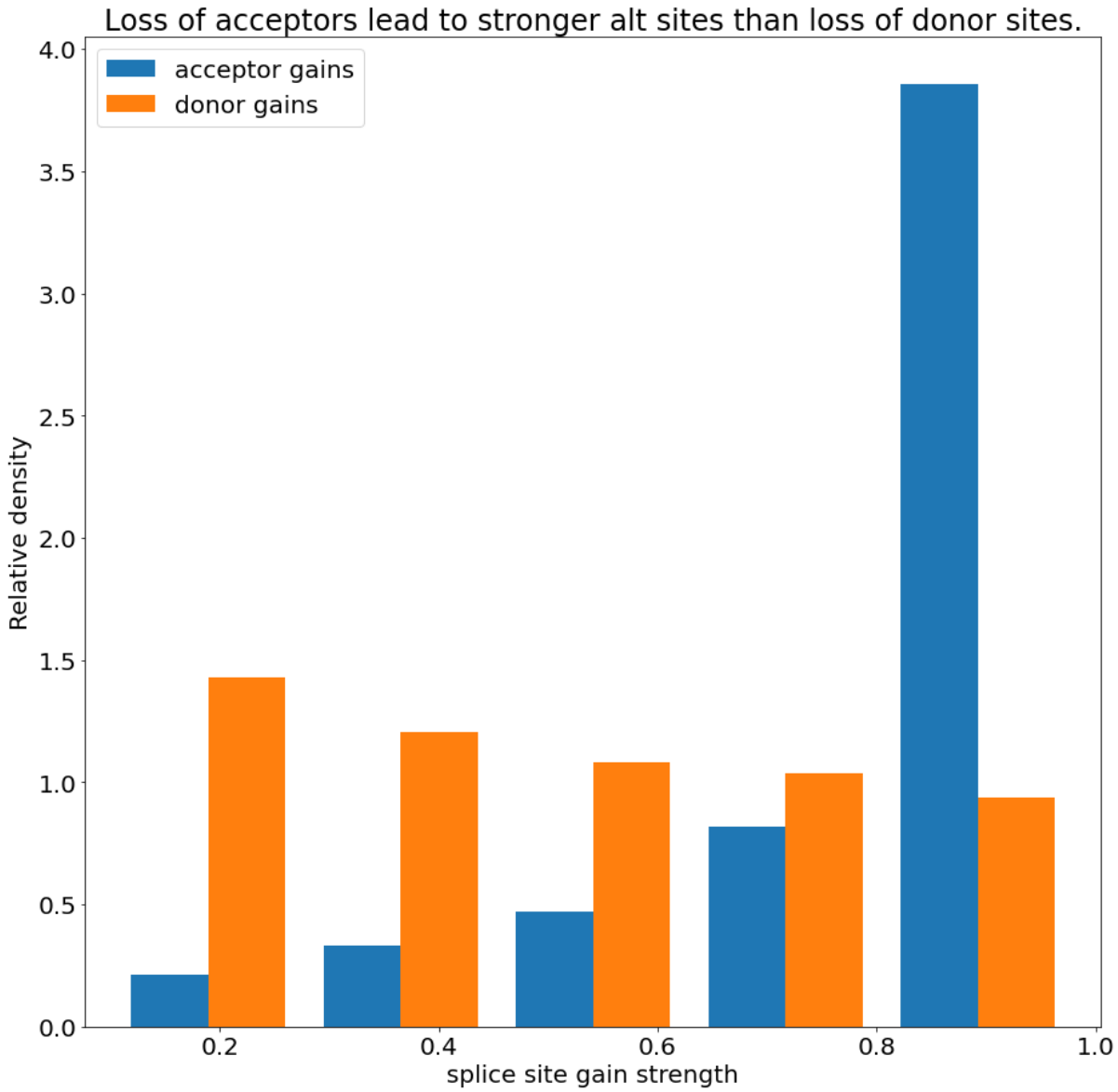


Figure 3.9: In this figure we compare predicted alternative splice site events defined by the model not predicting exon-gain/loss or intron retention. We compare the model predictions when the variant is predicted to be acting on acceptor sites or donor sites. We observe for predicted gain of acceptor sites a more skewed score distribution than for gained donor sites. For gained acceptor sites the model has a clear preference whilst for gained donors it typically 'spreads the probability' without a clear preference.

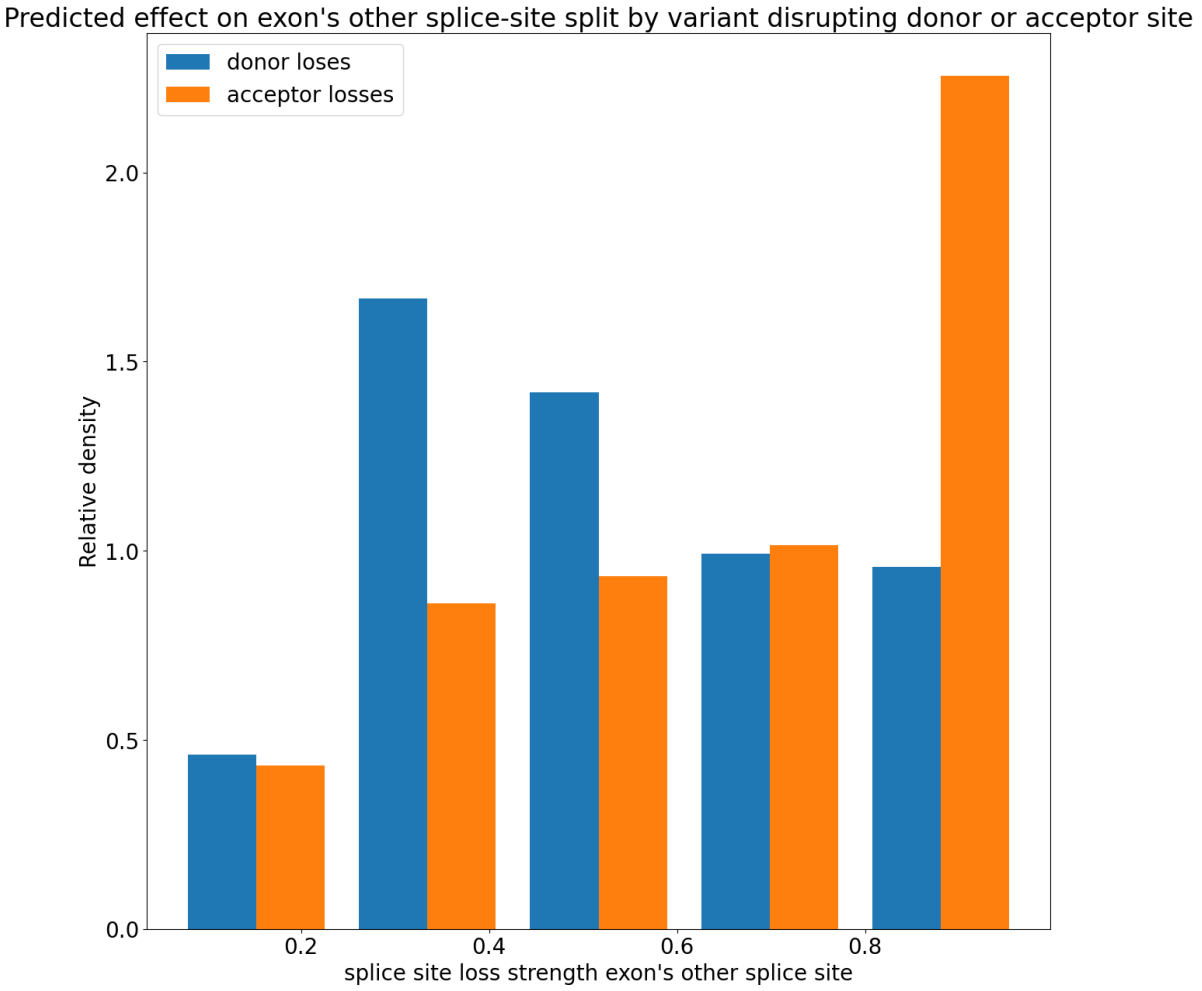


Figure 3.10: For predicted exon losses we split based upon whether the mutation is proximal to the donor site or acceptor site. When the acceptor site is predicted to be disrupted we typically observe a large downstream effect on the subsequent donor site. For predicted donor losses the model makes less strong predicted effects on the upstream acceptor site.

In conclusion of this brief section we obtain the following for SNVs:

1. If an acceptor site is disrupted, then either a clear alternative acceptor is gained, or a donor site is lost.

2. If a donor site is disrupted, then there is less clear cut interpretation whether an alternative donor is gained, a mixture of many donor sites gained, or an acceptor site is lost.

The second question is what factors influence the model prediction of an exon loss or an alternative splice site. SNVs which cause the loss of a donor/acceptor site can either lead to the loss of the corresponding exon, intron or to the creation of an alternative donor/acceptor site (or many alternative donor sites). The question we are left with is why the loss of a splice site sometimes leads to one scenario over another.

One metric to consider is the exon-length which we evaluate in the context of exon-skipping vs alternative splice site selection. When exon lengths are short the loss of either splice site tends to lead to predicted exon-skipping (figure 3.11 and 3.12). When exon lengths are longer the loss of the splice site tends to lead to an alternative site selection. It is striking how rare loss of exons are when acceptor sites are disrupted in long exons (> 150 base pairs).

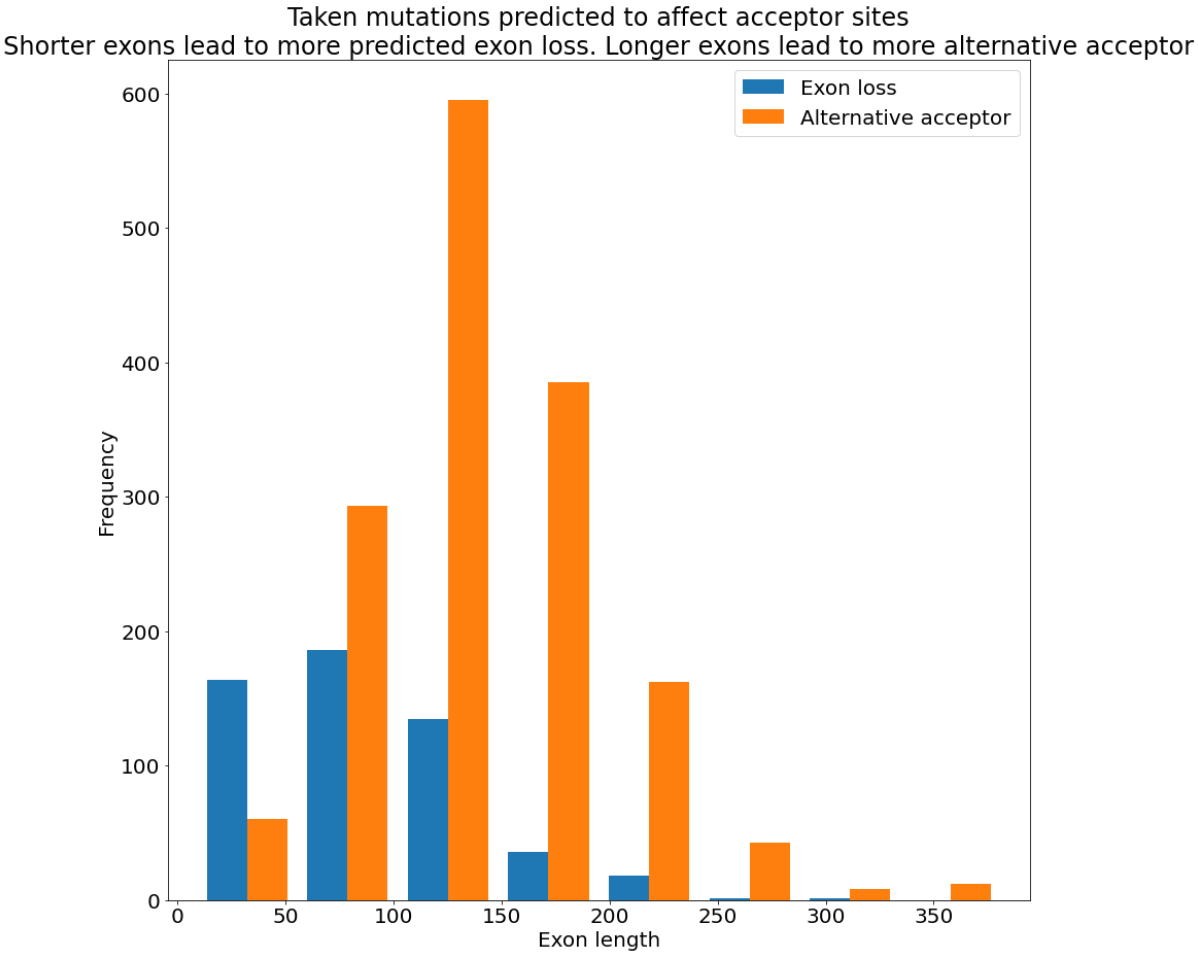


Figure 3.11: Taken over mutations predicted to lead to the loss of an acceptor site. We split the predictions by the length of the exon and by whether they are predicted to lead to an exon loss or an alternative acceptor site. Amongst the curated HGMD set we observe that exon-loss is more common when the exons are shorter in length. Exon loss is rare when the exons are longer.

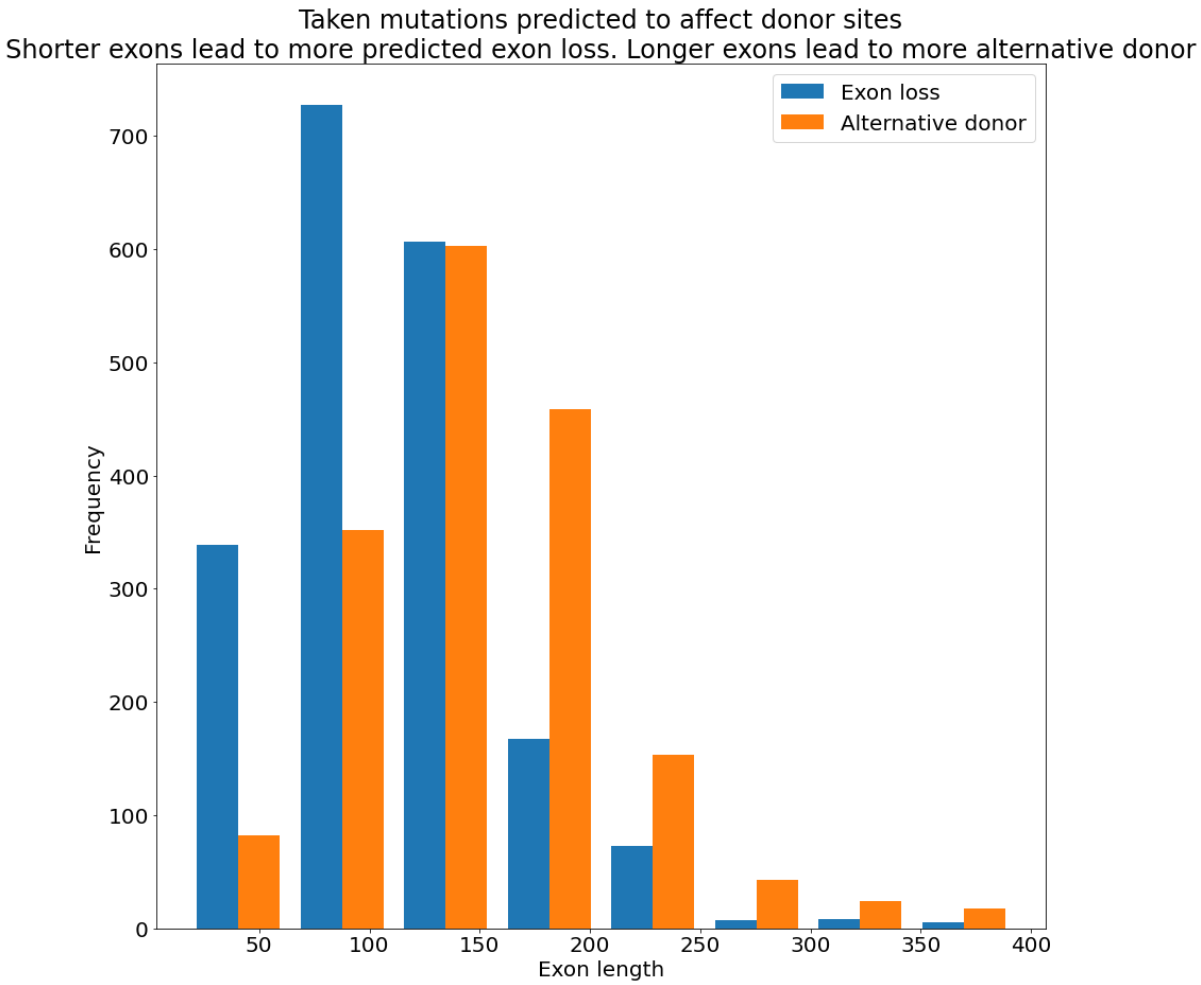


Figure 3.12: Taken over mutations predicted to lead to the loss of a donor site. We split the predictions by the length of the exon and by whether they are predicted to lead to an exon loss or an alternative donor site.

In conclusion, within the HGMD dataset we can utilise the model to go beyond the simple prediction of whether there is splicing disruption or not. The model makes additional predictions concerning this portion of the *gene-definition*. Typically, a mutation affecting an acceptor site will lead to exon-loss when the exon is short, and it will lead to an alternative acceptor gain when the exon is longer. A mutation affecting a donor site will lead to clear exon loss when the exon is short but to an unclear spread of alternative donor sites when the exon is longer.

3.7 Collaboration with disease diagnostics platforms

We formed a collaboration (lead by professor Jenny Taylor) which applied our model along with the next chapter's NMD model to examples found within Genomics England's datasets. Our part of the collaboration was to reanalyse selected SNVs, which due to family structure, were deemed of potential clinical interest.

The approach in this collaboration is to use three different methods to assess these novel mutations, of which our algorithm is one. Each SNV is assessed by our algorithm using methods outlined in this chapter. In addition, we add the nonsense mediated decay model which forms the basis of the next chapter. This is used for the annotation of the likely molecular outcome in the collaboration as it helped with subsequent validation steps. For instance, knowing that the predicted alternative transcript is likely to undergo nonsense mediated decay could change the expected results of RT-PCR analysis from mRNA.

This collaboration formed a driving motivation for our work. A key reason to focus on RNA-decay, in chapter 4, is due to the issues of validation. If a model predicts an alternative transcript which is not replicated by RT-PCR (or other protocols), should we negate the prediction? It is very exciting that, as this thesis is being written, we are submitting a preprint which collates our work in regards to this collaboration.

In this subchapter we will detail some of the interesting cases which have been described as splice-altering by our diagnostics pipeline.

The summary statistics from our pipeline is that 17 variants were identified as splice altering which was found within 16 of the 122 analysed cases. Of these variants, 15 had completed downstream validation, and of those 9 have been classified as pathogenic. Given that in HGMD we found relatively few deep-intronic labelled splice-altering variants, it is remarkable that 5 of these validated variants are found within deep-intronic regions.

Amongst the variants we have an interesting example of a potential missense variant (chr10:75222273:A->G) in *VDAC2* (figure 3.12). This was found for a patient with congenital thymic stromal defects and limb malformations. It was considered a strong candidate due to being a *de novo* mutation in the subject and prior knowledge surrounding the VDAC2 gene. The model predicts that the SNV would result in less usage of the canonical acceptor site by the creation of an alternative acceptor site within the exon, shortening it by three nucleotides. Sanger sequencing of the RT-PCR product confirmed this 3 base pair deletion. However, our model additionally predicted that there would be some residual usage of the canonical acceptor site which is also observed in the RT-PCR product.

A gene which was of interest for 4 patients with congenital erythrocytosis was *VHL*, via its known links to this condition. For these patients, two variants were predicted to lead to a gain of splicing. The first variant chr3:10143003:A->C was predicted to increase splicing at position chr3:10143021 which was confirmed by a minigene assay (data not shown). The second variant (found amongst 3 patients with this condition) chr3:10142957:T->C is predicted to increase splicing of an exon chr3:10142762-10143021 which again is confirmed by a minigene assay.

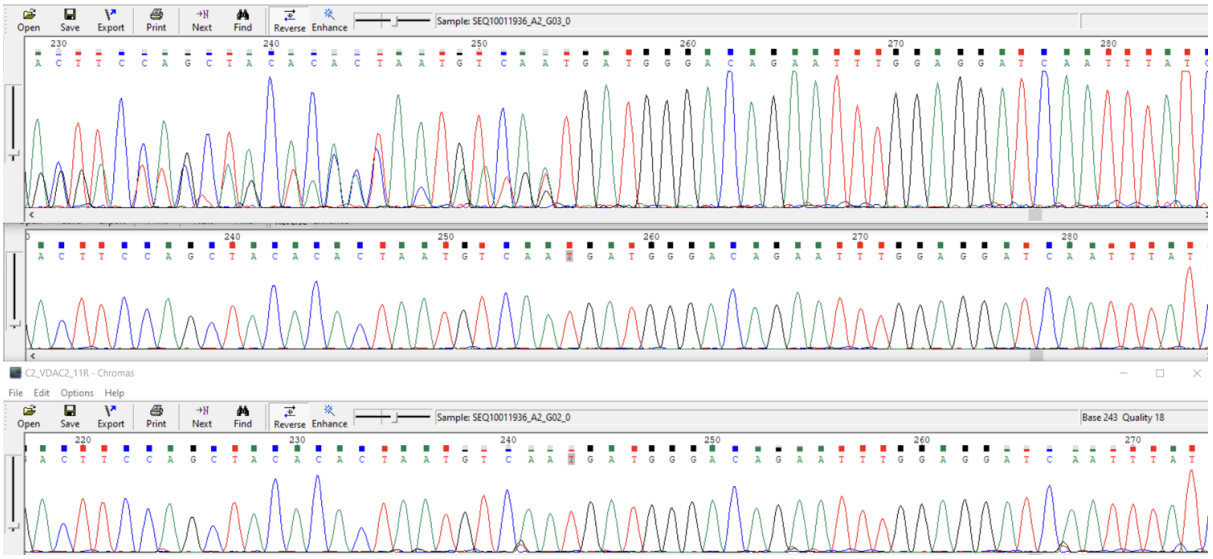


Figure 3.12: We observe sanger sequencing for the proband and two parents respectively for a mutation chr10:75222273:A->G in VDAC2. If no splicing change happened we would only observe a change A->G at position 253 in the diagram. The model predicted that the mutation leads to the alternative splicing of a new acceptor site 3 base pairs into the exon. It did not predict the entire removal of the original splice site. This is exactly what the Sanger sequencing reports with position 253 showing a mixture of 3 nucleotides A, T and G. The proceeding positions as further into the exon from 230-252 we observe 1 or 2 nucleotides consistent with a shift of 3 nucleotides.

Another illustrative example we found within our collaboration is a *de novo* variant found within a deep-intronic region. chr15:75885717:G->A in *UBE2Q2* is predicted to lead to the loss of a deep-intronic pseudo-exon. The hypothesis currently under investigation is that this would lead to increased expression of the canonical transcripts due loss of NMD.

Outside these variants, which formed the basis of our pending publication with the Taylor group, we have additionally been collaborating with them on other current unsolved candidate variants found with GEL. Many of the examples are near to known splice-sites, but of particular interest to us are three predicted gained pseudo-exons found within genes TSC1, TSC2 and MSH2. The first has been confirmed within GEL, the second is picked up in Mayer et al (2000) [33] and the third in Fulk at al [34]. A higher percentage of variants from both collaborations are deep-intronic than the ~300/15000 annotated splice-altering variants in HGMD. As these collaborations progress, it will be of great interest if this evidence for enrichment in deep-intronic regions continues to expand.

3.8 Conclusion

We found two simple approaches that lead to substantial improvements in accuracy when computing damage scores for putative SNVs. Combining separately trained models increased our model's accuracy over *SpliceAI* when comparing single method approaches. More substantial improvements came from the further addition of a XGboost based machine learning based approach to strengthen the classification of disease causing SNVs. This suggests that future better approaches to disease classification could be to combine an even wider variety of multiply trained models, along with genome annotations via aggregation of their outputs in a final XGboost classifier. We have achieved this by combining *SpliceAI* and our model, but for larger groups working on classifying splice variants there would be value in combining other approaches such as *MMSplice*. The general conclusion suggests that moving away from single model approaches via retrained models and distinct models to combining models through other trained networks could provide key benefits.

A limitation with the HGMD dataset is that it lacks deep-intronic SNVs. It still remains unclear what the general phenotypic consequences of gaining NMD target pseudo-exons are, as they are limited in the HGMD dataset. However, we have anecdotal examples from our clinical collaborations of gains of pseudo-exons which are being validated in terms of the mechanism and are likely to be disease causing. What we miss is the evidence for the prevalence of this mechanism amongst disease causing variation, which will be dependent on the generation of new appropriate datasets.

In addition, for genome-wide scans of variants we found that training an additional XGBoost machine learning model provided benefits in classifying disease causing variants. More examples of deep-intronic variants added to this set will benefit the classifier. It is possible that the gain of pseudo-exons have a subtly different character to loss of known exons. Adding more examples to a tree-based classifier, like XGBoost, would improve its performance.

In our clinical work we found that many of the deep-intronic and alternative splice predictions had nonsense mediated decay signatures. With this in mind, we became motivated to explore the interaction of alternative splicing and nonsense mediated decay. In relation to chapter 2, we have questions about its prevalence and effects on our model's accuracy, but in addition we have questions about its disease significance, which we will try to address in the subsequent chapters.

4 Finding Nonsense amongst our Predictions

"If I had a world of my own, everything would be nonsense. Nothing would be what it is, because everything would be what it isn't. And contrary wise, what is, it wouldn't be. And what it wouldn't be, it would. You see?" – Lewis Carroll

We saw in the first chapter that, despite the model achieving high sensitivity, there remains a relatively high level of apparent false positives. However, these are classed as false positives based on *observed* diversity of mRNAs in datasets such as GTEx. Therefore, the validity of these results really depends on how well these datasets represent the complete spliced transcriptome. However, some model predictions will be missing from these datasets, as some transcripts that the model predicts may be subject to RNA-decay [8], and such transcripts may be entirely missing from these standard RNA-seq datasets. Potentially, and more significantly, the model has no knowledge of transcript information, and so it has no means of distinguishing between splice sites which would lead to RNA-decay and those which would not.

In this chapter we refine the model by adding a component that predicts the effects of *Nonsense Mediated Decay* (NMD). NMD is the most prevalent form of RNA-decay, and acts by removing transcripts that contain a premature stop codon. Early estimates suggest that around 40% of alternative splicing events give rise to premature stop codons and induce NMD [1]. Its main function is thought to be the removal of faulty mRNA transcripts (due to, for instance, noisy splicing events), but this continues to be a debated topic [6]. In particular, it is contested how much NMD acts as a mechanism for gene regulation, as opposed to it being merely as a mechanism of reducing splicing noise [9]. In the last chapter, which applies

the model to analysing common diseases, we try to address this question from a different angle, by focusing on the potential disease impact of splice-alterations to AS-NMD sites.

If the transcripts undergoing NMD were the product of noisy splicing, they would serve no biological function, and so they would be either neutral or under negative selection and therefore not conserved. However, they would remain a credible source of human genetics, in particular for common traits which are typically under weak purifying selection. The outcome of the inclusion of a common variant that induces NMD would, in essence, be the same as other common trait mechanisms such as those that affect regulatory elements. In both cases, the effect of the variant is to change the productive output of the gene [6]. However, in the case of a variant that affects a gene regulatory element, the output of the gene promoter is changed, whereas in the case of an NMD inducing variant, the output of the promoter remains the same but the amount of transcript that contributes to the mRNA pool is changed. The end result is, therefore, the same and so it is important for the model to identify these as true positives, even if they are not evident in stable mRNA.
However, if NMD is in some instances an evolved mechanism to regulate the output of a gene, then such transcripts would be stable across individuals and species and so conserved [35]. Therefore, in this case the model should also predict them as true splicing events, as well as model the effect of variation upon them, even though they may not be evident in stable mRNA.

At present, the model is trained to only predict splicing *per se*, so we would expect the model to predict AS-NMD sites in a similar manner to non-NMD splicing events that are validated from mRNA; some of these predictions could represent real, but poorly validated

splicing. In this chapter we will add gene definition data to the model to predict splicing events that lead to NMD; thus we will distinguish these events as another class of splicing events and differentiate them from true false positives in the model.

We will show from our model predictions that clear patterns of bias indicate a significant proportion of our predictions (missing in UCSCv37 [36]) which are transcribed but unstable splice sites. We will further propose and apply an additional bioinformatic model which will show that these sites can be predicted and segregated bioinformatically. We will later apply the model to three different types of transcriptomic datasets, designed to reveal NMD decayed transcripts, to confirm whether our NMD-model is validated experimentally and, by proxy, to determine the most effective experimental approaches to detect such transient transcripts in the transcriptome.

4.1 Motivation

The motivation for developing a model for NMD is to help to more solidly build a hypothesis about disease aetiology [37] and genetics. The current models tell a clinician or a geneticist that we expect to see a collection of splice sites **A** replaced (or partially replaced) by a collection **B**. Knowing which splice sites found in **A** or **B** are likely NMD-targets can radically change what one would expect to observe in a typical RT-PCR or RNA-seq follow-up experiment. In the extreme scenario experimental evidence could incorrectly lead to the conclusion, in heterozygotes, that no splicing change had occurred if NMD were 100% efficient on the transcripts found in **B.**

However, the actual outcome of this, admittedly, extreme scenario, would be the complete loss of protein output from the affected allele.

Therefore, layering on top of the initial neural network a subsequent prediction of whether or not the transcript will be subject to NMD, will greatly help guide the appropriate experimental design: in particular, which pool of RNA, nascent or mature, should be used in the analysis. For example, in our clinical collaborations we now supply predictions, not only of which mutations lead to alternative splicing, but also which of the resultant transcripts are likely to be decayed or stable. Additionally, our reanalysis of publicly available datasets from experimental approaches to detect NMD decay can also suggest the most appropriate experimental protocols for validation experiments.

The NMD model will also help us to build better training data for the splice prediction model. Since NMD transcripts are missing from the GTEx data set, the associated true splice sites are not included in the training set, which runs the risk of confusing the model, and of making it less certain about the remaining predictions. In using the NMD model to illustrate the prevalence of such NMD-transcripts, we hope to motivate the construction of better training and validation datasets to train future splicing models.

4.2 Summary of mechanisms linked with NMD

It is useful to consider how sequences may affect gene definition. These included how losing an acceptor site could lead to acceptor site gains and subsequent exon or intron loss.

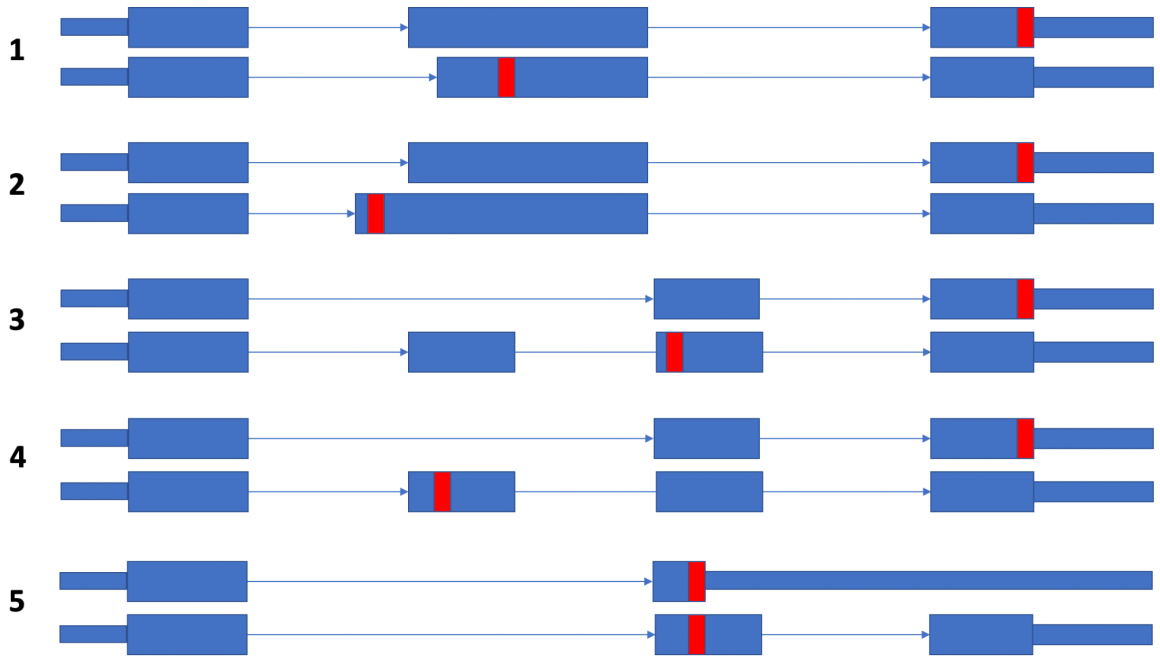


Figure 4.1: Here are listed alternative splicing scenarios which can modify a stable transcript into a NMD target. 1. Frame-shifting alternative splice site giving rise to in-frame stop-codon. 2. Direct inclusion of in-frame stop-codon from alternative splice site. 3. Additional inclusion of exon giving rise to inframe stop-codon. 4. Direct inclusion of in-frame stop-codon in the novel exon. 5. Alternative inclusion of intron leading to canonical stop-codon becoming a premature stop-codon.

In figure 4.1 we list the main scenarios in which an alternative transcript, through sequence variation,may create NMD-targets [38]. In each of the listed examples one would expect to see the canonical transcript more than the alternative transcript, if NMD is active, and this would lead to falsely underestimating the extent of a splice-alteration.

An interesting but unlisted scenario is the no-stop scenario [39]. This is where a frame-shift leads to no stop-codon being detected before the polyadenylation tail. This scenario we have not listed because, at present, our pipeline is not able to deal with this scenario, so we

cannot evaluate the prevalence of this mechanism but this must be noted as a current caveat to our approach.

4.3 Evidence amongst unannotated predicted alternative splice sites

In this section we look at novel splice sites - sites that are predicted by our model, but that are not annotated in existing transcript databases. We use UCSCv37 to classify our predicted sites as known or novel [36].

These novel splice sites fall into three categories:

1. Stably transcribed - splice sites missing in UCSCv37 but evidence for transcription is found in GTEx.
2. Transcribed but unstable - splicing occurs, but the resulting RNA transcript decays rapidly; no evidence in UCSCv37 or GTEx.
3. Not transcribed - a false positive of the splicing model.

At a model threshold of 0.3: novel predictions in category 1 can be directly identified. We separate these further into “rarely seen splice sites”, defined as having an empirical PSI < 1% in GTEx, and “commonly seen splice sites”, defined as those with an empirical PSI >= 5%. Of all novel splice sites, 19.3% (12772/66029) fall in the first category, with 7.7% (5084/66029) a high PSI >= 5% in GTEx.

It is more challenging to assign the remaining 80.6% (53257/66029) into one of the two remaining categories - splicing occurred but the transcript was degraded by NMD before it

was observed, or no splicing occurred. Here we aim to use the signature of NMD to estimate the fraction of true but unobserved splice events among our predictions.

Creating an acceptor or donor at a random location in a transcript will almost certainly induce a PSC, and by that reasoning NMD may be expected in the great majority of both true and false predictions. In this analysis we rely on a simpler but approximate signature of NMD, namely the tendency that many NMD transcripts include a frameshift compared to the functional transcript.

Since the model has no access to the reading frame, we expect that false positive predictions (category 3) cause frameshifts (and therefore very likely PSCs) according to chance, i.e. in about two-thirds of cases. However, true predictions that resulted in a splicing event but which were not observed because of RNA decay (category 2) are expected to follow to a large extent the NMD pathway, and of those that do, many will include PSCs because of induced frameshifts. This means that the degree to which in-frame alternative splicing events are depleted from the novel predictions is related to the fraction of true positive predictions. Specifically, if we conservatively assume that all true but unobserved splice events lead to complete NMD, as a result of splicing leading to a frameshift resulting in a PSC, then for all of the category 2 splice events we would observe a frame shift. The depletion in predictions of in-frame transcripts among category 2 and 3 splice predictions is therefore a conservative estimate of the relative proportion of true (category 2) predictions.

We can further break down our set of apparent false positive splice predictions into more manageable categories, which can be defined bioinformatically:

1. Alternative acceptors

2. Alternative donors
3. Predicted inclusion of exon in annotated intronic region

These categories are not fully comprehensive; for example, the possibility of mutually exclusive exons as well as predicted sites that could be considered as belonging to more than one of the above categories are not covered in this analysis. Nevertheless, the above three categories comprise the majority of novel predictions without read evidence and forms the basis of the analysis in this section.

4.3.1 Alternative acceptors

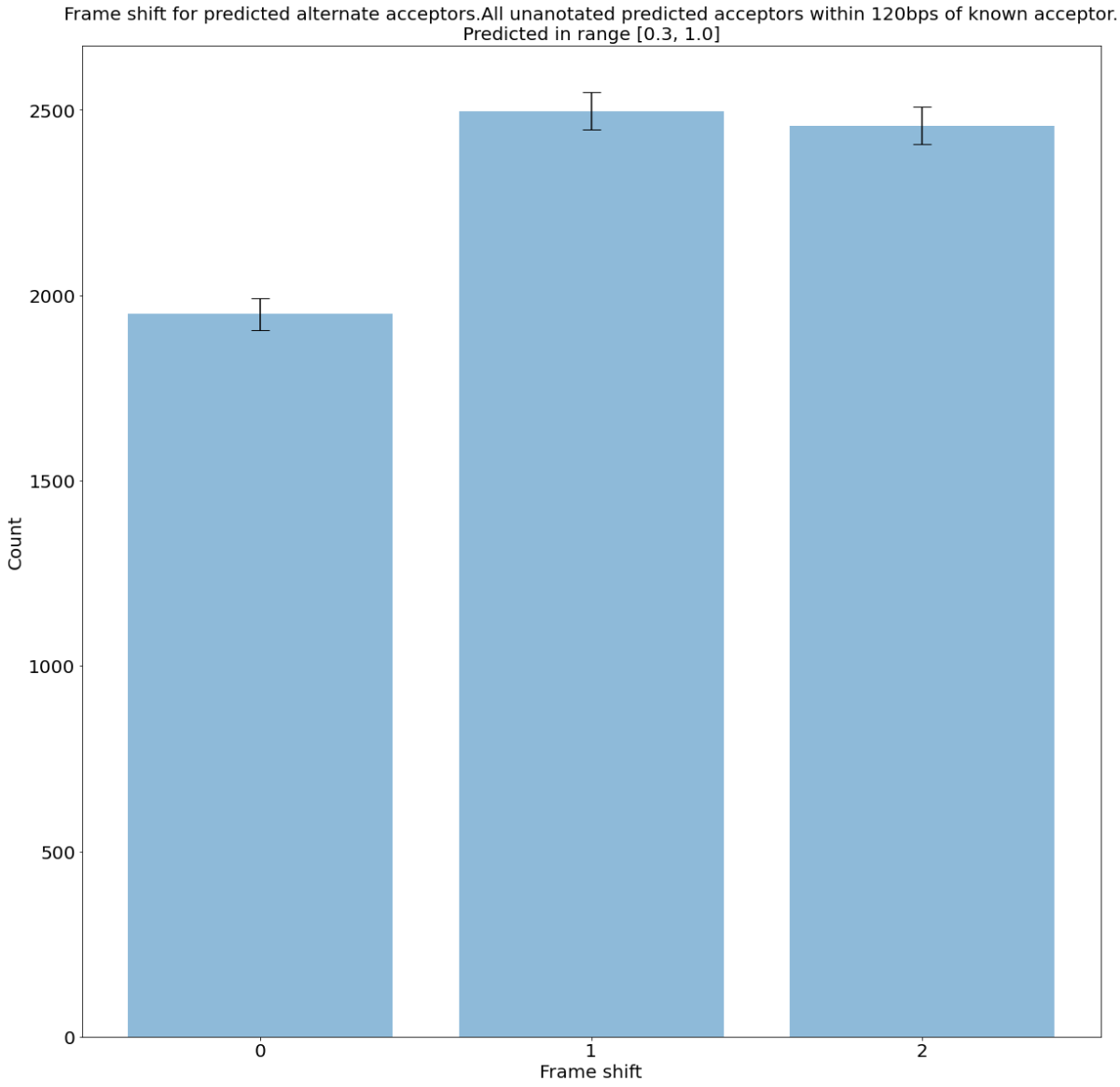


Figure 4.2: Novel predictions found with PSI <= 0.01 in GTEx. Consistent counts for frameshifting alternative acceptors but we find significantly fewer novel predictions which preserve the frame. These counts suggest that selection pressure is reducing frame-preserving alternative acceptors by 19.95% (95% confidence interval [15.18%, 24.45%])

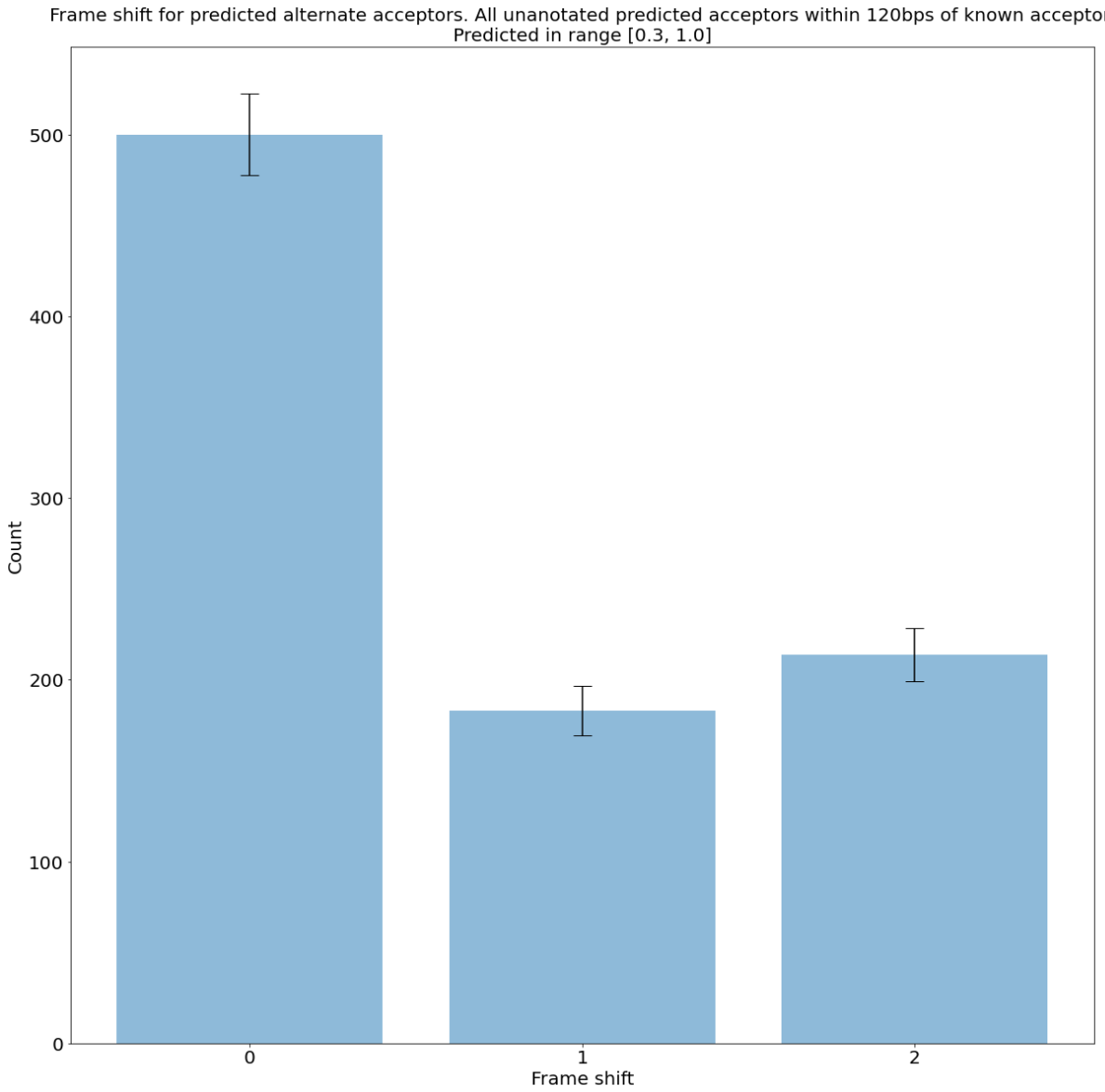


Figure 4.3: Novel alternative acceptor predictions (unannotated in UCSC V37) with an empirical PSI >= 5% in GTEx show converse bias. Strong bias is observed for alternative acceptors to preserve the frame. This suggests the novel predictions the model is picking up are real with the caveat that they occur at lower frequency than predicted.

Alternative splice sites are the easiest to analyse because of the presence of a nearby exon, which helps to define the reading frame due to it being able to easily utilise an existing exon to determine the frame. Of the 66029 predictions in categories 2 & 3, 38.2% (25211) are

acceptor sites. Of these acceptor sites, 11045 are within 120 base pairs of a known exon. Our analysis is restricted to 6915 of these 11045 sites where the existing transcript annotations mean that the frame information is unambiguous.

Figure 4.2 shows the result of analysing alternative acceptors predicted to have a PSI in range [0.3, 1.0] but to have empirical PSI < 0.01. Of these predicted novel but weak splice sites, more induce a frame shift than one would expect by chance alone. This can be explained by supposing that, as NMD is less able to remove transcripts that do not induce frame shifts, mutations leading to such transcripts can result in expression that can reduce fitness, thus bringing evolutionary pressure to remove these mutations from the population. We find no bias in the frequency frameshifting novel predictions (labelled as frameshift 1 or 2 in figure 4.2). We can compute an estimate for the relative depletion by using the counts of each frameshifting category as the expected count for the non-frameshifting category. From these we would expect 2477 sites with a 0 frameshift but we find 1949, which represents a 21.3% depletion. This therefore suggests that a large proportion of our frameshifting novel predictions are not false positives, or else we would not expect such a depletion.

To further validate the analysis above with the reassurance that there is not a general bias against frame-preserving alternative acceptors, we performed the same analysis with unannotated alternative acceptors with GTEx empirical PSIs >= 0.05. As shown in figure 4.3, we recover a reversed bias towards frame-preserving alternative acceptor sites. This shows that there is not a general bias towards frame-shifting alternative acceptors in general but their relative abundance is contingent on low observed GTEx PSIs.

This means that, amongst predicted alternative acceptor sites unannotated in UCSC, we found that novel predicted sites are:

1. Biased towards frame preserving if substantial read evidence support is found within GTEx.
2. Biased towards frame shifting if no substantial read evidence is found within GTEx.

Figure 4.3 replicates the known result that "unannotated splicing events in RNA-seq data … maintain reading frame" [8]. It additionally suggests the converse, namely that there are many unannotated splicing events which do not maintain the reading frame and which are depleted from standard RNA-seq data, presumably via NMD.

4.3.2 Alternative donors

We subsequently performed a similar analysis from the point of view of donor sites. Of the 66029 (novel) predictions in categories 2 & 3, 61.8% (40818) are donor sites. Of these, 25199 are found within 120 base pairs of known exons. Of these, 19735 which are found within gene-transcripts we can clearly analyse in terms of open-reading frame.

We find that the significant bias disappears for weak donor sites more quickly than it did for acceptors. The bias is still present at 12.6% depletion for 0 frameshift sites when donor sites are predicted at more than 0.3. However, when we choose a stronger model splice-strength of 0.5, this depletion increases to 20%.

For alternative acceptors we found similar rates of depletion, whether predicted at 0.3-0.5 or whether predicted strongly at 0.7-1.0. For alternative donors the rate of depletion increased based on how strong the predictions are. We additionally include analyses for donor sites restricted to predicted PSIs in [0.5, 1.0] where we obtain similar depletion rates.

This suggests that the relative split between false model predictions and predicted NMD-targets is more linked to the model predicted strengths than for acceptor sites.

Succinctly, amongst predicted alternative donor sites unannotated in UCSC we found that novel predicted sites are:

1. Biased towards frame preserving if substantial read evidence is found within GTEx.
2. Biased towards frame shifting if no substantial read evidence is found within GTEx.
3. Biases are more apparent for more strongly predicted novel sites.

We do not have a firm understanding as to what explains the slight difference between the results for novel donor sites and novel acceptor sites. It is similar to how we have observed distinct results in our mutagenesis chapter, whereby disrupting an existing donor site does not readily lead to one clear alternative as it tends to do for alternative acceptor sites.

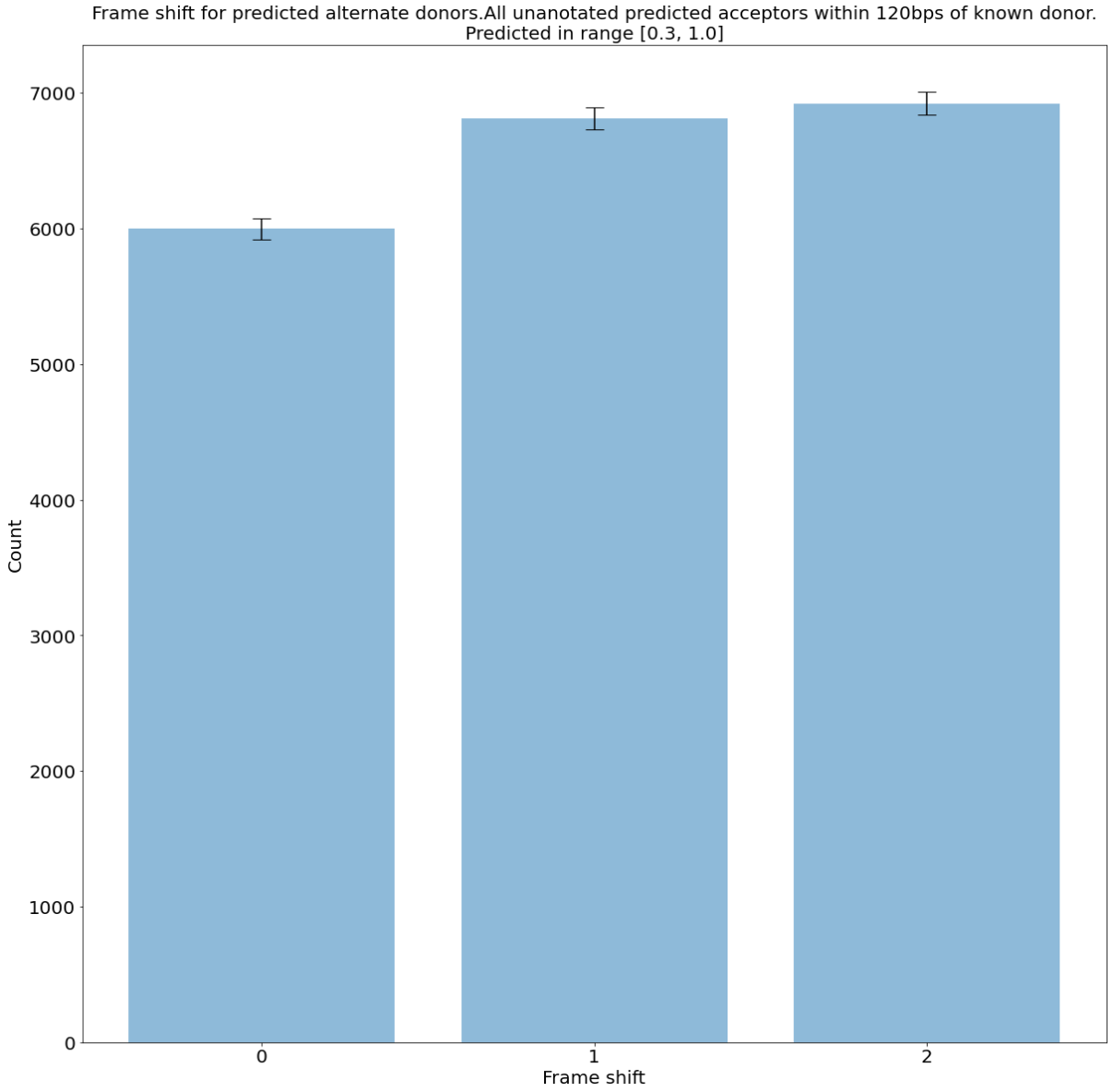


Figure 4.4: Novel predictions found with PSI <= 0.01 in GTEx. Consistent counts for frameshifting alternative donors but we find significantly fewer novel predictions which preserve the frame.

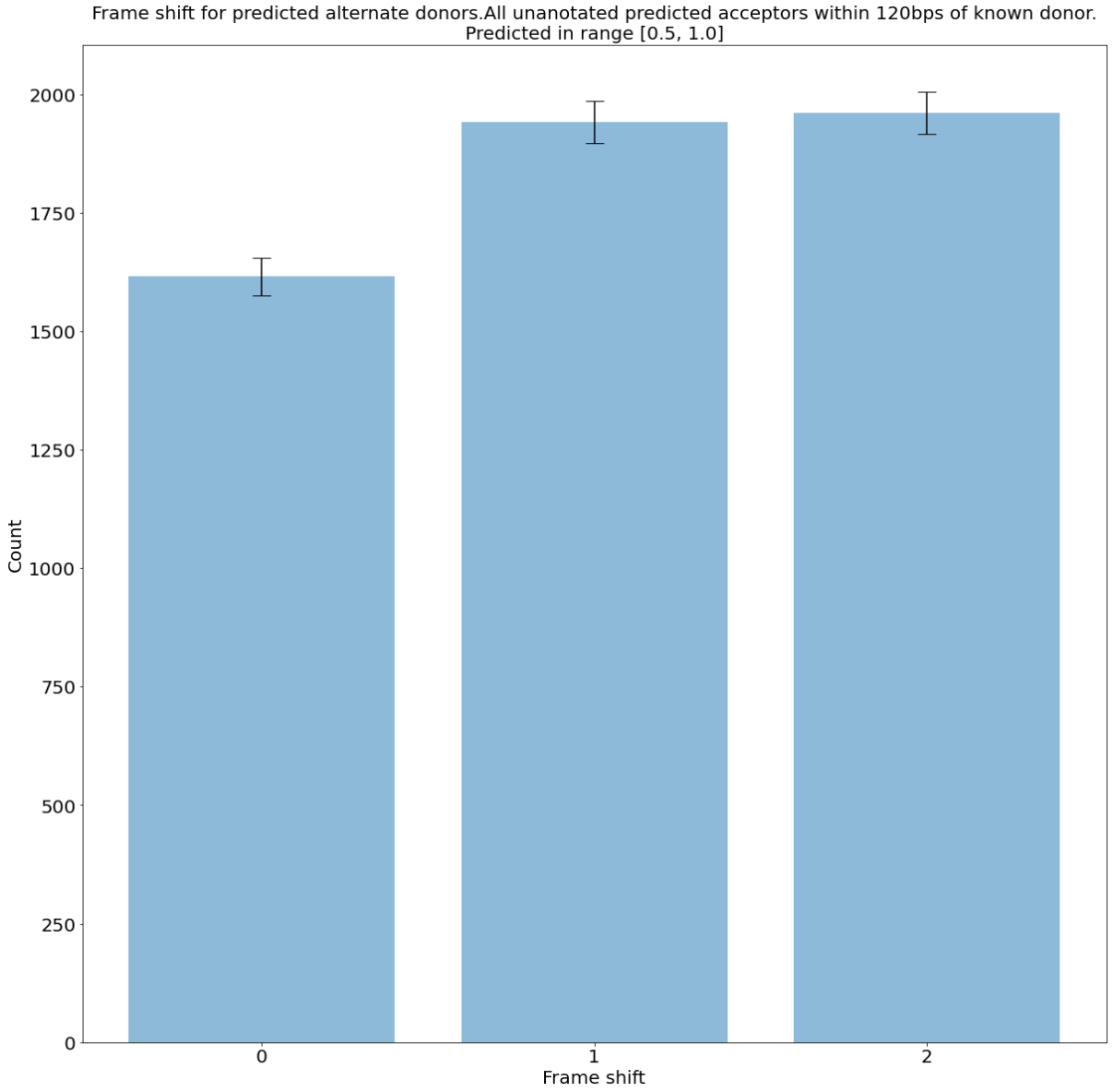


Figure 4.5: Novel predictions (predicted in [0.5, 1.0]) found with PSI <= 0.01 in GTEx. Consistent counts for frameshifting alternative donors but we find significantly fewer novel predictions which preserve the frame. The bias increases to 17% when considering only stronger donor site predictions.

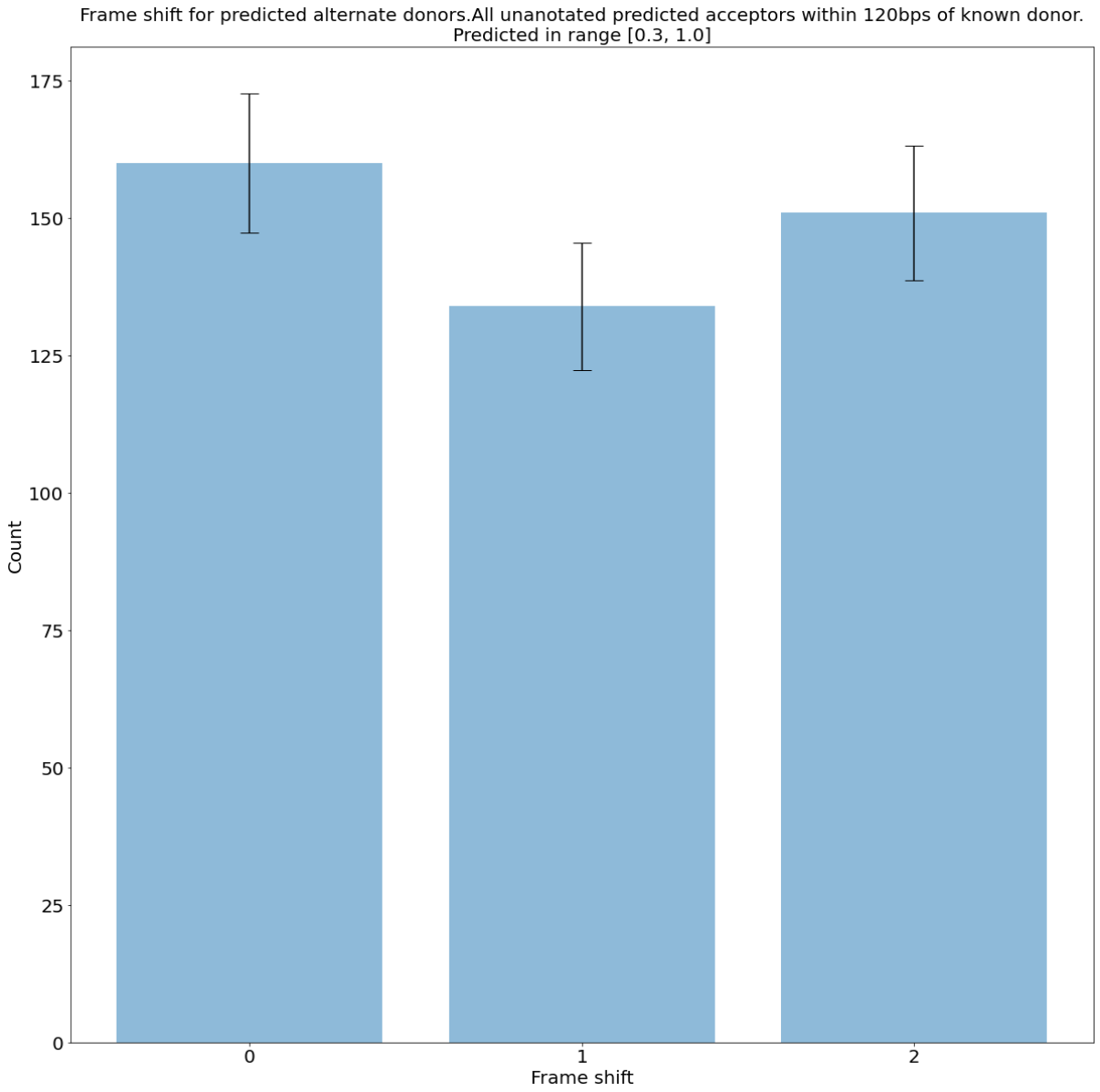


Figure 4.6: Novel alternative donor predictions (unannotated in UCSC V37) with an empirical PSI >= 0.05 in GTEx. We do not observe a bias towards frameshifting alternative donor sites as we do for unannotated sites found within GTEx.

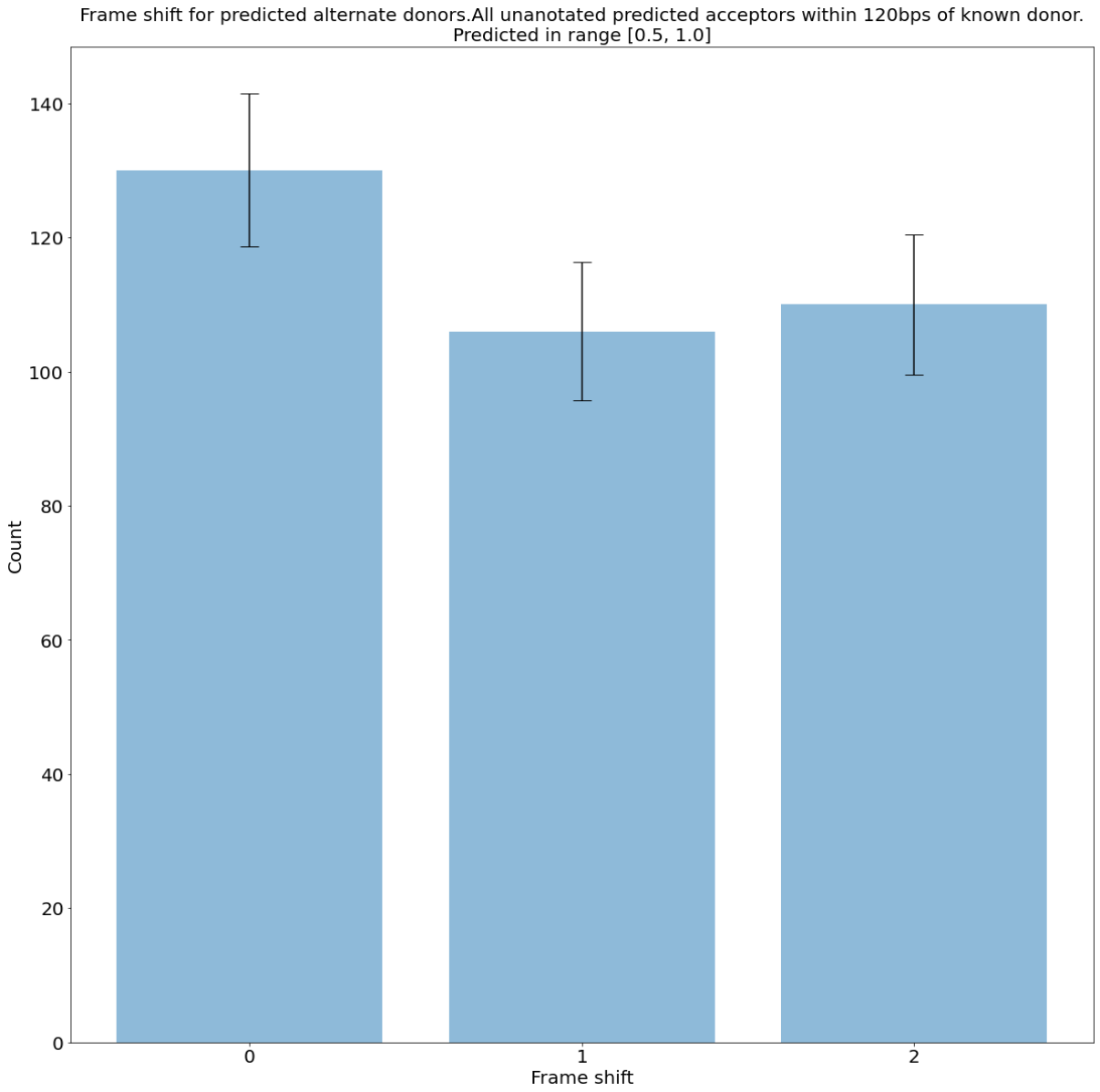


Figure 4.7: Same as figure 4.6 except considering only more strongly predicted novel donor sites. Unlike in figure 4.6 we start to observe a bias towards frame preserving alternative donor sites when evidence is found GTEx.

4.3.3 Predicted inclusion of exon in annotated intronic region

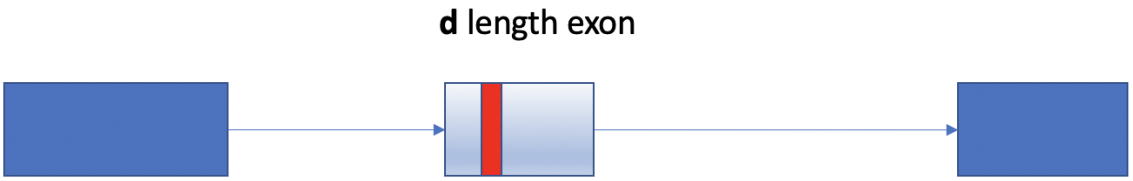


Figure 4.8: NMD can occur for novel exons based on exon length leading to a frameshift or the novel exon having an inframe stop codon

For the generation of novel exons, we analysed acceptor-donor pairs predicted by the model and found within annotated intronic regions. These are defined as model predicted pseudo-exons with the following properties:

1. Greater than 500 base pairs from known exons
2. Have empirical PSI <= 0.01 in GTEx
3. Acceptor-donor pairs both predicted at splice rate >= 0.5

The null hypothesis is that the model is incorrect in making a strong prediction. Precisely, the lack of empirical evidence for these sites is explained principally by these splice sites being false positives or very weak splice sites over any RNA-decay mechanism.

If our null hypothesis were the case, then one would expect little difference in the counts of 0, 1, 2 frameshifting pseudo-exons in these predictions. Precisely, one would expect 1/3 predicted exons to not shift the frame of the surrounding transcript.

The second relevant property to analyse is the relative position of stop-codons. Frameshifting exons almost always lead to nonsense mediated decay. However, frame-preserving exons require an in-frame premature stop-codon to lead to NMD. Under

our null hypothesis, one would not expect differences in frame-preserving and frame-shifting exon predictions.

In figure 4.10 we consider the bias in whether novel predicted exons have an in-frame stop-codon based upon whether they frame-shift or not. Of the sites which did not frameshift they were marginally more likely to contain an in-frame stop-codon than other predicted exons. 11% of non-shifting predicted exons didn't contain an in-frame stop-codon compared to 20% of predicted frame-shifting exons.

Based on this analysis we can reject the null hypothesis that these are not significant predictions with a p-value of 0.004. If we were to suppose that all these novel exon predictions were false predictions, then we would have no selection pressure on needing an in-frame stop-codon for non-frameshifting exons. This would mean that we should expect roughly 20% of non-shifting exons to not have any in-frame stop-codons. Conversely, if these novel predictions were all true predictions, despite missing annotations, then we should expect all the non-shifting exons to have an in-frame stop-codon and so we would expect a percentage closer to 0%.

The case for novel exons is essentially the same as for alternatively spliced acceptors/donors, in that we find a signature difference based upon selection. This suggests that a significant proportion of the predicted exons are real, as they can be acted on by evolution and so skew their relative occurrence in this analysis. This analysis is less adept at translating to estimates for prevalence as we will do for alternative acceptor or donor sites. For novel pseudo-exons our analysis is taken over clustered pairs of sites (donors and acceptors combined), which removes the majority of potential pseudo-exons where only one of the donor-acceptor pairs is strongly predicted, making prevalence estimates complex.

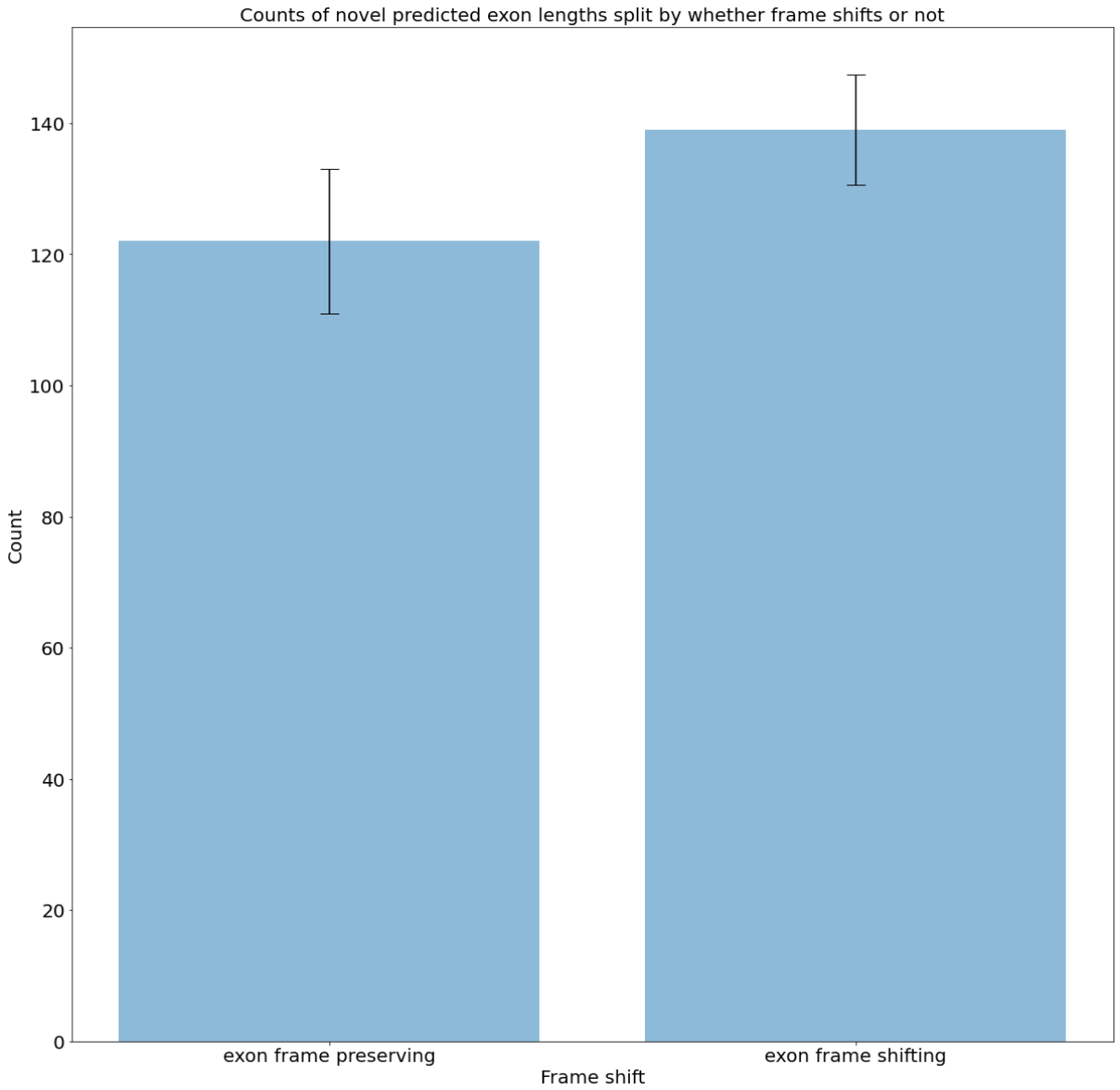


Figure 4.9: Novel acceptor-donor pair predictions exhibit small depletion in exon-frame preserving exons. Even with combining our exon-frame shifting counts we obtain overlapping error bars for the two. For clarity we took the mean counts for frame-shifting exons – so the actual count is double what is observed in the plot.

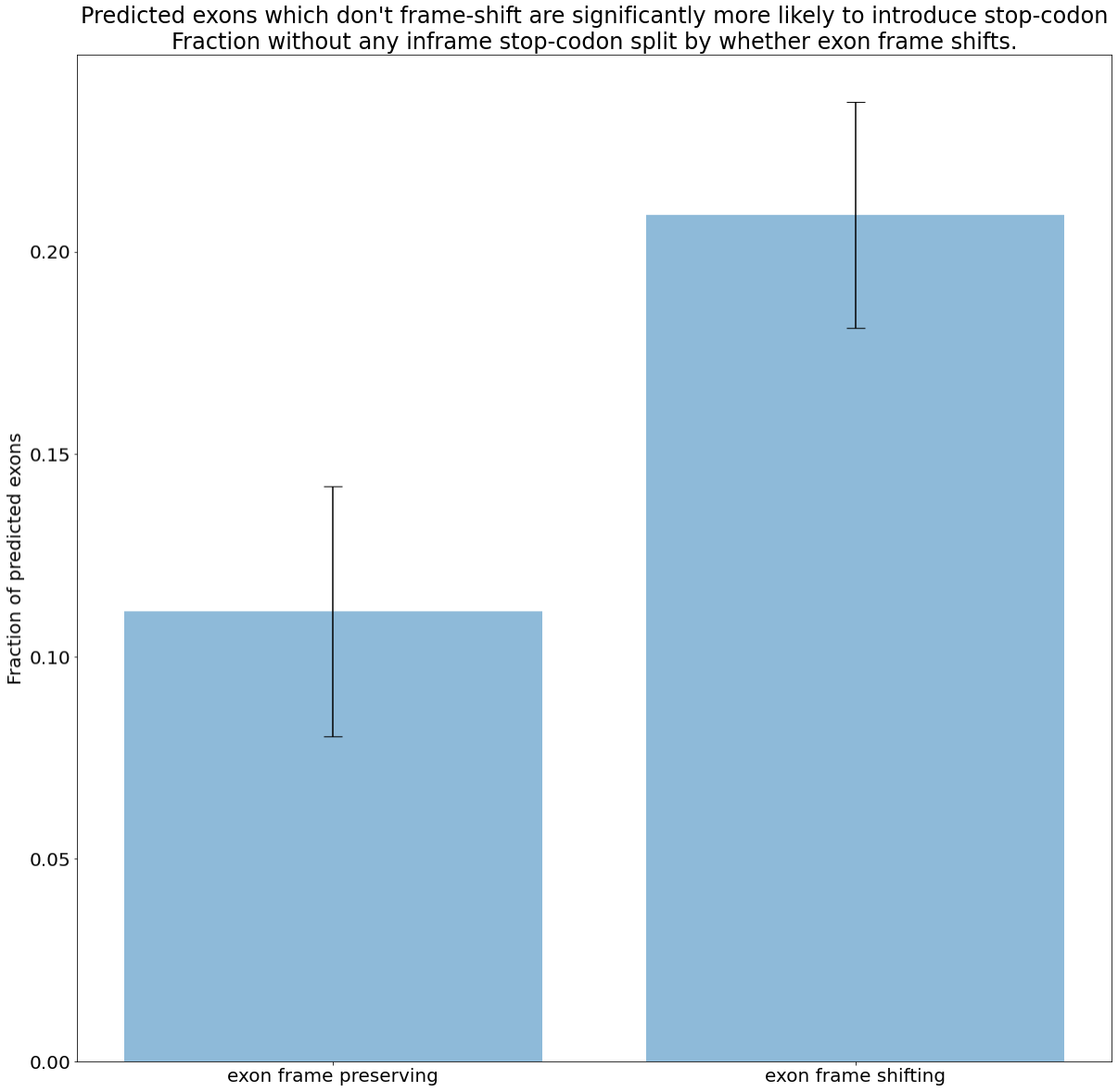


Figure 4.10: Novel exons which do not frameshift are more likely to introduce an inframe stop-codon. Due to stop-codon motifs being common it is rare for sequences to not introduce one. It is more rare for frame preserving novel predictions. This suggests that there is some selection pressure.

4.3.4 Using model to estimate percentage of strong splice sites are AS-NMD sites

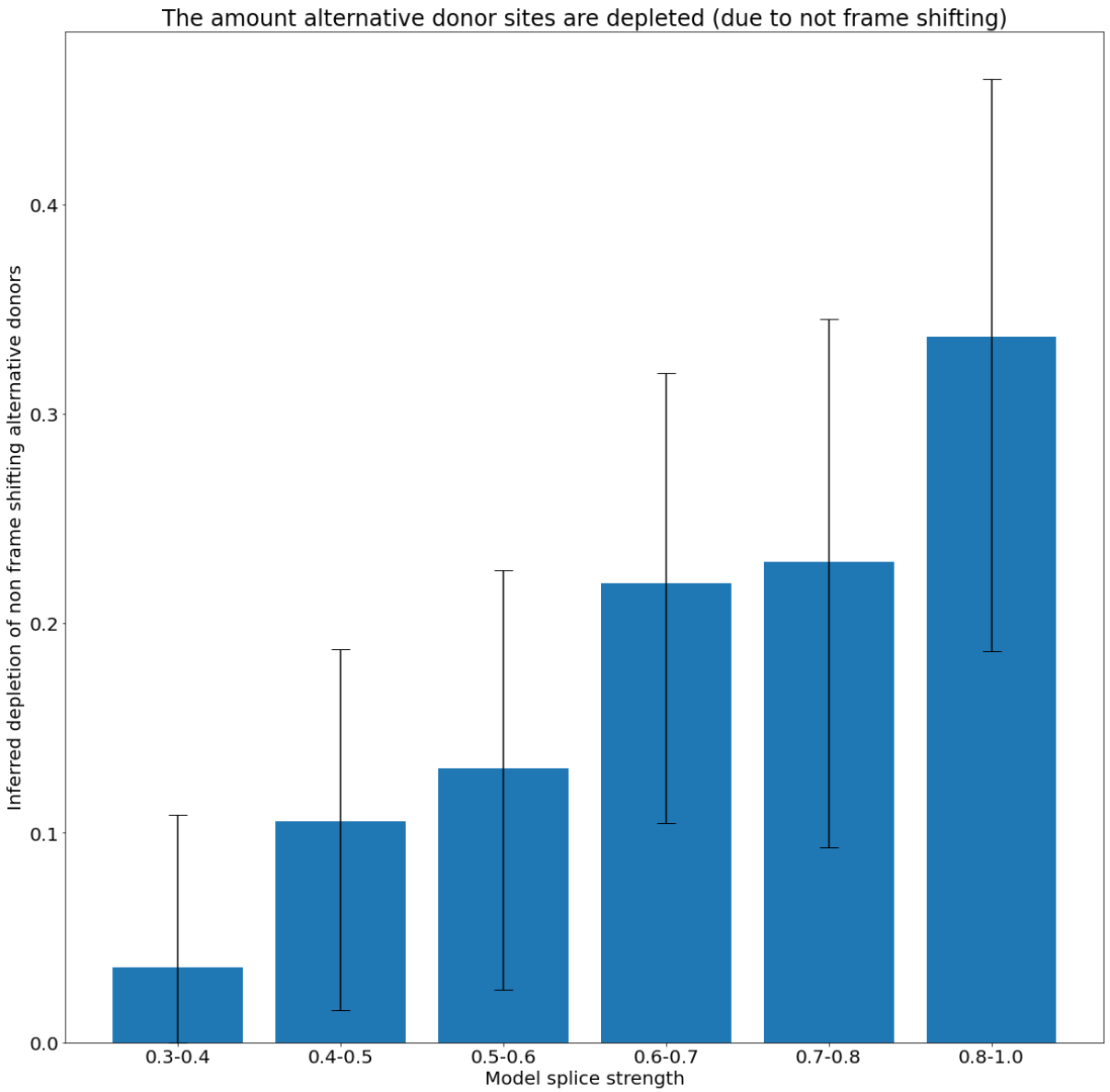


Figure 4.11: Depletion of in-frame alternative transcripts, by predicted strength of novel donor site. For low predicted alternative splice rates, the associated alternative transcripts have roughly expected proportions of in-frame vs. frame-shifted transcripts, consistent with low selection pressure to remove these occasional alternative and likely non-functional transcripts. For high predicted alternative splice rates, in-frame transcripts

are seen less frequently than expected by chance, consistent with selection acting to remove these transcripts.

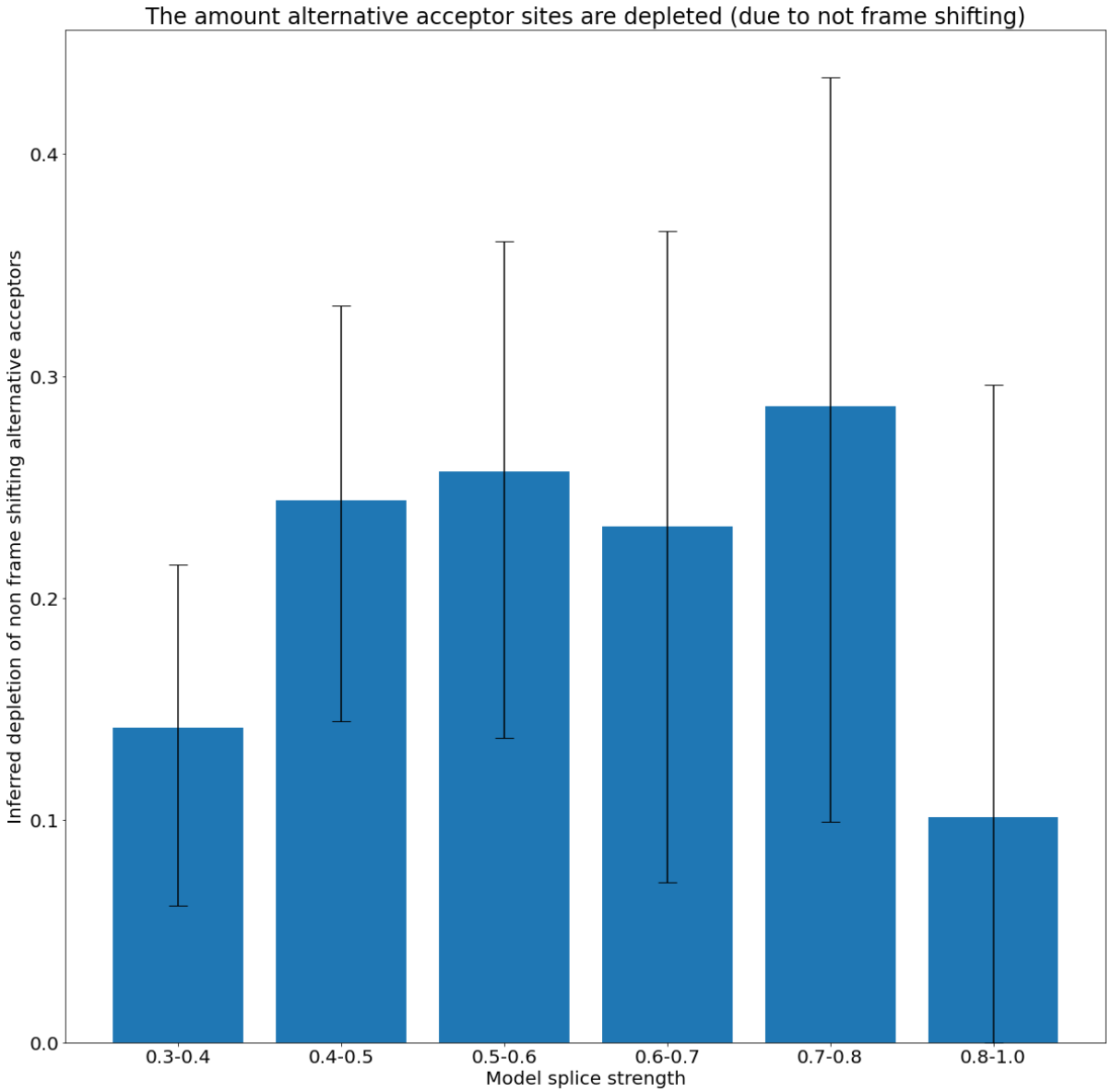


Figure 4.12. A property of these depletion analyses we found is donor sites were heavily dependent on the donor-site strength. On the other hand high depletion values are found for weakly scored acceptor sites. Our explanation for this phenomenon is that more of alternative splicing is captured at the acceptor sites compared to the donors.

We use the depletion rate distributions as computed by the analysis from the previous sections (figures 4.11 and 4.12). These tell us the depletion based on NMD-signatures found amongst novel AS predictions split by their model splice strength. This allows us to form a simple model to estimate the number of strong AS-NMD sites by combining these rates of depletion with the number of new predictions for each splice-site strength. The assumption of this analysis is that the amount of depletion is indicative of selection pressure to remove alternative splicing sites which do not lead to NMD. This provides a lower estimate of the amount of AS-NMD sites for a given splice strength.

Our method is the following:

1. Bin splice site strengths into bins from 0.3 to 1.0
2. For each bin compute the depletion rate amongst novel predictions
3. Multiply these two figures which estimate the number of novel predictions which are AS-NMD sites.
4. Sum these values over the bins

We compute 95% confidence intervals for each splice-site strength bin, which creates a very conservative confidence interval after summing the bins. Acceptor sites have more consistent depletion rates, so we felt more confident in combining our bins. This allows for tighter confidence intervals. Despite this caution, we observe that AS-NMD forms a high percentage of strong acceptor sites.

Based upon these we have a computation which leads to:

1. 5.5% (range [4.0%, 7.0%]) of strong predicted acceptors sites are AS-NMD sites

2. 5.01% (range [1.6%, 8.4%]) of strong predicted donor sites are AS-NMD sites

It should be noted that this estimate is providing something very distinct to existing AS-NMD abundance estimates. These alternative estimates do not take into account splice-site strength but rely on datasets where the relative abundance of AS-NMD and AS sites are not comparable. As such, their estimates are based upon absolute counts which do not adjust for splice-site strength. They can be highly biased by lowly transcribed isoforms which induce NMD. Instead, our estimates are based upon high model-strength AS sites.

4.3.5 Intermediate conclusion

A general picture has emerged which shows depletion of predictions unlikely to induce NMD compared to those likely. As discussed, we have some caveats on the estimates we find, but this depletion is around 20% and replicated across different categories of alternative splicing. To be clear, this does not provide estimates for the level of AS-NMD but tells us how much NMD-equivalent signatures lead to biases in our model predictions indicative of selection pressure.

However, although in aggregate we have a clear signal of functionality from this depletion analysis, this cannot tell us how well individual prediction strength from our model correlates with actual splice usage. Later in this thesis we will use experimental data from different studies to try and address this (see section 4.6).

4.4 NMD model pipeline

Nonsense-Mediated Decay is a well studied regulatory mechanism. A well quoted rule is the “50nt-model”, which seeks to explain the majority of NMD events based on the presence of a premature stop-codon [38]. This rule states that an in-frame stop-codon is likely to cause

NMD if found before 50 nucleotides upstream of the last donor-site. This is based on our understanding of the biology of the Exon Junction Complex model, where a pioneering round removes transcripts with such premature stop codons.

In this section we present a new NMD-aware predictive model which shows that the 50nt-rule partially explains the relative levels of expression of isoforms in GTEx. In the subsequent sections we will apply this model to a new nascent form of transcriptomics (POINT-seq) which is experimentally capable of capturing co-transcriptional splice events prior to the induction of NMD decay. We will confirm the NMD model's applicability and effectiveness, but additionally show the richness and variety of splicing events found in co-transcriptionally spliced RNA compared to stable transcriptomics.

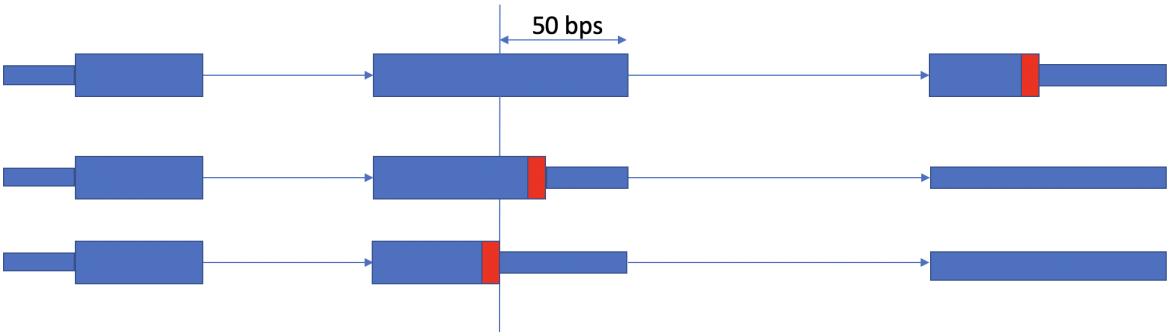


Figure 4.13: Bioinformatic pipeline which annotates first two transcripts as not inducing NMD with the third one as inducing NMD. A 50 base pair cutoff is applied in the penultimate exon. As a result only the 3rd of the diagram transcripts would be predicted to induce NMD.

Our model works by implementing the 50nt-rule adapted to take the output of our neural network model.

A necessary working assumption is that splicing changes are local to the exon. This means that we assume that the inclusion or exclusion of an exon would not affect distal exon inclusions within the gene. Precisely, our new model assumes that a novel exon will mostly be found as an addition to existing transcripts or as a replacement for an existing exon. This means that we cannot address cases of mutually exclusive exons or other co-dependent scenarios.

Essentially, our model partially refines a simpler annotation which records the frameshift by additionally finding premature stop-codons from sequence inclusion. To confirm that the model is picking out genuine NMD sites we ran the NMD model on unannotated novel-predicted splice-sites at a model threshold of 0.8. Precisely, these are strongly predicted sites with no annotation in UCSC37 meaning our prior belief should be that they have limited expression in GTEx. The motivation for comparing unannotated splice sites is that we utilise the annotations in the 50nt-model to form the wider gene-transcripts. Therefore, this analysis removes any bias which arises from using annotated splice sites as well.

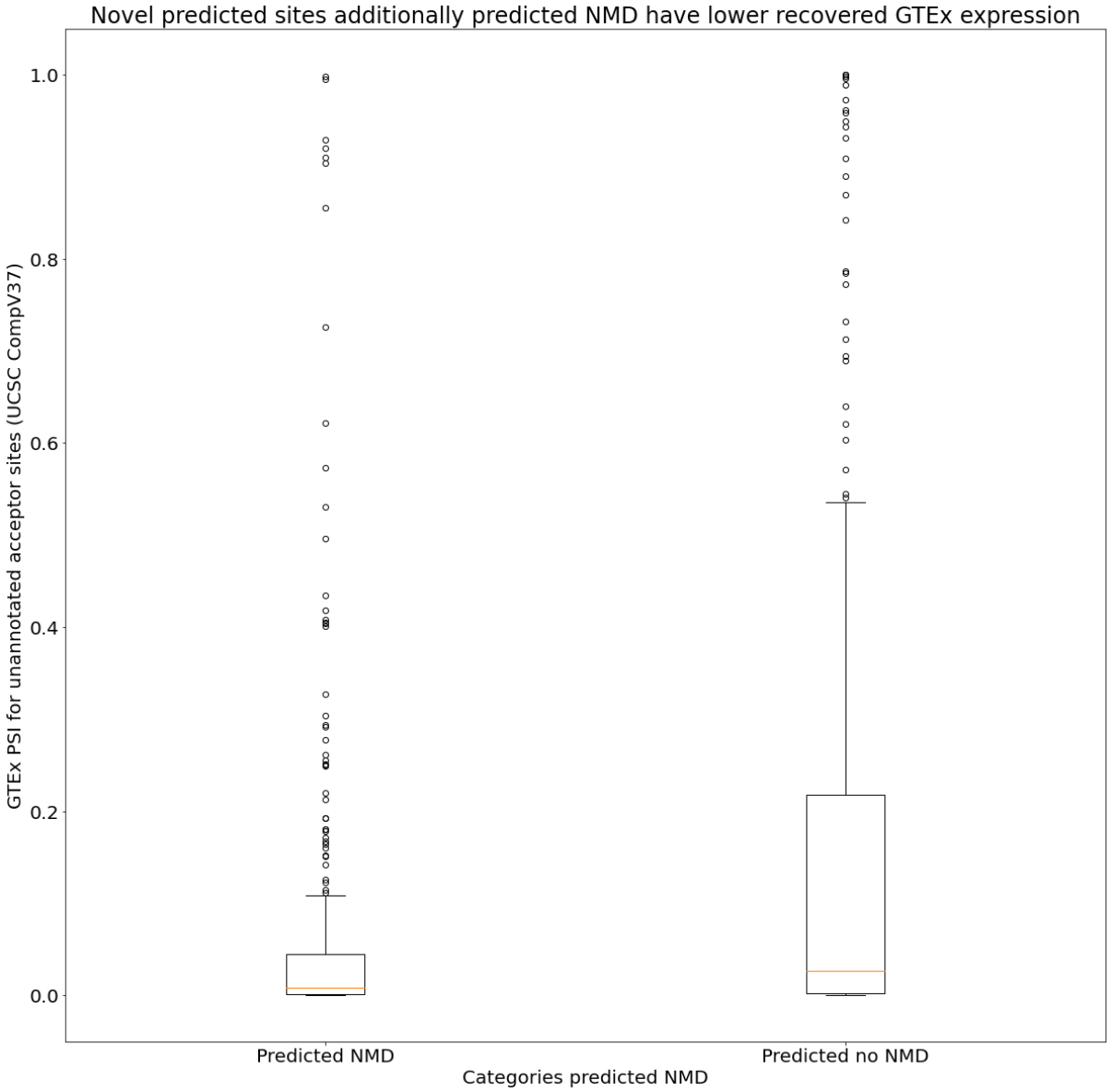


Figure 4.14: Amongst unannotated splice sites those predicted to induce NMD have a mean GTEx PSI of 6% compared to a mean PSI of 18% predicted to not induce NMD. This taken over splice sites with predicted splice strength >= 0.8

Given that these predicted splice sites are not annotated in UCSCv37, it is unsurprising that their empirical PSIs in GTEx are relatively small compared to their predicted PSI. However, we find that there is a 3-fold increase in PSI between the two datasets. We see that predicted non NMD splice sites have a mean GTEx PSI three times higher than the

predicted NMD splice sites (18% vs 6%). The ratio of their medians is also 3-fold (2.7% vs 0.9%). Decreasing our prediction threshold from 0.8 to 0.5 maintains this ~3-fold difference between the two datasets. The predicted not-NMD splice sites have a mean GTEx PSI three times higher than the predicted NMD splice sites (8.8% vs 2.8%). The ratio of their medians is also ~3-fold (0.7% vs 0.3%). In summary, these results suggest that the implemented NMD model is a good predictor as to whether we expect to find evidence in GTEx.

4.5 Position of open reading frame

Splicing related nonsense-mediated decay has been shown to occur due to alternative splicing which alters the coordinates of the open reading frame (ORF), so we needed to add predictions of ORF to the model. This would require the additional complexity of accurate predictions of the start-codon, which is challenging to determine because of the complexity of annotated coding and non-coding splice isoforms.

Therefore, we opted instead to focus on splicing changes occurring within the ORF of a gene which can be reasonably and reliably extracted from gene level annotations as found in UCSC V37. A caveat to this approach is that some complex potential splicing patterns will be missed. An example of this, which could induce NMD, would be the creation of an intron in the 3' UTR after the stable stop-codon. Although nothing changes within the open reading frame, a previously stable transcript becomes potentially subject to decay in this setting. We see later on an example within the TRIP13 gene which exhibits this phenomena.

4.6 Application of model to NMD datasets

So far we have presented indirect evidence that a fraction of splice site predictions without support in transcript databases is in fact correct. In this section, we provide direct evidence for this by using a recently published dataset made with a new transcriptomic protocol called POINT-seq (for mammalian polymerase intact nascent transcript sequencing, [40]). This protocol captures nascently transcribed portions of RNA by immunoprecipitation of the RNA polymerase II subunit and, because of the kinetics of splicing, it can capture splice events co-transcriptionally. Our assumption is that RNA-decay mechanisms mainly occur post-transcriptionally, so that true splice events whose resulting transcripts are subsequently degraded by NMD are expected to be visible in a POINT-seq experiment.

Furthermore, the lack of presence of splicing events in mature transcripts that can be detected in POINT-seq suggests that we should obtain more realistic empirical estimates for the relative proportion of these events to the total output of the gene, thereby providing a more quantitative validation set for our model to test the relationship between strength of prediction and splice usage.

In this setting we have obtained two replicates made on HELA cells [40]. We analyse in this subchapter the extent to which the model predictions better reflect these co-transcriptional datasets over our in-house mature RNA-seq datasets.

Moreover, utilising our NMD model and the relative depletions found amongst predictions which have a NMD prediction we can estimate the extent that this decay mechanism “hides” the underlying splicing in standard RNA-seq. We additionally include comparisons between

the two technologies (POINT-seq and RNA-seq) when our NMD prediction is uncertain or predicts no NMD. In this way, consistency between the empirical PSIs for the two datasets in this context can be used to strengthen the case for our ability to predict NMD inducing splice sites.

4.6.1 Worked gene examples

Here we start with some examples which are found principally in POINT-seq and our predictions but not in RNA-seq.

Here we present visual examples of NMD transcripts which have high expression in POINT-seq. We will show examples that exhibit the different mechanisms in which alternative splicing can result in a NMD transcript from alternative acceptors, alternative donors, gained exons and gained introns after the canonical stop-codon.

Interestingly, some of these examples exhibit high levels of associated sequence conservation, in addition to very high levels of inclusion in the POINT-seq datasets. As previously discussed, this association with conservation points to conserved functionality for these NMD associated transcripts. Although the genes have been selected in an unbiased manner based on high expression in POINT-seq and low expression in our RNA-seq dataset, we find an over-representation splice-regulating genes (16 of the 35 examples in table 4.15) replicating a hypothesis of AS-NMD being a mechanism of autoregulation for splicing factor genes [41].

Gene	Chrom	Position	Predicted PSI	POINT-PSI	RNA-PSI	Known Splice factor Gene
DHX9	chr1	182855576	0.865022	0.665025	0.119489	Yes
DHX9	chr1	182855713	0.854301	0.658163	0.187829	Yes
SRRM1	chr1	24647078	0.401683	0.427586	0.038966	Yes
SRSF11	chr1	70230618	0.347009	0.697872	0.009375	Yes
SRSF11	chr1	70231094	0.678059	0.778723	0.030208	Yes
SRSF11	chr1	70231203	0.185604	0.469388	0.016878	Yes
SRSF11	chr1	70231858	0.82045	0.622449	0.040084	Yes
SRSF11	chr1	70231975	0.849243	0.926441	0.228814	Yes
ZNF638	chr2	71393645	0.933486	0.477419	0.095238	No
ZNF638	chr2	71395246	0.983441	0.477419	0.096346	No
SKP2	chr5	36181817	0.988976	0.924051	0.137993	No
BRD2	chr6	32969352	0.855294	0.316667	0.024575	No
SRSF3	chr6	36599820	0.810893	0.323529	0.028956	Yes
CTTN	chr11	70421469	0.618867	0.929134	0.305857	No
KTN1	chr14	55663954	0.805722	0.546392	0.181818	No
ATXN2L	chr16	28836490	0.21259	0.3125	0.058336	No
RHOT2	chr16	671114	0.179264	0.30137	0.048868	No
SF3B3	chr16	70550678	0.900779	0.300699	0.083742	Yes
SPG7	chr16	89540943	0.64813	0.581169	0.167273	No
SRRM2	chr16	2760008	0.685459	0.405333	0.066996	Yes
SRRM2	chr16	2768504	0.845538	0.314917	0.08295	Yes
KIF1C	chr17	4999046	0.383594	0.377358	0.026316	No
LUC7L3	chr17	50750746	0.645944	0.88172	0.086604	Yes
MPRIP	chr17	17180606	0.872239	0.4	0.036532	No
MPRIP	chr17	17180669	0.728354	0.360656	0.036026	No
MSI2	chr17	57674917	0.722914	0.338983	0.100719	No
SPECC1	chr17	20156236	0.972733	0.98583	0.006739	No
U2AF2	chr19	55661770	0.534718	0.369863	0.051534	Yes
U2AF2	chr19	55661830	0.240909	0.386667	0.062389	Yes
CSE1L	chr20	49096175	0.677472	0.359375	0.022733	No
NOP56	chr20	2657372	0.310297	0.317241	0.038799	No
BACE2	chr21	41241939	0.783555	0.515337	0.128631	No
SON	chr21	33572630	0.979621	0.75	0.159168	Yes
TTC3	chr21	37157160	0.503383	0.347826	0.104478	No
TTC3	chr21	37157214	0.443991	0.318182	0.083969	No

Table 4.15: Examples selected where predicted splice sites have POINT-seq PSI >= 0.3 and found with PSI 3 times larger in POINT-seq to RNA-seq. Taken amongst sites with read coverage larger than 50 reads in both datasets. Genes annotated as to whether in known splicing factor genes.

4.6.2 Gained NMD exon examples

4.6.2.1 SRSF11

This example gene serves as a good example of evidenced exons in POINT-seq. SRSF11 is known to be involved in pre-mRNA-splicing factors.

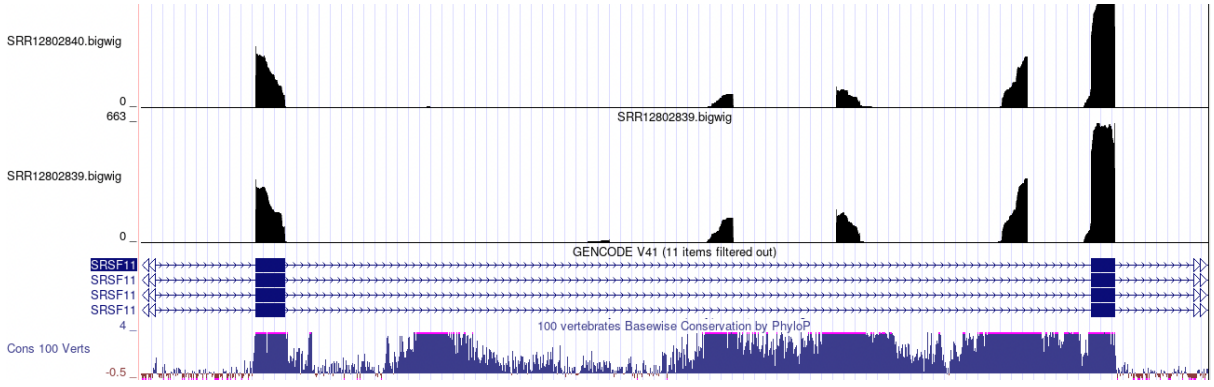


Figure 4.16: Large quantities of evidence for exons found in POINT-seq replicates within splicing factor gene SRSF11 within intron chr1:70228555-70232267. We observe many regions of high conservation coinciding with where there is evidence for extra exons.

In this example, the canonical intron (chr1:70228555-70232267) is spliced in 78% of transcripts in RNA-seq but it is found in only 7.8% of transcripts in POINT-seq. This is due to internal exons being commonly spliced in POINT-seq reducing the splicing of the main intron. Our model and POINT-seq both suggest that these spliced-in exons are likely to cause NMD. It is also curious that high levels of conservation are observed within the intron.

Curiously, this gene SRSF11 is believed to be implicated in pre-mRNA-splicing itself. It has been otherwise stated that conserved AS-NMD is typically found within genes known to involve splicing [41], [42]. This example is one of many examples where there is evidence of high splicing inclusion rate within POINT-seq, high conservation and predicted NMD.

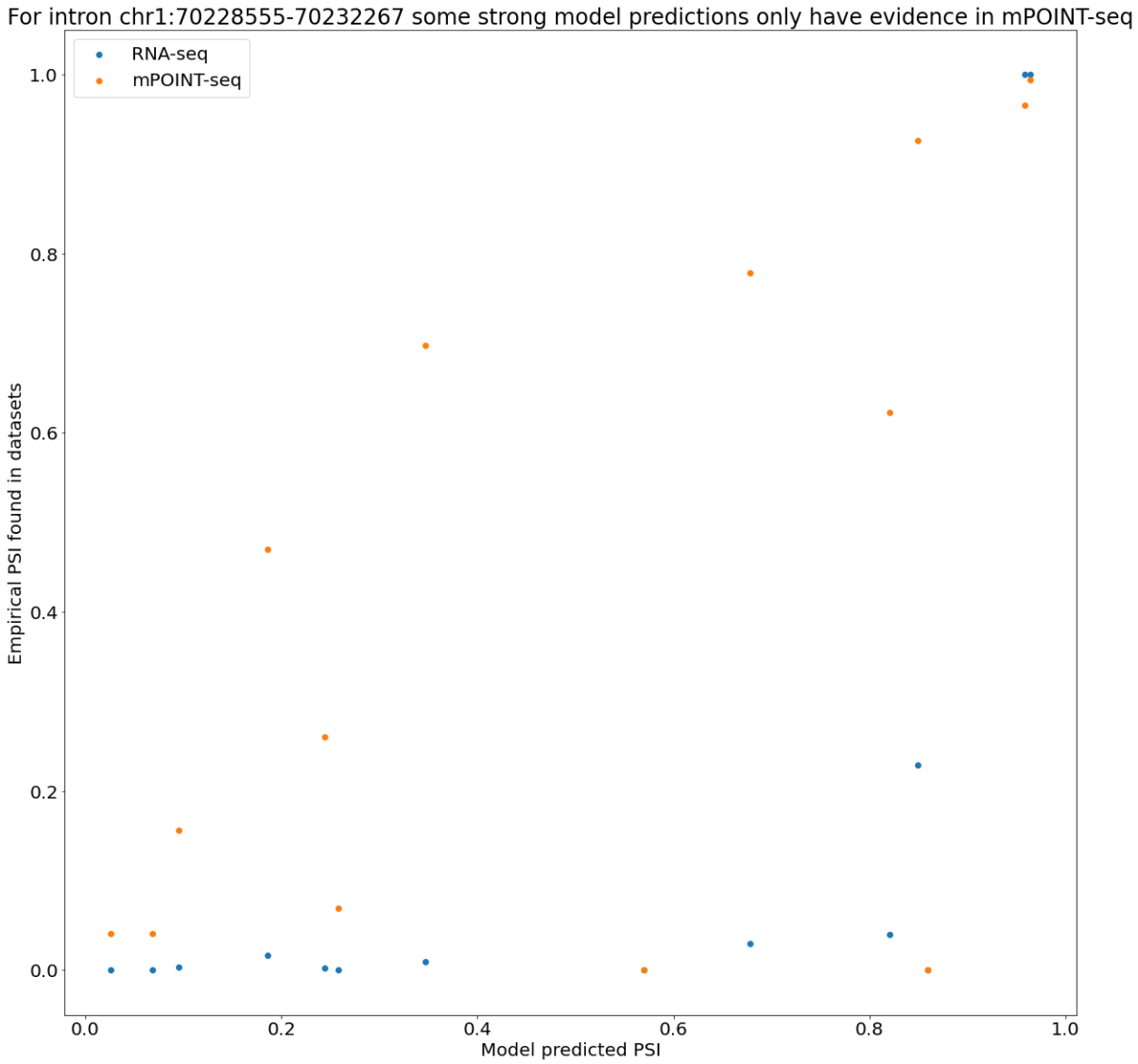


Figure 4.17: Analysis comparing RNA-seq and POINT-seq PSIs against model predicted PSI for SRSF11 intron chr1:70228555-70232267. We can see for this region our model predictions are incredibly inaccurate when compared against our RNA-seq dataset. However, when compared against POINT-seq we find that there is abundant evidence for many of the sites.

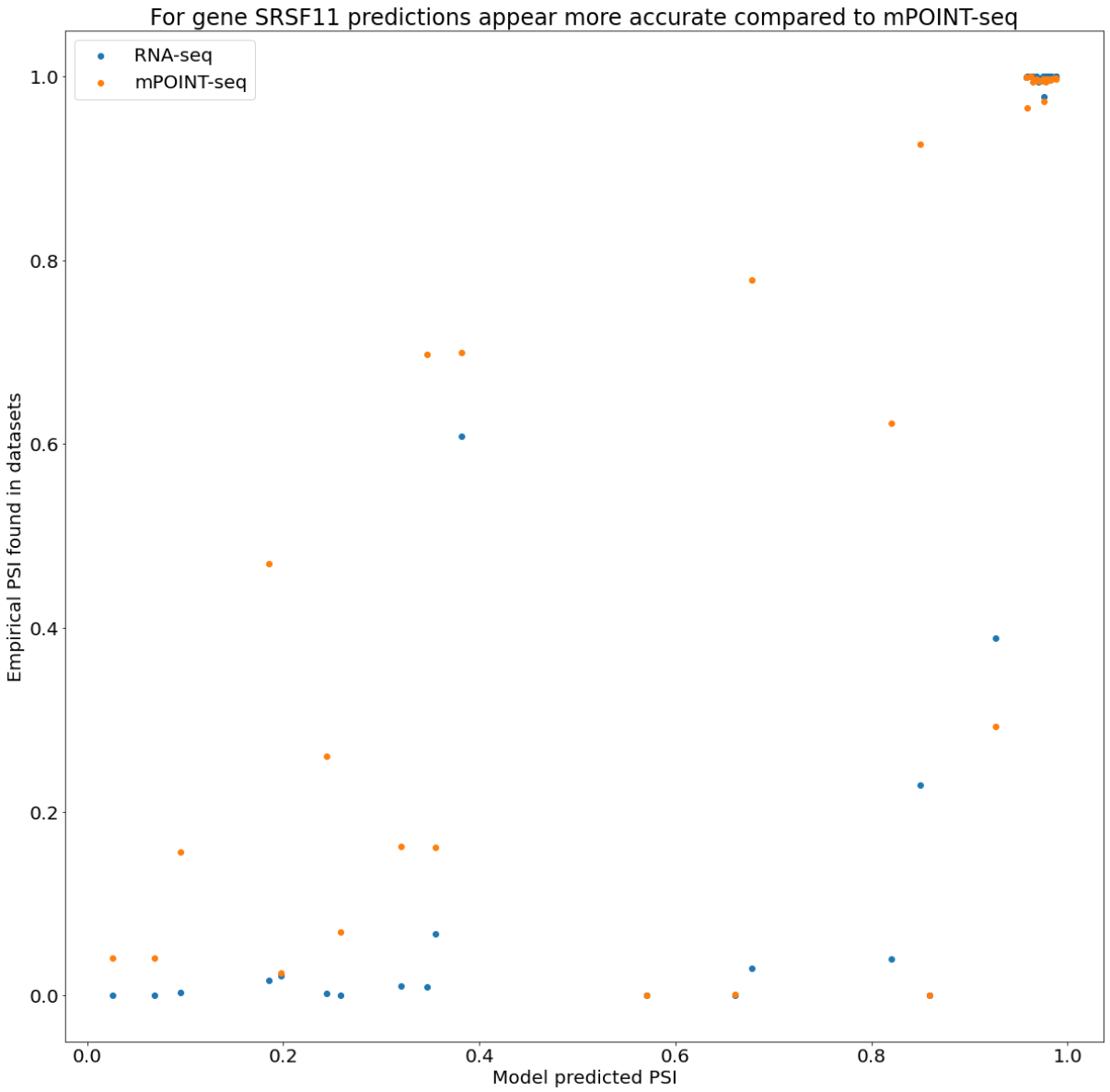


Figure 4.18: Analysis comparing RNA-seq and POINT-seq PSIs against model predicted PSI for full gene SRSF11. Apart from 3 splice-sites which we don't find evidence in POINT-seq the majority of SRSF11 predicted splice-sites with no RNA-seq evidence have POINT-seq evidence.

Due to an abundance of predicted exons in the we compute the empirical POINT-seq PSI for all splice sites within the intron chr1:70228555-70232267 in figure 4.17. There are 2 sites which have PSIs almost 1 for both RNA-seq and POINT-seq. These are the donor and acceptor sites defining the intron. Otherwise we observe far greater empirical PSIs for

POINT-seq and RNA-seq. In future 4.18 we observe a similar trend for the whole gene. This kind of analysis forms the backbone of the next section although we do not find genome-wide it is as clean as for this gene.

SON

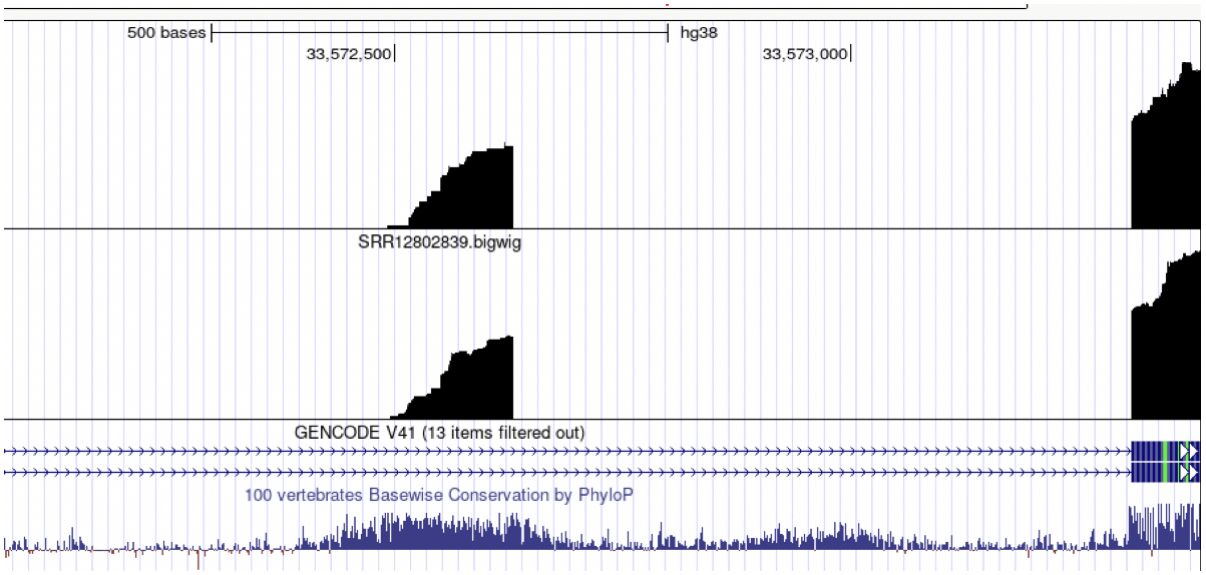


Figure 4.19: Conserved exon found in POINT-seq replicates within splicing factor gene SON. Another example of a well expressed and conserved pseudo-exon which is not an annotated exon.

Like in other examples of well-conserved and expressed AS-NMD, we find that this example is found within a splicing factor gene. Our models both predict its expression and that it induces NMD.

4.6.2.2 DHX9

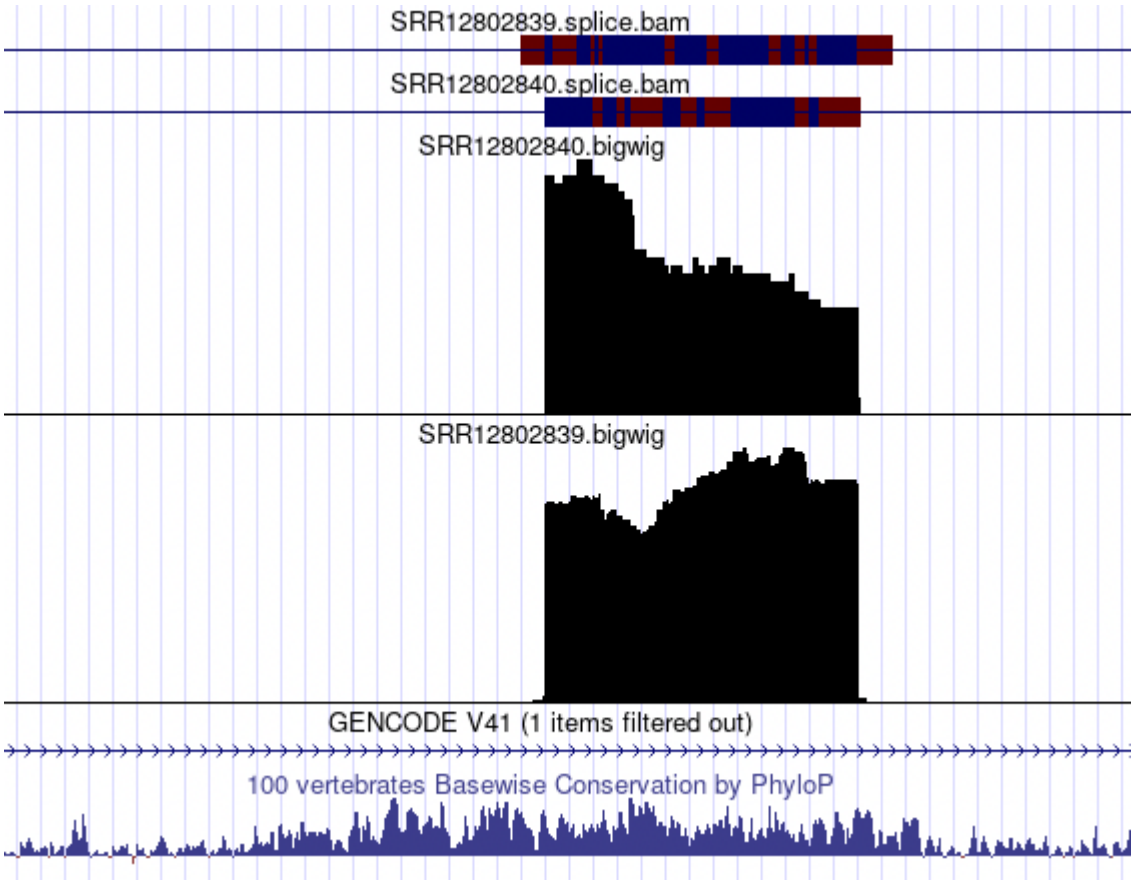


Figure 4.20: Conserved exon found in POINT-seq replicates within splicing factor gene DHX9. This is another example of a conserved exon which we found highly expressed in POINT-seq but near an annotated exon. Conservation pattern follows likely location for acceptor and donor site definitions.

DHX9, like other found examples, is known to regulate splicing by having a role in spliceosome assembly. The conservation pattern suggests conservation of the acceptor locus sequence along with the donor locus.

U2AF2

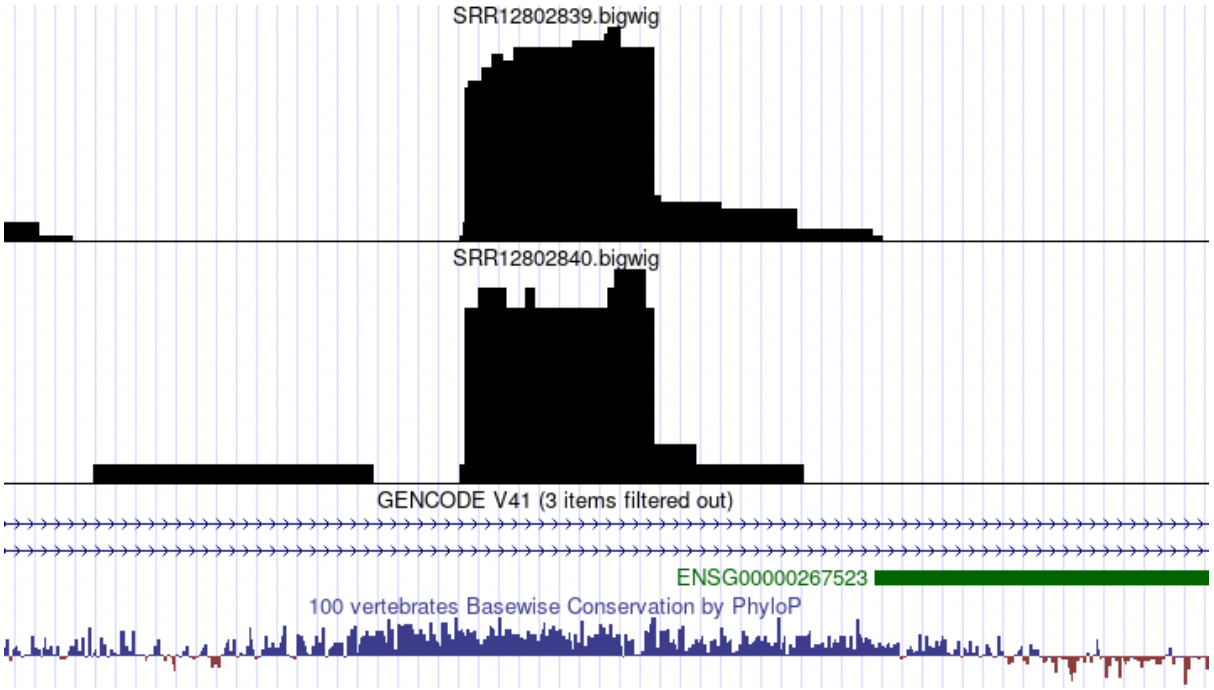


Figure 4.21: Conserved exon found for POINT-seq replicates within splicing factor gene U2AF2. This example is included in 37% of transcripts of POINT-seq and predicted to be included in ~50% of transcripts. We observe that no annotated exon is located nearby but high levels of conservation exist.

This is also involved with the splicing factor U2 snRNP specifically related to the binding of U2 snRNP to the branch point locus. We find it curious that there is high conservation surrounding the unannotated exon.

4.6.3 Alternative acceptors

4.6.3.1 ADAM15

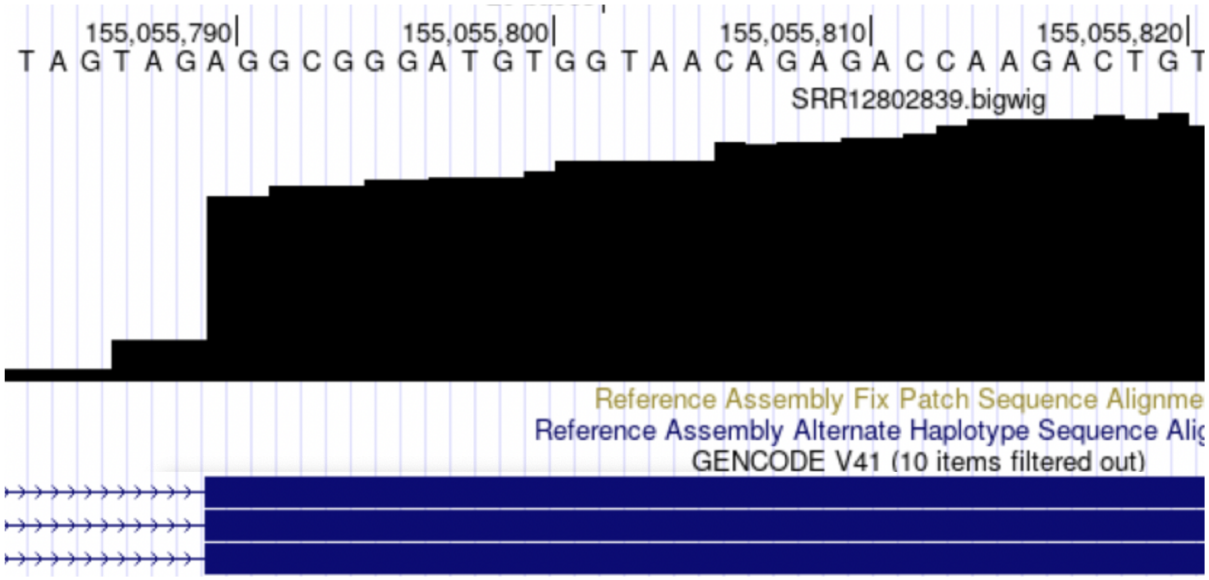


Figure 4.22: Read support for POINT-seq at exon 7 of ADAM15. Model predicts the presence of additional AS-NMD acceptor site with PSI of 0.66. Two competing acceptor sites predicted 3 base pairs apart with the minor one leading to an inframe stop-codon. Non-decaying acceptor site predicted with PSI of 0.39.

For exon 7 of ADAM15, the model predicts 2 competing acceptor sites due to a *TAGTAG* repeat preceding the exon's acceptor site. The model preference is for the first TAG forming the acceptor site with less strength provided for the second. When splicing occurs at the first *TAG,* the second *TAG* forms an in-frame stop-codon.

Using POINT-seq datasets we find that this unannotated acceptor site has 12.0% support which reduces to 2.8% in mature RNA-seq datasets.

4.6.4 Alternative donors

4.6.4.1 OBSCN

In complex gene OBSCN the model predicts an additional unannotated NMD-exon recovered in POINT-seq. This example is interesting because according to our POINT-seq a large percentage of the transcription will use the alternative, unannotated donor site which we predict to be subject to NMD.

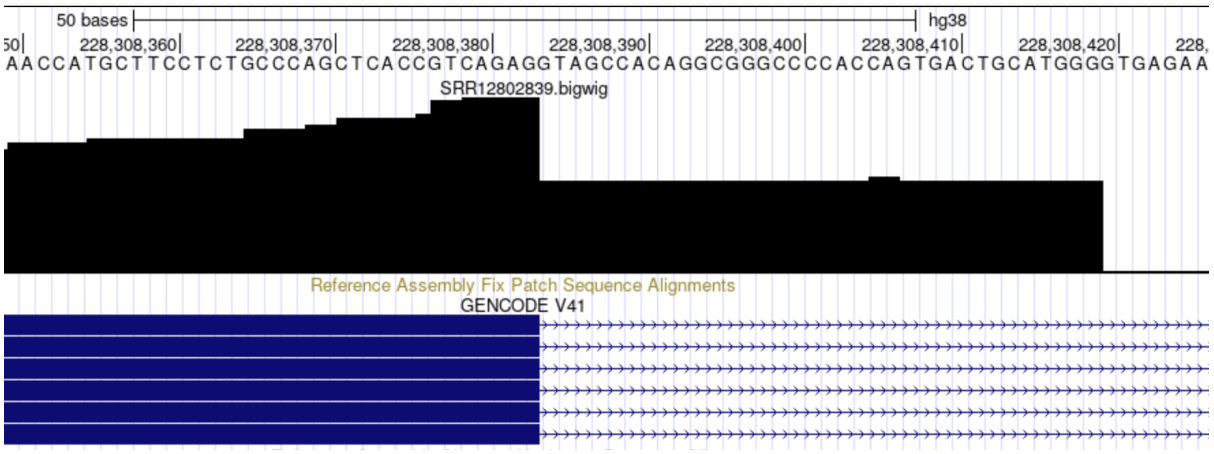


Figure 4.23: Read support in POINT-seq for presence of additional alternative donor site in gene OBSCN. Model predicts two donor choices to both be spliced alternatively with only one annotated.

4.6.5 Gained intron

4.6.5.1 TRIP13

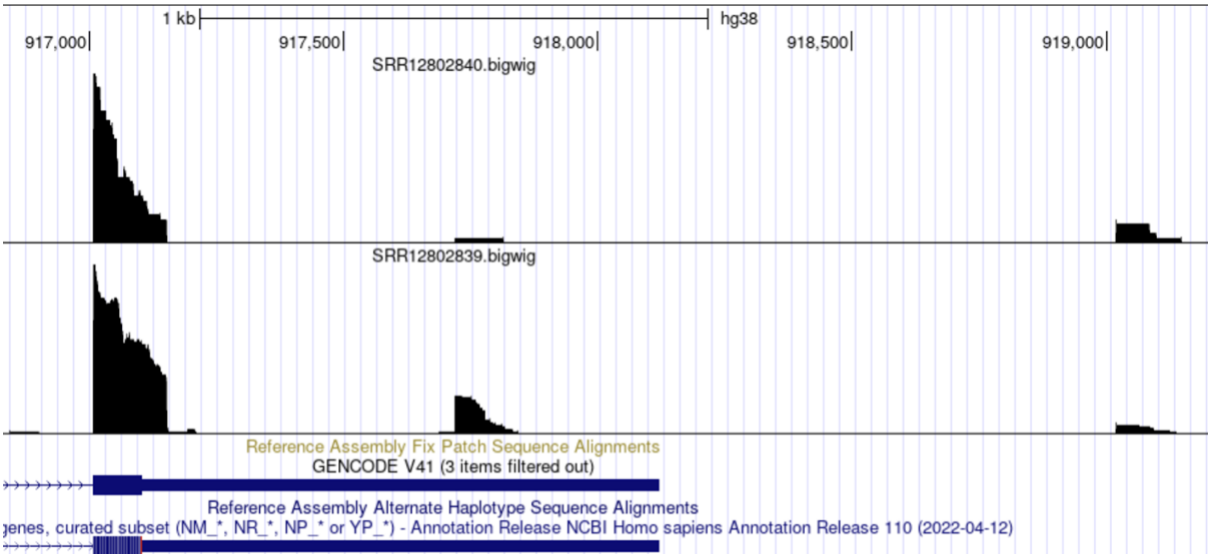


Figure 4.24: Read support in POINT-seq for presence of additional alternative intron within the 3' UTR. This example showcases the less obvious mechanism for NMD where a stop-codon becomes 'premature' due to an additional intron.

Our model predicted this curious example due to the additional prediction of an acceptor at position chr5:917718 in the 3' UTR of the Trip13 gene. This leads to the creation of an intron after the stop-codon which is hypothesised to induce nonsense mediated decay. This is illustrated by category 5 in figure 4.1. While this was initially highlighted by the prediction of our initial splicing neural network however, as previously discussed, it is not currently explicitly modelled by our NMD-model. It does, however, show a concrete example of this mechanism and highlights the need for its inclusion in our predictive framework.

4.6.6 Alternative initial donor site

EIF3K

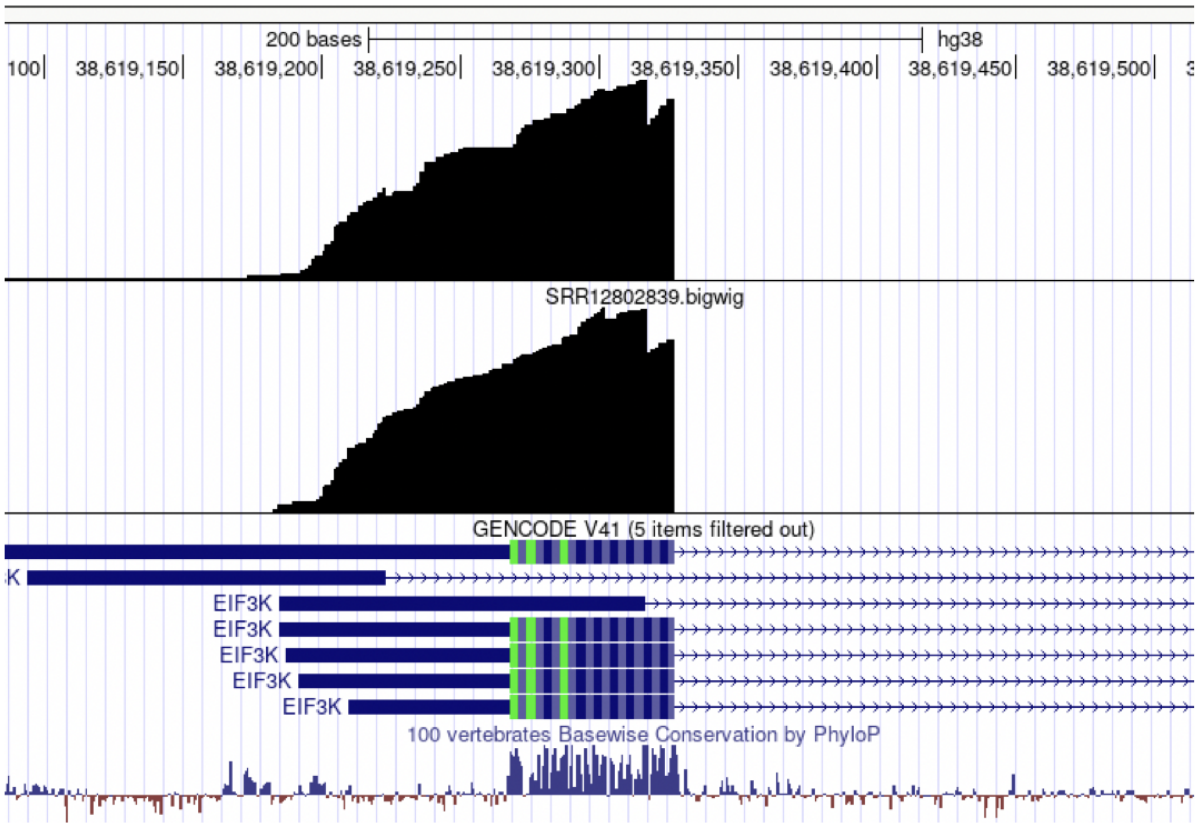


Figure 4.25: Secondary initial donor site found in 19% of transcripts in POINT-seq and strongly predicted by our model. Found in only 2.3% of transcripts in RNA-seq sequence. Despite being annotated in UCSC it is not annotated in any listed coding transcripts.

This example is different to many of the other listed examples, as the concerned splice site is annotated in UCSC as part of a non-coding transcript. It is curious in our setting, as it is an example of a change to the initial donor site giving rise to NMD and additionally reduced expression within RNA-seq libraries.

PRPF38B

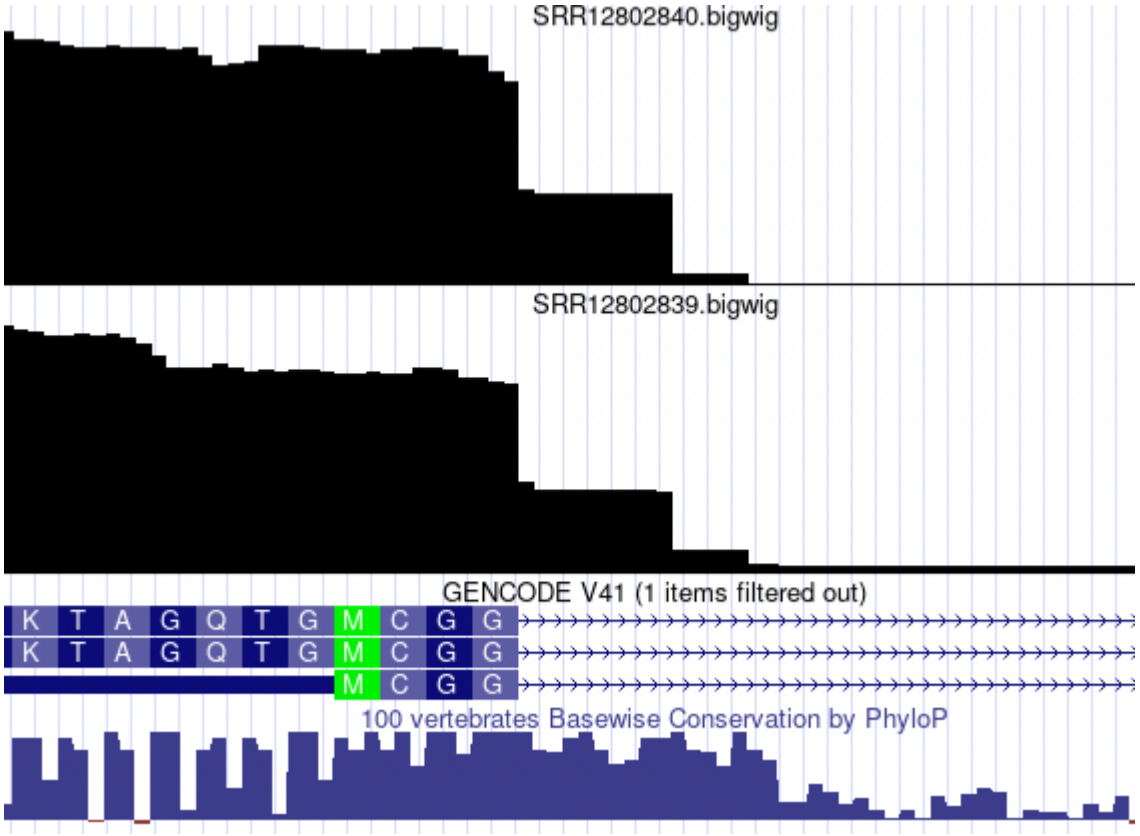


Figure 4.26: This example is an alternative donor site immediately after the canonical donor site. Our model predicts similar inclusion rates (~40%) to what is found in POINT-seq with an additional prediction of NMD.

This is another similar prediction for another splicing-factor coding-gene. We observe high conservation beyond the first annotated donor site. Our model predicts the presence of the alternative donor site and that it would lead to NMD.

4.6.7 Estimates of NMD reducing expression

In the previous section we listed motivating examples of splice sites predicted to be AS-NMD splice sites with varying levels of splice-strength. These are predicted to lead to a range of consequences on the transcript level, from alternative splice-sites and exon-inclusion to intron-inclusions. Each of these transcript-level consequences result in a variety of

alternative transcripts and associated opportunities to induce NMD. Moreover, by varying thresholds we can find many more examples which are in splicing factor genes and are apparent AS-NMD sites.

In this section we make use of the POINT-seq data to provide evidence for two related claims:

1. NMD explains discrepancies between predicted and RNA-seq-based empirical PSI estimates; and
2. The combination of the neural network model and the NMD model identifies true splice sites that result in transcripts targeted by NMD.

In our analysis, we do not have matched cell types with POINT-seq and RNA-seq data due to the novelty of the POINT-seq method and to the subsequent rarity of POINT-seq datasets. This will be rectified by the future generation of matched POINT-seq and mRNA-seq datasets. Such datasets are underway in my host laboratory. In the interim, we mitigate this issue by performing the analysis across multiple co-expressed genes, however, this feature of our data is expected to cause some disagreement with empirical PSIs at many sites.

In section 2.5 we presented our model predictions split by the empirical PSIs. This showed a broad trend where bins of splice sites with high empirical PSIs have high model predictions, alternatively spliced sites are predicted as such but with a broad variance, etc.

We now replicate this analysis but on sites with both RNA-seq and POINT-seq coverage to see if POINT-seq better reflects our predictions than RNA-seq. In section 2.5, the empirical

data may have exaggerated our observation that strong predictions are more constitutive and relatively weak predictions more variably used. Our hypothesis is that this is partly due to biassed representation of canonical transcripts in mRNA-seq. In the schematic figure 4.27 we summarise how the presence of RNA-decay can result in a canonical transcript being more constitutive, whilst splice-sites in decaying transcripts appear as less strong in mRNA-seq datasets.

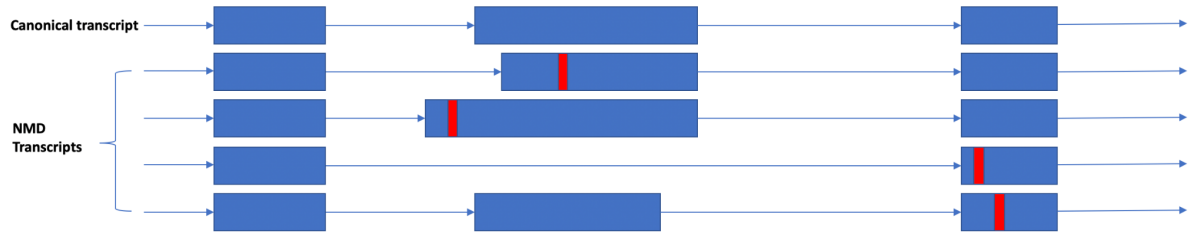


Figure 4.27: Schematic illustration showing how a canonical transcript can appear to be constitutively spliced through the removal of competing NMD-targeted transcripts. Consequently, compared to predictions, we do not only expect reduced usage of splice sites involved in NMD transcripts, but also increased usage of constitutive splice-sites.

Generally, in figures 4.28 and 4.29, we find that predicted low PSI splice-sites in POINT-seq data tend to have even lower PSI in mature RNA-seq data. Similarly, we find that predicted high PSI sites in RNA-seq tend to have even higher PSI in POINT-seq. This suggests that NMD is indeed reducing unannotated splicing transcripts to make canonical transcripts appear more constitutive. Moreover, we attempt to approximate by how much RNA-decay reduces the relative expression of certain splice-sites.

Figure 4.30 suggests a median value for how much decay occurs between the two datasets. We took splice-sites found in both datasets with at least 5 reads supporting them. Neural network-predicted minor transcripts have increased relative expression 2.5 times higher in POINT-seq compared to RNA-seq, whilst neural network-predicted canonical transcripts have the relative expression of alternative transcripts in RNA-seq 2.8 times lower than in POINT-seq. This shows how removal of alternative transcripts makes near-constitutive splice sites appear almost entirely constitutive. Our hypothesis is that this difference is due to NMD following the illustration in figure 4.27.

We present the same data as a scatter plot in figure 4.31. Although noisy, there is a trend that the relative expression of relatively rare transcripts is less in RNA-seq than in POINT-seq. The ratio of the two has a median of 0.4, which indicates a substantial average depletion.

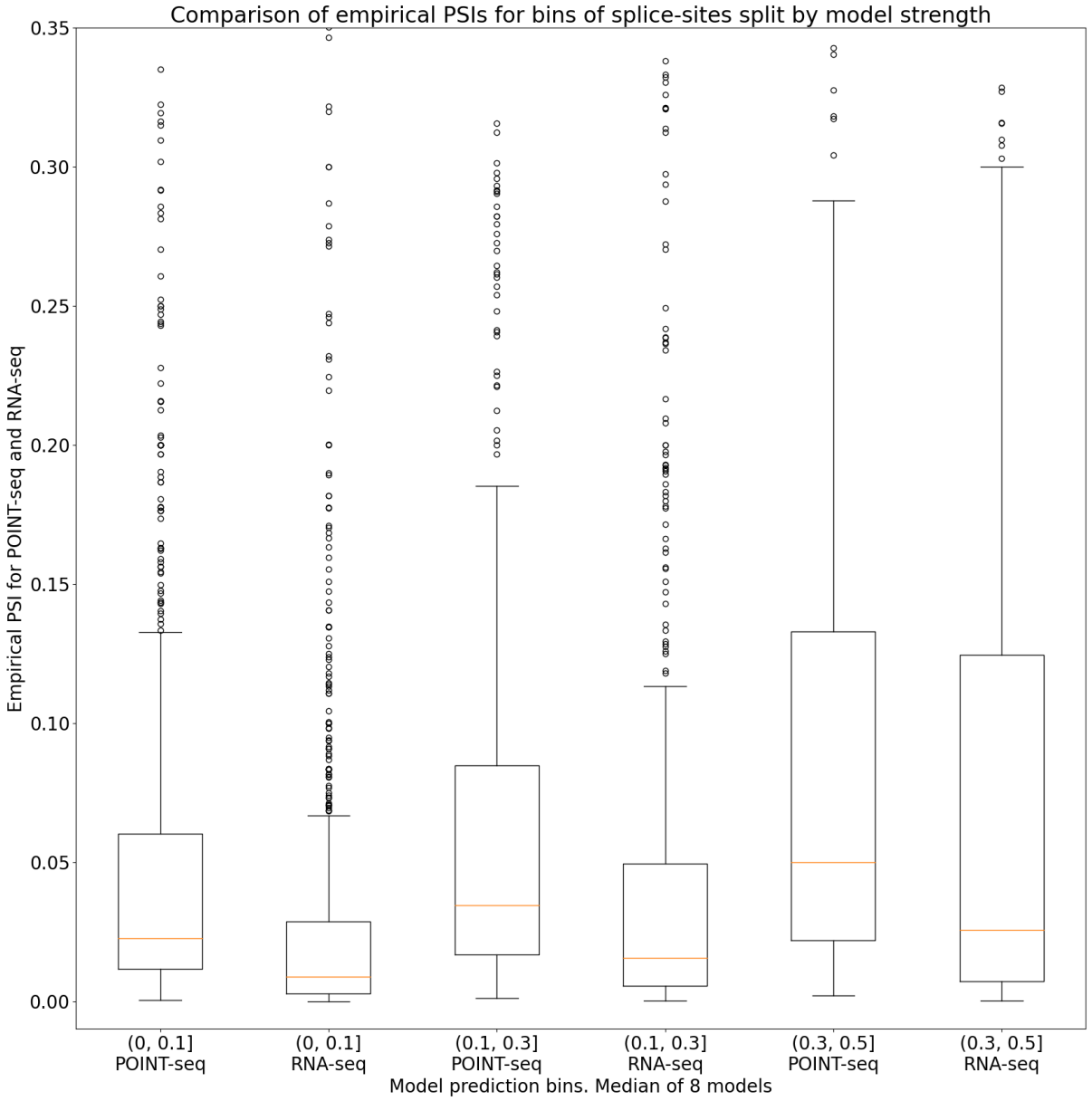


Figure 4.28: Empirical RNA-seq and POINT-seq PSI by model prediction, for low predicted splice rates. Predicted PSI are fairly consistent with PSI measured by POINT-seq, but PSI measured by RNA-seq show a strong downward bias, consistent with the action of NMD. This analysis taken over splice-sites predicted by our model and evidenced in both datasets (>= 2 reads support).

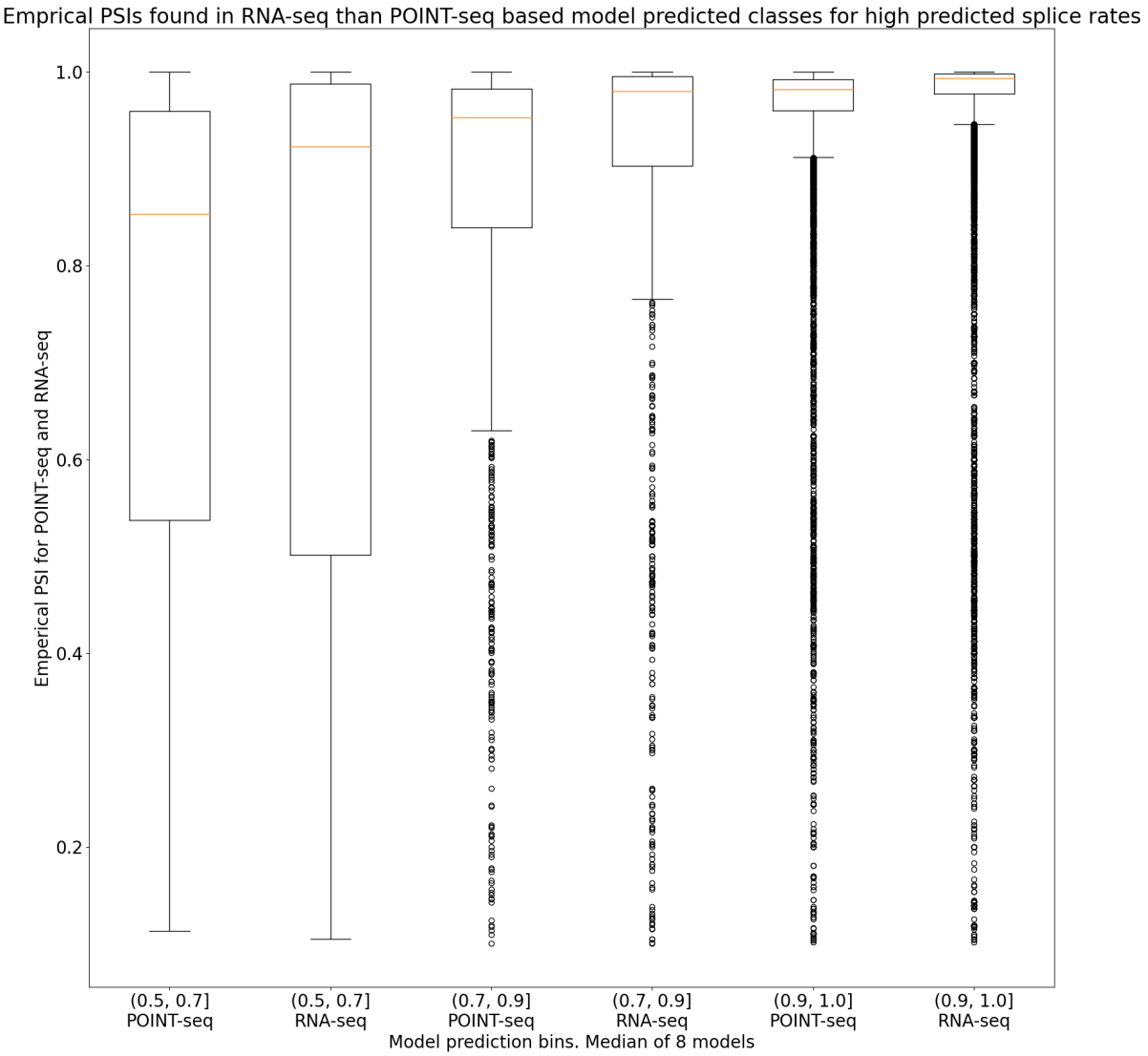


Figure 4.29: Empirical RNA-seq and POINT-seq PSI by model prediction, for high predicted splice rates. Predicted PSI are fairly consistent with PSI measured by POINT-seq, but PSI measured by RNA-seq show an upward bias, consistent with the action of NMD removing other transcripts. This analysis taken over splice-sites predicted by our model and evidenced in both datasets (>= 2 reads support).

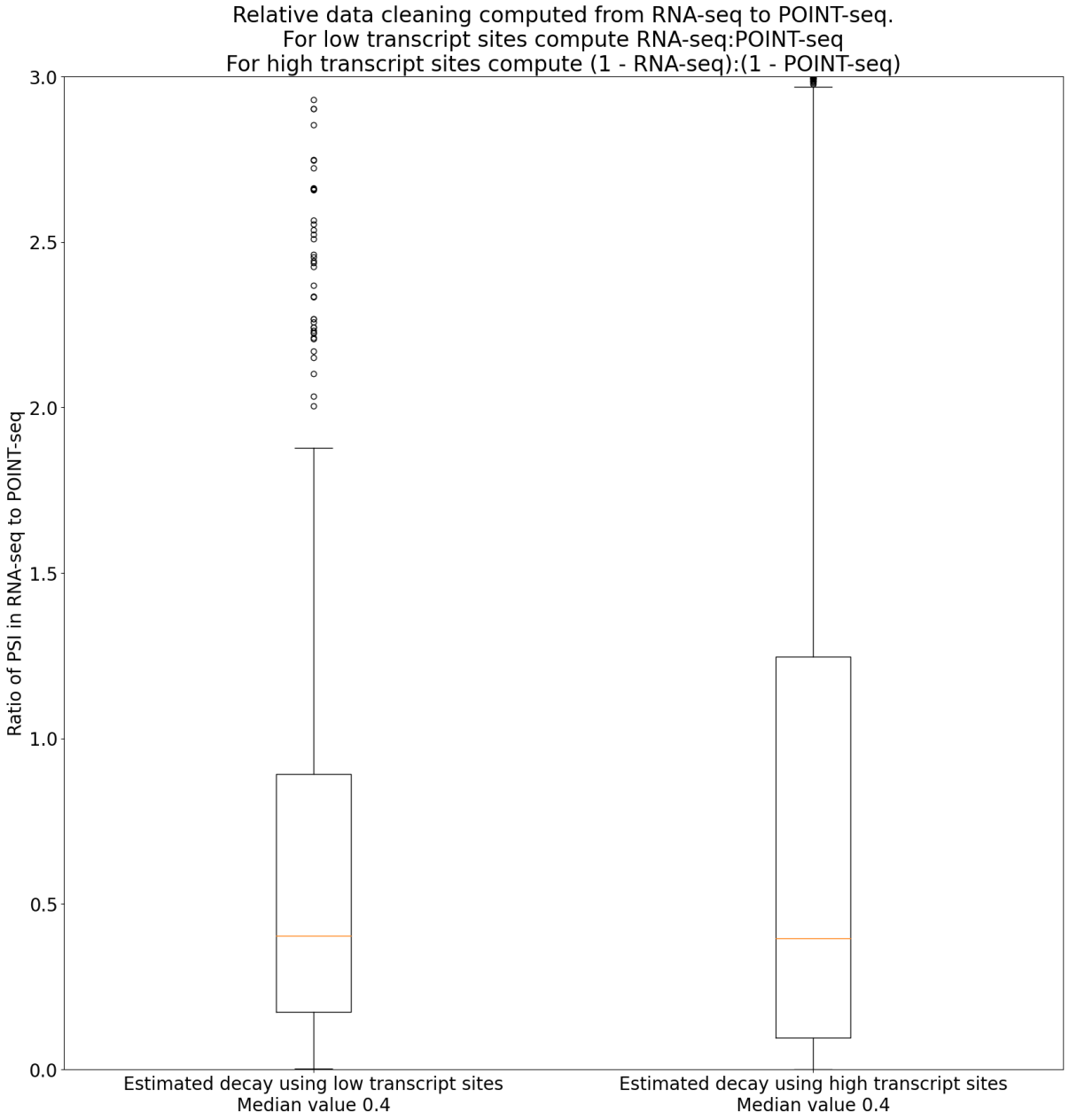


Figure 4.30: Estimates for the overall rate of downward bias for low predicted splice sites and the rate of upward bias for high predicted splice sites. Compute the ratio of RNA-seq:POINT-seq for low predicted splice sites. A ratio below 1 is consistent with the action of NMD. Compute (1-RNA-seq):(1-POINT-seq) for the high predicted splice sites. A ratio below 1 is consistent with the action of NMD degrading other transcripts.

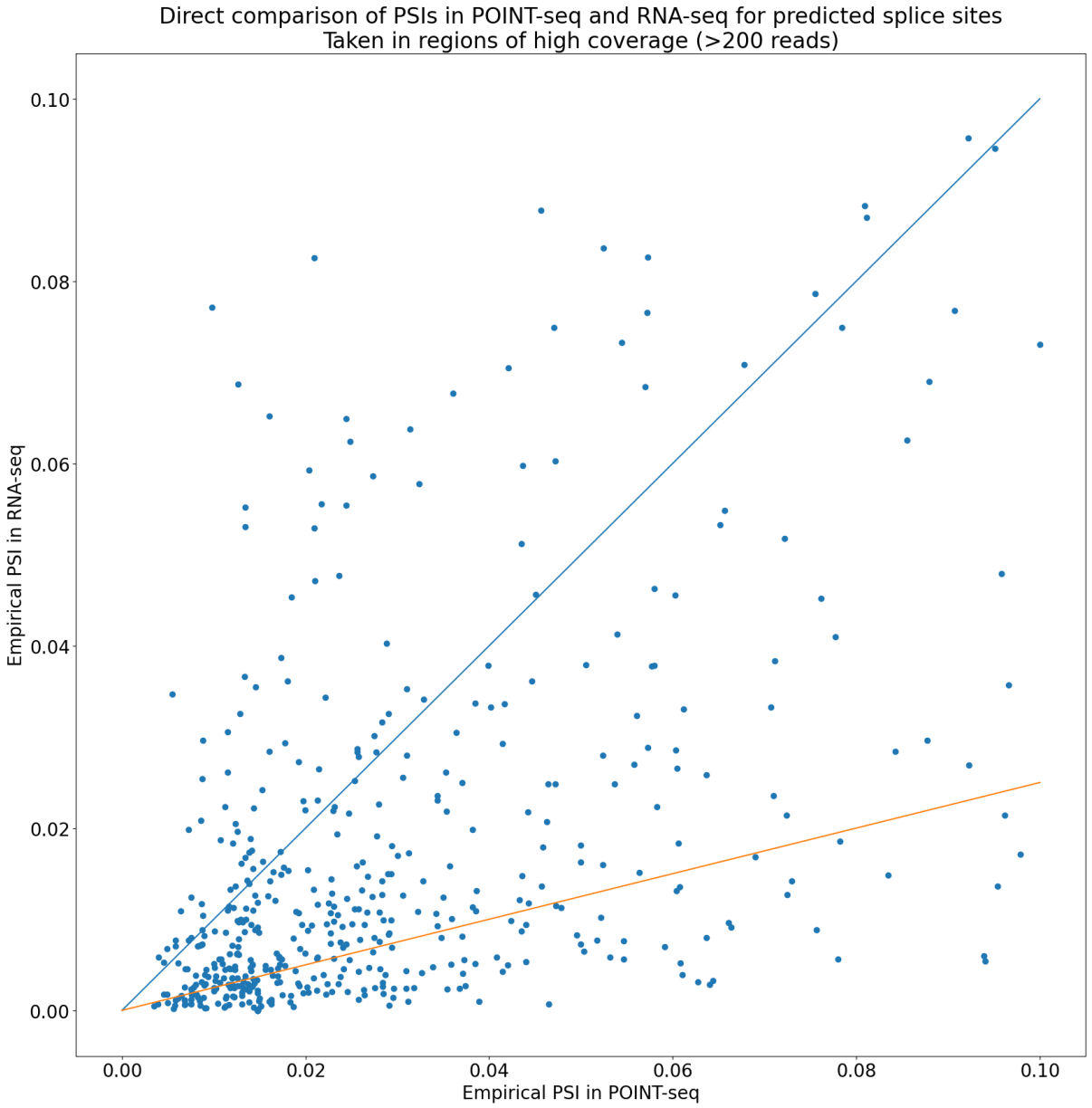


Figure 4.31: Scatter plot comparing empirical RNA-seq and POINT-seq PSIs for low predicted splice rates. It is possible that there are two populations with one where POINT-seq and RNA-seq are roughly equal and another where POINT-seq is roughly 3 times RNA-seq. The latter could reflect transcripts which are subject to RNA-decay.

4.6.8 Application of NMD model to POINT-seq

We next applied the NMD model to our neural network predictions to analyse their differing behaviours in POINT-seq and RNA-seq data. The data is split into three categories, namely:

1. Predicted NMD
2. Predicted no-NMD
3. Unsure NMD status

The third category arises due to either our NMD-model not being able to determine the alternative transcript, or it providing multiple alternative NMD predictions.

When the NMD model predicts that a transcript does not undergo NMD, the ratio of RNA-seq PSI to POINT-seq PSI has an interquartile range [0.999, 1.027] with a median of 1.008. In contrast, when the NMD model does predict that a transcript undergoes NMD, the same ratio has an interquartile range [0.142, 0.942] with a median ratio of 0.381. This means that when no NMD is predicted the empirical PSIs are broadly comparable between the two technologies, but that when NMD is predicted there is a marked reduction of PSIs in RNA-seq compared to POINT-seq. When our NMD pipeline is unclear whether NMD or no-NMD should be predicted, the ratios are clustered around 1.007 with an interquartile range [0.993, 1.030]. This means that the uncertain NMD category is clustering well with the no-NMD category.

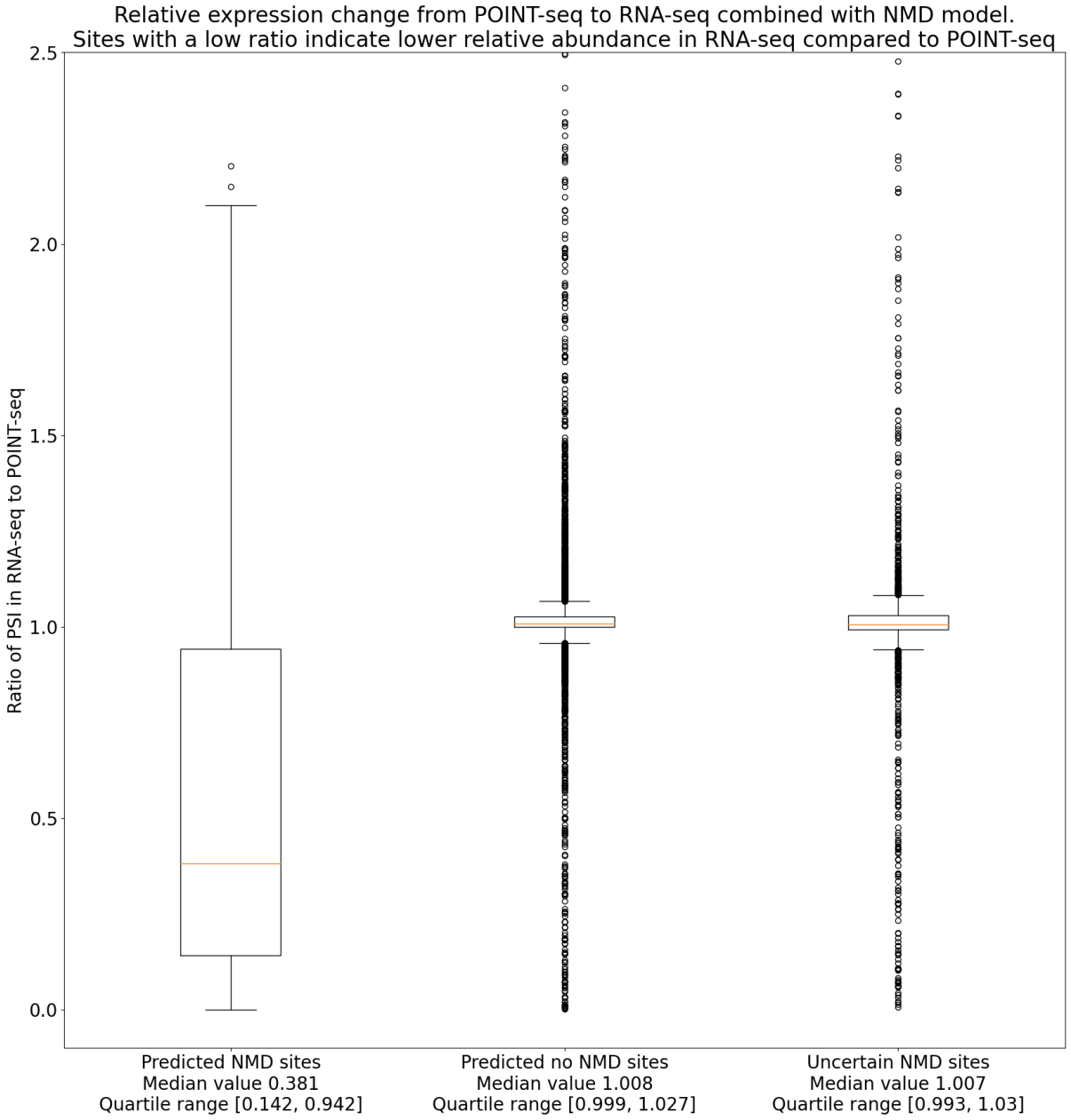


Figure 4.32: NMD model applied to sites with POINT-seq PSI >= 0.05. When reapplying the bioinformatic NMD model to our POINT-seq vs RNA-seq data the dataset separates with predicted NMD sites showing significant loss of expression in RNA seq. Sites not predicted as NMD have a ratio centred around 1. Sites predicted NMD are typically found to have a ratio greater than 1.

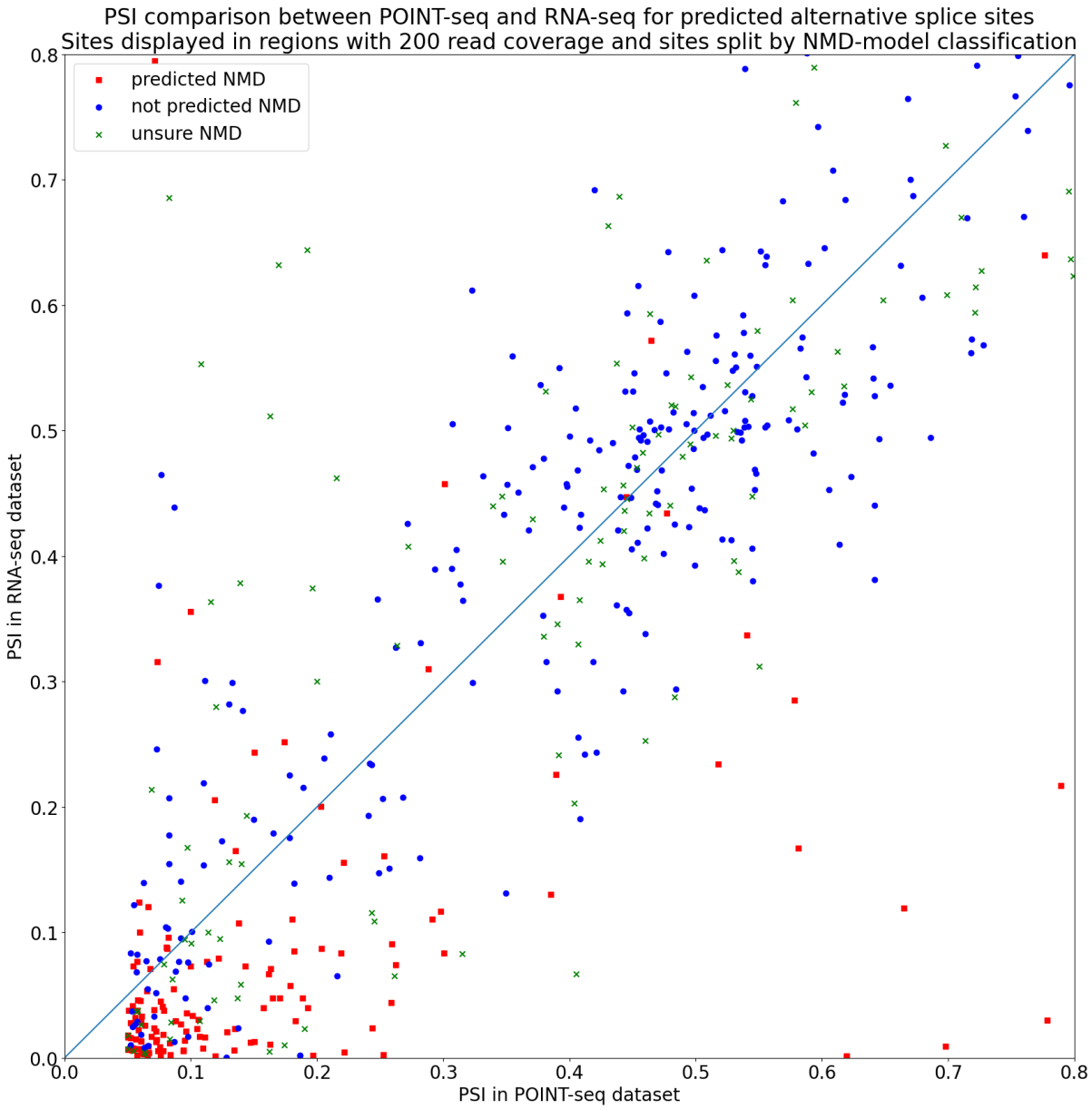


Figure 4.33: The same data as figure 4.32 (in regions of high expression) plotted as a scatter plot shows that despite these overall trends there is a large degree of variation in the level of loss of relative expression. Uncovered within this plot are some exceptions to the overall NMD trend. Some outliers would be explained by a more sophisticated NMD model which took into account combinations of alternative splicing.

Figures 4.32 and 4.33 are based upon analysing the evidence-predicted splice sites (in RNA-seq and POINT-seq) partitioned by our NMD model. There are a few outliers which are

due to our simple NMD model not assuming an exon-inclusion is potentially a multi-exon event – some alternative exons are included in tandem. We have additionally found outliers which are due to our model not handling gain of intron events perfectly. This shows that, in order to cover these classes of splicing outcomes, more sophisticated NMD models, informed by these validation results, will be required.

However, despite this, figure 4.32 shows that, when NMD is not predicted, there are equivalent PSIs found between the two technologies. However, when the NMD model picks out a splice site as inducing NMD (our analysis finds another matching splice site to form this analysis), there is a large difference between the datasets. This confirms that the model has predictive power to identify NMD inducing splicing events.

The majority of the splice sites in the above plot are constitutive splice sites. 10688 of 12653 the analysed sites have a PSI in POINT-seq greater than 0.9. For practical reasons of sensitivity and noise, we excluded analysis on sites with low PSIs <= 0.05. Of the 1965 alternatively spliced sites, 787 have a ratio (POINT-seq:RNA-seq) consistent with RNA-decay, with 381 confidently identified by the NMD-model. Of the total analysed splice-sites this is a relatively small figure due the abundance of constitutive splice sites. However, when considered from the point of view of gene output, of the 386 analysed genes these 381 AS-NMD sites are found within 173 genes. This is a total of 45% of all genes in the gene-set, suggesting that a large proportion of transcription is subject to NMD. We note that this relatively high percentage of genes is obtained despite requiring at least 5% PSI in POINT-seq and, therefore, we have not obtained this high gene percentage through the inclusion of rare transcripts.

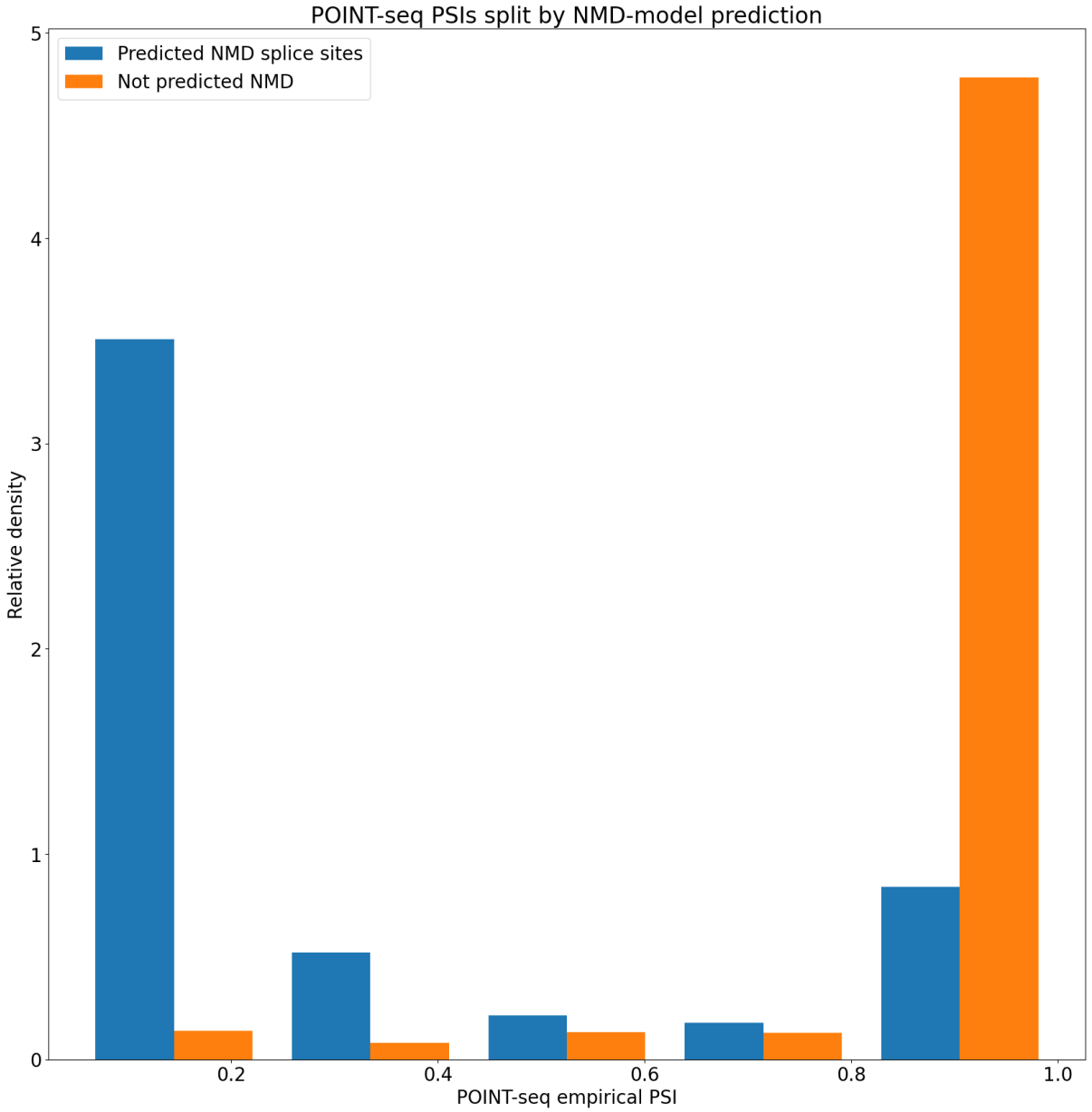


Figure 4.34: Relative POINT-seq PSI distributions based upon NMD prediction. Amongst selected splice sites (empirical PSI >= 0.05) there is greater weighting towards low PSIs even in POINT-seq. This suggests that a mixture of the two forms of AS-NMD are present with strong regulatory AS-NMD sites forming a small but significant proportion of AS-NMD sites.

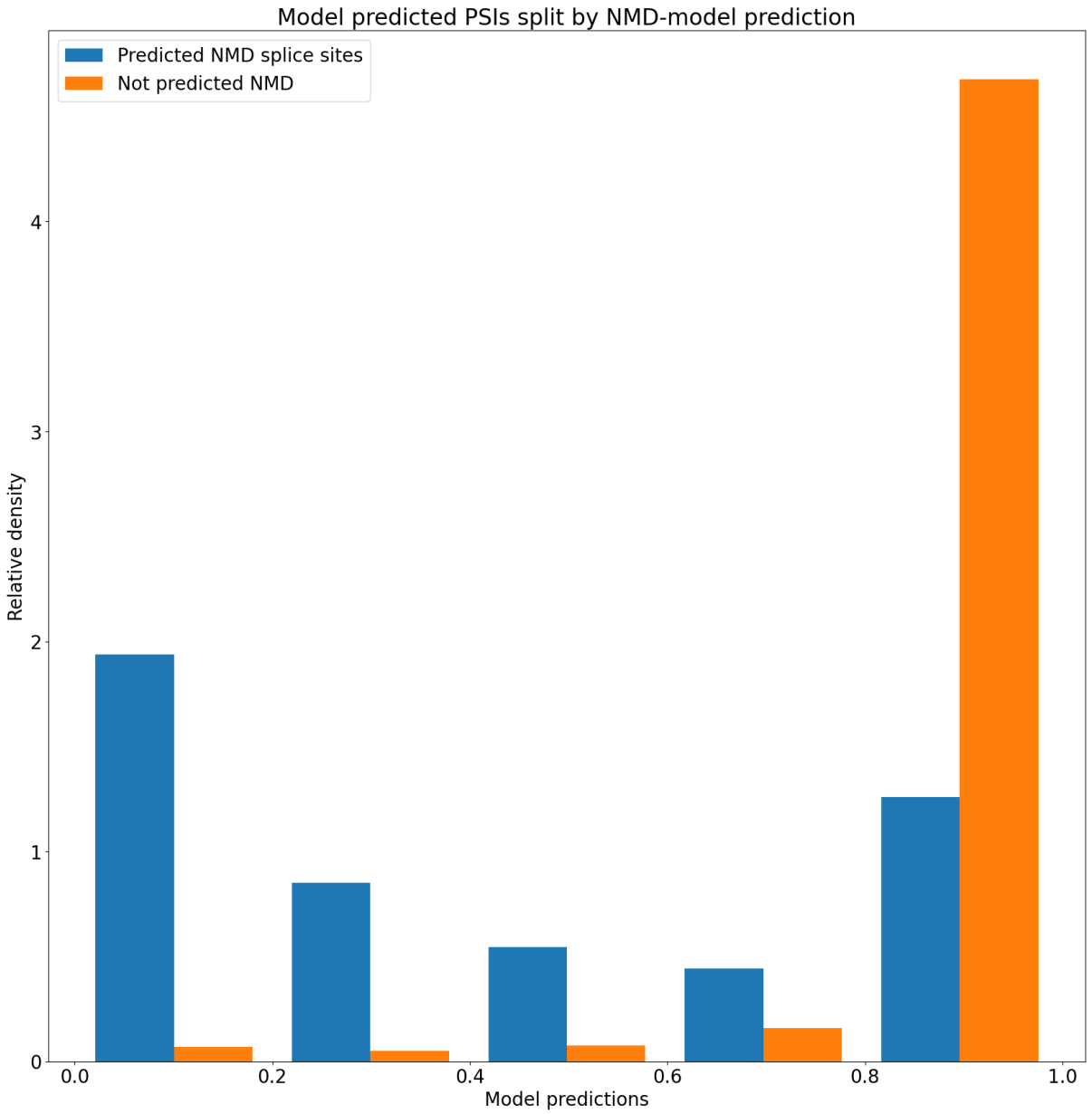


Figure 4.35: Relative model predicted PSI distributions based upon NMD prediction Similar analysis to figure 4.34 but we instead plot the PSI distributions based upon our neural network predictions. Neural network model predicts the same phenomena: our model is more likely to predict a NMD-splice site as having a low PSI.

We next ask whether AS-NMD sites follow the same PSI distribution as non-NMD AS sites. For instance, how much are they clustered towards low PSIs? In figures 4.34 and 4.35 we

measure this in two different manners, with 4.34 using the empirical POINT-seq PSIs and 4.35 using the neural network predicted PSIs. Sites predicted to not induce NMD are skewed towards high PSI scores in both POINT-seq and by our model predictions. Conversely, there is a different distribution for sites predicted to induce NMD with a larger percentage found in POINT-seq to be alternatively spliced.

Finally, we attempt to estimate the effect on the overall false-positive rates. This is defined as the percentage amongst a class of predictions which have an empirical PSI <= threshold. Amongst predicted sites where our model predictions are >= 0.9, we have:

Empirical threshold	RNA-seq FPR % (N)	POINT-seq FPR % (N)
0.01	0.97% (73)	0.6% (45)
0.05	1.57% (119)	0.9% (68)
0.1	1.89% (143)	1.06% (80)
0.5	4.8% (243)	3.23% (144)
0.9	10.49% (793)	10.02% (758)
0.95	16.69% (1262)	20.14% (1523)

Table 4.36: We compute what percentage of predictions made with threshold 0.9 have an empirical PSI below different thresholds. In RNA-seq datasets nearly 1% of strongly made predictions are found with a PSI less than 1%. In POINT-seq this is reduced to 0.6%. We also note that fewer POINT-seq has fewer constitutive sites amongst strongly predicted sites. 20% of predicted sites have an empirical PSI <= 95% compared to 17% in RNA-seq.

We see from table 4.36 that, for low thresholds, POINT-seq reduces our model's false positive rate. However, we slightly increase our false positive rate when restricted to

empirically constitutive sites. This is consistent with the discussed dual action of NMD in reducing in RNA-seq the relative expression of lowly expressed sites whilst increasing the expression of highly expressed sites (see figure 4.30). Our model by no means becomes perfect when analysed against POINT-seq data – there are still 0.6% of strong predictions which have a PSI <= 0.01 in POINT-seq. However, for each of the smaller thresholds our approach is highly effective at removing apparent false positives (~40%).

In summary, we find that POINT-seq and RNA-seq broadly provide the same gene-definitions amongst the genes co-expressed in both datasets and in the absence of NMD prediction from our models. When NMD is predicted, there is a significant increase in relative expression found within the POINT-seq dataset. This shows that our NMD-model is correctly segregating POINT-seq and RNA-seq differences, and that the simplest solution is that these differences have arisen due to NMD. While performance is still not perfect, an obvious route to further improving this performance is to explicitly train our models on POINT-seq data, when sufficient datasets are generated.

4.6.9 Analysis of other experimental approaches to detect NMD associated splicing events.

For POINT-seq we found that our NMD-model broadly reflected the ratio of PSIs to RNA-seq. In this section we briefly compare to contrasting methods which have been used to reveal NMD-transcripts. Kovalak et al [8] provide a method to reveal nascent transcripts using RNA immunoprecipitation of exon junction complex components. The goal was to reveal AS-NMD splice sites by capturing nascent splicing events prior to RNA-decay. Karousis et al [43] use an alternative approach by applying chemical inhibitors of critical NMD proteins to block NMD and reveal AS-NMD transcripts. In addition to our reanalysis of the previously described

POINT-seq data, during the course of this thesis we also reanalysed these datasets to judge their utility in detecting NMD inducing splicing events.

4.6.10 Kovalev et al Exon Junction Complex analysis

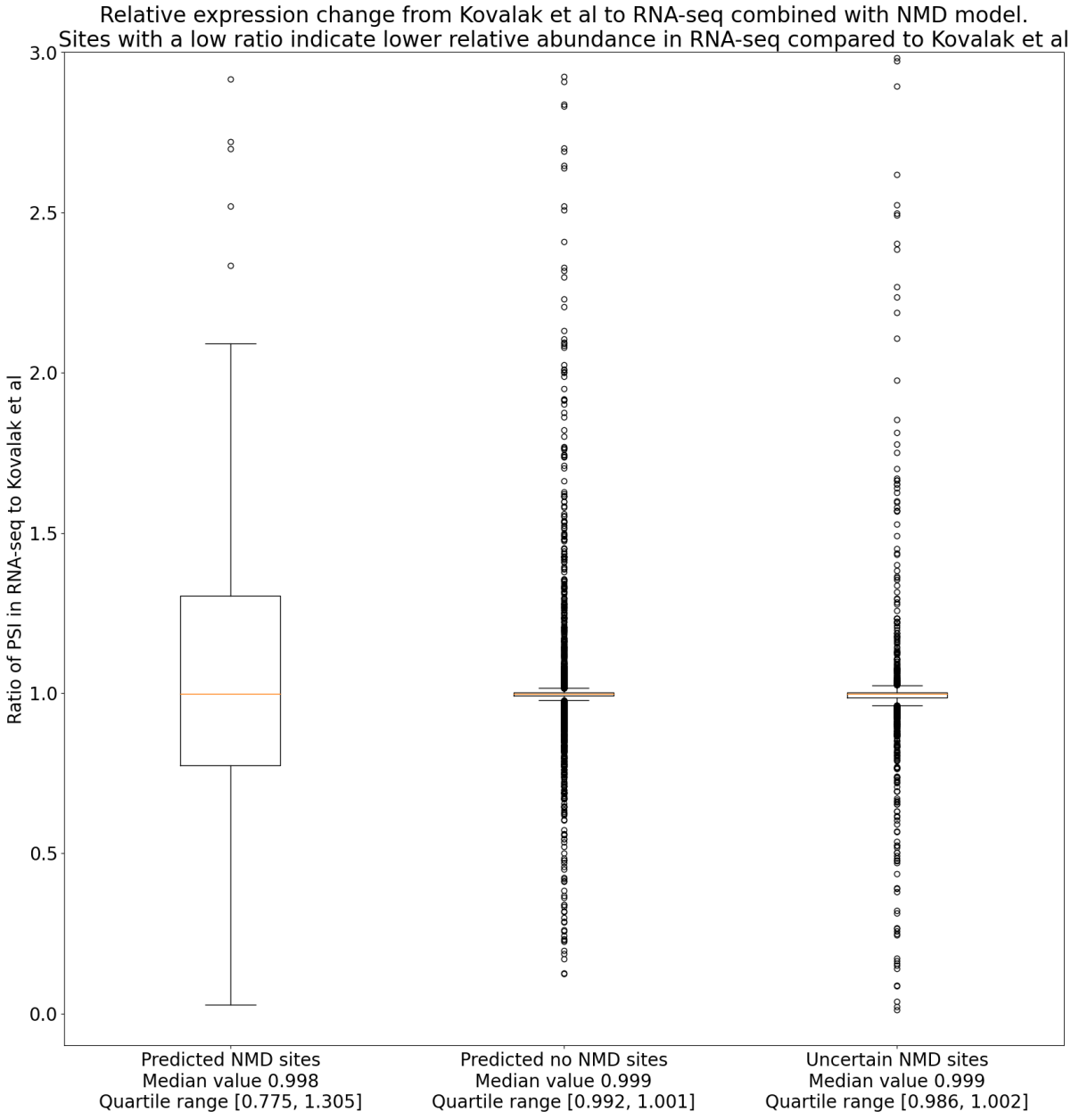


Figure 4.37: NMD model applied to sites with Kovalak et al PSI >= 0.05. Reapplying the bioinformatic NMD model to our Kovalak et al vs RNA-seq data the dataset shows no separation between the classes. All classes have empirical ratios centred around 1.0.

The most informative analysis aims to determine whether we believe that these methods are capable of detecting the NMD transcripts, which are predicted by our model in a manner analogous to POINT-seq.

However, in the data from both these methods we find no evidence to suggest that NMD transcripts have increased representation in their datasets as compared to standard RNA-seq datasets. Unlike for POINT-seq, where for our predictions combined with our NMD model we find a clear separation, for the Kovalak et al (Figure 4.37) and Karousis (Figures 4.37 & 4.38) datasets we do not find any separation. It is unclear why these approaches are not as sensitive at detecting such nascent splicing events, however, from these analyses we empirically decided to concentrate on the POINT-seq methods to detect and validate our predictions going forward.

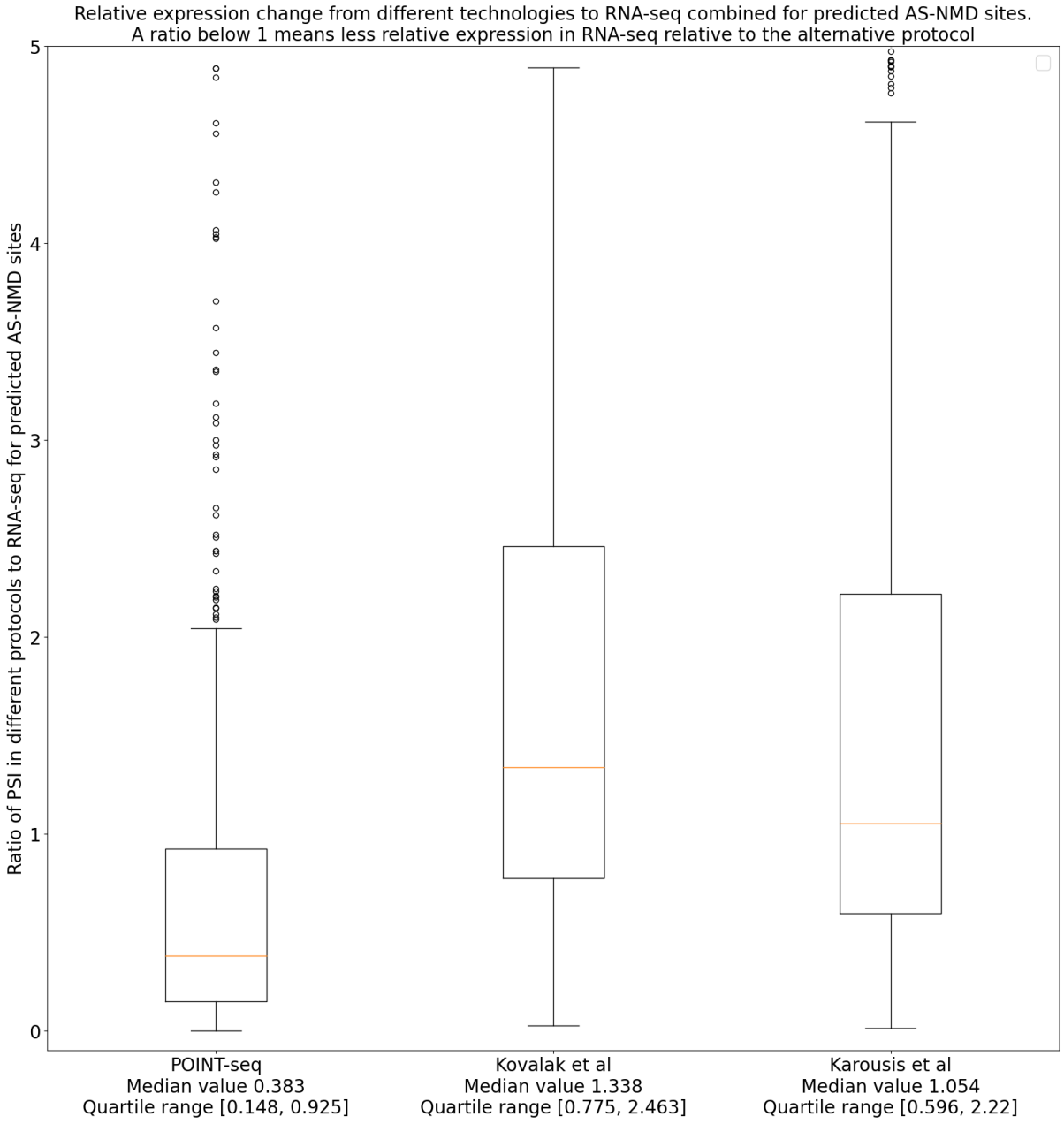


Figure 4.38: Ratio analysis of empirical PSIs applied to predicted AS-NMD sites. Sites are both predicted by the neural network and NMD-model. In addition restricted to splice-sites where there is read coverage (>=50) for all protocols. Under this methodology POINT-seq appears to be best reflecting nascent transcription.

4.7 Discussion

Despite being trained on mature RNA-seq which is depleted of transcripts that undergo Nonsense Mediated Decay (NMD), the neural network predicts from genome sequence many sites at elevated PSI compared to RNA-seq data, at loci where splicing would result in NMD. This suggests that the model may be picking up genuine but unannotated splicing signals. Biases in in-frame and out-of-frame splice sites are consistent with about 20% of predictions without sequence evidence being, in fact, true splice sites, and resulting in transcripts undergoing NMD and therefore not present in mRNA-seq data. Direct evidence from POINT-seq data is consistent with this interpretation. We built a bioinformatic NMD model, which allowed us to refine our predictions, and in particular to make predictions on the effect of a mutation on the final transcript. We observe that, when we predict that NMD is not involved with the usage of alternate splice sites, there is good agreement between empirical PSIs measured from the POINT-seq and RNA-seq datasets, and that the ratio of the computed PSIs between the two datasets is concentrated at 1. However, when NMD is predicted to affect the resulting alternate transcript, there is a marked difference between the empirical PSIs obtained from the two experimental protocols. The PSI determined from POINT-seq data is typically ~2.5-fold higher than PSI determined from RNA-seq data. Additionally, we find that in these genes, near-constitutive splice sites appear to be *more* constitutive in RNA-seq datasets than they do in POINT-seq data. We interpret this as the result of reduced “competition” from alternate transcripts in RNA-seq data in the calculation of PSI, because transcripts eventually undergoing NMD are not visible in these data. From this we can conclude that, indeed, NMD splice-sites are fairly common within many gene bodies (found in 44% of analysed genes at PSI level of >= 5%), and that they can be partially separated by using downstream bioinformatic algorithms subsequent to the neural network.

Additionally, we saw examples of genes for which NMD-transcripts are nearly canonical, by which we mean that in these examples we have datasets where the most prevalent transcript is subject to NMD and not readily seen in RNA-seq. Consequently, even a model trained to model-canonical transcripts (e.g. *SpliceAI*) will be limited due to ignoring these transcripts. Our contention is that any neural network should attempt to model the nascent transcripts, as these RNA-decay mechanisms are orthogonal in their nature. In short, the spliceosome has no knowledge of the reading frame. An effective model should model both splicing and the effect of splicing on the production of stable RNA.

An alternative approach could be to build a training library that utilises POINT-seq to enrich the GTEx dataset. This could either work by not penalising the model on AS-NMD sites, or by adjusting the values at those sites. A final approach could be to additionally train our model on a POINT-seq dataset to retrain existing parameters. POINT-seq, by design, picks out co-transcriptional splicing, which means it could miss splicing signals to the extent that non co-transcriptional splicing occurs. When building training datasets this should be considered carefully, which is why we suggest an approach of enriching the datasets. As there are no shortage of splice sites to train a neural network, an approach could be to use POINT-seq as a means to label sites in which the loss function does not penalise the network in the training stages. Essentially, AS-NMD sites would become free learning sites.

Returning to the question concerning the significance of AS-NMD, our analysis on POINT-seq, along with the NMD model, suggest that, many AS-NMD sites are strong splice-sites, they are more heavily weighted towards weaker splice-sites than non-NMD alternative

splice-sites. This is not inconsistent with a model in which they are mostly aberrant splice-sites being removed by selection. However, what is interesting is that they still amount to a significant proportion of the alternative splice-sites at ~40%.

It could be argued that this is placing too much weight on a small percentage of total splice-sites. Most splice-sites with empirical PSIs > 10% are additionally found in RNA-seq. However, it is precisely these alternatively splice sites that are so difficult to model. We recall that in our first chapter splice sites which are empirically constitutive are invariably predicted as such. The complexity lies in discerning the additional features which lead to a splice site being alternatively spliced. Therefore, there is great value in these empirical PSIs being as accurate as possible, and we believe that there is a clear issue with upwards of 20% being missing. In addition to this, we saw how POINT-seq typically had reduced PSIs for near-constitutive splice sites. The increased presence of many minor AS-NMD splice sites leads to reduced splicing at near-constitutive splice sites. Inaccurate modelling of the minor splice sites causes inaccurate modelling the major splice sites.

We saw that our bioinformatic model indeed picks out AS-NMD sites as defined by the POINT-seq data when compared to RNA-seq. NMD is not the only known type of RNA-decay, and indeed our pipeline does not account for all methods of inducing NMD. An enriched POINT-seq dataset would not exhibit a bias to only considering NMD and it would not be limited, as we are, to certain regions where NMD is computable.

5 Application to common human variation in genetics

To assess whether our model can be useful to interpret the impact of common variants on gene expression through their impact on splicing, we applied the model to 1000 Genome SNPs and assessed the predictions using RNA-seq data available for our Oxford Biobank of 50 samples. Each sample was sequenced using 10x technology with its reads aligned using Long Ranger (version 2.2.2) [44]. Each of these samples have their variants called using octopus v0.7.0 [45].

These datasets are used to determine the quality of the model's predictions of the effect of mutations. In this section we have restricted our analysis to common human variation, as these variants are typically found in multiple samples. This enables us to obtain enough read coverage across multiple samples to confirm the effect of the mutations.

5.1 Read evidence supporting

Read evidence is computed from the RNA-seq data using the pysam 'find_introns' method (version 0.20.0 pysam.readthedocs.io). This method extracts reads which have a gap in alignment produced by the STAR aligner (version 2.7.3a) [46]. For a putative splice-site we compare the number of reads supporting the splice site compared to the total number of overlapping reads.

To assess the quality of the model we consider the read evidence for different predicted outcomes (alternative acceptor, alternative donor, exon-skipping, and exon-gain). We further split these outcomes into categories based on whether we would expect NMD, no NMD, or no definite prediction as to whether NMD is expected or not ("unsure"). The latter would

happen for instance when an exon could be part of multiple transcripts leading to different reading frames. The splitting of our predictions into these categories uses the bioinformatic model outlined in section 4.4 of the previous chapter.

In a similar manner to previous sections on NMD, we test the hypothesis that NMD and NMD-related signatures explain some of the lack of read evidence. For example, for alternative-splice-site SNPs we ask the following questions:

1. Given that a wildtype splice-site is supported and the model predicts an alternative splice-site is used: do we see evidence for this alternative acceptor in the samples with the variant?
2. Is the presence or absence of evidence consistent with frame-shifts (for alternative acceptors)?
3. Is the presence or absence of evidence consistent with predictions from the NMD model?
4. Is the extent that the alternative splice-site is not found consistent with our previous estimates for NMD-depletion?

The aim of these analyses is to understand the potential for the model to explain the molecular phenotype of the variant and any genetics associated with it.
Often acceptor-switching mutations are within the vicinity of a known acceptor which is being lost. A sensible approach would therefore be to look at the relative change in PSI for that acceptor. If the alternate acceptor leads to NMD, then this might be missing in the RNA-seq datasets, thus leading to the potentially incorrect conclusion that no splicing

alterations are in fact occurring. The NMD model has the potential to turn that conclusion on its head – it tells us that observing no read evidence supporting the alternate acceptor is, in fact, consistent with a splicing change. We could therefore conclude that absence of evidence is not evidence of absence in this case, and that additional evidence should be sought.

In the case of alternative splice-sites we condition on the known presence of the wildtype splice-site. The reason for this restriction is that it allows us to use the gene-annotations with confidence, thus confirming that the predicted lost splice-site is actually real.

We ask a similar set of questions for the loss and gain of exons: the presence of read evidence is correlated with whether we predict the presence of NMD or not. However, this evidence of the model's correctness is less strong, as whether a site is predicted to have NMD is very dependent on the gene-annotations in this case.

5.2 Results of NMD suppression

We consider four separate classes to assess our predictions when both models are joined together: the raw neural network model combined with the NMD analysis.

The four classes are the following:

1. Model predicts switch of acceptor site.
2. Model predicts switch of donor site.
3. Model predicts loss of exon
4. Model predicts gain of exon

5.2.1 Alternative acceptor site results

Without the presence of an RNA-decay mechanism, one would expect that a loss in PSI for the original acceptor would coincide with a gain of PSI for the novel acceptor. We find this is indeed the case, and in figure 5.2 we observe that the read data aligns adequately with the neural network predictions when NMD is not predicted. Additionally, the empirical PSI predictions broadly reflect the predicted PSI predictions for these alternative acceptors.

There is a potential concern of bias due to the NMD predictor. It utilises existing transcript information thereby it could be biased towards known splice-sites, as some of the gained splice sites are already annotated. To avoid this, we reanalyse the alternative acceptor case, where we had sufficient examples, by splitting into categories independent of gene annotations.

From a switch of predicted alternative acceptor sites, we can split into 3 categories: frameshift (due to either gain or loss of sequence); gain of sequence without frameshift; and loss of sequence without frameshift. Only in the last category we would expect NMD to be impossible, as the other two categories provide opportunities for stop-codons to arise.

Figure 5.1 shows that, when all sites are analysed *en masse,* the empirical evidence does not match the predictions well. However, when each category is analysed separately, only in the category where we explicitly do not expect NMD (loss of sequence and no frame shift) the empirical evidence reflects the predictions well. Therefore, the empirical evidence for splicing matches well when there is no frameshift and no NMD, but matches poorly when there is a frameshift and so NMD is predicted. As all the technical biases for the detection of

both frameshifting and non-frameshifting exons in RNA-seq are the same, this suggests that there is a systematic reason for this loss of evidence for frameshifting exons, which we would suggest is evidence for the induction of NMD.

Figure 5.2 shows that the analysis from 5.1 can be well summarised by the NMD model. Predicted no-NMD sites have read evidence, as expected by the neural network predictions. Predicted NMD sites have limited read evidence. When the model is unsure there is a spectrum between the two scenarios.

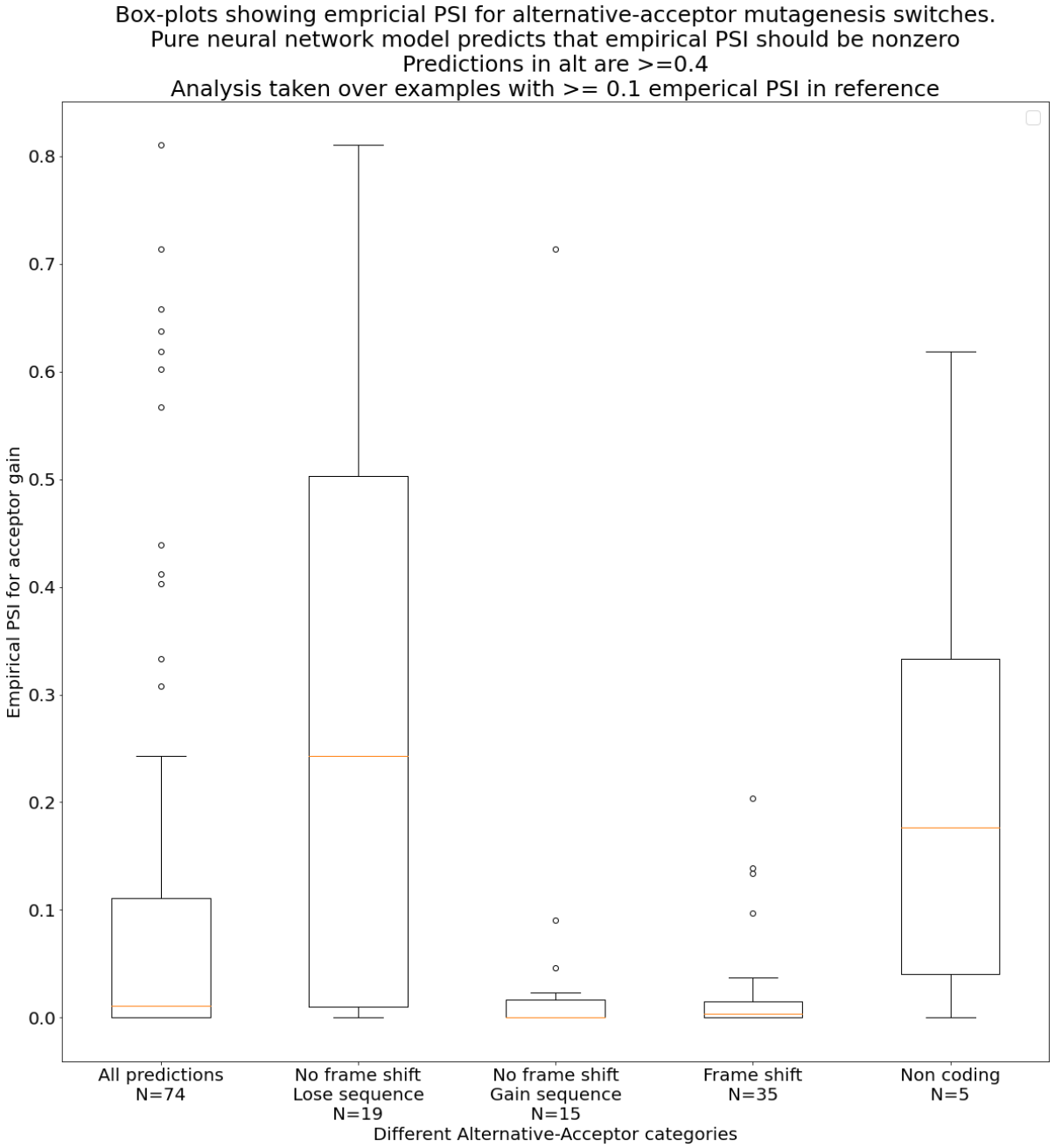


Figure 5.1. Empirical PSI levels for predicted gained alternative acceptors as found in our RNA-seq biobank. Predicted alternative acceptor sites split by relative frame information. When the alternative acceptor results in a shorter exon with no frame shift we observe a high empirical PSI in our biobank.

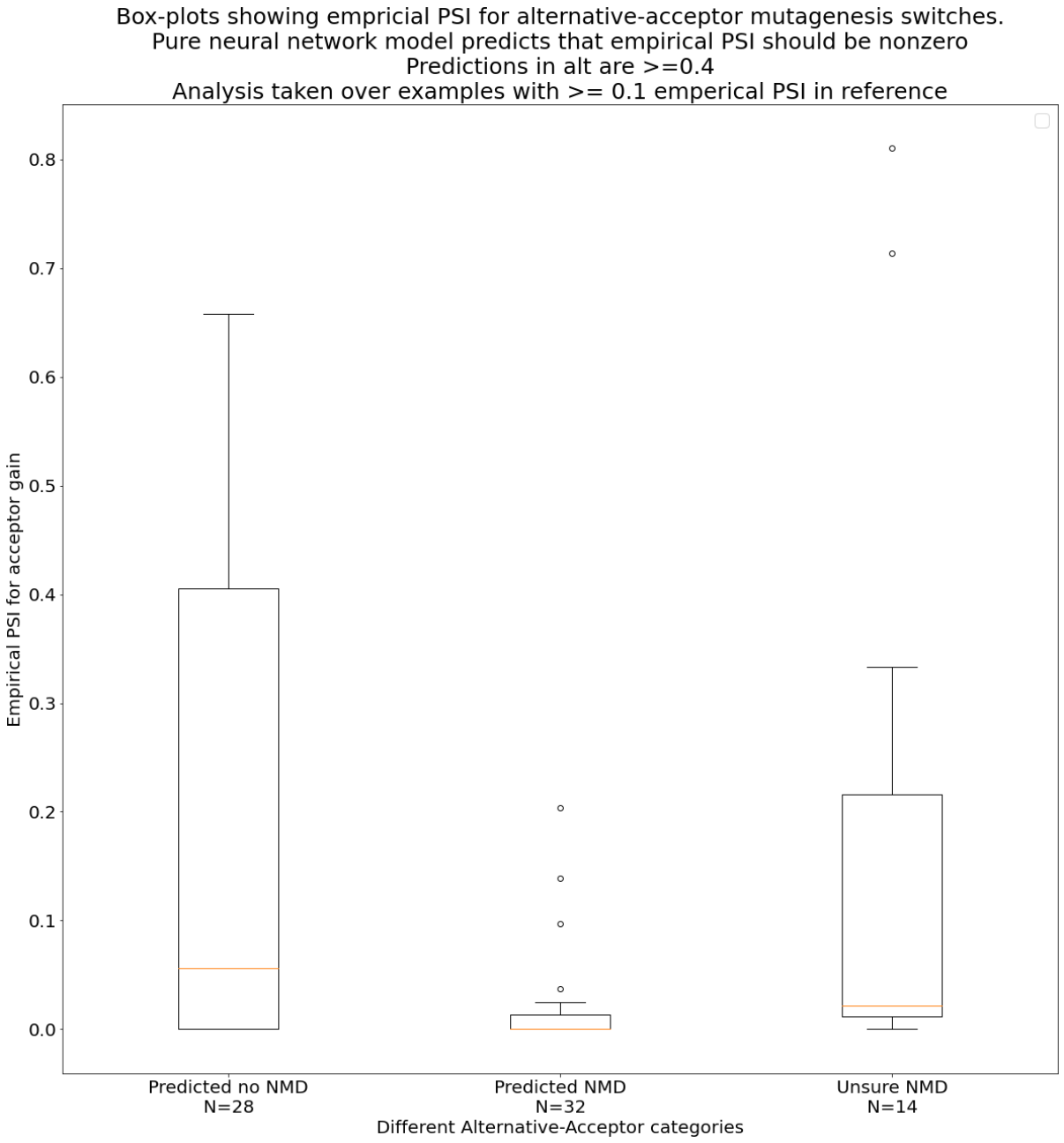


Figure 5.2: Empirical PSI levels for predicted gained alternative acceptors as found in our RNA-seq biobank. Predicted alternative acceptor sites split by additional NMD model. When no NMD is predicted we observe high empirical PSIs. When NMD is predicted we observe low empirical PSIs.

5.2.2 Alternative donor site results

We find similar results for predictions in the other categories. For alternative acceptor sites the NMD predictor acts as a refining tool to predict whether there is empirical PSI. Due to a

smaller dataset, we cannot deduce any significance from these results, but crucially the trend is consistent with what we have previously found.

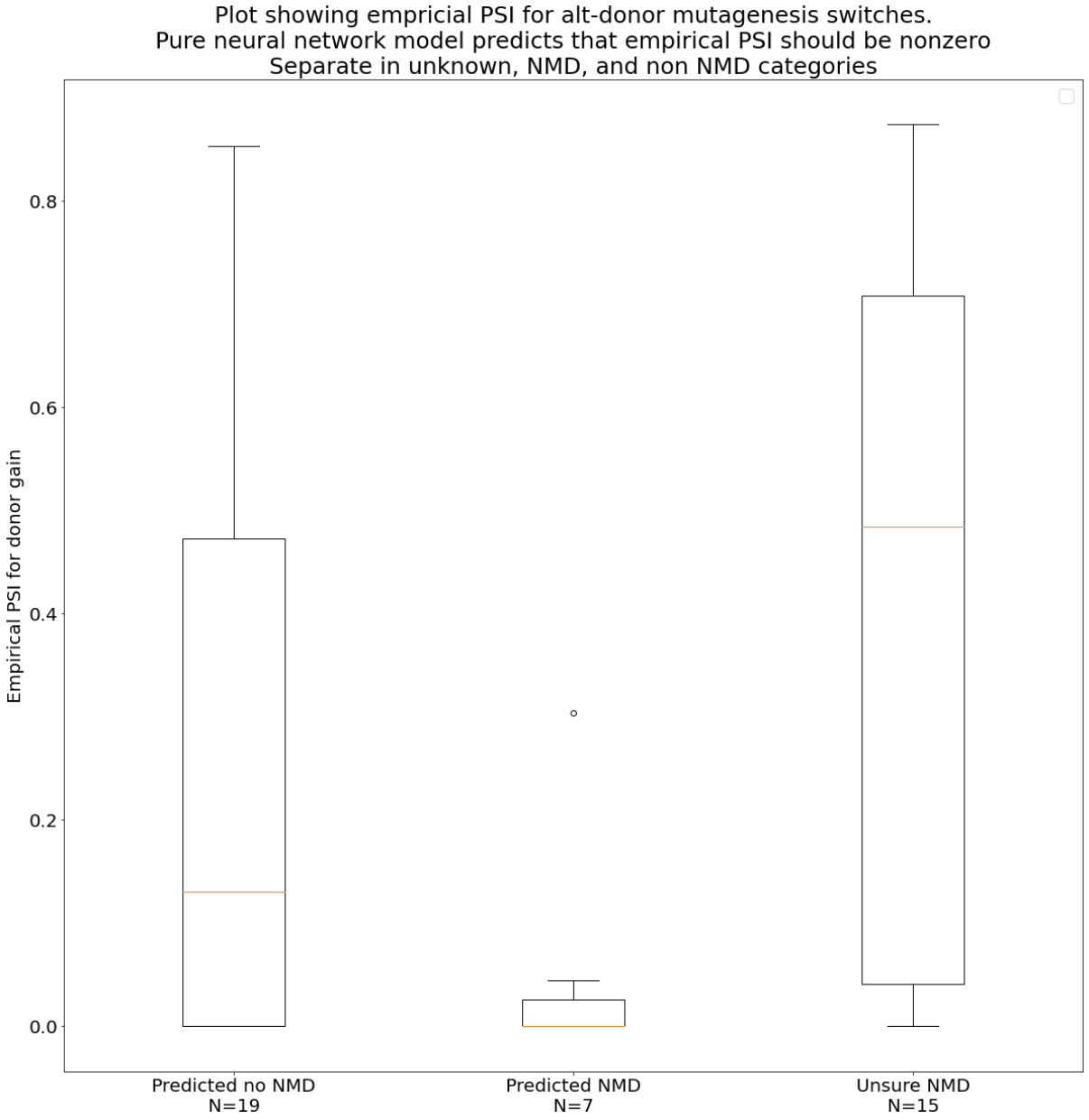


Figure 5.3: Empirical PSI levels for predicted gained alternative donors as found in our RNA-seq biobank. Predicted alternative donor sites split by additional NMD model. When no NMD is predicted we observe high empirical PSIs. When NMD is predicted we observe low empirical PSIs.

5.2.3 Model predicts gain of exon

These last two categories show some additional interesting aspects. As before, they have read evidence when the model does not predict NMD. However, unlike alternative splice sites, for a given exon they should not be constrained to be near known annotated exons. This is indeed what we find. In figure 5.4 we observe that gained/skipped exons that have NMD signatures are enriched in deep-intronic regions compared to those not annotated NMD. Importantly, this provides evidence of gene-regulation, via splicing, occurring in deep-intronic regions due to the loss or gain of local exons. Due to the action of NMD, these splicing events will be difficult to detect in standard RNA-seq datasets and so this currently represents a mechanism of altering gene expression that is difficult to detect.

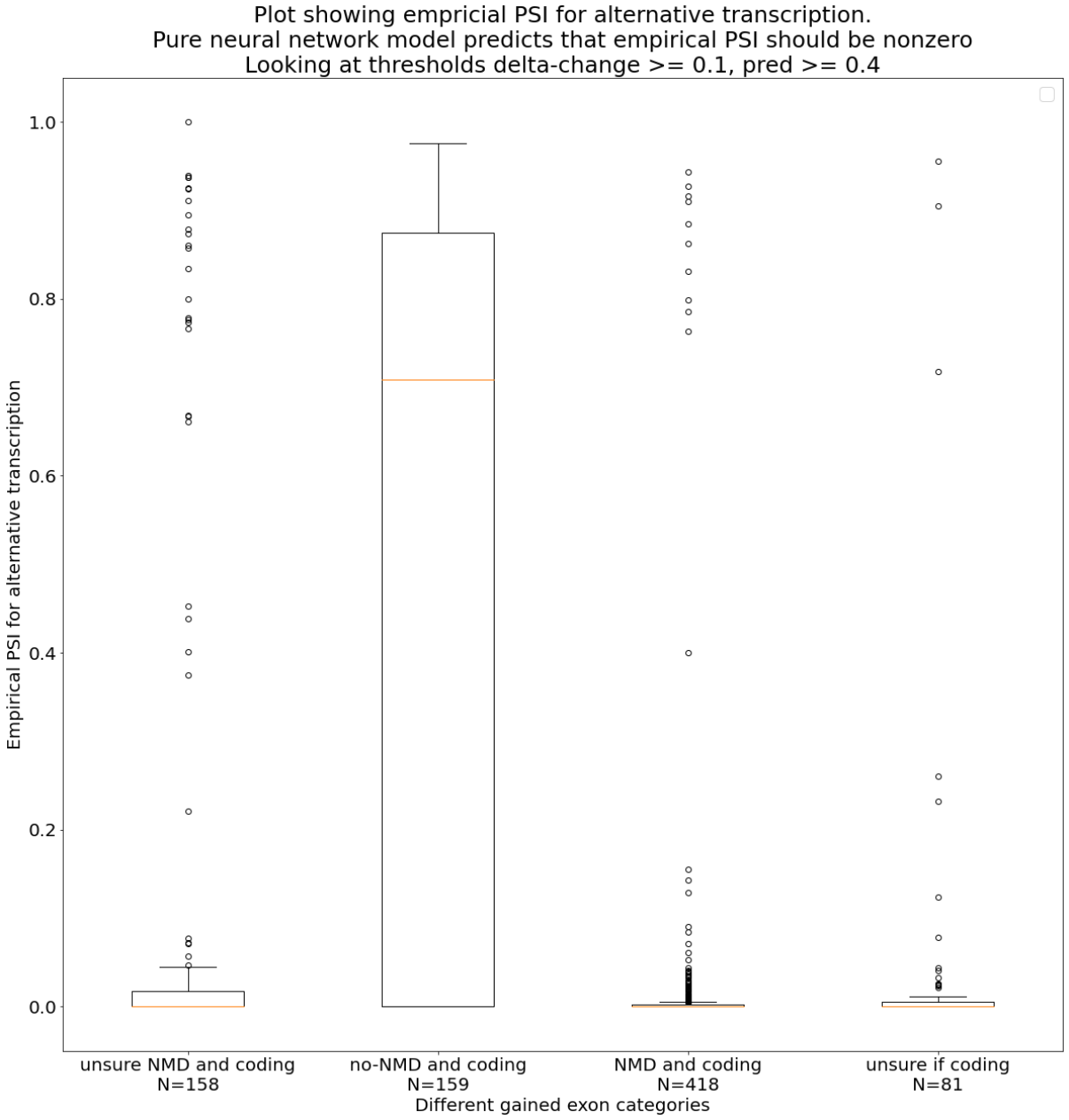


Figure 5.4: Empirical PSI levels for predicted gained exons as found in our RNA-seq biobank. Examples split into categories based upon NMD-model annotation. Predicted gained exons which are additionally not predicted to induce NMD exhibit high empirical PSIs.

5.3 Conclusions

In this section we found that the biobank validates our mutagenesis results in certain scenarios. It supports our belief that the neural network model has learnt which sequence signatures lead to splicing or not.

In our biobank datasets we find read evidence matching our predictions if and when no-NMD is predicted. For the different categories of alternative acceptor sites, alternative donor sites and alternative exons we obtain similar results: if NMD is additionally predicted, then the read evidence does not match the prediction of splicing.

What is unexpected is how little evidence is found for predicted alternative NMD sites in the standard RNA-seq from our biobank. In Figure 5.1 the median value for predicted NMD sites is a PSI of 0. This seems extreme considering that our work analysis of POINT-seq to RNA-seq datasets suggested a depletion for predicted NMD splice sites with a median fold depletion of ~2-3. This suggests that figures 5.1 and 5.2 are exaggerating the NMD effect compared to what has been previously found.

This is likely due to low read-coverage. For many examples, the combined effects of not being a constitutive site, having NMD and being found in only a few samples means that observing 0 supporting reads is not unsurprising. This means that obtaining a median empirical PSI of 0 is less surprising.

To address this we can boost our read coverage by combining the splice sites categorically (NMD, no-NMD) across samples of the same genotype. By doing this we find that the predicted sites within the no-NMD category have a combined PSI of 34%. The combined PSI for the NMD category now has a PSI of 2.3%. This is in line with our estimates from our POINT-seq and RNA-seq analysis with a combined RNA-seq PSIs of 39% compared to 4%. This would suggest a real benefit is to be had in sequencing these biobank datasets to much greater depth for these types of analyses, or alternatively in using POINT-seq directly.

What this does illustrate is how we could easily discount these predictions based on absence of any RNA-seq evidence, despite it being consistent with NMD, POINT-seq and other potential validationary approaches. Our main conclusion from this chapter is that absence of evidence in RNA-seq should be treated cautiously and that orthologous types of transcriptomics that capture pre-NMD transcripts are critical for these analyses.

6 Applications in common disease fine mapping

In this chapter we use the networks we tested and validated in previous chapters to ask whether there is an enrichment for predicted splice-altering SNPs in GWAS catalogue as compared to background SNPs. To the extent that SNPs have the potential to alter gene expression and so to affect phenotypes through splicing, such an enrichment would be expected if splicing was a frequent mechanism in GWAS genetics. It is difficult, however, to get a direct estimate of the fraction of SNPs that affect phenotype in this way, because of the ascertainment biases found in annotating GWAS SNPs. Typically, a GWAS catalogue SNP is annotated as splice-altering only if it is found near a known annotated exon-intron boundary. This leaves out potential mutations creating cryptic splice sites, particularly deep-intronic sites. It must be noted that, although instances of splicing in GWAS signals have been reported and validated in the literature, it is currently considered a minor mechanism. It is, however, possible that this is an underestimate because of the difficulty in detecting such events in standard RNA-seq in the presence of NMD, as discussed in the previous chapters. Therefore, in this chapter we aim to re-address this by performing an enrichment analysis across the GWAS catalogue, using the approaches developed in the previous chapters to come to a better understanding of the prevalence of splicing linked to NMD in GWAS genetics.

We saw in the previous chapter that a significant proportion of SNPs gave rise to NMD transcripts or removed NMD transcripts. Moreover, we have confidence in these predictions due anomalies in their detection arising from our frame-shift analysis. Therefore, in this analysis we will use the same approach to ask whether such SNPs are enriched in GWAS

datasets. This would not only provide further support to the predictions of the mechanism, that these are indeed genuine splice-sites, but it would also identify likely causal GWAS SNPs and the genes they affect.

Why does it provide evidence for the model? As discussed, we typically don't find read evidence supporting lost or gained exons within deep-intronic regions. Our hypothesis is that these exons do not have supporting read evidence due the predicted presence of NMD or another RNA-decay pathway. If our hypothesis were not correct, then we would not expect there to be an enrichment in the GWAS catalogue compared to a similarly sampled background set of 1000G SNPs.

Why are these sites interesting in terms of fine mapping? SNPs near known exons are somewhat easier to analyse. If they don't have an NMD hypothesis, then simple analysis of RNA-seq data would obtain all the evidence one needs. Conversely, AS-NMD sites with potential phenotypic effects are difficult to observe by both using gene-annotations or read-evidence. Even for splice-sites not predicted to be NMD targets, validating a splice-change may be difficult due to the gene being expressed in an inaccessible tissue.

Finally, we are interested in two related questions. Firstly, what level of predicted splice-change can be detectably enriched in GWAS? This will tell us about the sensitivity of the model to detect subtle splicing changes which have a molecular phenotype based on genetic linkage.

Secondly, what is the overall burden of GWAS SNPs marked as enriched by our splicing model? This will provide us with a robust lower estimate for how many GWAS signals are

caused by splicing alterations. It is an underestimate due to our GWAS SNPs potentially not being the causal SNPs, as we do not have the capacity to impute the complete GWAS catalogue, and due to our background 1000G SNPs potentially having effects on splicing. We can then compare the enrichment found in GWAS between near exonic and deep intronic regions.

What should be emphasised here is how we are using the GWAS catalogue in two ways in this chapter. The obvious use is the direct attempt to link aberrant splicing to the GWAS SNPs. The more subtle use is to treat the GWAS SNPs as a means to validate the model's predictions in difficult regions, such as deep-intronic regions. If for a region of SNPs the model shows an enrichment for GWAS SNPs, then this enrichment serves as a means to validate the model for this region.

A final remark concerns what we might consider the genomic mutagenic space for splicing. If there are indeed many deep intronic pseudo-exons (such as those hidden by NMD, low PSI inclusion, or zero-length recursive splicing exons), then the mutational target for splicing is larger than it appears based on just the splice sites of known transcripts. Our model can serve as an unbiased method to measure the relative proportion of GWAS SNPs in different genomic locations, e.g. deep-intron vs near-exonic.

6.1 Enrichment analysis introduction

Obviously, for an enrichment analysis the validity of the background we compare is very important. We, therefore, performed the enrichment analysis using two different approaches to the generation of the background. One approach is to create a balanced set

of control SNPs which are contained within the same LD-window. This allows us to be confident that the control SNPs and GWAS catalogue SNPs derive from similar genomic locations spatially.

The next approach is to form a control set from the entire 1000G SNPs found within genic regions, then down-sample based upon allele frequency and relative gene locations (proximity to known exon-intron boundaries). This means that our analysis would have similar frequencies of control SNPs in near-exonic, exonic, deep-intronic loci. The approach is to divide the control SNPs into separate bins based upon their relative gene location, followed by subsampling so they match the GWAS catalogue's distribution for these bins. There is a natural bias within GWAS (relative to the background) for exonic and near-exonic SNPs, which this approach accounts for.

In addition, this allows separate analysis where we only consider SNPs which are near to known exon-intron boundaries and those found in deep intronic regions. This can be used to compute the enrichment, and relative quantity of enriched GWAS SNPs based upon their genomic location to test the contribution of deep intronic changes to GWAS genetics.

6.2 Initial look at predicted splice altering GWAS SNPs

Figure 6.1 shows that other categories such as exonic missense variants are picked up by our model as likely to be additionally disrupting splicing. Moreover, equivalent numbers of synonymous variants are annotated as splice altering by the model as those annotated missense. This does not invalidate the labelling of an exonic variant as splicing, but rather it

strongly suggests that splicing mechanisms should also be considered when labelling exonic variants.

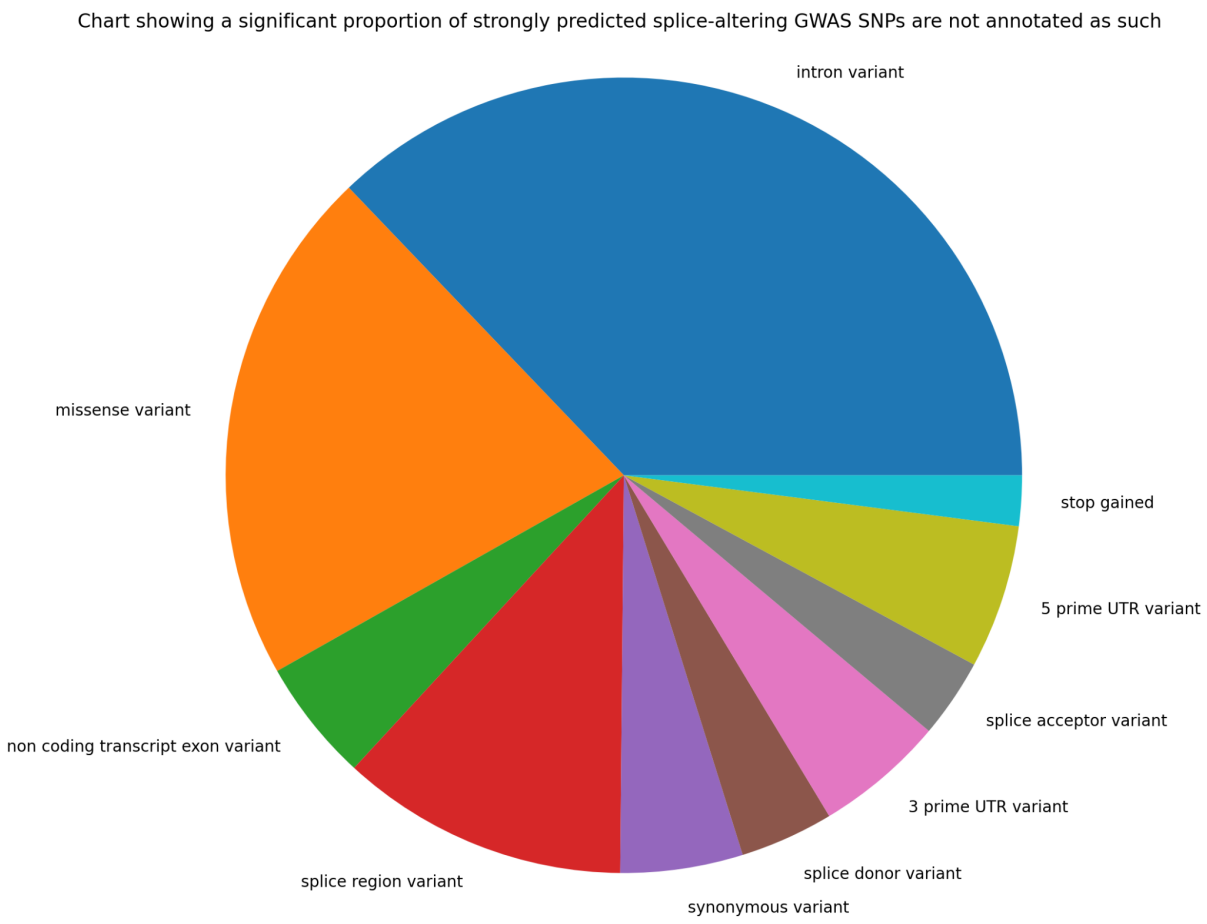


Figure 6.1: GWAS catalogue lead-SNPs with strong predictions (> 0.2 splice change) by our model. This accounts for 348 of the 69,120 analysed lead SNPs using context annotations in GWAS catalogue database. Amongst these SNPs 64 are annotated as splice-altering.

Next we look at the shape of our model's predictions for GWAS SNPs based upon their annotated context. In figure 6.2 it is clear that annotated splice-altering SNPs are skewed to high predictions by our model with 44.5% (167 / 375) obtaining predicted scores >= 0.05 and 91% (341 / 375) obtaining scores above 0.008. This shows that existing splice-altering annotations for GWAS SNPs are detected by the model with high scores.

Figure 6.3 shows that GWAS SNPs unannotated as splice-altering are still more likely to have higher model scores than typical 1000G background. 3.6% (2519 / 69120) have model scores >= 0.05 compared to 2.25% (81188 / 3600932) found amongst 1000G backgrounds (see figure 6.6). They are not as skewed to high scores as annotated splice-altering variants consistent with their missing splice-altering annotations.

Figures 6.4 and 6.5 show that annotated missense and synonymous GWAS SNPs have a large percentage with high model scores. 14.9% (595 / 4003) have model predicted scores >= 0.05. It could be argued that this is not surprising, as these SNPs are found within exons meaning predictions that they affect splicing are coincidental. Later we ask this question by performing an enrichment analysis over SNPs in such exonic regions. Figure 6.6 shows the distribution one obtains for splice-scores from the 1000G background SNPs. In the next section we will carefully balance this set against the GWAS SNPs in an enrichment analysis.

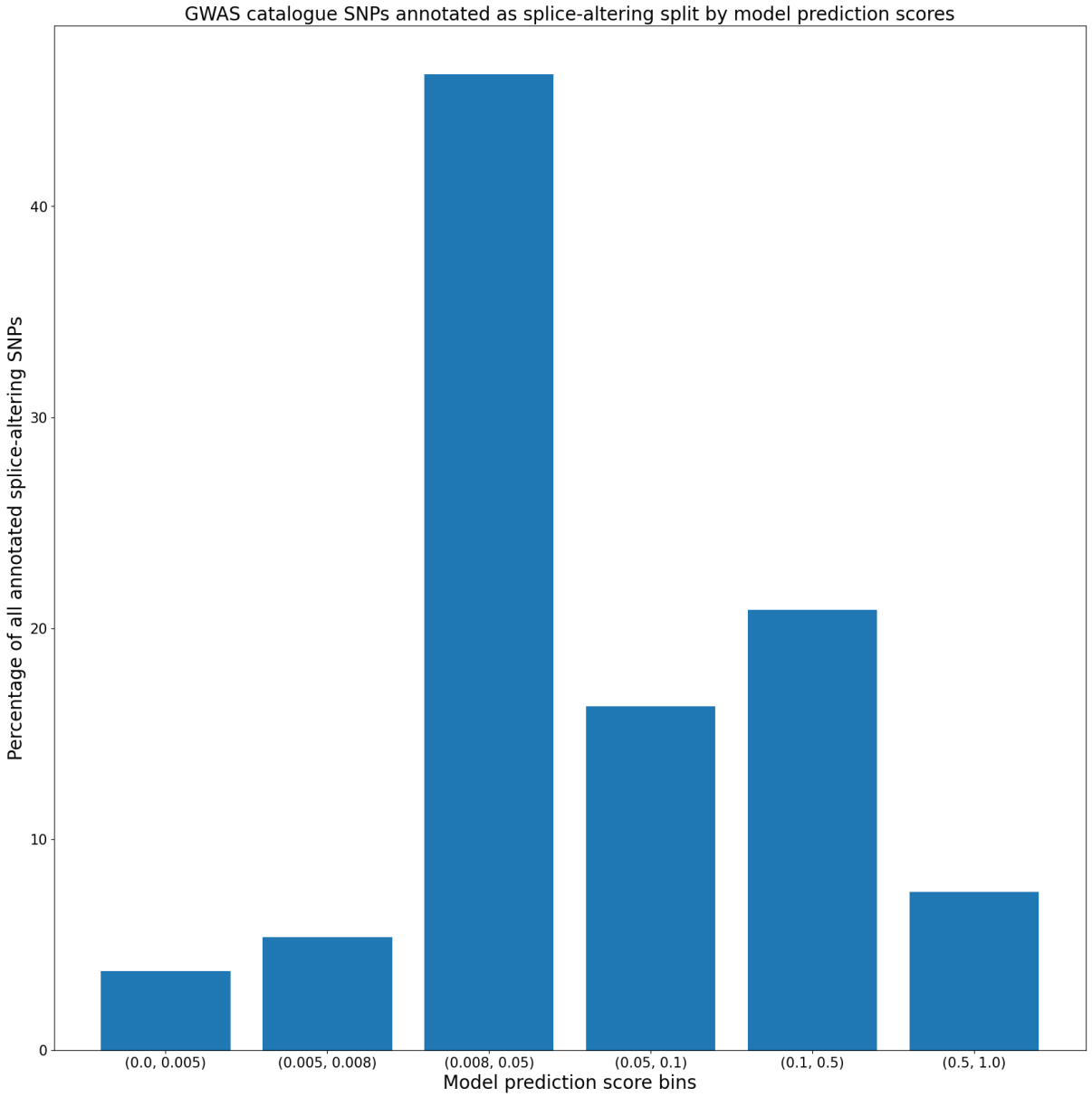


Figure 6.2: Histogram of model predictions for annotated splice-altering lead SNPs in the GWAS catalogue. Very few SNPs have predictions less than 0.008 (32 / 375) with 44.5% obtaining scores >= 0.05 (167 / 375). This plot shows that nearly all splice altering GWAS SNPs would be identified by our model at a 0.008 threshold.

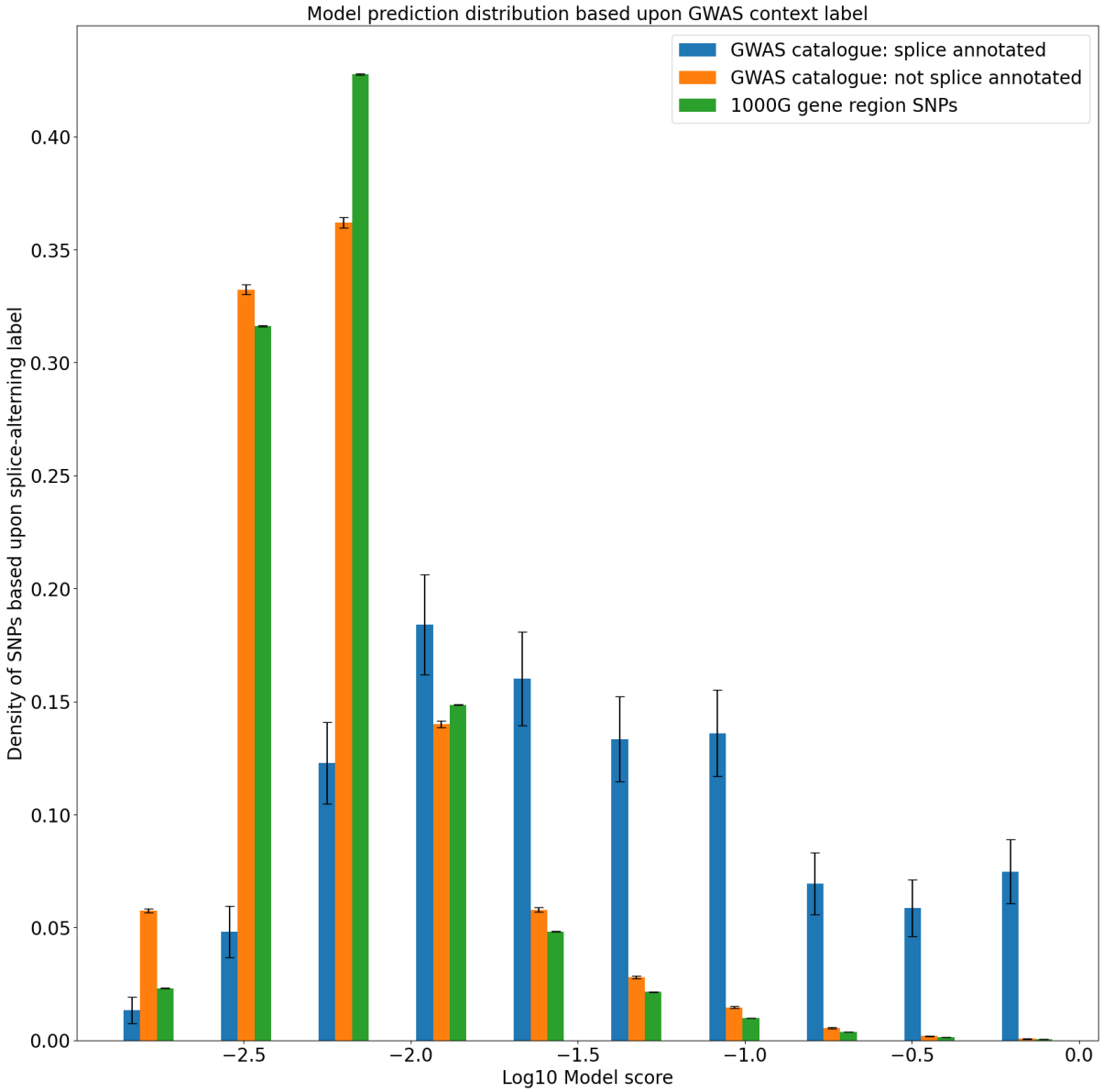


Figure 6.3: Direct comparison of the distribution of model predictions for GWAS splice-variants, GWAS non-splice-variants and background 1000G SNPs. We observe that annotated splice-variants have distribution weighted to higher model scores. Additionally, we observe a modest enrichment for higher scores amongst GWAS SNPs not annotated as splice-altering.

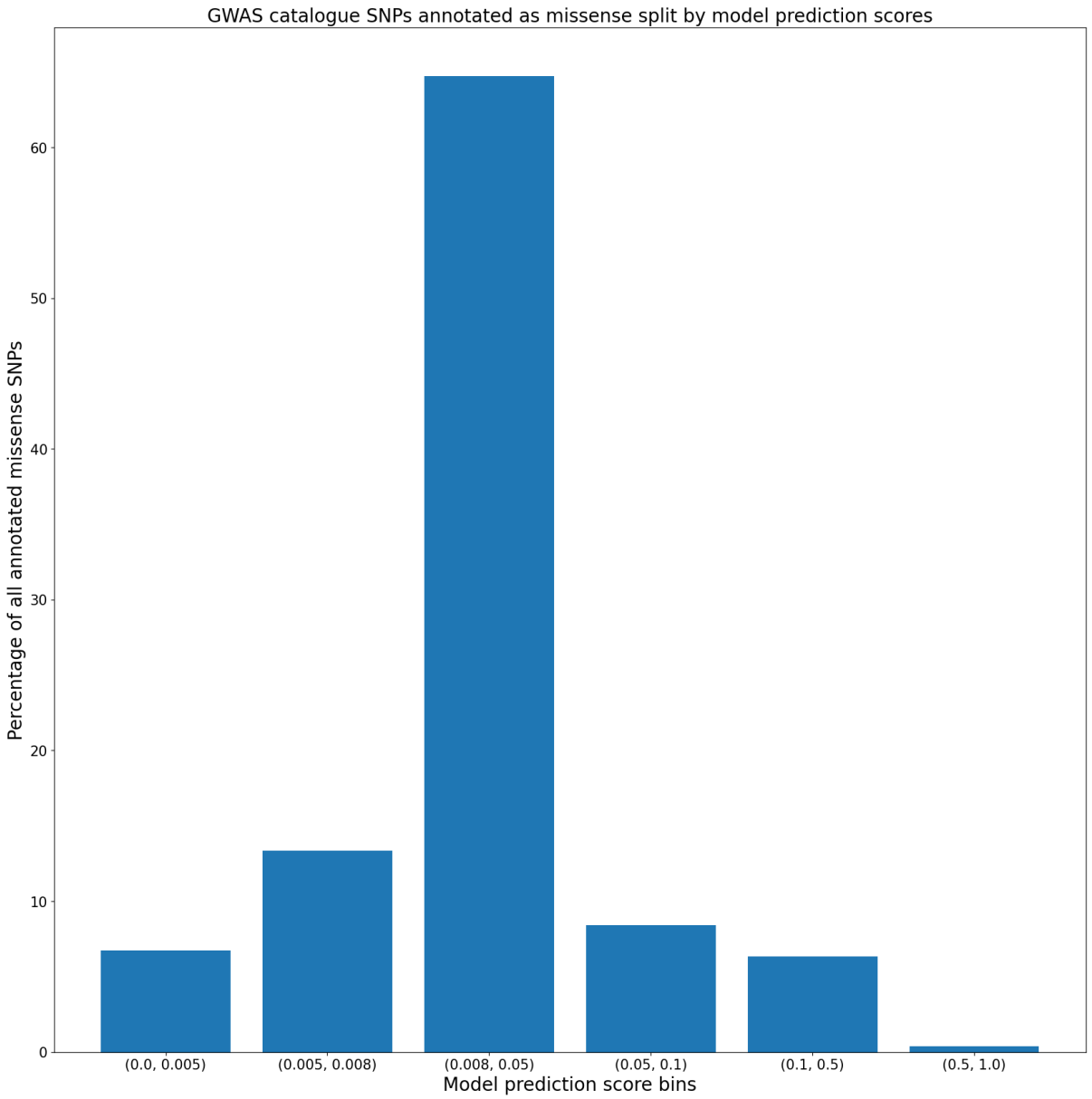


Figure 6.4: Histogram of model predictions for annotated missense lead SNPs in the GWAS catalogue. 15.2% (452 / 2979) of SNPs have scores >= 0.05 and 20.1% (599 / 2979) scores less than 0.008. This analysis shows that according to our model many missense variants are predicted to have a small non-trivial effect on splicing.

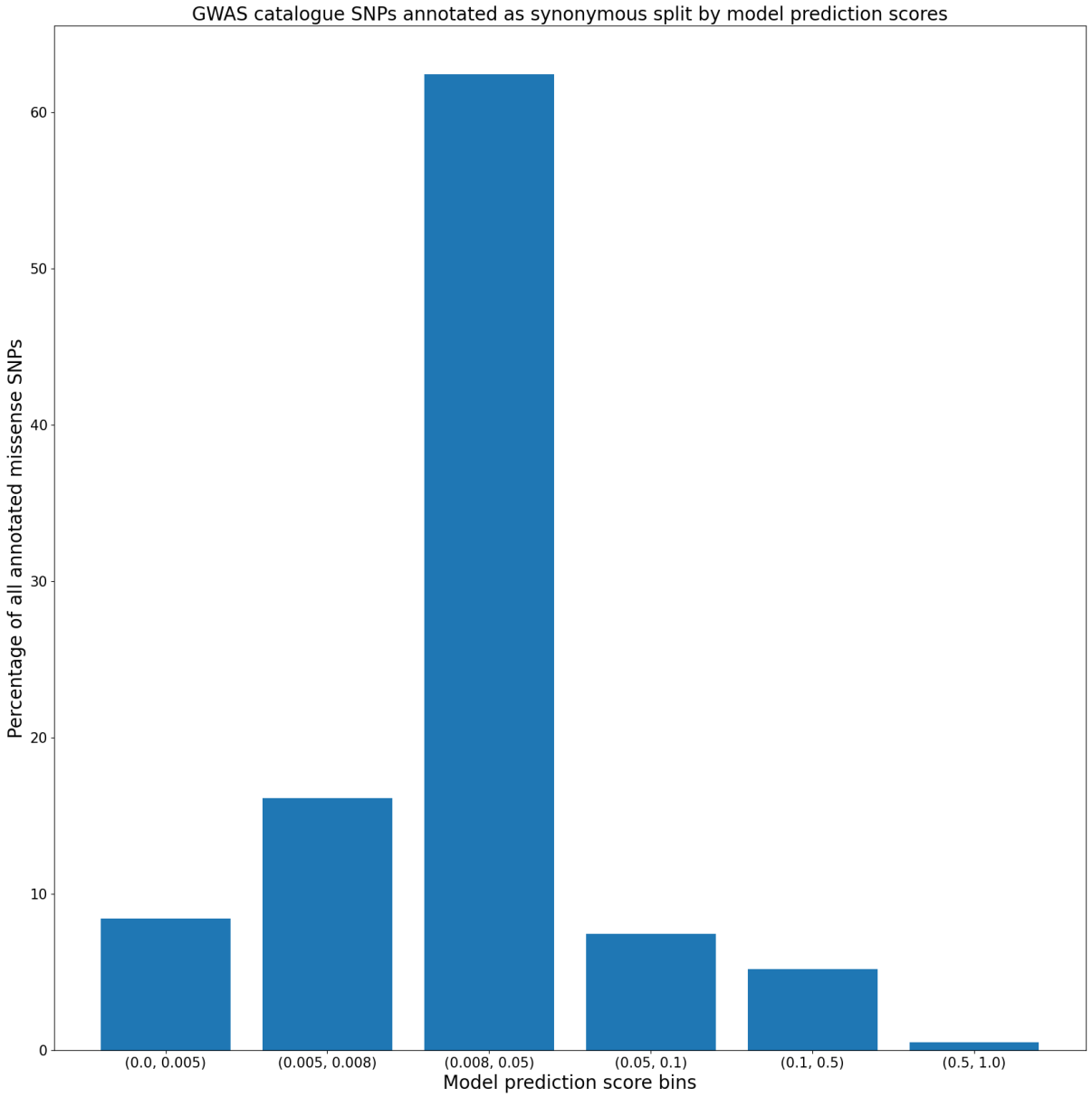


Figure 6.5: Histogram of model predictions for annotated synonymous lead SNPs in the GWAS catalogue. 13.1% (143 / 1024) of SNPs have scores >= 0.05 and 24.5% (251 / 1024) scores less than 0.008. This analysis shows that according to our model many synonymous variants are predicted to have a small non-trivial effect on splicing.

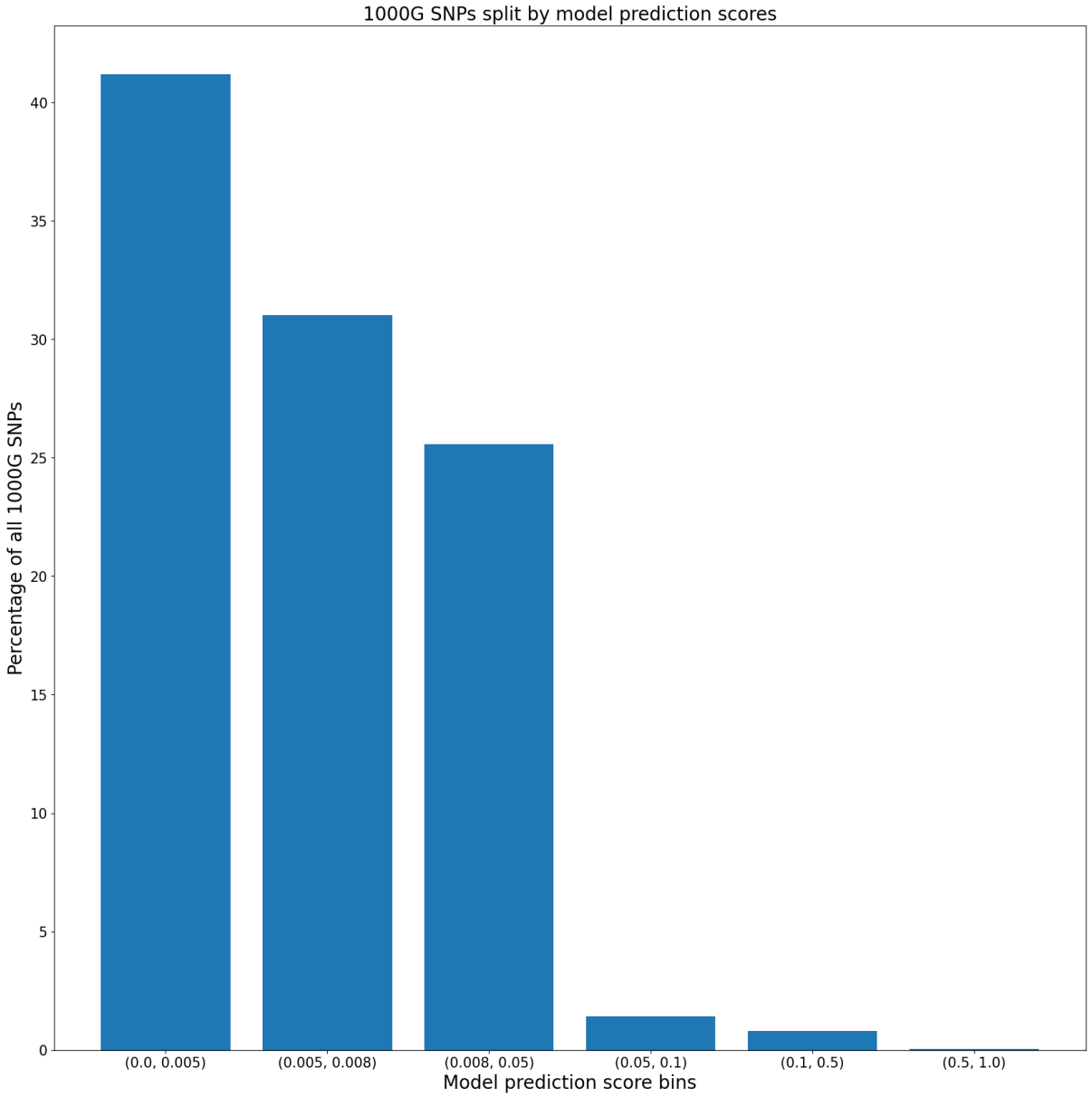


Figure 6.6: Histogram of model predictions for 1000G SNPs found in genic regions. 2.2% (81188 / 3600932) of SNPs have scores >= 0.05 and 72.1% (2599154 / 3600932) scores less than 0.008. This analysis suggests that according to our model only a small percentage of 1000G background SNPs have a predicted effect on splicing.

6.2.1 Balanced Genomic region enrichment analysis

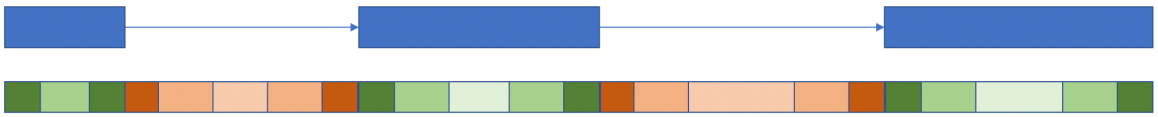


Figure 6.7: Pictorial explanation of how background SNPs are rebalanced for enrichment analyses. A gene is partitioned into equivalent regions with respect to exon locations. This allows us to balance the background SNPs to reflect the GWAS catalogue SNPs in regards to their genomic locations. For example the ratio of exonic GWAS SNPs to background SNPs will be the same as for deep-intronic SNPs.

It is possible to expect that SNPs in the GWAS catalogue show a spatial bias because they are more likely to be found around gene regions, and so may as a side effect be enriched within exons or proximal to intron-exon boundaries. For our enrichment analysis it is therefore necessary to account for this potential bias.

We, therefore, annotate GWAS SNPs with reference to their genomic location, by building interval trees for the gene locations, exon locations, and open reading frames. The distance to boundaries is computed by taking the smallest distance for a SNP to its nearest intron/exon boundary. A SNP is additionally annotated based upon its position in an intron or exon. Our definitions of these intron-exon boundaries are obtained from UCSC comprehensive gene annotations (GENCODE CompV37). These form our basis for defining genomic loci.

In this analysis we balance the GWAS and 1000G controls on two axes. The first axis is whether they fall within exons or introns (we only consider SNPs in genic regions), and we stratify SNPs into bins based on their proximity to exon/intron boundaries. The second axis in which we match the GWAS SNPs and controls is on their European allele frequencies.

The motivation for matching spatial location and allele frequency in control SNPs is to create a conservative estimate for how much splicing is mechanistically a function for the GWAS signals, rather than a by-product of any skewed genomic distribution. For instance, just by being adjacent to an intron-exon boundary we expect higher splice-scores, therefore, by matching the spatial location of control SNPs we avoid that the enrichment of GWAS SNPs near intron-exon boundaries generates spurious associations. We saw from figures 6.4 and 6.5 that exonic GWAS SNPs have this skew towards higher model scores. The purpose of our enrichment analysis is to remove the possibility that this is merely a coincidental effect of many GWAS SNPs being found within exonic or near exonic regions. Essentially, we upweight the background 1000G SNPs to have a balanced quantity of exonic SNPs as we find in the GWAS catalogue.

We divide both the GWAS catalogue SNPs and 1000G background SNPs into bins based upon their European allele frequency. For each bin we take the ratio of SNPs between the GWAS catalogue and 1000G SNPs. Then we downsample the background SNPs so that they all have the same ratio.

For the genomic locations we perform a similar analysis to what we did for European allele frequencies. We label SNPs based upon their proximity to exons into 6 categories:

1. Exonic and less than 30 base pairs to intron/exon boundary
2. Exonic and greater than 30 base pairs to intron/exon boundary
3. Intronic and less than 30 base pairs to intron/exon boundary
4. Intronic and less than 100 base pairs to intron/exon boundary
5. Intronic and less than 1000 base pairs to intron/exon boundary
6. Intronic and greater than 1000 base pairs to intron/exon boundary

In a similar manner, we compute the ratio of GWAS SNPs to background SNPs for each bin. Then we downsample the background SNPs so that they are matching. In practice, this means that we heavily downsample deep-intronic background SNPs, as the GWAS catalogue has an enrichment for exonic or near exonic SNPs compared to 1000G.

Figure 6.8 shows how European allele frequency is balanced. The background 1000G has a higher percentage of SNPs with european allele frequencies near 0.01 or 0.99. The GWAS catalogue SNPs are biased towards medium allele frequencies relative to the background. The consequence of rebalancing the background is to remove 1000G SNPs with high or low allele frequencies. The background set of medium allele frequency 1000G SNPs is kept unchanged. This reweighting of the background 1000G SNPs allows us to confidently state that any found enrichment is not due to different allele frequency distributions.

Figure 6.9 shows that the GWAS catalogue lead SNPs are skewed towards exonic regions relative to the background 1000G SNPs. The consequence of downsampling is a large reduction of deep-intronic background SNPs. There is a co-coincidence that GWAS SNPs are locationally biassed towards (near) exonic regions, as are splice-altering SNPs. This spatial

reweighting allows us to conclude that any found enrichment is not an artefact of this co-coincidence.

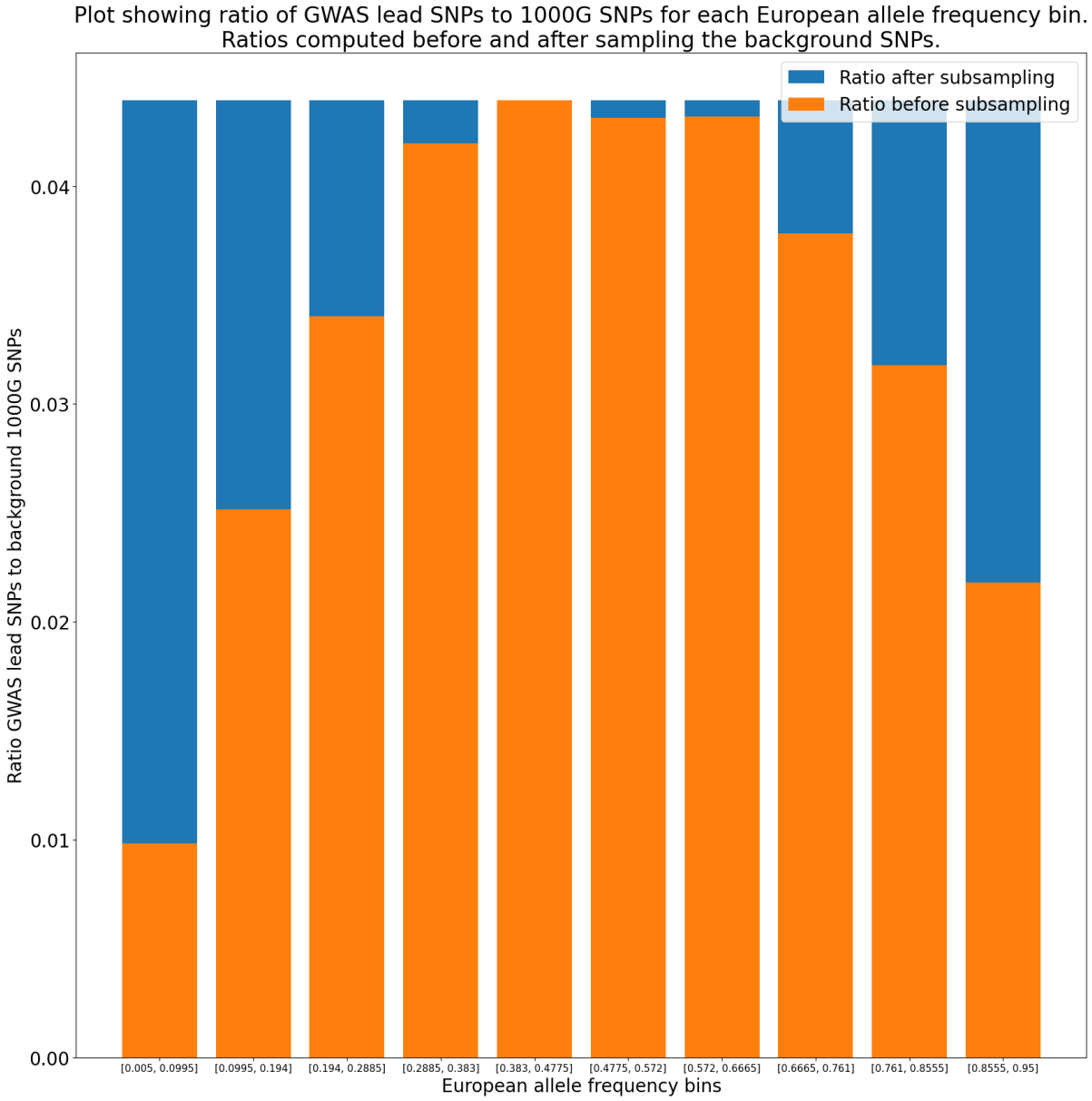


Figure 6.8: Size of background SNP classes subsampled based upon European allele frequency. Classes with a smaller ratio of GWAS catalogue SNPs to 1000G SNPs are downsampled to create a balanced dataset. Heaviest downsampling happens for low and high European allele frequency bins. Background SNPs in the medium bins are largely unchanged.

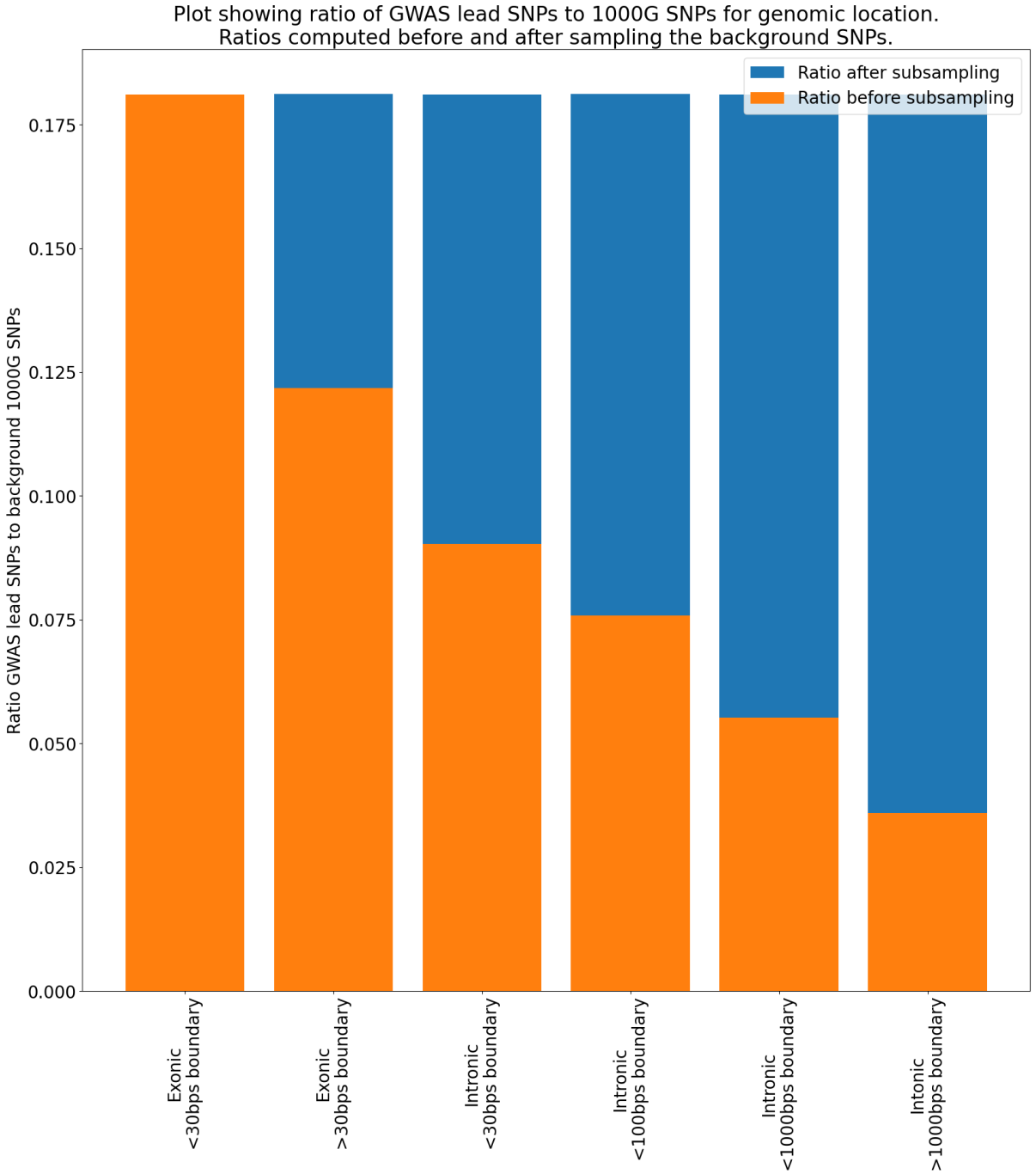


Figure 6.9: Size of background SNP classes subsampled based upon genomic location. Classes with a smaller ratio of GWAS catalogue SNPs to 1000G SNPs are downsampled to create a balanced dataset. Heaviest downsampling happens for deep intronic SNPs. Background SNPs in the exonic locations are largely unchanged. There is a relative abundance of GWAS catalogue SNPs in genomic locations near to exons relative to the background 1000G SNPs.

As such, these enrichments have the potential to underestimate the significance of splicing in relation to common disease. The GWAS catalogue SNPs themselves are not all likely to be disease-causing, but could be merely in LD to disease-causing SNPs. In addition, another concern is that we largely reduce the number of deep-intronic SNPs. This is necessary to build confidence in the enrichment analysis, but it could also be argued to compute enrichments lower than the true enrichment.

Unsurprisingly, higher splice scores lead to larger enrichments. Less expected was that SNPs with smaller scores (>0.008) are also found to be enriched in GWAS. An issue with scores in the range [0.008, 0.05] when fine-mapping an individual SNP is the enrichment is only 1.11, (1.10-1.12, 95% CI) which is too low to interpret individual SNPs, but we can make a statement about a class of GWAS SNPs within this range.
The result that there is a non-trivial enrichment for small scale splicing changes suggests that splicing is a factor in GWAS at various levels of predicted change.

Using the enrichments along with the total counts of SNPs within each bin, we obtain estimates for the total proportion of GWAS catalogue SNPs caused by splicing being 3.72% (3.4%-4.0%, 95% CI) of the GWAS catalogue. This, therefore, suggests that splicing is a ubiquitous mechanism in common disease.

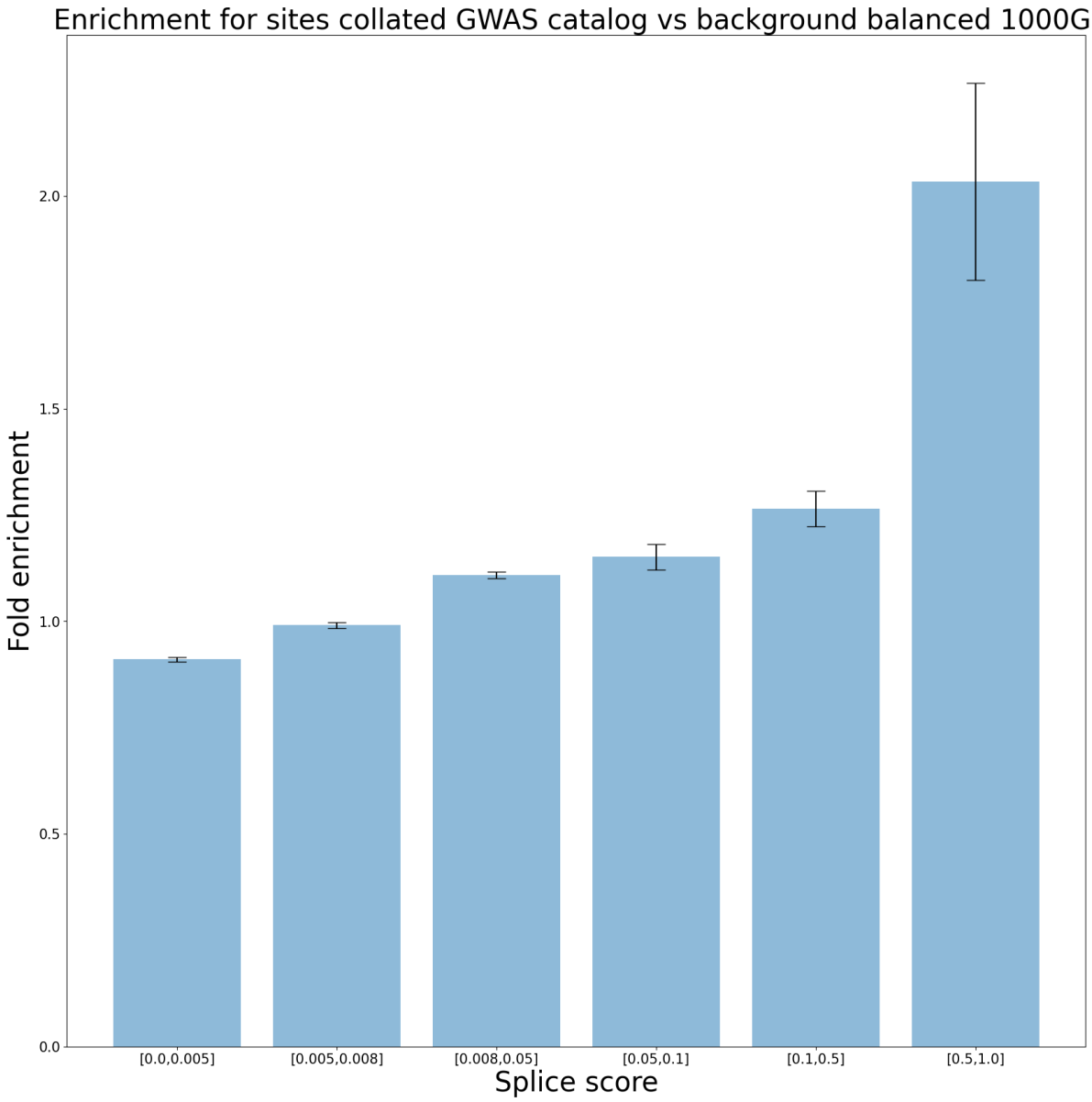


Figure 6.10: Enrichment analysis for GWAS catalogue SNPs against a rebalanced 1000G SNP background. Background balanced on allele frequency and gene location. Find a significant positive enrichment in all bins with splice-score difference greater than 0.008 (see table 6.12). For all smaller splice-score bins we find a significant depletion in GWAS SNPs compared to the background.

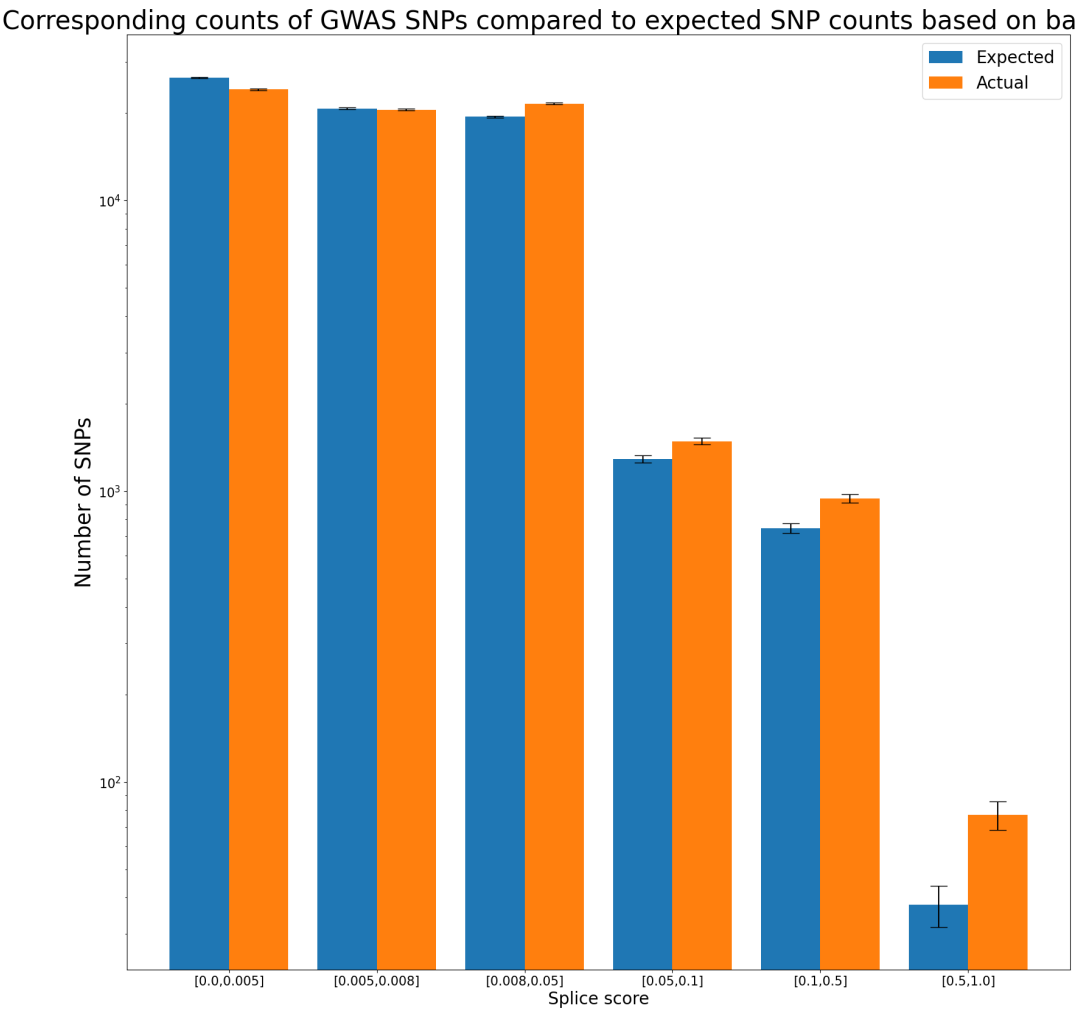


Figure 6.11: From enrichment analysis the expected vs actual counts for different splice-score bins. Enrichment is smaller when splice-scores are small but due to magnitudes more SNPs the total number of SNPs is larger. As expected we find very few SNPs with strong splice scores in both GWAS catalogue and in background.

Prediction	enrichment	expected	actual	diff	95% CI enrich	odds-ratio	F-E p-value
[0.0,0.005]	0.910531	26431.8	24067	-2364.83	(0.905, 0.916)	1	1
[0.005,0.008]	0.991037	20696.5	20511	-185.502	(0.984, 0.998)	1.08842	2.00827e-16
[0.008,0.05]	1.10903	19425.1	21543	2117.9	(1.101, 1.117)	1.218	6.56921e-83
[0.05,0.1]	1.15156	1290.42	1486	195.582	(1.122, 1.181)	1.26472	5.44468e-15
[0.1,0.5]	1.26494	746.283	944	197.717	(1.224, 1.306)	1.38923	4.82235e-18
[0.5,1.0]	2.03394	37.8576	77	39.1424	(1.802, 2.266)	2.23379	1.9898e-08
[0.005,1.0]	1.05604	42196.2	44561	2364.83	(1.051, 1.061)	1.15981	3.41682e-66
[0.008,1.0]	1.11862	21499.7	24050	2550.34	(1.111, 1.126)	1.22854	1.62039e-95
[0.05,1.0]	1.20845	2074.56	2507	432.441	(1.184, 1.233)	1.32719	7.50846e-33
[0.1,1.0]	1.30206	784.141	1021	236.859	(1.261, 1.343)	1.43	2.04288e-22

Table 6.12. From enrichment analysis the raw counts for actual GWAS catalogue SNPs found within each splice-model prediction bin. Expected counts computed using the general proportion of GWAS catalogue SNPs when no model is used. Separately, Fisher Exact odds-ratios with p-values computed for the odds-ratio. Comparison made to predictions in the [0, 0.005] bin.

Briefly, we would like to add a caveat to these results in consideration of the biases which may still remain in our analysis. We have a clear negative bias, as not all catalogue lead SNPs are causal. Many of the control SNPs in and near-exons will inevitably affect splicing but are not annotated in the current GWAS catalogue.

A further mechanistic complication also exists for the predicted splice altering missense variants within exons, as it may be the change in splicing or coding or a combination of the two that is driving the genetics, which would be difficult to untangle.

It is also interesting to consider the effect of the strength of the splice predictions distribution in the GWAS catalogue, in order to try and understand what level of splice change might constitute a detectable GWAS signal. Taking only SNPs with high splice-scores, (>0.1) we find that these only account for 0.35% (0.29%-0.40%, 95 CI) of GWAS SNPs. This is consistent with our observation that most annotated splice-altering SNPs are recovered by

our model with relatively high splice-scores. However, only considering these highly-scored evidently splice-altering SNPs misses much of the total enrichment we detect. This suggests that much of the enrichment we see comes from small predicted changes far away from annotated splice sites, such as deep-intronic. Therefore, even taking the discussed biases into account, this suggests that focusing on evident splice-altering SNPs will miss most splice-altering GWAS SNPs, and that highly sensitive and quantitative transcriptomic methods will be needed to validate such GWAS splicing signals.

In section 6.4 we will break down this analysis to ascertain the relative enrichments found in exonic and intronic regions.

6.3 Effect of combining multiple neural network models

This subsection relates to technical aspects which result from the stochastic nature of neural network training. Ideally, a single trained model would successfully evaluate a SNP's splicing significance. As previously discussed, we found that for our HGMD analysis there was a benefit arising from combining the output of multiple trained instances of the same model. In this approach, we combined the splice-scores by taking the mean of the splice-scores for each independently trained model. In this section we revisit this technical aspect to evaluate whether it is similarly beneficial in our GWAS analysis.

Repeating the GWAS enrichment analysis using combined models, we now find bigger enrichments for each splice-score bin, and that a greater percentage of GWAS hits are found as enriched for splicing under the combined model.

Figure 6.13 shows how each included extra model increases the number of enriched SNPs in the range [0.008, 1.0]. Figure 6.12 shows that this is largely caused by a greater depletion arising in the range [0.003, 0.008]. This shows that combining models, up to around 12

models, incrementally increases splice-scores for GWAS SNPs whilst relatively reducing the background scores.

Based on the assumption that GWAS arises polygenically from minor changes across multiple GWAS signals, we would expect there to be few SNPs with strong effect sizes. Indeed, we see that combining models reduces the number of GWAS SNPs in the range [0.8, 1.0] by 60%. This is distinct to what we found for HGMD SNVs, where we found strong effect mutagenesis to persist through combining models. In short, combining models reduces the number of GWAS SNPs with very high splice-scores, but it significantly increases the number of SNPs with moderately low splice-scores. This supports both the use of combined models and the idea that GWAS signals tend to subtly affect gene expression at individual loci.

An additional consideration is how to most effectively combine the outputs of each individual model. When considering the HGMD cohort, we combined outputs by using a mean of model scores. The GWAS scenario is quite different, as we are more concerned with constraining small scores. Combining by using the mean consistently increases the fraction of all SNPs with scores greater than 0.008. Combining by using the median keeps this fraction stable. One consequence is that the total percentage of background SNPs with scores above this threshold is almost 20% smaller, making fine-mapping of individual SNPs more robust.

In summary, this shows that, as with our HGMD analysis, the analysis of GWAS predictions benefits from combining the scores of multiple individual trained models. However, it also shows that the way in which the scores of these models are combined can be tuned to either more robustly detect large changes in splicing or to weaker changes closer to noise.

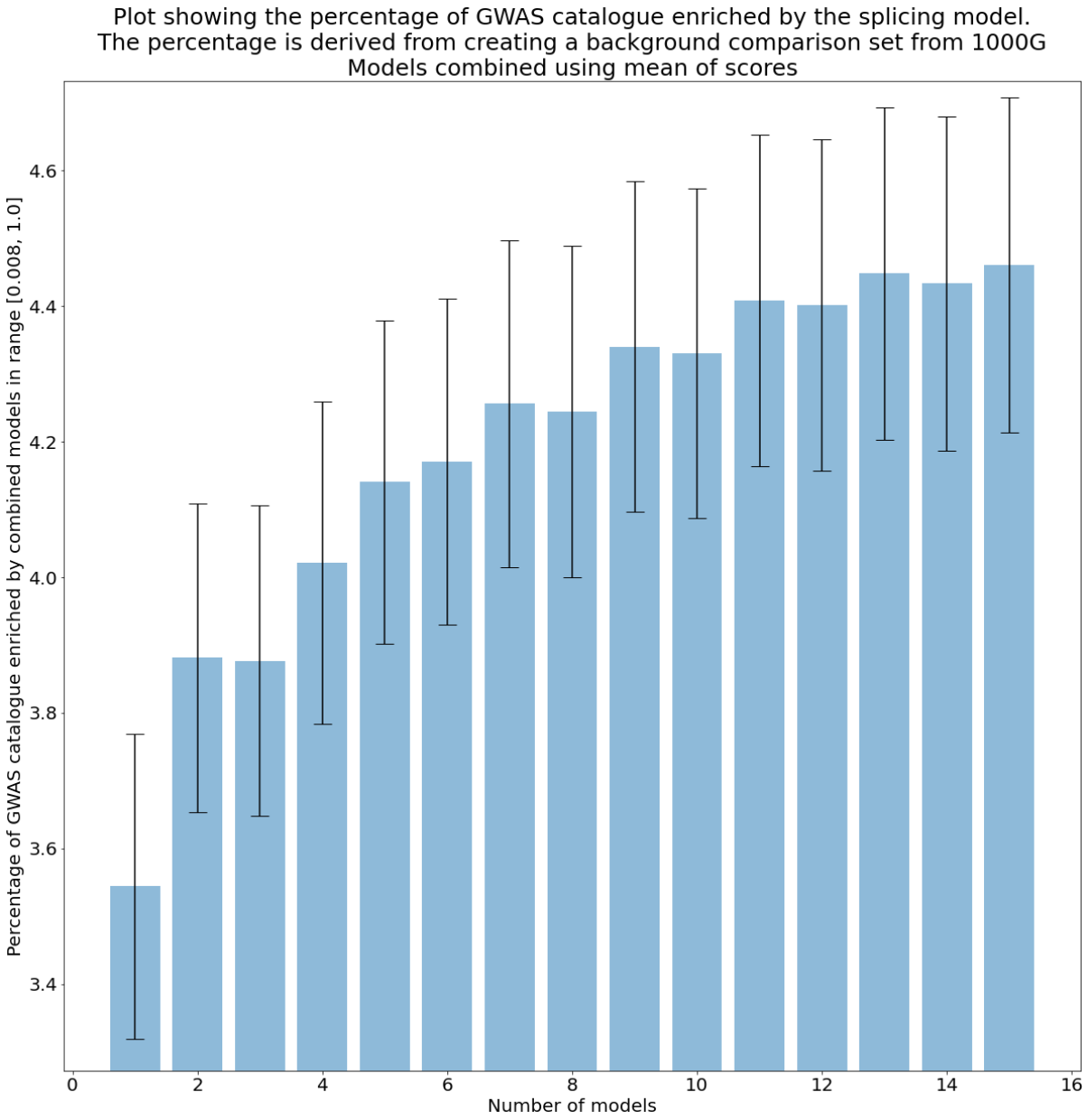


Figure 6.13: Analysis of GWAS enrichment percentage based upon number of combined models (using mean). Adding further models increases the enrichment for GWAS catalogue. Each additional model increases the difference between the number of GWAS catalogue SNPs compared to the expected number of GWAS SNPs based on the background.

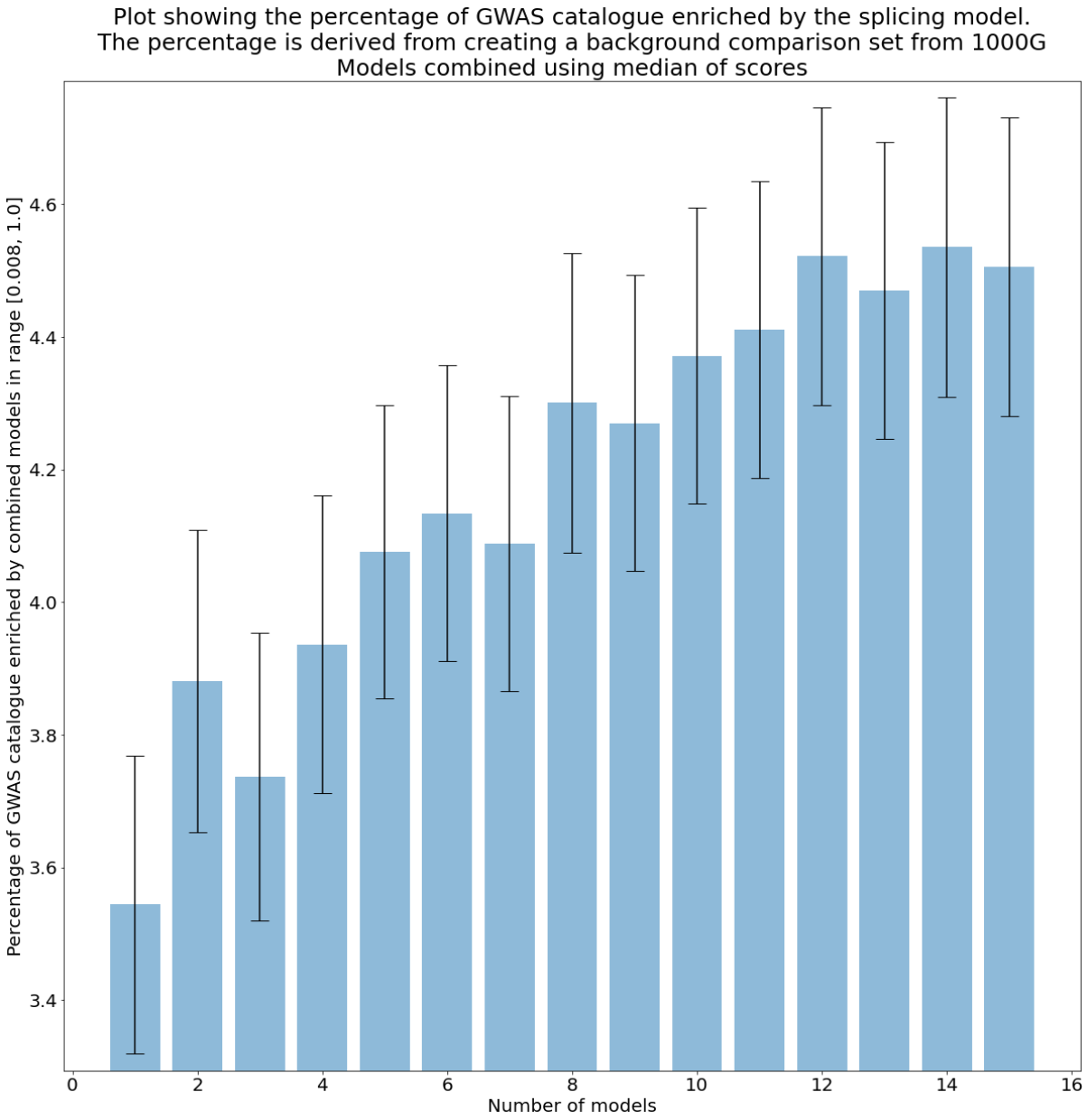


Figure 6.14: Analysis of GWAS enrichment percentage based upon number of combined models (using median). Adding further models increases the enrichment for GWAS catalogue. Each additional model increases the difference between the number of GWAS catalogue SNPs compared to the expected number of GWAS SNPs based on the background.

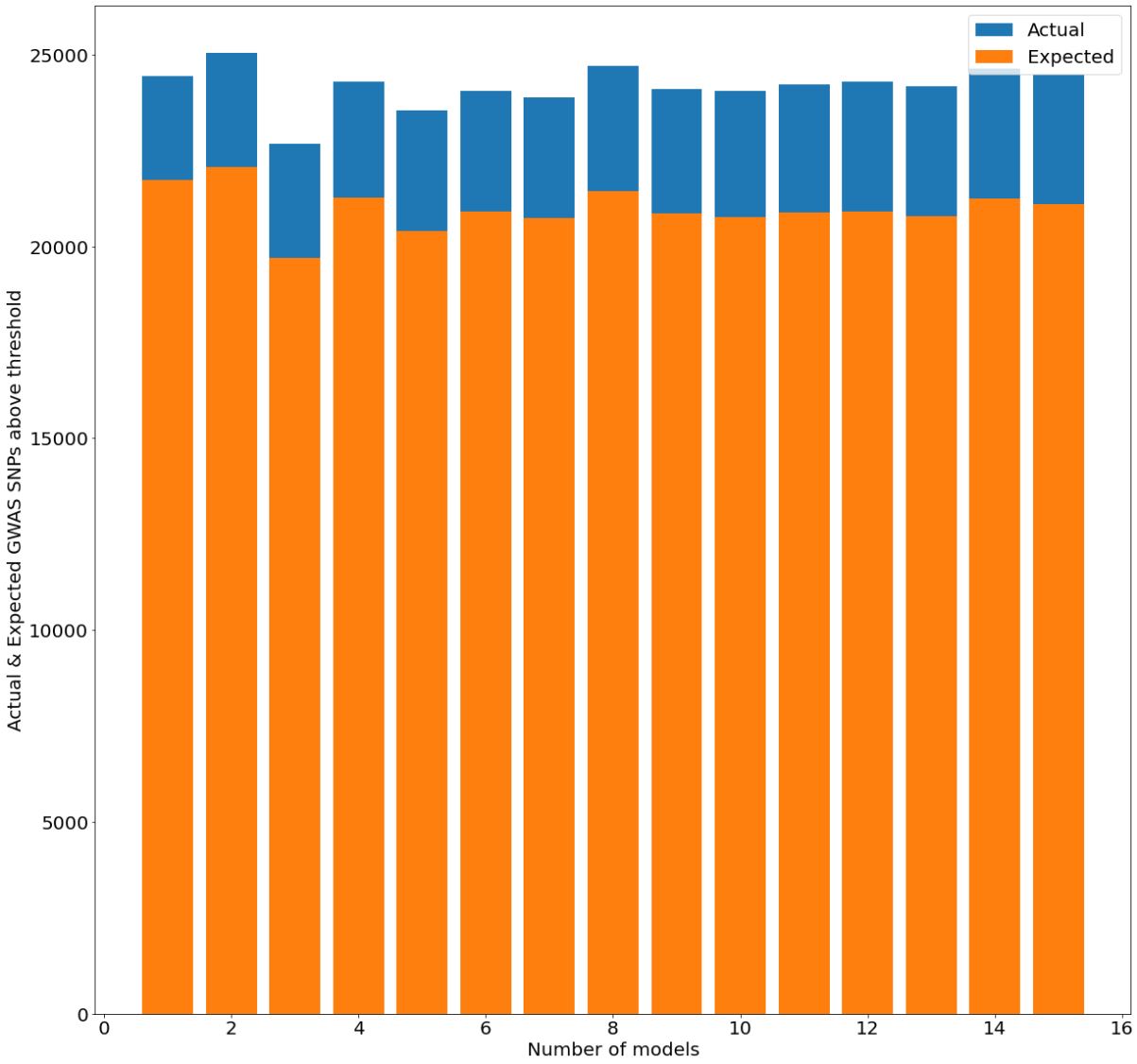


Figure 6.15: Number of SNPs above significant threshold of 0.008. Expected counts based upon 1000G background and the actual counts from GWAS. Combining models utilising a median score does not increase the overall number of annotated SNPs through increasing the number of models. This means you obtain greater accuracy with no cost to specificity.

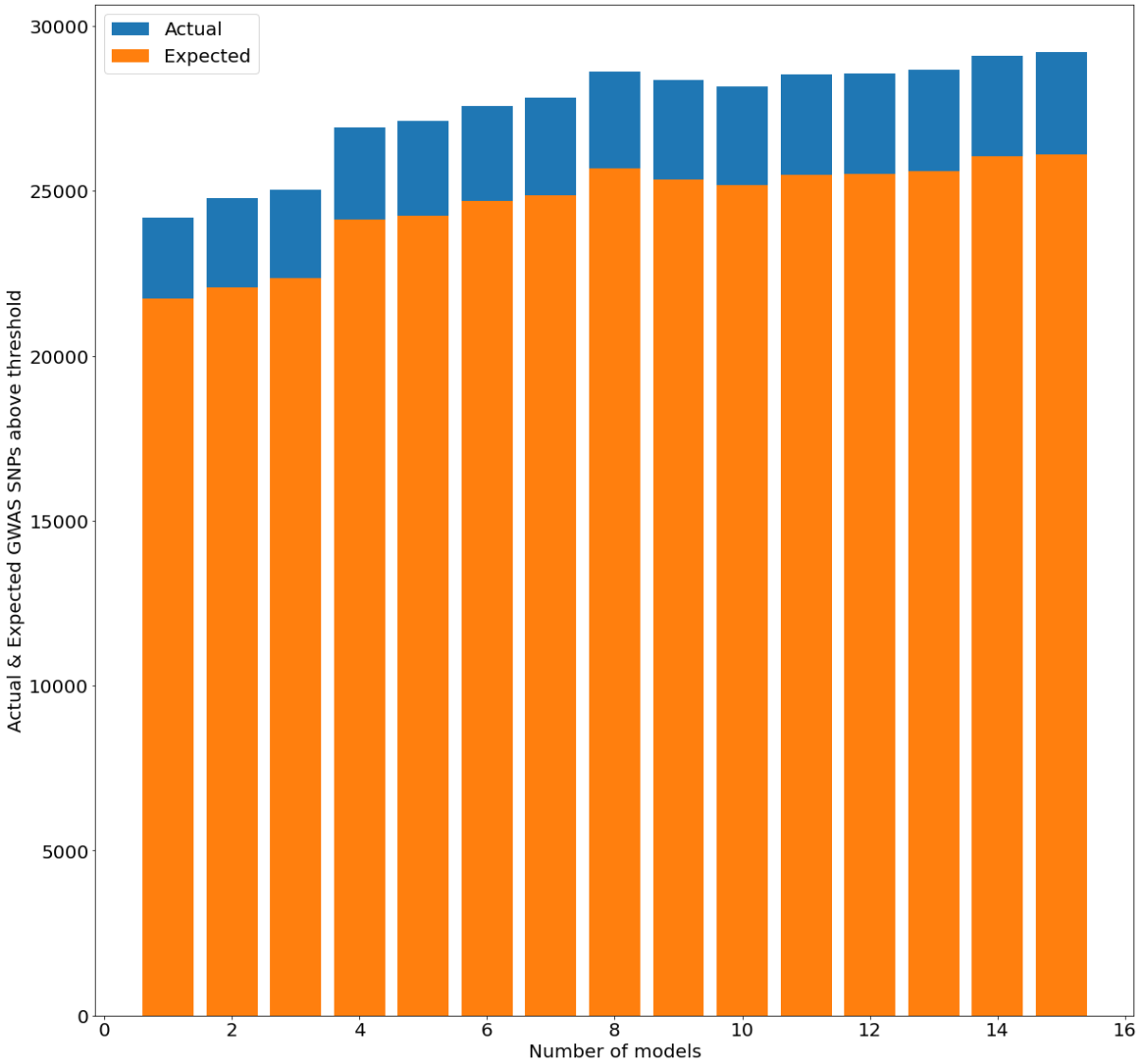


Figure 6.16: Number of SNPs above significant threshold of 0.008. Expected counts based upon 1000G background and the actual counts from GWAS. Combining models utilising a mean score does increase the overall number of annotated SNPs through increasing the number of models. This means you obtain greater accuracy with a cost to specificity.

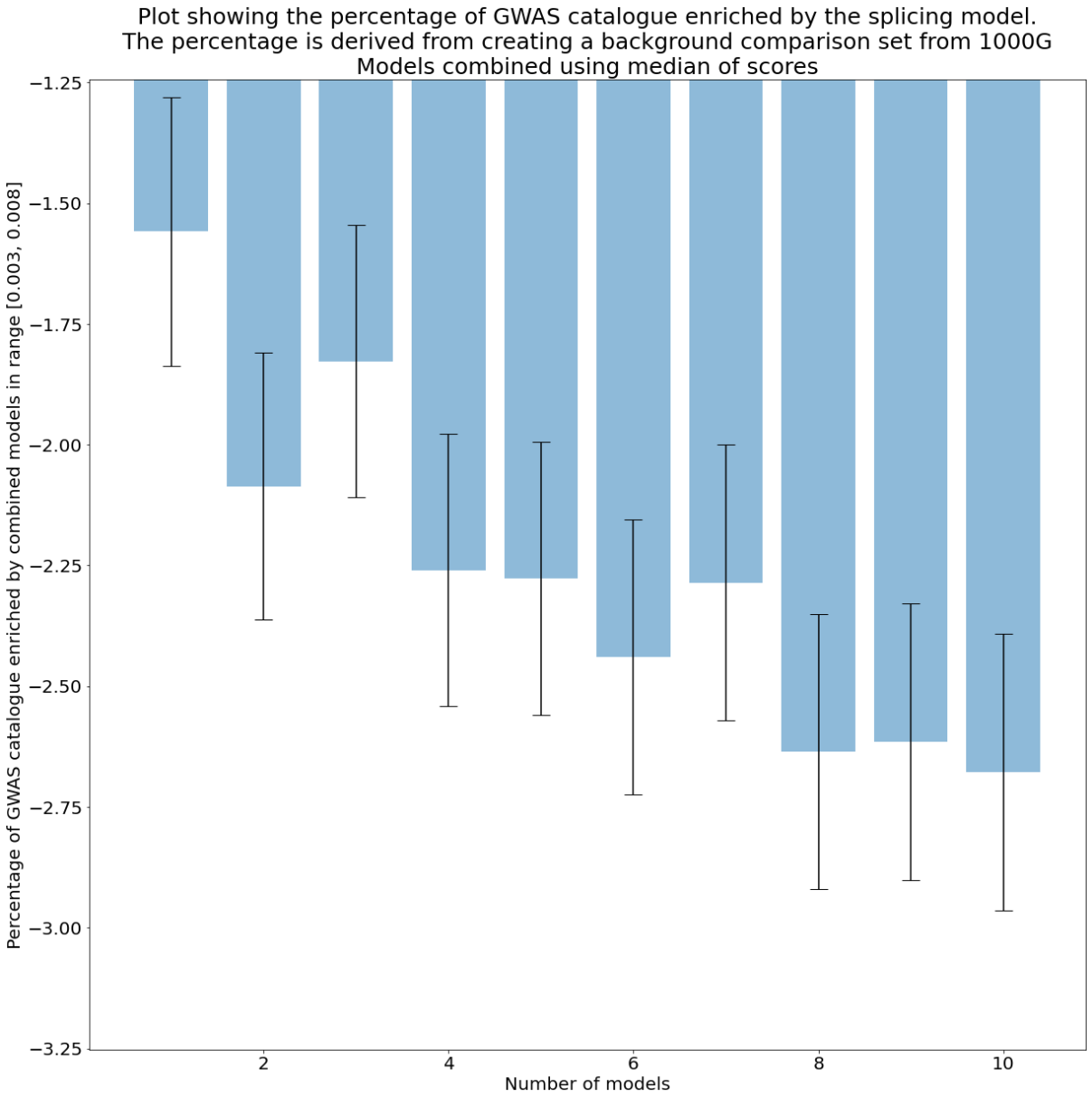


Figure 6.17: Analysis to determine why increasing the number of models increases the enrichment. We observe many background SNPs previously found with scores >= 0.008 end up within this bin [0.003, 0.008] through combining models. Reducing statistical noise is key in getting more precise enrichments. Further improvements will be as a result of reducing the background noise.

In summary, we find that a significant proportion of GWAS is enriched using our model, including many sites enriched by relatively weak splice-scores. Our confidence in the validity

of this approach is increased by the result that combining models further improves this enrichment. This suggests that the model is slowly picking out genuine splice-altering signals above the background noise.

6.4 Enrichment in or near annotated exons

We can restrict the genome-wide analysis to regions formed by padding exons by 100bps. These are the regions where we would expect to find splice-altering SNPs based upon genomic annotations alone. This allows us to compare the size of our found enrichments to deep-intronic regions. Do we find a greater excess of GWAS SNPs in near-exonic regions or deep-intronic regions?

Within this 100bp window we find an excess of GWAS SNPs predicted to affect splicing (change in predicted PSI >= 0.008) as (667, (8153 - 7486)). However, despite it being a small minority of the excess of the GWAS SNPs in all gene-regions, it is higher as a percentage. Of this excess, (479, (5911-5432)) are found in purely exonic regions, with (409, (4263 - 3854)) more than 30 base pairs from known exon-intron boundaries. This means that our model believes that a substantial proportion of exonic GWAS SNPs, far from known splice sites, should be considered to have an additional splicing mechanism.

A larger percentage of GWAS catalogue SNPs are seen as splice-altering by our model in exonic regions, compared to the genome-wide analysis. However, despite this larger percentage, in total counts we find more excess SNPs in deep-intronic regions than in near-exonic regions. We explore this in the next subsection.

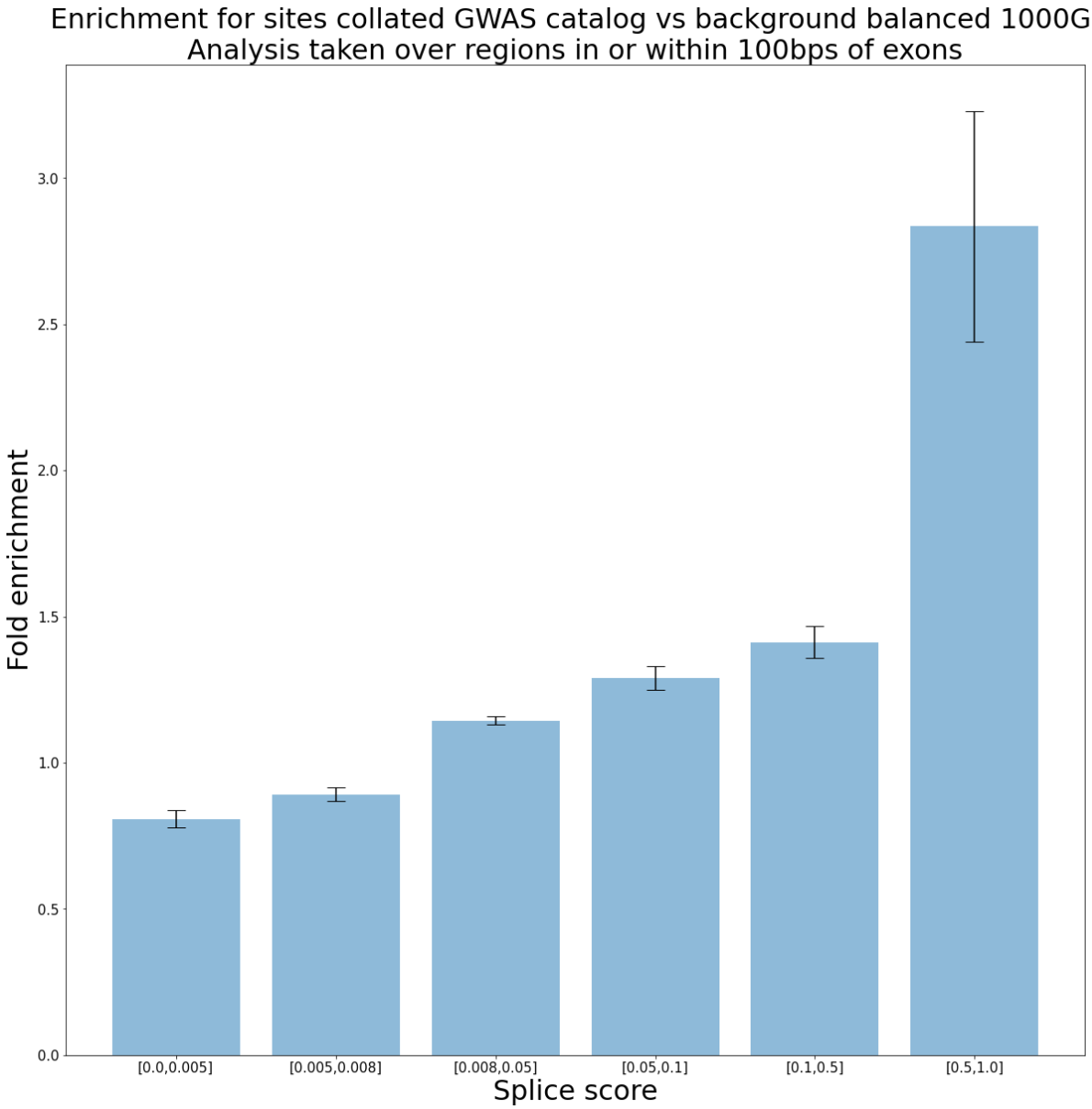


Figure 6.18: Enrichment analysis for near exonic GWAS catalogue SNPs against a rebalanced 1000G SNP background. Background balanced on allele frequency and gene location. Enrichment analysis restricted to GWAS catalogue and background 1000G SNPs found within 100bps of known exons. Larger fold enrichments found for near exonic SNPs compared to the genome-wide analysis as seen in figure 6.10.

6.5 Enrichment in deep intronic regions

Akin to the previous enrichment analysis, here we directly compare the enrichment found in deep-intronic regions to near intron-exon boundary regions. In this case, as the analysis is constrained to the same precise genomic locations, we only need to balance our test and control sets based on European allele frequency.

Surprisingly, we find a greater excess of GWAS SNPs predicted to affect splicing (change in predicted PSI >= 0.008) in deep-intronic regions (1884, (15897 -14013)) compared to in-regions near known splice sites (667, (8153 - 7486)). Here we are defining deep-intronic as regions greater or equal to 100 base pairs from a known exon.

We have two plausible explanations for this excess of deep-intronic predictions. Either these are deep-intronic SNPs predicted to have distal effect on a known annotated splice-site, or they are predicted to result in the loss or gain of unannotated splice-sites. Of the 15897 deep-intronic GWAS SNPs predicted to affect splicing, only 755 are predicted to affect the nearest annotated splice site. This means that the majority of the excess must be due to splicing-alterations of unannotated splice-sites in regions classified as intronic.

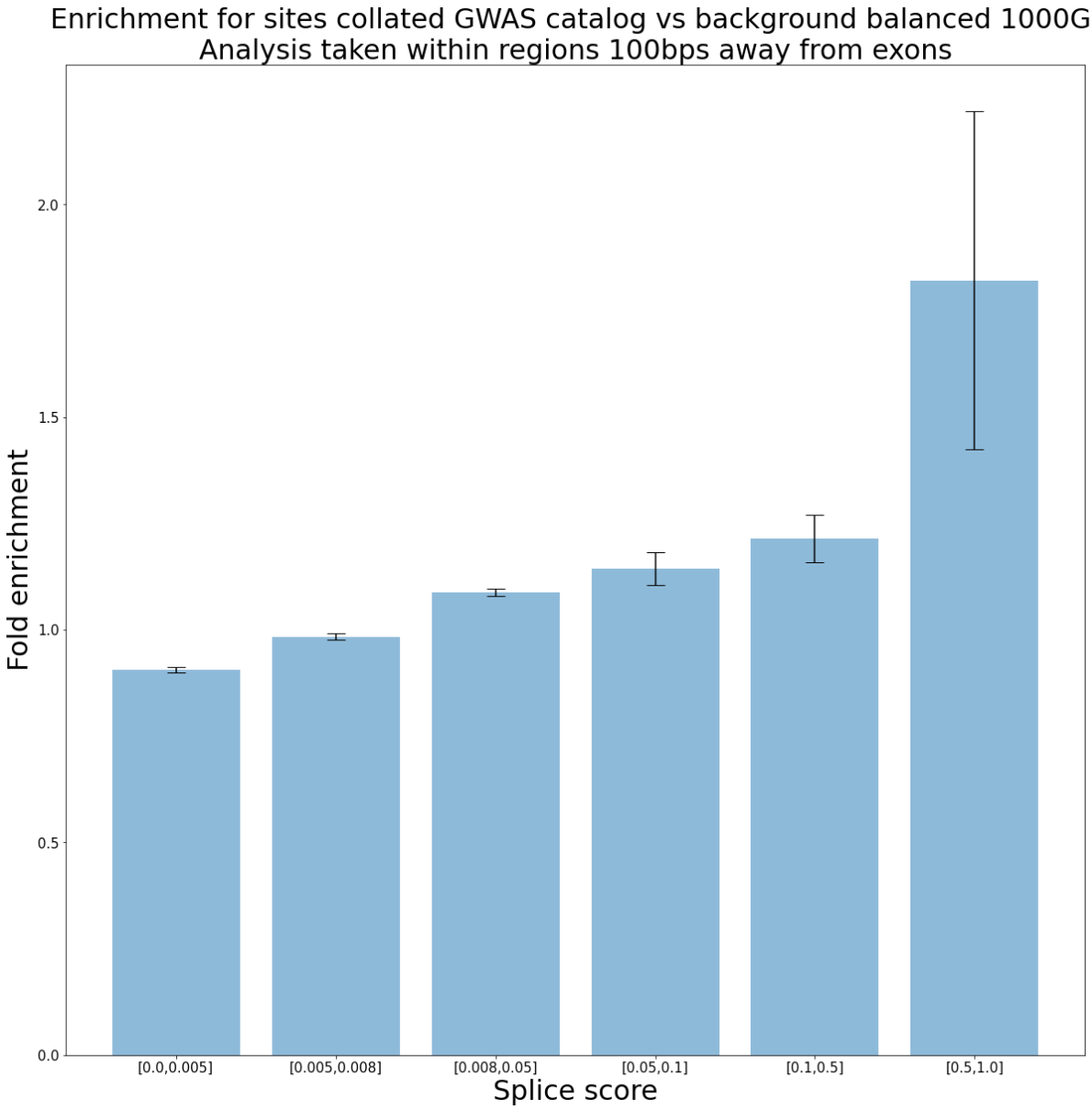


Figure 6.19: Enrichment analysis for deep intronic GWAS catalogue SNPs against a rebalanced 1000G SNP background. Background balanced on allele frequency and gene location. Enrichment analysis restricted to GWAS catalogue and background 1000G SNPs found 100bps away from known exons. Smaller fold enrichments found for deep intronic SNPs compared to the genome-wide analysis as seen in figure 6.10.

One plausible explanation we tend to return to is that these are sites which additionally have RNA-surveillance, and that AS-NMD is a common feature of gene regulation and, by extension, of GWAS signals. An issue we find in testing this theory is that, due to this excess, GWAS SNPs typically have weak splice-scores. This makes confident computational identification of predicted gained or lost exons difficult. Consequently, we can compute whether we can predict NMD only for a subset of these predicted splicing-changes defined by having a splice score of >= 0.05, predicted both at a donor and acceptor site. Doing this we obtain an enrichment of 1.05 (1.02-1.07, 95% CI) for predicted AS-NMD sites compared to an enrichment of 1.16 (1.08-1.22, 95% CI) for deep-intronic predicted sites which are not predicted to induce NMD. For both categories we find a similar excess of predicted SNPs within the GWAS catalogue: (45, 294 - 248) for non-NMD sites and (78, 1275 -1196) for NMD sites.

Based on this analysis we can, therefore, claim that deep-intronic regions have a greater excess of GWAS SNPs predicted to affect splicing compared to exonic and near-exonic regions. We can claim with certainty that a large proportion of these excess GWAS SNPs are affecting intronic splice-sites (in annotated intronic-regions, as opposed to distally affecting known splice-sites). Despite the conclusion that a majority of such SNPs are predicted to affect NMD-exons (65159 out of 78713), we cannot confidently discern whether these are the cause of the excess of predicted GWAS SNPs due to the difficulty in applying the NMD model to such small splicing changes individually.

6.6 LD-bin Enrichment analysis

As previously described, to test the robustness of the enrichment results discussed above, we also used an alternative method to construct a further set of control SNPs. Rather than matching spatial location and allele frequency directly, we decided to do so indirectly by choosing, for each GWAS lead SNP, a 1000G SNP that had an R^2 value in the range [0.6, 0.9] (linkage between SNPs was computed using plink version 1.07 [47]). When more than one such SNP was available, the one in closest proximity to the lead SNP was chosen. In this way, the linkage disequilibrium structure of control and GWAS SNPs is similar, and it is known that SNPs in large LD blocks have a higher probability of becoming the lead SNP in a GWAS study, a fact exploited by LD score regression [48]. By matching on R^2, control SNPs are also indirectly partially matched for both spatial location (to within the LD block) and allele frequency (because only SNPs with similar allele frequency can have high R^2). Requiring R^2 < 0.9 ensures that the lead SNP itself will not be chosen as its own control.

This analysis acts as an orthogonal check for other analyses. In brief, we find that the results of the previous enrichment analysis are replicated, and that they are so robust in terms of the control set used and showing such a similar result that enrichment becomes present when splice scores are greater than 0.01.

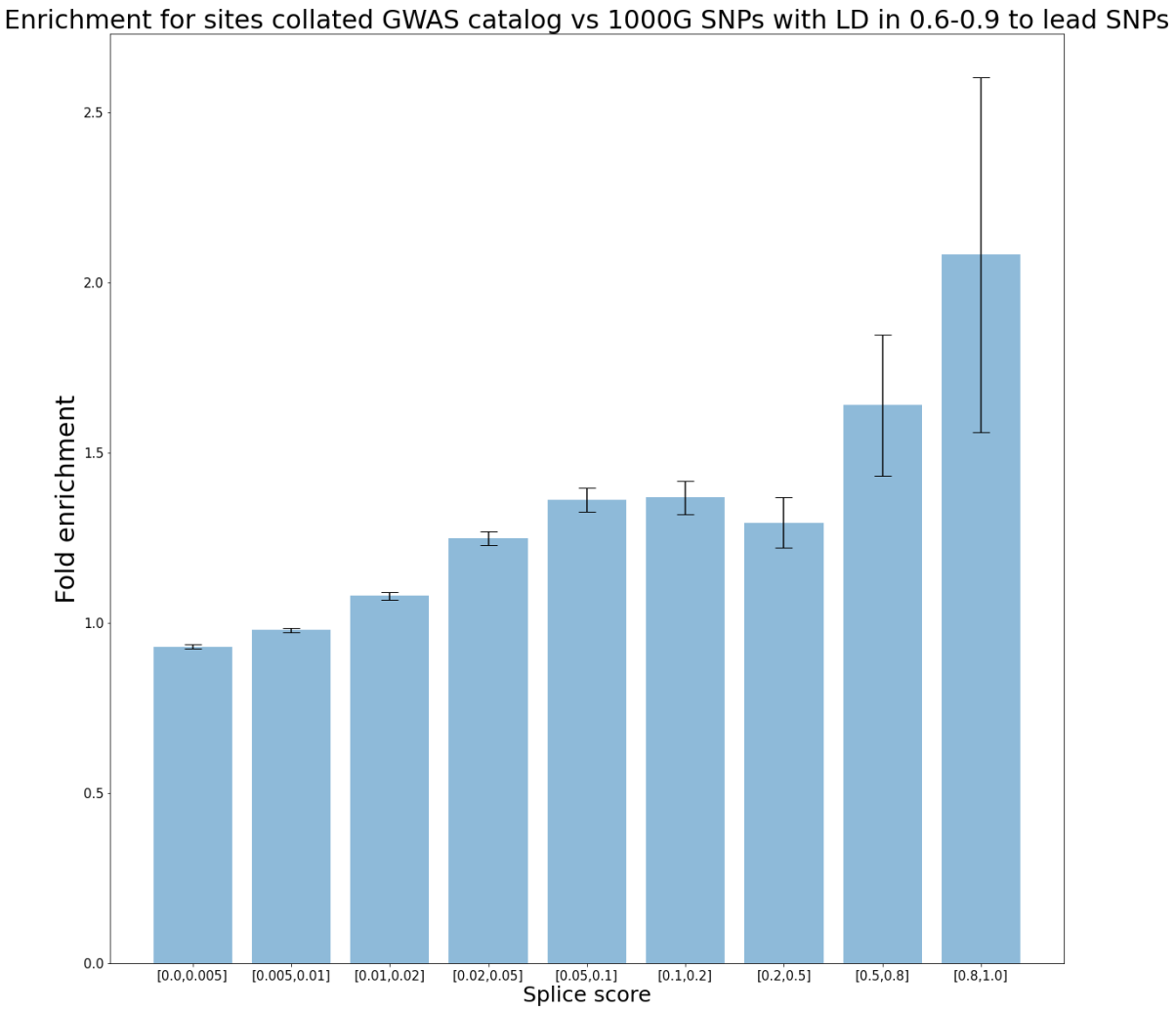


Figure 6.20: Enrichment analysis for deep intronic GWAS catalogue SNPs against a rebalanced 1000G SNP background. Background set obtained from 1000G by obtaining matching SNPs based upon LD to GWAS SNPs. Similar enrichment distribution found as found in figure 6.10. Acts as a control analysis for the results found by rebalancing on genomic location and allele frequency.

6.7 Examples of predicted NMD splice altering GWAS catalogue polymorphisms.

In this section we showcase three illustrative examples of GWAS catalogue SNPs in which our model predicts a AS-NMD mechanism. In the first example we obtained sufficient read-evidence in our standard RNA-seq biobank to support our claim for this example. The next two examples require the usage of our model predictions combined with the POINT-seq datasets to validate. We include two interesting mechanisms, one an example of a predicted lost pseudo-exon, and the other a predicted loss of an intron in the 3' tail.

6.7.1 SPATA1

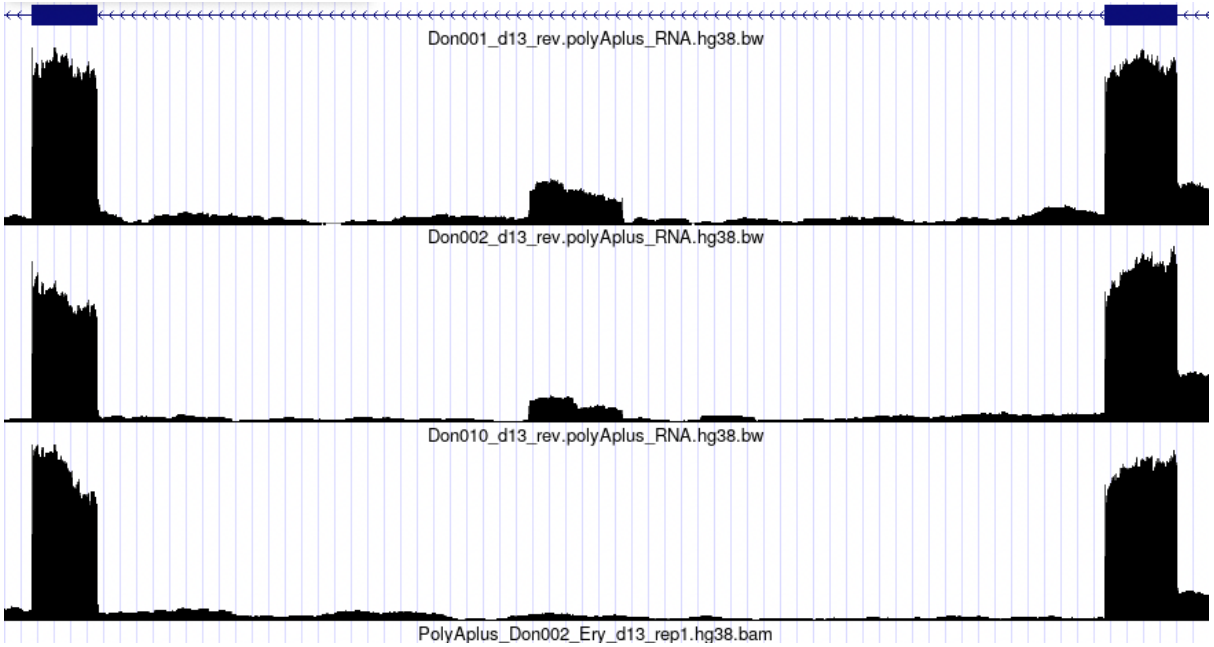


Figure 6.21: Context within gene where evident that expression is far lower than surrounding exons. These three tracks show RNA-seq evidence for three samples Don001, Don002 and Don010 from our biobank. The three tracks are for samples which are homozygous reference, heterozygous and homozygous alternative for the GWAS SNP chr1:158650687:C->G. We observe more read evidence for the pseudoexon in Don001 homozygous reference, followed by Don002 which is heterozygous. Read evidence is absent from Don010. The psuedo-exon includes in-frame stop codons for all frames.

The polymorphism chr1:158650687:C->G is predicted to lead to the loss of a pseudo-exon. This exon fits the NMD hypothesis well, given that it has stop-codons in all frame positions. Our model predicts a strong but alternative splice site with the SNP removing entirely the pseudo-exon. It is marked in GWAS catalogue as just an “intron variant”, due to the absence of any genome annotations for this pseudo-exon.

Surprisingly, despite clear stop-codons and NMD model-prediction, we still have relatively high expression for this locus in mature RNA (PSI ~25% in homozygous samples). This is consistent with our predictions combined with the evidence that NMD targets appear to be decayed to a varied degree.

This polymorphism is obtained from two GWAS papers: Astle et al [49] and Chen et al [50], which study red cell width distribution and mean platelet volume. Our biobank samples are obtained from blood samples, and it is not surprising that these two GWAS studies are on blood traits. Interestingly, they report different risk alleles (G for Astle et al and C for Chen et al). It is a gene that is expressed, but whose known biology is spermatogenesis ([51], [52]).

6.7.2 UBP1 example

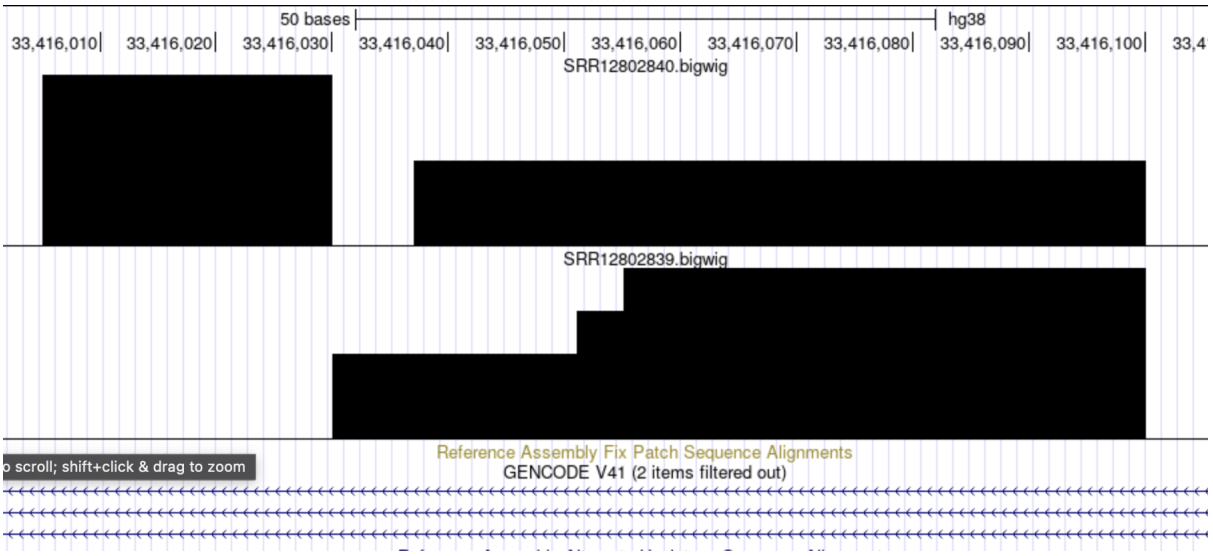


Figure 6.22: Read evidence for pseudo-exon found in datasets for two replicates obtained using POINT-seq on Hela cells [40]. We combined this with our model's prediction of pseudo-exon loss and NMD. The SNP is found at chr3:33416001 and is reported in three separate GWAS studies.

This SNP at position chr3:33416001 is labelled as a "splice_region_variant" in the GWAS catalogue and our model suggests the mechanism for this has the disruption of a cryptic NMD-exon.

This polymorphism is obtained from 3 separate GWAS studies: Sakaue et al [53], Pazoki et al [54] and Han et al [55], in order to study traits C-reactive protein levels, liver enzyme levels and Serum albumin levels respectively. UBP1 is a transcription factor known to regulate multiple genes [56].

Our proposed mechanism is that the risk allele of the polymorphism leads to a predicted loss of the pseudo-exon in figure 7.2, evidenced in POINT-seq. Ideally, if we were to have

POINT-seq for our biobank samples, we would be able to validate this claim on those samples directly.

6.7.3 TCF19 gene

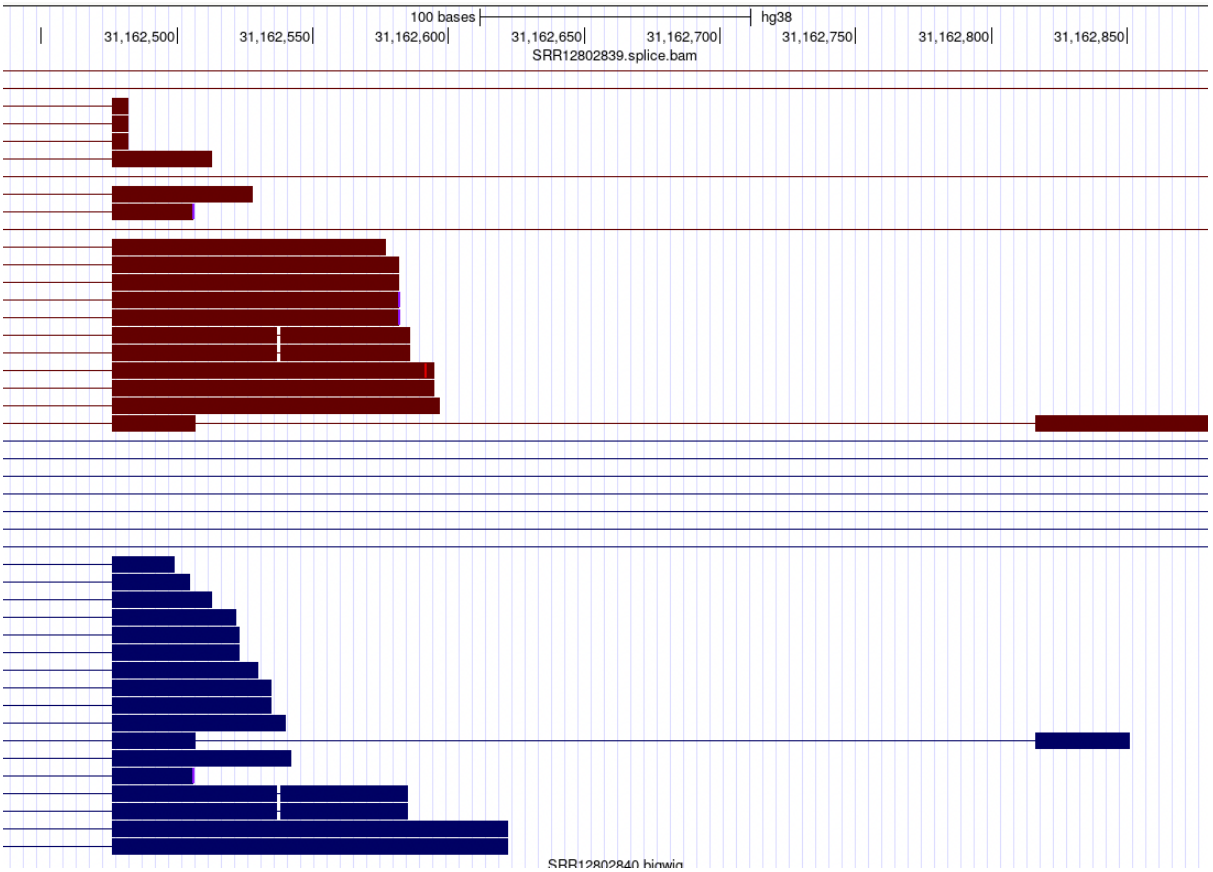


Figure 6.23: Model predicts presence of cryptic intron in last exon of TCF19 gene. Presence of SNP at chr6:31162816 (rs1419881) leads to predicted loss of cryptic intron. Shown in these tracks is the read evidence found from two POINT-seq Hela cell replicates which show presence of UTR intron [40].

This is an interesting example of an unusual mechanism which would be predicted to induce NMD due to the creation of an intron in the last 3' UTR exon. It is associated with two separate GWAS traits within the GWAS catalogue, identified as a "3_prime_UTR_variant". Our analysis suggests as a mechanism that an existing alternatively spliced intron, present in the reference genome, is lost in the presence of the SNP. Our hypothesis for the genetic

mechanism would, therefore, be increased productive expression of the TCF19 gene due to loss of NMD.

This polymorphism is identified in two separate GWAS studies: Kim et al [57] and Jiang et al [58], both for Chronic hepatitis B infection. It is noted in Kim et al that this SNP rs1419881 obtained their largest signal in relation to the studied trait, and that TCF19 is a plausible causal gene due to its regulation of cell cycle progression. It is also noted that TCF19 has been identified as a causal gene in studies concerning type 1 diabetes [59]. However, the mechanism is not fully discussed with only the fact it is found within the 3' UTR mentioned. The current annotation would lead investigations to either chase a mechanism related to binding of proteins, or mRNA to 3' UTR, or possible effects on polyadenylation. However, our analysis shows clearly that this affects the splicing of this region via the removal of a RNA-decayed transcript.

The combined evidence of an additional cryptic NMD-inducing intron found within POINT-seq data, along with our model prediction of its loss, suggests a mechanism which would result in increased gene expression. Again, we have formed our hypothesis based upon the analysis of POINT-seq data obtained from Hela cells. Having POINT-seq data on our biobank samples would enable a more direct validation of this molecular mechanism.

6.8 Discussion

The conclusions of these analyses partially fit the suspected biology of GWAS traits. We find that a large excess of near exonic SNPs are predicted by our model. We also find that relatively few SNPs have strong splice-scores, suggesting that likely causal splicing effects in GWAS are generally subtle in nature.

However, our analysis suggests new and interesting potential facets of GWAS genetics. From our analysis it appears plausible that many of the excess SNPs causing the detected enrichment may have more than one potential mechanism. We observed a high fraction of annotated missense variants having a predicted alternative hypothesis of splicing. Our predictions do not rule out that coding changes are causal, but it suggests that both splicing and coding changes should be considered in untangling the underlying mechanism.

More surprisingly, we find a large enrichment for deep-intronic GWAS SNPs predicted by our splicing model, many of which are predicted to be AS-NMD transcripts.

This finding has important implications in how the functional fine-mapping of such intronic SNPs should be carried out. Variants in gene introns have two main potential explanations, namely that they affect splicing or that they affect the function of gene regulatory elements, such as in enhancers, which are also frequently found in introns.

A standard approach would be to rule out splicing as a mechanism using *SpliceAI* scores, however, these are typically masked out within deep-intronic regions [60]. Our analysis therefore suggests that, in a large number of cases, this will lead to incorrectly ruling out splicing as a mechanism and that it will lead to wasted effort trying to experimentally follow a enhancer based hypothesis.

Similarly, we showed that there is an enrichment even for small predicted splicing scores, presumably reflecting small changes in splicing. While the enrichment for these small changes is highly evident in a meta-analysis of the complete set of genetics, the robust prediction of such small predicted splicing-changes at individual loci is bioinformatically challenging. Similarly, finding a greater number of excess GWAS SNPs in deep-intronic regions suggests that many genomic regions beyond near-exonic regions should be

considered with splicing as a mechanism for fine-mapping GWAS SNPs. For such regions we will need to test both the removal of a NMD-target transcript or a low level non-decaying transcript as hypotheses.

In conclusion, based on our model predictions and their preliminary validation, we find that splicing is a prevalent mechanism in GWAS signals and that these are enriched in regions not normally identified with a splicing mechanism. This has ramifications for pipelines which seek to interpret GWAS results, especially if the pipeline uses an *in silico* model to argue that splicing is not a key mechanism. Furthermore, the excess GWAS SNPs predicted to be NMD-target transcripts suggests that it is not sufficient to use standard RNA-seq libraries to negate a splicing-mechanism hypothesis. Therefore, sensitive transcriptomics methods that can quantitatively detect splicing events before NMD are needed to validate and to train a next generation of NMD aware splice prediction algorithms. From the analyses in chapter 4, POINT-seq seems a likely candidate methodology, or at least an experimental starting point.

7 Discussion

At a high level our ambition is to understand the phenotypic variation which arises from a particular haplotype. Broadly, to understand the consequences of a particular haplotype we need to better understand the gene-regulation, coding-variation, structural variation, UTR sequences and lastly splicing.

7.1 Why is the prediction of splicing and the effects of sequence variation important?

We detailed why an *in silico* model for *alternative splicing* is important in different chapters across this thesis. We saw in chapter 2 how such an *in silico* model allows one to test hypotheses concerning the underlying biology of splicing. For instance, by using the *in silico* model we are able to provide insights into the mechanism in which splice sites are selected, and the extent to which an exon-definition model explains joint acceptor-donor splice site selection.

In chapter 3 we used an extensive mutation database to show that splicing is a common mechanism underlying mendelian disease, and that our model can be applied to provide better genetic diagnoses. In chapter 6 we applied the same model to common genetic variants, and showed that our model led to better understanding of splicing as a mechanism of common disease. Increasing our understanding of the subtleties of splicing will be helpful in providing guidance for drug development. After all, splice-disrupting variation, like coding-disrupting variation, has the key property shared to make the causal gene is evident: the predicted disrupted splice-site automatically tells the analyst which gene is affected.

We showed examples of how GWAS SNPs could be fine-mapped using a mixture of an *in silico* model and POINT-seq data. Moreover, we included an example found within a 3’ UTR

which leads to the variable inclusion of an intron in the 3'UTR to showcase the complexity of mechanisms that alternative splicing when linked with RNA-decay can generate.

7.2 What was lacking in existing approaches? What motivated our work?

Our analysis in chapter 2 shows clearly how splicing is not an all-or-nothing affair. Many sites targeted by the spliceosome, as evidenced in GTEx, are genuine splice sites but far from canonical. The standout existing approach to modelling splicing in recent years has been *SpliceAI*. However, by design, this model treats all putative splice sites in the same manner, regardless of how they are *alternatively spliced*.

We can observe how this does not readily apply to our work on GWAS. In chapter 6 we observed that a large burden of our predicted splice-altering signal is found in deep intronic regions. By definition, these will be genomic regions with no constitutive splicing and not annotated as canonical splicing. Our contention is that, in order to better model splicing for the purposes of both common disease and for clinical diagnostics, a model must take care to integrate all forms of *alternative splicing* and not just modelled as constitutive splice sites.

Finally, as observed in chapter 4, splicing signals will not determine transcript abundance alone due to RNA-surveillance.

Previous approaches have not satisfactorily attempted to integrate these related and interconnected components of transcript processing into a satisfactory model predictive of gene transcriptional output. In our clinical collaborations it has been important to understand whether the alternative transcript is likely to be expressed or decayed, and to form more correct hypotheses of mechanism and design validation experiments.

Likewise, in our work on common variation a likely hypothesis is that much of this deep-intronic splice-variation is coupled to RNA-decay.

A caveat which is important to state in existing approaches (including our approach) is the almost complete predominance of stable mRNA, measured using mRNA-seq, as training data. Our approach to training a splice model is predicated upon the assumption that the PSIs found within our training examples broadly reflect the reality of nascent transcription, and the prevalence of nonsense-mediated decay and recursive splicing show that this is an important approximation.

7.3 Technological advances

We observed in chapter 3 that our model represents an advancement in terms of diagnostic accuracy. Through the combination of multiple separately trained neural networks the model is able to obtain better accuracy than what is obtained using *SpliceAI* alone. However, even from an optimistic perspective we still observe that this improvement from combining the output of multiple models of the same type has diminishing returns. The improvement in accuracy between our model and *SpliceAI* would only be observable very rarely and does not fundamentally alter a clinical pipeline handling a limited amount of samples.

What helped us substantially was the work in chapter 3. The use of a new model, subsequent to the neural network, that now incorporates other information (such as gene definition and gene tolerance scores) showed clearly improved predictive accuracy. In fact, the improvements through better classification in this regard far exceeded the improvements that we were able to make using the neural network alone.

7.4 Biological of splicing

Nonsense-mediated decay (NMD) is a process that confounds the modelling of splicing, as stable mRNA is used for training. In the vast majority of cases, a predicted splice site without evidence for splicing would decay through the NMD pathway, and could therefore be "explained away". This undermines simplistic validation of any splicing model.

We saw from the analysis on POINT-seq data how these AS-NMD sites are frequent for predicted *alternatively spliced* sites. This means that any model which seeks to identify these *alternatively spliced* transcripts will come up against this dilemma. If AS-NMD is common amongst less utilised transcripts, based upon its true splice rate, then the model will appear inaccurate in these predictions. Currently, our model is more accurate when validated on POINT-seq despite being trained on RNA-seq. Ignoring NMD, our model predicts a fair number of apparent false-positive splice sites, where we do not find evidence in RNA-seq data. A good fraction of these do have evidence in POINT-seq data, implying that POINT-seq is more reflective of the underlying nascent splicing.

This is similar to what we observed in chapter 2 in relation to recursive splicing. A splice-site subject to recursive splicing is not visible in our training data, meaning that, in theory, the model is biassed against learning this signal. Despite this, the model readily learnt a signal consistent with recursive splicing, in particular predicting exon-skipping. This suggests that partial or complete recursive splicing is an underexplored mechanism when attempting to understand alternative splicing.

Many of the common disease-associated variants produce alternate transcripts which are predicted to undergo nonsense mediated decay. Much of the enrichment we found is in deeply intronic regions. This leads to several questions. How much of the genome away from known exons is important for splicing regulation? How much of splicing regulation occurs via these decay mechanisms?

Instead of considering the predictions found within these intronic regions as model noise, our enrichment analyses suggest that a proportion are true splicing events driving the genetic associations. This opens up questions on how model predictions within these locations should be utilised. As with most groups, our laboratory had historically utilised masked files of splicing scores to discount splicing as a mechanism in deep intronic regions, which is subsequently validated by the absence of any evidence of splicing in RNA-seq. Our analysis on intronic enrichments suggests that deep intronic regions do harbour a number of invisible but active splice signals, opening the possibility that mutating these will lead to phenotypic consequences.

An additional consideration is the significance of weak splicing. A large focus is naturally applied to canonical splice sites and their prediction. Intuitively, this is a sensible approach when focusing on applications to rare variants. Again in our GWAS work we found that a greater percentage of GWAS SNPs were enriched for small splice changes disrupting weak splice sites. In the context of GWAS, this is not unreasonable where effect sizes need not be large to produce a significant association.

This is consistent with an understanding of splicing as a stochastic process where mostly minor transcripts are cleaned using RNA-surveillance. In the context of GWAS, these transcripts may have some phenotypic impact, not due to the production of aberrant proteins, but rather because a change in production of AS-NMD transcripts decreases or increases the productive output of the gene. In time, a better understanding of this complex picture will emerge. We hypothesise that splicing in deep intronic regions contributes to GWAS results, and that it is currently under-appreciated.

7.5 Future perspectives

We believe that new assays that give access to a better approximation of the nascent transcriptome, such as POINT-seq, will provide more useful datasets to train and validate splicing models, as it appears to reflect pre-mRNA splicing better than GTEx or our internal RNA-seq libraries. This suggests that a really useful addition to our work would be to develop, train and test a neural network model to predict splicing not on RNA-seq but additionally on POINT-seq. Many models currently being developed or found in the literature utilise GTEx like we did. This is because there is a significant variety of cell types of good read-depth which our example POINT-seq datasets do not have. With a fuller POINT-seq library, perhaps for a variety of tissues, we could really start to learn how different signatures change the levels of nascent splicing.

Small changes in the model's results could be directly linked to a better understanding of splicing. Any small change could easily be evaluated in terms of its ability to predict the nascent RNA. Moreover, without this unobserved secondary process (RNA-decay) we can be far more confident that the differences in empirical PSI levels between splice sites or between tissue types are a reflection of the underlying splicing.

Our contention is that our model is very adept at modelling constitutive splice sites. We observe that the model predicts very few false positives. Moreover, this is also true for *SpliceAI* which reports very impressive metrics in their paper. However, there remains a large proportion of alternative splicing which no model is adept at predicting. Our attempt to rectify this automatically hits the stumbling block of nonsense-mediated decay. If 40% of the splice sites are subject to NMD, then no model will be able to pick out this signal without

being heavily penalised in training. Such expansive datasets will change entirely the parameters of this problem as a learning task. We also saw that some apparent constitutive splice sites, which form a part of canonical transcripts, are actually less strong in POINT-seq, but that due to other sites experiencing NMD they are inflated. This suggests that benefits in *learning alternative splicing* will not just be at the low level splicing but also in accurately learning moderate splicing.

With such a dataset our model could be directly retrained on this enriched data. This would enable not only the development of neural networks to predict nascent splicing but also of a more comprehensive model to discern RNA-decay. In this thesis we presented a simple model to only predict one of the known mechanisms of RNA-decay. Such a comprehensive dataset, more truly reflective of nascent transcription and splicing, would enable one to fully develop a comprehensive package combining not just nonsense-mediated decay but also no-stop decay, no-go decay and potentially recursive splicing.

In addition, over the last 4 years we tried many techniques to augment our models for which we did not obtain evidenced improvements. It is known that the RNA-secondary structure has an effect on splicing, that antisense splicing has an effect, that splicing can be tissue-specific and that the location of the polyadenylation tail is important. With this in mind, we designed our modelling architecture to be able to additionally include tissue specific aspects, the secondary structure etc. However, the consistent issue was validation, which helps the modeller know whether these small changes are improving the models or not. By including data that more accurately reflects antisense splicing, RNA-secondary structure, or other features we would obtain even more predictive models for splicing

events influenced by these factors. However, often these transcripts were not readily evidenced in GTEx, thereby making it hard to discern whether our understanding of splicing is improving or not. It would therefore be interesting to further investigate the utility of methods such as POINT-seq to detect such events, when more comprehensive datasets become available.

With a dataset better reflective of nascent splicing the modeller would be able to rapidly integrate such modelling experiments, thereby bringing even further the modelling accuracy. In turn, this would enable significant advances in our ability to confidently apply *in silico* splicing models genome-wide, and by combining them with the POINT-seq datasets to enable us to mechanistically solve more GWAS signals and clinical variants.

8 References

[1] E. Park, Z. Pan, Z. Zhang, L. Lin, and Y. Xing, 'The Expanding Landscape of Alternative Splicing Variation in Human Populations', *Am. J. Hum. Genet.*, vol. 102, no. 1, pp. 11–26, Jan. 2018, doi: 10.1016/j.ajhg.2017.11.002.

[2] J. Cheng *et al.*, 'MMSplice: modular modeling improves the predictions of genetic variant effects on splicing', *Genome Biol.*, vol. 20, no. 1, p. 48, Dec. 2019, doi: 10.1186/s13059-019-1653-z.

[3] C. Rowlands *et al.*, 'Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders', *Sci. Rep.*, vol. 11, no. 1, p. 20607, Dec. 2021, doi: 10.1038/s41598-021-99747-2.

[4] K. Jaganathan *et al.*, 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, vol. 176, no. 3, pp. 535-548.e24, Jan. 2019, doi: 10.1016/j.cell.2018.12.015.

[5] A. Chlebowski, M. Lubas, T. H. Jensen, and A. Dziembowski, 'RNA decay machines: The exosome', *Biochim. Biophys. Acta BBA - Gene Regul. Mech.*, vol. 1829, no. 6–7, pp. 552–560, Jun. 2013, doi: 10.1016/j.bbagrm.2013.01.006.

[6] S. Nasif, L. Contu, and O. Mühlemann, 'Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression', *Semin. Cell Dev. Biol.*, vol. 75, pp. 78–87, Mar. 2018, doi: 10.1016/j.semcdb.2017.08.053.

[7] C. J. Wright, C. W. J. Smith, and C. D. Jiggins, 'Alternative splicing as a source of phenotypic diversity', *Nat. Rev. Genet.*, vol. 23, no. 11, pp. 697–710, Nov. 2022, doi: 10.1038/s41576-022-00514-4.

[8] C. Kovalak, S. Donovan, A. A. Bicknell, M. Metkar, and M. J. Moore, 'Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved alternative splicing nonsense-mediated decay pathways', *Genome Biol.*, vol. 22, no. 1, p. 132, May 2021, doi: 10.1186/s13059-021-02309-y.

[9] B. Saudemont *et al.*, 'The fitness cost of mis-splicing is the main determinant of alternative splicing patterns', *Genome Biol.*, vol. 18, no. 1, p. 208, Dec. 2017, doi: 10.1186/s13059-017-1344-6.

[10] J. G. Conboy, 'Unannotated splicing regulatory elements in deep intron space', *WIREs RNA*, vol. 12, no. 5, Sep. 2021, doi: 10.1002/wrna.1656.

[11] C. R. Sibley *et al.*, 'Recursive splicing in long vertebrate genes', *Nature*, vol. 521, no. 7552, pp. 371–375, May 2015, doi: 10.1038/nature14466.

[12] E. R. Gamazon and B. E. Stranger, 'Genomics of alternative splicing: evolution, development and pathophysiology', *Hum. Genet.*, vol. 133, no. 6, pp. 679–687, Jun. 2014, doi: 10.1007/s00439-013-1411-3.

[13] B. J. Blencowe, 'Alternative Splicing: New Insights from Global Analyses', *Cell*, vol. 126, no. 1, pp. 37–47, Jul. 2006, doi: 10.1016/j.cell.2006.06.023.

[14] M. H. Çelik *et al.*, 'Aberrant splicing prediction across human tissues', Bioinformatics, preprint, Jun. 2022. doi: 10.1101/2022.06.13.495326.

[15] E. Agirre, A. J. Oldfield, N. Bellora, A. Segelle, and R. F. Luco, 'Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition', *Nat. Commun.*, vol. 12, no. 1, p. 682, Dec. 2021, doi: 10.1038/s41467-021-20979-x.

[16] S. Schafer, K. Miao, C. C. Benson, M. Heinig, S. A. Cook, and N. Hubner, 'Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI)', *Curr. Protoc. Hum. Genet.*,

vol. 87, no. 1, Oct. 2015, doi: 10.1002/0471142905.hg1116s87.
[17] R. Brown, ‘Deep learning for alternative splicing’, University of Oxford, 2019.
[18] H. Bretschneider, S. Gandhi, A. G. Deshwar, K. Zuberi, and B. J. Frey, ‘COSSMO: predicting competitive alternative splice site selection using deep learning’, *Bioinformatics*, vol. 34, no. 13, pp. i429–i437, Jul. 2018, doi: 10.1093/bioinformatics/bty244.
[19] The GTEx Consortium *et al.*, ‘The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans’, *Science*, vol. 348, no. 6235, pp. 648–660, May 2015, doi: 10.1126/science.1262110.
[20] J. Ule and B. J. Blencowe, ‘Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution’, *Mol. Cell*, vol. 76, no. 2, pp. 329–345, Oct. 2019, doi: 10.1016/j.molcel.2019.09.017.
[21] G. M. Venkatesh, Y. G. Naresh, S. Little, and N. E. O’Connor, ‘A Deep Residual Architecture for Skin Lesion Segmentation’, in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, vol. 11041, D. Stoyanov, Z. Taylor, D. Sarikaya, J. McLeod, M. A. González Ballester, N. C. F. Codella, A. Martel, L. Maier-Hein, A. Malpani, M. A. Zenati, S. De Ribaupierre, L. Xiongbiao, T. Collins, T. Reichl, K. Drechsler, M. Erdt, M. G. Linguraru, C. Oyarzun Laura, R. Shekhar, S. Wesarg, M. E. Celebi, K. Dana, and A. Halpern, Eds. Cham: Springer International Publishing, 2018, pp. 277–284. doi: 10.1007/978-3-030-01201-4_30.
[22] H. Huang *et al.*, ‘UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation’, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 1055–1059. doi: 10.1109/ICASSP40776.2020.9053405.
[23] Y. Weng, T. Zhou, Y. Li, and X. Qiu, ‘NAS-Unet: Neural Architecture Search for Medical Image Segmentation’, *IEEE Access*, vol. 7, pp. 44247–44257, 2019, doi: 10.1109/ACCESS.2019.2908991.
[24] F. Nazem, F. Ghasemi, A. Fassihi, and A. M. Dehnavi, ‘3D U-Net: A voxel-based method in binding site prediction of protein structure’, *J. Bioinform. Comput. Biol.*, vol. 19, no. 02, p. 2150006, Apr. 2021, doi: 10.1142/S0219720021500062.
[25] J. Valcárcel and M. R. Green, ‘The SR protein family: pleiotropic functions in pre-mRNA splicing’, *Trends Biochem. Sci.*, vol. 21, no. 8, pp. 296–301, Aug. 1996, doi: 10.1016/S0968-0004(96)10039-6.
[26] A. Corvelo, M. Hallegger, C. W. J. Smith, and E. Eyras, ‘Genome-Wide Association between Branch Point Properties and Alternative Splicing’, *PLoS Comput. Biol.*, vol. 6, no. 11, p. e1001016, Nov. 2010, doi: 10.1371/journal.pcbi.1001016.
[27] M. K. Sakharkar, V. T. K. Chow, and P. Kangueane, ‘Distributions of exons and introns in the human genome’, *In Silico Biol.*, vol. 4, no. 4, pp. 387–393, 2004.
[28] L. Blazquez *et al.*, ‘Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing’, *Mol. Cell*, vol. 72, no. 3, pp. 496-509.e9, Nov. 2018, doi: 10.1016/j.molcel.2018.09.033.
[29] Y. Strauch, J. Lord, M. Niranjan, and D. Baralle, ‘CI-SpliceAI—Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites’, *PLOS ONE*, vol. 17, no. 6, p. e0269159, Jun. 2022, doi: 10.1371/journal.pone.0269159.
[30] 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', *Hum. Genet.*, vol. 133, no. 1, pp. 1–9, Jan. 2014, doi: 10.1007/s00439-013-1358-4.

[31] S. T. Sherry, 'dbSNP: the NCBI database of genetic variation', *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.

[32] K. Budholiya, S. K. Shrivastava, and V. Sharma, 'An optimized XGBoost based diagnostic system for effective prediction of heart disease', *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.

[33] K. Mayer, W. Ballhausen, W. Leistner, and H. Rott, 'Three novel types of splicing aberrations in the tuberous sclerosis TSC2 gene caused by mutations apart from splice consensus sequences', *Biochim. Biophys. Acta*, vol. 1502, no. 3, pp. 495–507, Nov. 2000, doi: 10.1016/s0925-4439(00)00072-7.

[34] K. Fulk, M. Turner, A. Eppolito, and R. Krukenberg, 'RNA sequencing uncovers clinically actionable germline intronic MSH2 variants in previously unresolved Lynch syndrome families', *BMJ Case Rep.*, vol. 15, no. 4, p. e249580, Apr. 2022, doi: 10.1136/bcr-2022-249580.

[35] D. A. de Lima Morais and P. M. Harrison, 'Large-Scale Evidence for Conservation of NMD Candidature Across Mammals', *PLoS ONE*, vol. 5, no. 7, p. e11695, Jul. 2010, doi: 10.1371/journal.pone.0011695.

[36] D. Karolchik, 'The UCSC Genome Browser Database', *Nucleic Acids Res.*, vol. 31, no. 1, pp. 51–54, Jan. 2003, doi: 10.1093/nar/gkg129.

[37] S. M. Ghiasi and G. A. Rutter, 'Consequences for Pancreatic β-Cell Identity and Function of Unregulated Transcript Processing', *Front. Endocrinol.*, vol. 12, p. 625235, Mar. 2021, doi: 10.3389/fendo.2021.625235.

[38] R. G. H. Lindeboom, F. Supek, and B. Lehner, 'The rules and impact of nonsense-mediated mRNA decay in human cancers', *Nat. Genet.*, vol. 48, no. 10, pp. 1112–1118, Oct. 2016, doi: 10.1038/ng.3664.

[39] K. T. Powers, J.-Y. A. Szeto, and C. Schaffitzel, 'New insights into no-go, non-stop and nonsense-mediated mRNA decay complexes', *Curr. Opin. Struct. Biol.*, vol. 65, pp. 110–118, Dec. 2020, doi: 10.1016/j.sbi.2020.06.011.

[40] R. Sousa-Luís *et al.*, 'POINT technology illuminates the processing of polymerase-associated intact nascent transcripts', *Mol. Cell*, vol. 81, no. 9, pp. 1935-1950.e6, May 2021, doi: 10.1016/j.molcel.2021.02.034.

[41] A. Sureau, 'SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs', *EMBO J.*, vol. 20, no. 7, pp. 1785–1796, Apr. 2001, doi: 10.1093/emboj/20.7.1785.

[42] L. F. Lareau, A. N. Brooks, D. A. W. Soergel, Q. Meng, and S. E. Brenner, 'The Coupling of Alternative Splicing and Nonsense-Mediated mRNA Decay', in *Alternative Splicing in the Postgenomic Era*, vol. 623, B. J. Blencowe and B. R. Graveley, Eds. New York, NY: Springer New York, 2007, pp. 190–211. doi: 10.1007/978-0-387-77374-2_12.

[43] E. D. Karousis, F. Gypas, M. Zavolan, and O. Mühlemann, 'Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells', *Genome Biol.*, vol. 22, no. 1, p. 223, Dec. 2021, doi: 10.1186/s13059-021-02439-3.

[44] P. Marks *et al.*, 'Resolving the full spectrum of human genome variation using Linked-Reads', *Genome Res.*, vol. 29, no. 4, pp. 635–645, Apr. 2019, doi: 10.1101/gr.234443.118.

[45] D. P. Cooke, D. C. Wedge, and G. Lunter, 'A unified haplotype-based method for

accurate and comprehensive variant calling', *Nat. Biotechnol.*, vol. 39, no. 7, pp. 885–892, Jul. 2021, doi: 10.1038/s41587-021-00861-3.
[46] A. Dobin *et al.*, 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
[47] S. Purcell *et al.*, 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.
[48] A. Dehman, C. Ambroise, and P. Neuvial, 'Performance of a blockwise approach in variable selection using linkage disequilibrium information', *BMC Bioinformatics*, vol. 16, no. 1, p. 148, Dec. 2015, doi: 10.1186/s12859-015-0556-6.
[49] W. J. Astle *et al.*, 'The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease', *Cell*, vol. 167, no. 5, pp. 1415-1429.e19, Nov. 2016, doi: 10.1016/j.cell.2016.10.042.
[50] M.-H. Chen *et al.*, 'Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations', *Cell*, vol. 182, no. 5, pp. 1198-1213.e14, Sep. 2020, doi: 10.1016/j.cell.2020.06.045.
[51] K. Giesecke, H. Sieme, and O. Distl, 'Infertility and candidate gene markers for fertility in stallions: A review', *Vet. J.*, vol. 185, no. 3, pp. 265–271, Sep. 2010, doi: 10.1016/j.tvjl.2009.07.024.
[52] L. Zhang *et al.*, 'Mouse spermatogenesis-associated protein 1 (SPATA1), an IFT20 binding partner, is an acrosomal protein', *Dev. Dyn.*, vol. 249, no. 4, pp. 543–555, Apr. 2020, doi: 10.1002/dvdy.141.
[53] S. Sakaue *et al.*, 'A cross-population atlas of genetic associations for 220 human phenotypes', *Nat. Genet.*, vol. 53, no. 10, pp. 1415–1424, Oct. 2021, doi: 10.1038/s41588-021-00931-x.
[54] R. Pazoki *et al.*, 'Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes', *Nat. Commun.*, vol. 12, no. 1, p. 2579, May 2021, doi: 10.1038/s41467-021-22338-2.
[55] X. Han, J.-S. Ong, J. An, A. W. Hewitt, P. Gharahkhani, and S. MacGregor, 'Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration', *Eur. J. Epidemiol.*, vol. 35, no. 2, pp. 139–146, Feb. 2020, doi: 10.1007/s10654-019-00598-z.
[56] A. Taracha, G. Kotarba, and T. Wilanowski, 'Neglected Functions of TFCP2/TFCP2L1/UBP1 Transcription Factors May Offer Valuable Insights into Their Mechanisms of Action', *Int. J. Mol. Sci.*, vol. 19, no. 10, p. 2852, Sep. 2018, doi: 10.3390/ijms19102852.
[57] Y. J. Kim *et al.*, 'A genome-wide association study identified new variants associated with the risk of chronic hepatitis B', *Hum. Mol. Genet.*, vol. 22, no. 20, pp. 4233–4238, Oct. 2013, doi: 10.1093/hmg/ddt266.
[58] D.-K. Jiang *et al.*, 'Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B', *Hepatol. Baltim. Md*, vol. 62, no. 1, pp. 118–128, Jul. 2015, doi: 10.1002/hep.27794.
[59] Y. H. Cheung, J. Watkinson, and D. Anastassiou, 'Conditional meta-analysis stratifying on detailed HLA genotypes identifies a novel type 1 diabetes locus around TCF19 in the MHC', *Hum. Genet.*, vol. 129, no. 2, pp. 161–176, Feb. 2011, doi: 10.1007/s00439-010-0908-2.
[60] D. J. Downes *et al.*, 'Identification of LZTFL1 as a candidate effector gene at a

COVID-19 risk locus', *Nat. Genet.*, vol. 53, no. 11, pp. 1606–1615, Nov. 2021, doi: 10.1038/s41588-021-00955-3.