Genomic determinants of tissue-specific splicing in the human genome

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Aaryan

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

March, 2024

Supervisor: Prof Sureshkumar Balasubramanian

School of Biological Sciences, Monash University, VIC 3800, Australia

Expert: Dr Kalika Prasad

Associate Professor, Biology, IISER Pune, INDIA

Certificate

This is to certify that this dissertation entitled "Genomic determinants of tissue-specific splicing in the human genome" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents work carried out by Aaryan at Monash University, Australia under the supervision of Prof. Sureshkumar Balasubramanian, School of Biological Sciences, Monash University during the academic year 2023-2024.

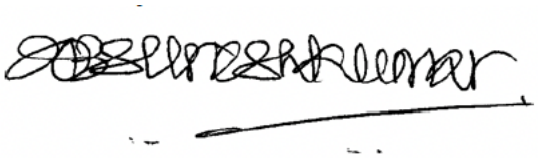
Aaryan
Student

Prof. Sureshkumar Balasubramanian
Supervisor

Declaration

I hereby declare that the matter embodied in the report entitled "Genomic determinants of tissue-specific splicing in the human genome" is the result of the work carried out by me at the School of Biological Science, Monash University, Australia, under the supervision of Prof Sureshkumar Balasubramanian and the same has not been submitted elsewhere for any other degree. Wherever others contributed, every effort has been made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Aaryan
Student



Prof. Sureshkumar Balasubramanian
Supervisor

Acknowledgements

I want to express my sincere gratitude to my supervisor, Prof. Sureshkumar, for his constant guidance, support in every aspect, and expertise throughout this thesis journey. Your encouragement, helpful feedback, and patience have been invaluable in shaping both my work and personal growth. I consider myself extremely lucky to have you as my supervisor. I also thank Dr. Kalika Prasad for consistently supporting me throughout my thesis. I am truly grateful for his mentorship, which has been invaluable in my academic journey. I also want to thank SKB and GATC lab members, past and present, including Sridevi, Jordy, James, Michael, Aishwarya, Sourav, Stefan, Craig, Anshuman, Rucha, and Param, for their friendship, insights, valuable feedback and collaborative spirit. Their diverse perspectives and lively discussions have made my research experience enjoyable and fruitful.

I sincerely thank Dr Kalika Prasad, Prof. Uptal Nath, Dr Pranay Goel, Dr Arthur Sherman, and Dr Deepak Barua for their invaluable support throughout the different projects I undertook during my degree. Their guidance laid a strong foundation for my academic path. I am also grateful to my PhD mentors, Vijina, Anurag, Raghav, Srijan, and Dhrubo, for their invaluable assistance throughout those projects.

I thank all my teachers, especially Gaba Sir, Raghuvanshi Sir, and Gaurav Sir, for their guidance during critical times, which helped me reach this point. I thank my parents and my sister for always being my pillars of strength and for standing by me every step of the way. I couldn't have done it without them. I'm deeply grateful to my family for their endless love, encouragement, and support. Special thanks to Paru Bhai for always being there to help. And, of course, I would like to thank all of my friends at IISER. They played a massive role in my personal growth and made my time there special. I'm so thankful for them!

I acknowledge the Kishore Vaigyanik Protsahan Yojana, DST, for providing me with a scholarship throughout my degree.

Abstract

Cellular differentiation into various tissues is a critical developmental feature of multi-cellular organisms. One of the key gene regulatory processes that play a vital role in tissue differentiation is RNA splicing. Splicing differs between tissues, generating a multitude of mRNA isoforms that encode diverse proteins across tissues. Despite a substantial understanding of RNA splicing, very little is known about the genomic determinants of tissue-specific splicing. Until recently, there have been no systematic ways to analyse splice-sites that are regulated in a tissue-specific manner at a scale that would allow deciphering genome-wide patterns. This is primarily due to methods that focus on mRNA isoforms or splicing events as opposed to analysis at the level of splice sites, whose selection primarily determines the regulation of splicing. This thesis makes an attempt to exploit splice-site level analysis to address determinants of tissue-specific splicing. Here, we quantified the usage of individual splice sites for genes expressed in the human testis from GTEx data, revealing significant variation between individuals. Using splice-site strength as a phenotype, we conducted over 130,000 Genome-Wide Association Studies (GWAS) to identify genetic variation associated with differential splice-site usage. We compared our results from testis with previous analyses carried out in the heart to identify and catalogue splice sites that are utilised in a tissue-specific manner. By motif-enrichment analysis, we reveal several genomic motifs that are enriched among splice sites that are used in a tissue-specific manner both in heart and in testis. This thesis presents these findings in both tissue-specific and evolutionary contexts of genomic determinants of splicing variation. Our studies suggest that tissue-specific splicing is a function of tissue-specific gene expression. We also reveal that genetic variation can affect splicing in a tissue-specific manner, which has potential implications for human disease.

Table of Contents

Chapter 1

Introduction

1.1 Introduction

In molecular biology, the flow of information is described by the central dogma. DNA of protein-coding genes are transcribed into messenger RNAs (mRNA), which undergo translation to produce proteins that perform diverse functions in a cell (Clancy, 2008).

However, eukaryotic RNA transcripts undergo several processing steps before becoming functional. After RNA synthesis from the DNA template, precursor mRNA goes through several processing steps, including 5’ capping, splicing, and addition of the poly-A tail. During RNA 5’ capping, a modified guanine nucleotide is attached to the 5’ end of the RNA molecule, which prevents its degradation and instability (Hocine *et al.*, 2010). Poly-adenylation occurs by adding a chain of adenosine nucleotides at the 3’ end of the mRNA molecule, enhancing mRNA stability and facilitating its transport from the nucleus to the cytoplasm. This process also significantly contributes to translation initiation (Manning and Cooper, 2017).

1.2 mRNA Splicing

Most of most eukaryotic coding genes, consist of alternating exons and introns. Introns are transcribed gene regions that are removed from the pre-mRNA molecule and are absent in the final mRNA molecule and thus will not contribute to the final protein produced from that transcript. Most human protein-coding genes are made up of both exons and introns. Through the process of splicing, these introns are removed, and the exons are joined (Figure 1.1).

Splicing is regulated by multiple features of pre-mRNA, predominantly dictated by *cis*-sequences that signal intron-exon junctions and *trans* elements facilitating the splicing

process. The primary *cis* motifs containing the essential information for exon recognition during splicing encompass the donor (5' end of intron) and acceptor (3' end of intron) splice sites, the branch point, and a poly-pyrimidine tract before the 3’ end of the intron (Manning and Cooper, 2017). Typically, introns begin with the dinucleotide GT present at the 5′ end(donor) and end with AG at the 3′ end(acceptor). The branch point, positioned approximately 20 to 50 nucleotides upstream from the 3′ end of an intron, consistently has an adenine, although the sequences surrounding the branch point are generally variable (Clancy, 2008). The schematic representing the splicing reaction is shown in Figure 1.1.

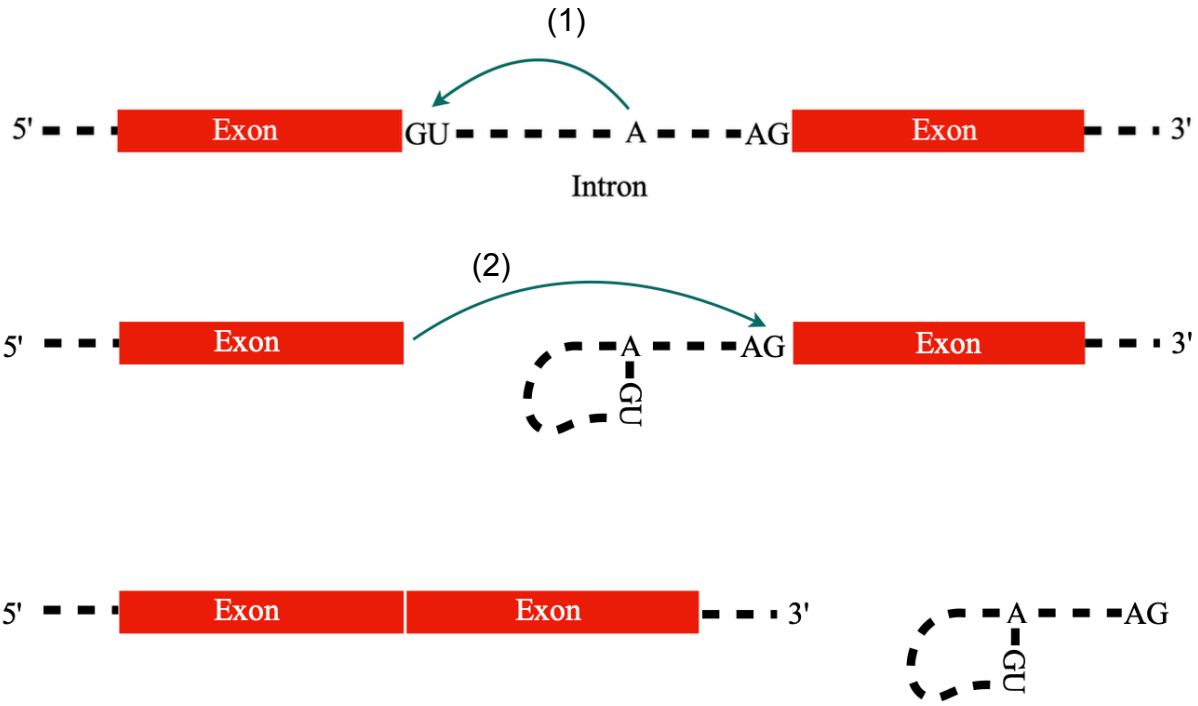


Figure 1.1: Shows the process of splicing reaction. (1) The 2'OH of a branchpoint adenosine in the intron acts as a nucleophile, attacking the first intron nucleotide at the donor site to form a lariat intermediate. (2) In the second step, the 3'OH of the released 5' exon acts as a nucleophile, attacking the first nucleotide after the intron at the acceptor site, which joins the two exons and releases the intron lariat.

Splicing patterns vary among cells, tissues, and individuals. Differential usage of splice sites creates multiple mRNA isoforms from a single RNA transcript, a phenomenon known as alternative splicing (Early et al. 1980). Alternative splicing serves as one of the major mechanisms in controlling gene expression. Controlling gene expression has two components - first, by production of diverse proteins with different functions (Nilsen and Graveley, 2010) and by triggering nonsense-mediated mRNA decay to cause differences in the levels of gene expression. Alternative splicing may vary depending on factors such as tissue type (Rice et al., 2019), developmental stage (Su et al., 2018), sex (Moschall et

al., 2019) and external conditions (Anduaga et al., 2019). More than 90% of human genes undergo alternative splicing, producing a variety of transcript isoforms (Wang et al., 2008). Various types of alternative splicing are shown in Figure 1.2.

The regulation of alternative splicing typically involves a variety of *cis*-elements and *trans*-factors. These elements assemble into intricate interaction networks capable of offering significant regulatory flexibility (Matera and Wang, 2014).

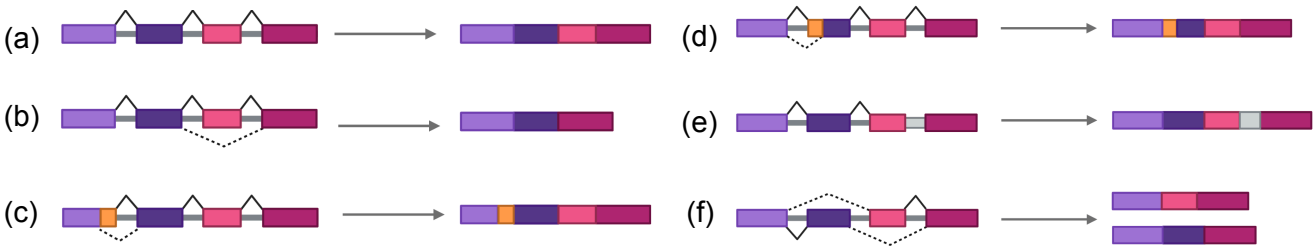


Figure 1.2: Six modes of alternative splicing: (a) Constitutive splicing, (b) Exon skipping, (c) Alternative 5' splice site, (d) Alternative 3' splice site, (e) Intron retention, and (f) Mutually exclusive exons.

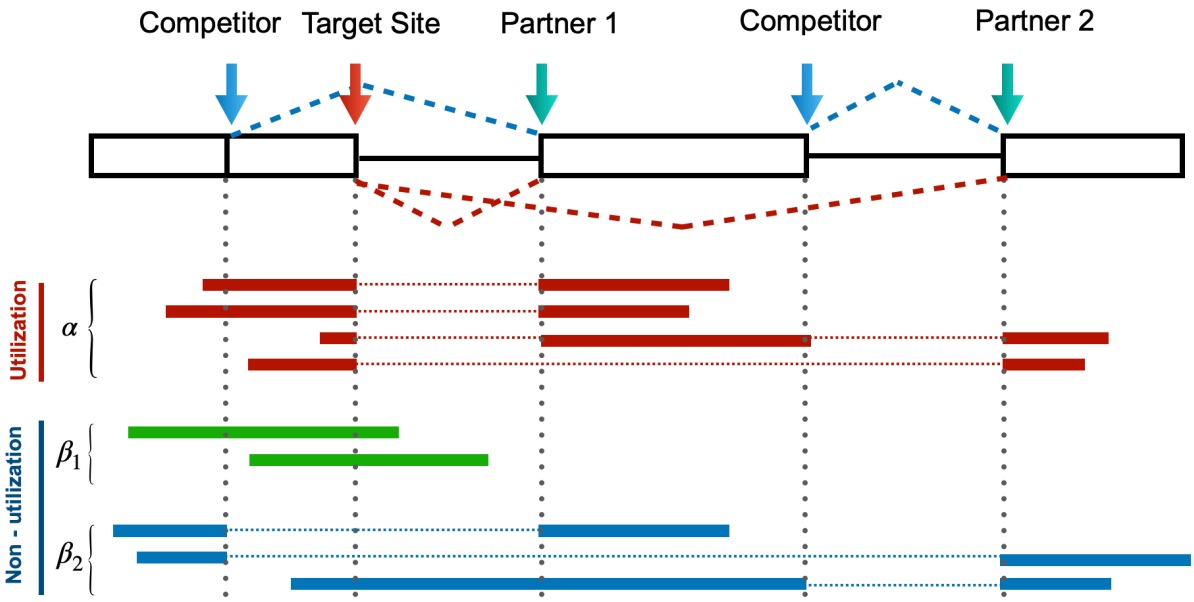
The protein complex responsible for splicing is called the spliceosome, and it recognises specific sequences at the exon-intron boundaries to remove introns and join exons. A spliceosome is a dynamic complex comprised of RNA and proteins, including uridine-rich small nuclear ribonucleoprotein (U snRNPs), like U1, U2, U3, U4/6, and U5 snRNPs (Fica and Nagai, 2017). Typically, introns undergo U2 type splicing, wherein the GU donor 5' splice-site motif attracts U1 snRNP, while U2 auxiliary factors (U2AF2 and U2AF1) recognise the polypyrimidine tract. Furthermore, the branch point is identified by SF1 (splicing factor 1), which, along with U2AF, facilitates the recruitment of U2 snRNP to the branch point (Fu and Ares, 2014). Upon recruitment to the donor site and the branch point, U1 and U2 snRNPs can interact, leading to the RNA looping and subsequent removal of the intronic region.

Despite significant progress made through biochemical and genetic investigations elucidating numerous proteins and mechanisms involved in splicing, a key knowledge gap remains in understanding how splice-sites are selected among competing splice-sites and how genetic variation affects splicing. It is also unclear how tissue-specific is achieved and

what are the underlying mechanisms. This gap is partially due to challenges in measuring or quantifying splicing accurately and precisely.

1.3 SpliSER - Splice-site Strength Estimate from RNA- seq

Splicing is commonly studied using RNA-seq data, typically through two main methods: quantifying splice isoforms or analysing specific splicing events. However, these approaches face challenges, such as the lack of distinct regions to differentiate isoforms, the lack of a comprehensive isoform catalogue, and the complexity of splicing events, which are descriptive rather than biological in nature. Splicing regulation primarily



(Figure adapted from Dent *et al.*, 2021)

Figure: 1.3 Overview of SpliSER and Splice-site Strength Estimation - To calculate the splice-site strength estimate for a given splice site (shown by a red arrow - target site), the proportion of reads that give evidence for the splice site's use to the (α reads; red reads) to those that show total potential usage (i.e.,α reads; red reads and non-utilization (β1 and β2 reads; green and blue rectangles that span over that site) is calculated.

occurs at the level of selection of splice-sites. However, there has been less systematic effort to measure splicing based on the usage of individual, independent splice sites. The Splice-site Strength Estimate from RNA-Seq (SpliSER) (Dent et al., 2021) is a

bioinformatics tool designed by our lab to quantify the usage of individual splice-sites using RNA-Seq data. An overview of estimating splice-site usage using SpliSER is described in Figure 1.3.

From RNA-Seq data, we obtain several reads that align to a specific genomic region. These reads can provide insights into the usage of a splice site if they span across an intron. As shown in Figure 1.3, α reads (red) are split-reads with a gap in mapping that ends at the splice site, indicating utilisation of the target site. The other splice-site in this read is considered a Partner of the target site. β1 reads (green) map on either side of the splice site but shows no gap in mapping at this position, indicating the target site has not been utilised. β2 reads (blue) show known Partners of the given site utilising another splice site, which has outcompeted the target site. β2 reads directly indicate non-utilization of the target site. SpliSER calculates the Splice-site Strength Estimate (SSE) for each splice site using these read counts. SSE is calculated as the ratio of splice site utilisation (α reads) to the total potential utilisation of the splice site (the sum of α, β1, and β2 reads).

Equations (1) and (2) give the formula for calculating SSE.

$$SSE = \frac{Reads\ that\ give\ evidence\ for\ the\ splice-site\ usage}{Evidence\ for\ total\ potential\ splice-site\ usage} \quad (1)$$

$$SSE = \frac{\alpha}{\alpha + \beta_1 + \beta_2} \quad (2)$$

This approach provides a quantitative measure for the splice-site's usage/strength for all the sites individually. SpliSER was employed to quantify the usage of each splice site across individuals in the genome for testis tissue. This quantitative data could then serve as a phenotype for identifying genetic variants associated with variation in splice-site usage through Genome-Wide Association Studies (GWAS).

1.4 Genome-wide Association Study (GWAS)

Genome-wide association studies (GWAS) aim to identify genetic variants statistically associated with specific traits such as diseases (Uffelmann *et al.*, 2021). In GWAS studies, DNA samples are genotyped from a group of individuals and analysed to identify single nucleotide polymorphisms (SNPs) across the genome. At the population level, GWAS relies on the concept of linkage disequilibrium (LD), which refers to non-random segregation of alleles due to shared ancestry. LD is influenced by mutation, drift, natural selection, and recombination, with stronger LD observed between loci that are physically closer on a chromosome. The underlying principle of GWAS is that for a common disease or trait, a particular disease-causing allele will be more prevalent at a specific site, allowing GWAS to identify genetic variants strongly associated with the trait or disease (Visscher *et al.*, 2012).

1.5 Genome-wide association study of splicing using SpliSER-GWAS

SpliSER provides a quantitative measure of splice-site usage/strength. This measure can be further used as a phenotype to quantify variation (which occurs due to the differential usage of splice sites in different individuals) in splice-site strength among individuals, which can then be associated with specific genomic variants by GWAS. In GWAS, variation in a phenotype can be attributed to genomic variants that explain a significant portion of the phenotypic variance. Our lab has developed a pipeline that uses SpliSER to quantify splice-site usage (SSE) for all splice sites from the transcriptomes of the individuals in a given population. Based on previous studies on Arabidopsis and Drosophila (*Balasubramanian, personal communication)*, we suggest that SSE represents a molecular phenotype with high heritability. This SSE is then used as a phenotype for a GWAS, termed SpliSER-GWAS, where each splice site undergoes its own GWAS to identify the genomic variant most strongly associated with variation in the usage of that splice site (Dent et al. 2021). These associations are represented graphically through Manhattan plots.. An example of a Manhattan plot created using SSE as a phenotype for site Chr6:71304391 is illustrated in Figure 1.4. The most strongly associated SNP was identified at position Chr6:7130434.

It is important to note that GWAS has limitations, as it can only detect common variants present in the tested population, and the identified variants are not necessarily causative. The SNPs could be either causative or they could be in LD with some other polymoprphism that is causative. However, if the mapped SNP is really close to the splice-site the chances of that being causal is really high. This is reflected in the high precision of mapping and the fact that the vast majority of associated variants are located near the splice sites of the measured phenotype (Dent *et al*., in preparation)

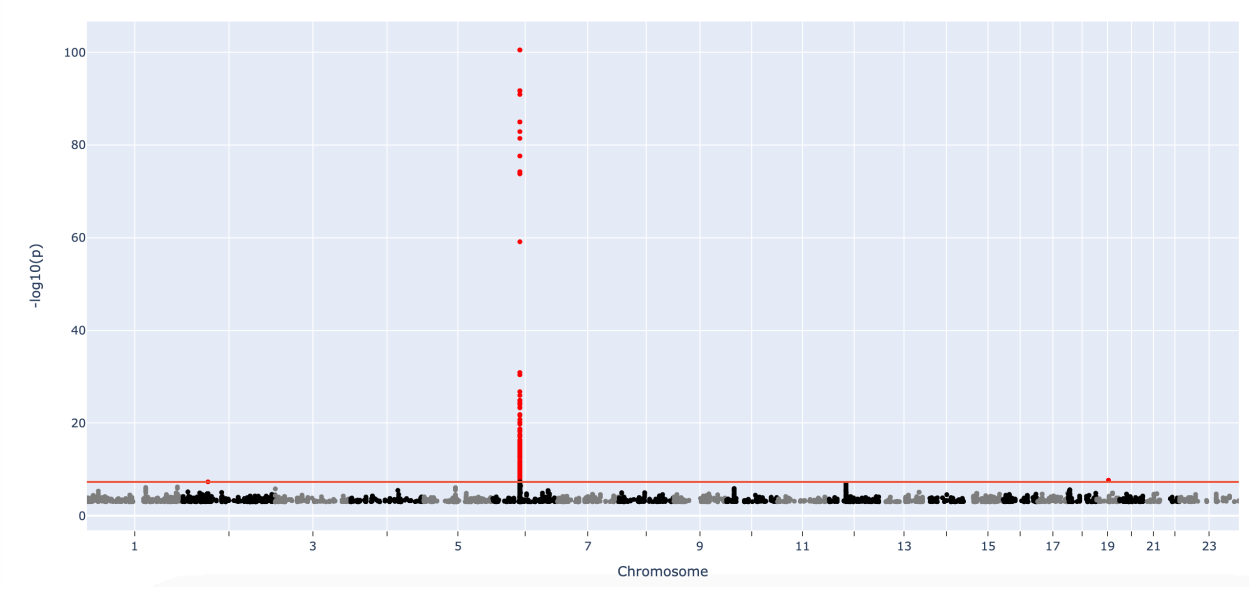


Figure 1.4: A Manhattan plot for the SpliSER-GWAS analysis of splice sites at genomic position Chr6:71304391 is presented. The highest associated SNP is situated at position Chr6:7130434. The Bonferroni threshold is indicated by a red line.

1.6 Tissue-specific splicing:

Alternative splicing plays a crucial role in generating proteins specific to different tissues (Wen *et al.*, 2010; Wang *et al.*, 2009). Some studies have revealed that 42% of examined cassette exons show differential expression in at least one of 48 human tissues (Castle *et al.*, 2008). Tissue-specific splicing is regulated by a combination of splicing factors that are specific to particular tissues as well as those that are expressed ubiquitously (Wang *et al.*, 2008). These factors interact with sequences around the splice sites, influencing spliceosome assembly and, consequently, the isoforms produced. Various splicing factors can either activate or repress splicing in different cellular contexts. This differential usage

of sites across tissues leads to distinct alternative splicing patterns and, consequently, the production of different protein isoforms.

Many studies on tissue-specific splicing primarily focus on looking at differences in mRNA isoforms and splicing events, which complicates understanding the biology underlying this process. Since splicing machinery operates at the level of individual splice sites, it becomes important to investigate tissue-specific splicing mechanisms by looking at the level of individual splice sites. SpliSER provides a quantitative measure of splice site strength for all the sites in the genome, allowing us to compare splice site strength for a given site across different tissues. Using the splice site-based approach, this thesis aims to answer the following questions.

1.) Can we detect splice-sites that differ in their usage between tissues? If yes,
2.) How many genes harbour splice sites that exhibit tissue-specific splicing?, and last
3.) What controls tissue-specific splicing? Can we identify potential regulators?

.

Chapter 2

Materials and Methods

2.1 Choice of tissue and RNA-Seq data

This thesis used publicly accessible data from the Genotype-Tissue Expression (GTEx) consortium database (GTEx Consortium, 2020), a comprehensive gene expression database for various human tissues and genotyping information. The version 8 dataset comprises RNA-sequencing samples from 49 normal tissues across nearly 838 post-mortem donors. Transcriptome data was downloaded as RNA-Seq alignment/BAM files, and genotype data as VCF (Variant Call Format) for analysis.

Our lab has previously done SpliSER-GWAS analysis for heart atrial tissue and found that there are roughly 14,700 analysable genes in the dataset. To assess the overlap between heart atrial tissue and other tissues in terms of gene expression, we employed the Jaccard Similarity Coefficient (JSC) (Tanimoto and T.T., 1958). This metric, calculated as the ratio of the intersection to the union of two sets, quantifies the degree of overlap, with a higher JSC indicating greater similarity. For example, denoted by equation 3, *J(A, B)* represents the JSC, measuring the overlap between two sets *A* and *B*. We utilised median TPM values from the GTEx study for gene expression across tissues to compute the Jaccard Similarity coefficient for comparing gene overlap with heart atrial tissue. Subsequently, we identified the tissue (testis) with the lowest JSC value (JSC=0.67) with heart atrial tissue for further analysis. To cover maximum genes, we chose to analyse testis tissue due to its unique gene expression profile compared to heart tissue. This analysis will cover many genes that were not included in the previous study because they are not expressed in the heart.

Jaccard Similarity coefficient:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

2.1.1 Binary Alignment Map (BAM) files

In this project, we used SpliSER version 1.8 (Dent et al., 2021). To run SpliSER, it requires a BAM (Binary Alignment Map) file containing mapped RNA-seq reads, an index file for these BAMs, and a BED (Browser Extensible Data) file containing a list of splice junctions identified in the alignment file. The BAM files generated by the GTEx project (GTEx Consortium, 2020) were downloaded. These BAM files were generated by aligning the reads to the human reference genome (GRCh38/hg38) using STAR v2.5.3a. BAMs were downloaded with authorised NIH (National Institutes of Health) credentials. At first, JSON (JavaScript Object Notation) files or an index file of all available GTEx BAM files were downloaded. Along with this, the GTEx IDs of individuals whose testis tissue had been analysed were also downloaded. Using these GTEx IDs, a trimmed JSON file for the BAMs and BAI (BAM Index file) of testis tissue was created. The JSON file was used to download BAMs and BAIs from the repository using a module named gen-3-client.

2.1.2 Browser Extensible Data (BED) files

SpliSER needs a gap junction file (Browser Extensible Data - BED) containing splice junction information. RegTools (Cotto *et al.*, 2023) was used to create the BED files. It is a command-line tool that needs a BAM file, minimum anchor length, minimum intron length, maximum intron length, and information on the library preparation's strand specificity for RNA sequencing.

The RegTools command used to generate the BED files is as follows:

```
regtools junctions extract -m 20 -M 16000 -a 6 -s 0 -o <SAMPLE>.bed <SAMPLE>.bam
```

The RegTools' junction extract command was used to get the exon-intron junction information from the BAM files. The "-a" flag was used to select junctions where a read has at least a 6bp overlap on both ends. Additionally, the "-m" and "-M" flags were used to specify the minimum and maximum intron lengths, set to 20 bp and 16000 bp, respectively. The "-s" parameter was set to zero to indicate the unstranded nature of the RNA library preparation (Illumina TruSeq protocol with poly-A selection). Finally, these parameters were followed by a suffix for the output file and corresponding BAM.

2.1.3 Annotation file - GTF file

The annotation file or Gene Transfer Format (GTF) file holds information about gene structure. It was downloaded from the 26th release of the GENCODE gene annotation project. The GTF file was sourced from open-access data available on the GTEx portal. Although SpliSER does not rely on annotations, it can use an annotation file to determine the corresponding gene for each quantified splice site.

2.1.4 Variant Call Format file - VCF File

A VCF file contains single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variation (SVs) found in a cohort. The GTEx-VCF file was obtained from the GTEx portal using NIH credentials. It contains details such as the genomic position of the variants, the reference and alternative allele, and the variant genotype information for each individual. This file also helps determine the frequency of each variant within the cohort, indicating whether it is common or rare. This genotype data is crucial, as it allows us to examine for associations with the phenotype of interest (Splice-site Strength Estimate in this case).

2.2 SpliSER pipeline

The process of using SpliSER involves three main steps: SpliSER-process, SpliSER combine, and SpliSER-output.

1. SpliSER-process:

The SpliSER process command requires the BAM file, BAI file, BED files, annotation file in GTF format, and the maximum intron size used in the alignments (16000 bp). It must be

executed separately for each BAM file. The execution of the SpliSER Process command for all 311 BAM files was accomplished using a parallel job running in the M3 cluster. This command generates a .tsv file for each GTEx individual, containing SSE values for all splice sites detected in the alignment file. The SSE information for all 311 samples can then be combined using the next step, SpliSER Combine.

The command used to run the SpliSER process is as follows:

```
python SpliSER_Script.py process -B <BAM_file>.bam -b <bed_file>.bed -o
<output_folder/out_Prefix> -A <Annotation_File>.gtf
```

2. SpliSER combine:

In SpliSER Combine, data from all samples is aggregated for every splice site. Only sites within a protein-coding region were selected for this step based on a list of protein-coding genes derived from the GTF file. Additionally, to ensure sufficient evidence for each site, a splice site was considered only if there were more than 10 RNA-seq reads indicating potential usage. Furthermore, splice sites detected in at least five separate samples with an SSE level exceeding 0.05 were included in further analysis. If a splice site was not found in the sample of interest but was identified in other samples, and there were more than ten reads indicating potential usage, the SSE value of the splice site was set to zero. This approach ensured that information on the usage of each splice site was comparable across all samples with sufficient gene expression for accurate quantification.

The command used to run SpliSER combine is as follows:

```
stdbuf -oL -eL python SpliSER_Script.py combineShallow -S <samplesFILE> -g
<GENE> -m 5 -r 10 -e 0.05 -o <outFILE>
```

3. SpliSER output:

Following the SpliSER combine step, the resulting files were divided into individual files for each splice site using the SpliSER output option. These files now can be used as phenotype files for the subsequent GWAS analysis.

The command used to run SpliSER output is as follows:

```
stdbuf -oL -eL python SpliSER_Script.py output -t GWAS -S <samplesFILE> -C
<combineFILE> -g <GENE> -r 10 -m 1 -o <outFILE>
```

2.3 Variance Filter

Following the SpliSER output, the phenotype (SSE) data for all sites becomes available for GWAS. But, for a phenotype to be suitable for GWAS, it must exhibit variation within the population. Consequently, only highly variable splice sites were opted for subsequent analysis. The variance of each splice site was calculated to determine its variability among individuals. All the detected splice sites were sorted according to their variance among individuals, and the top 25%, encompassing those with the highest quartile of variance, were tentatively classified as “highly variable splice-sites” and chosen for GWAS analysis.

However, mapping these sites requires a minimum number of individuals. Based on our previous analysis, we determined that at least 100 data points may be needed to detect an association reliably. Consequently, after selecting the highly variable splice-sites, an accession filter was applied to select “GWASable” sites. Only sites detected in at least 100 individuals were considered for SpliSER-GWAS.

2.4 GWAS

Kinship Matrix:

One needs to make sure that the associations captured by GWAS are not purely a result of genetic relatedness, and thus needs to take care of this relationship among individuals under study while carrying out the GWAS analysis. A kinship matrix can be created that represents the degree of relatedness among individuals participating in a study. It aims to mitigate associations that may arise solely due to ancestral genetic similarities due to population structure. The kinship matrix was constructed using PLINK v1.9 (Purcell *et al.*, 2007).

The initial step involves creating a genotype file containing genetic data for all individuals in the population. These genotype files are generated in binary format, with extensions ".bed" (contains genotype information), ".bim" (contains genetic variant information), and ".fam" (contains individual ID and phenotype). These files contain essential genetic information required for performing a Genome-wide Association Study (GWAS) using the GEMMA (Genome-wide Efficient Mixed Model Association) module to examine associations between genetic variants and phenotype (in this case, SSE).

The command used to create the kinship matrix is as follows:

```
plink --vcf <path/to/vcf_file.vcf > --maf 0.05 --make-bed --out GTEx plink --
distance-matrix square --bfile GTEx
```

Genome-wide Efficient Mixed Model Association (GEMMA):

GEMMA (Genome-wide Efficient Mixed Model Association) (Zhou and Stephens, 2012) is a command-line tool designed to uncover relationships between genetic variation and specific traits, such as SSE, in this context. It utilises PLINK genotype and phenotype files (previously created) as input to produce association files. Only variants with a minor allele frequency (MAF) exceeding 5% are considered for the association study. Typically, variants with an MAF below 5% are omitted from genetic association analyses for reasons such as reduced statistical power, rarity, insufficient understanding, or potential for generating false positive associations from data noise. However, disregarding these variants might also hinder the detection of rare variations.

The gemma command to generate the association files is as follows:

```
gemma -bfile <FILE_PREFIX> -maf 0.05 -k plink.mibs -lmm 1 -o <FILE_PREFIX>
```

2.5 Manhattan Harvester

A program called Manhattan Harvester (Haller *et al.*, 2019) was used to extract significant peaks from GWAS summary files. This tool can detect peaks in GEMMA association files and calculate parameters that describe various aspects of individual peaks. The

Manhattan Harvester also provides a General Quality Score (GQS) for each peak on a scale of 1-5, which indicates how good a peak is.

The command used to generate Manhattan Harvester outputs is given below:

```
harvester -chrcolumn 1 -l column 3 -pcolumn 12 -header yes -file
<GEMMA_output>.txt -out <PREFIX>.mhv.out
```

2.6 Top SNP Calling

An in-house Top-SNP-calling pipeline was used to identify significant peaks from GWAS plots. This pipeline was designed to call the significant peaks and extract information from the associated top SNP. It relied on three main parameters: the Bonferroni threshold, Manhattan Harvester peak detection, and the estimated noise threshold. The Bonferroni threshold, a standard statistical cut-off, adjusts the p-value to correct for multiple testing (It is calculated by dividing the nominal p-value (0.05) significance threshold by the total number of SNPs considered in the testing model). This approach helps control false positive rates in GWAS.

2.7 Generating the Allele Table/SNP-table

The peak with the lowest p-value was selected as the top SNP; in cases where more than one SNP exhibited the lowest p-value, the SNP closest to the splice site was chosen for further analysis. SNPs located within one megabase of the splice site were classified as "cis" associations, while those further away were classified as "trans" associations. The allelic change in splice site strength (SSE), indicated by delta SSE, was calculated as the difference between the average SSE for the major and minor alleles.

The SpliSER-GWAS pipeline:

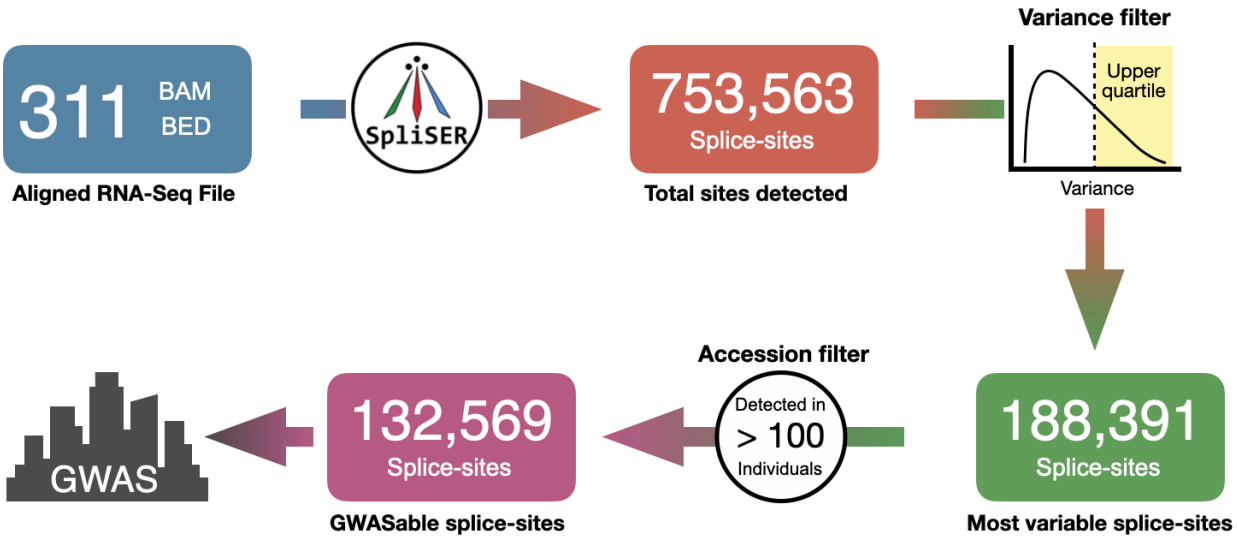


Figure 2.1: SpliSER-GWAS pipeline. It begins by processing BAM and BED files to quantify splice-site usage across the genome. Subsequently, only the sites that exhibit variability in terms of SSE among populations, with data from at least 100 individuals, are selected for analysis using SpliSER-GWAS.

2.8 Tissue-specific Splicing

To identify splice sites spliced in a tissue-specific way, splice site strength (SSE) information was extracted from one individual for all sites for both the tissues (Heart Atrial tissue and Testis tissue). Only genes expressed in both tissues were considered to eliminate the possibility that differences in splice site strength were due to differences in gene expression. To identify sites preferentially used in one tissue over another, sites with SSE close to zero (<0.1) in one tissue and greater than 0.5 in the other were sought. After identifying these sites, it was ensured that there was not more than a 10-fold difference in read coverage between the two tissues, as a big difference in read coverage could compromise the comparison.

Once we identify the sites that are spliced in a tissue-specifc way, the sequence information was extracted for a 600-base pair window centred around the site. Subsequently, motif enrichment was done separately for donors and acceptors using XSTREME - Motif Discovery and Enrichment (Grant and Bailey, 2021). XSTREME looks

for motifs enriched in the test sequence compared to the provided control. For control, the sites that are equally utilized in both tissues were extracted - with an SSE of at least 0.75 in both tissues, ensuring that the difference in SSE between the two tissues does not exceed 0.1. Adequate read depth coverage in the RNA-Seq for the sites was ensured. Sequence information for a 600 BP window centred around the site was extracted to use as a control. Four different batches of 1500 sites were randomly chosen from the total filtered sites for controls.

XSTREME motif enrichment was performed using the following command:

```
xstreme --oc Result --time 240 --streme-totallength 4000000 --meme-searchsize
100000 --dna --dna2rna --evt 0.05 --minw 5 --maxw 12 --align center --meme-mod
zoops --sea-noseqs --m "<path/to/known/motifs/Ray2013_rbp_Homo_sapiens.meme>" --
p "<path/to/testFASTA_file.fasta>" -n "<path/to/controlFASTA_file.fasta>"
```

2.9 Checking for mutation in the region:

To verify whether a mutation in the motif region changes the SSE for the corresponding site, sites containing the enriched motif of interest in their vicinity were sought. Subsequently, the position of the motif relative to the site was extracted. Individuals with a mutation in the motif region were searched using the VCF file. Once we obtained the individuals with a mutation in the motif region, we extracted the SSE value for the corresponding splice site. We compared it with the average SSE of individuals who do not have a mutation in the motif region.

2.10 Comparative analysis GWAS for tissue-specific effects

Once the results for SpliSER-GWAS from testis were obtained, they were compared with the SpliSER-GWAS result for the previously conducted analysis on heart arterial tissue. Sites that crossed the accession and variance filter in both tissues were considered for the analysis to ensure that GWAS analysis was conducted on those sites using SSE values from the corresponding tissues. The SNP table for both tissues was then examined, and the sites that yielded a peak (crossing the GQS score of 3.5) in one tissue but not in the other were selected. Subsequently, sequence information was extracted for a seven bp window centred around the top SNP. ATtRACT (Giudice *et al.*, 2016) was utilised to

determine if any motif was present in that small window around the top SNP, for which the corresponding binding factor is known.

2.11 Verification of Hexamers as Primary Determinants of splice-site Choice

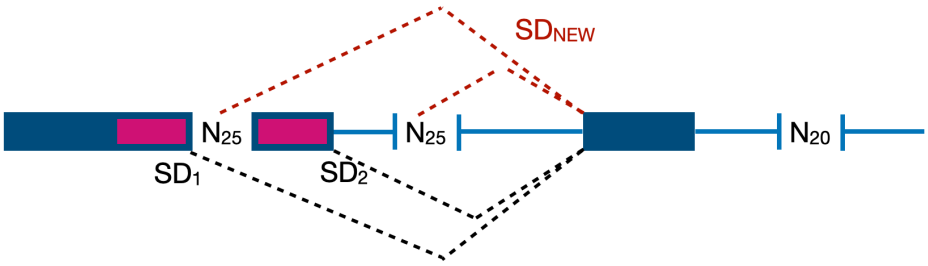
In our lab, we analysed how splice site strength (SSE) relates to nearby sequences of the splice sites. By studying different sequence sets around splice sites, we found that the hexamer sequence ($GT[N]_4$ & $[N]_4AG$) best explains SSE variation. To validate this further, we grouped all possible splice sites into different k-mer groups (e.g., GTNNNN with 256 sequence combinations) and calculated the average strength for each group by taking the average of SSE values for the sites in a group. We created a k-mer rank based on the average SSE for each group. Now, to check if the hexamer ($GT[N]_4$ & $[N]_4AG$) best explains the splice site choice. We extracted sequence information around the splice site from the Arabidopsis, Drosophila, and human reference genomes. We used the VCF file for all species to adjust for mutations, insertions, and deletions in the population's genome.

Then, we looked for other GT/AGs present near the splice site (in a 100 bp window centred around the site). We extracted the corresponding K-mers for these potential splice sites containing the canonical GT/AG. The number of occasions when the splice site had the strongest and unique k-mer among the k-mer for all GT or AG in that window was noted. This number over the total analysed sites to obtain a success rate. We also calculated the percentage of possible k-mers that are present in splice sites. We multiplied both these scores and divided them by 100 to obtain the percentage of splice site choices explained by the k-mer for all three species. We excluded splice sites that are within 100/200 bp of each other, eliminating competition between detected sites for this analysis.

2.11.1 Rosenberg et al. analysis

To validate the idea of hexamers, we re-analysed the RNA-Seq data generated by Rosenberg *et al.*, 2015. In the publication, the dataset comprises 265,137 unique minigene sequences. Each minigene construct contains fixed splice donors, SD1 and SD2, each with a defined sequence. A pair of random 25-nucleotide (N25) sequences is incorporated, corresponding to a 20-base pair long barcode, N20, as seen in Figure 2.2. It is important to

note that the entire intron sequence remains consistent, except for the two N25 sequences



and the N20 barcode.

(Figure adapted from Rosenberg et al., 2015)

Figure 2.2: A schematic of the alternative 5' library, showing SD1 and SD2 as default donors. N_{25} shows the regions with random nucleotides where new potential splice sites originate. N_{20} shows the region for a barcode that is unique to each fragment.

Within this extensive dataset are default splice donors (SD1 and SD2), alongside emerging novel splice sites within the N25 region and the remaining intronic area. The primary focus of the analysis revolves around cases where at least two donors compete with each other, investigating whether splice site selection can be explained by hexamer ranking.

We started by processing a FASTQ file containing sequence information for all minigenes. This involved segregating reads based on barcodes, which we achieved using FASTQ-Multx. We then created a reference file for each FASTQ file and aligned the reads using the STAR alignment tool. This alignment process produced BAM files. Subsequently, we generated additional essential files (.bai and .bed) using samtools and regtools to facilitate further analysis. Using SpliSER, we analysed the 265,137 BAM files, obtaining crucial information regarding splice site positions and their corresponding SSE values for all minigene constructs. After this analysis, we compiled data on splice site positions, SSE values, and the hexamers for these splice sites. We kept only the minigenes with a minimum of two splice donors where the hexamer starts with a 'GT' at the splice-site position. We calculated a total of 4,095 unique hexamer pairs for consideration. To refine the data, we excluded pairs that appeared less than ten times across all files, resulting in 535 unique hexamer pairs. Since each pair is recurrent in these files, we calculated the percentage of cases where our ranking explains the winning hexamer.

All the aforementioned analyses were carried out using custom Python scripts.

Chapter 3

Results - Section - A - GWAS

3.1 Testis tissue covers 85% of the genes in the genome.

Our lab has previously performed a SpliSER-GWAS analysis for heart atrial tissue. To continue our previous analysis, we chose testis tissue based on a comparative analysis of gene expression profiles in heart atrial tissue against all tissues. A total of 19,278 protein-coding genes were found in the human genome, according to the information available in the GTF file. In heart tissue, 14,733 protein-coding genes are expressed, accounting for 76.4% of the total protein-coding genes. However, this misses a significant number of genes. In contrast, 17,257 protein-coding genes are expressed in testis tissue, which accounts for 89.51% of the total protein-coding genes in testis, which includes many new genes that were not covered in our previous analysis (Figure 3.1).

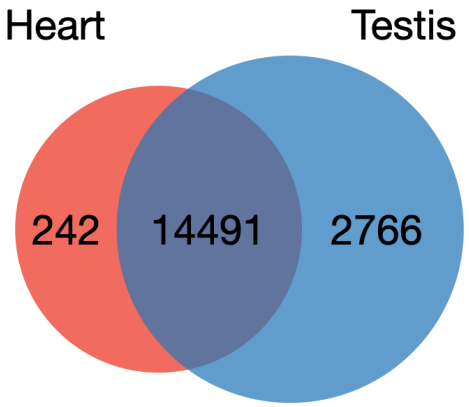


Figure 3.1: Illustrates the number of genes expressed in the two tissues and the overlap between them. Testis tissue was found to have the least overlap with heart atrial tissue in terms of gene expression.

Jaccard Similarity coefficient (JSC) was used for comparative gene overlap analysis, revealing that testis tissue had the least overlap with heart atrial tissue in terms of gene expression with a JSC of 0.67. A list of the JSC for all the tissues with the heart atrial

tissue can be found in Table 1. This indicates that the analysis of testis tissue will cover most of the genes that were not covered in the previous analysis. The data for testis tissue was available from 311 individuals, providing us with good statistical power. We restricted our analysis to protein-coding genes to enhance the detection of splicing variation with functional impacts.

3.2 There is extensive variation in splice-site usage between individuals in testis transcriptome

To study splicing variation, we quantified splice-site usage for all the sites from all the 311 individuals using SpliSER. We identified 753,563 splice sites in testis tissue, representing 17,257 protein-coding genes. Next, we calculated the variance in the usage of every splice site and the difference between the maximum and minimum SSE for each splice site detected. Around 63% (472,794 out of 753,563) of the sites showed a 20% difference in the maximum and minimum SSE value detected among individuals, as shown in Figure 3.2. This shows an extensive variation in splice-site usage in humans.

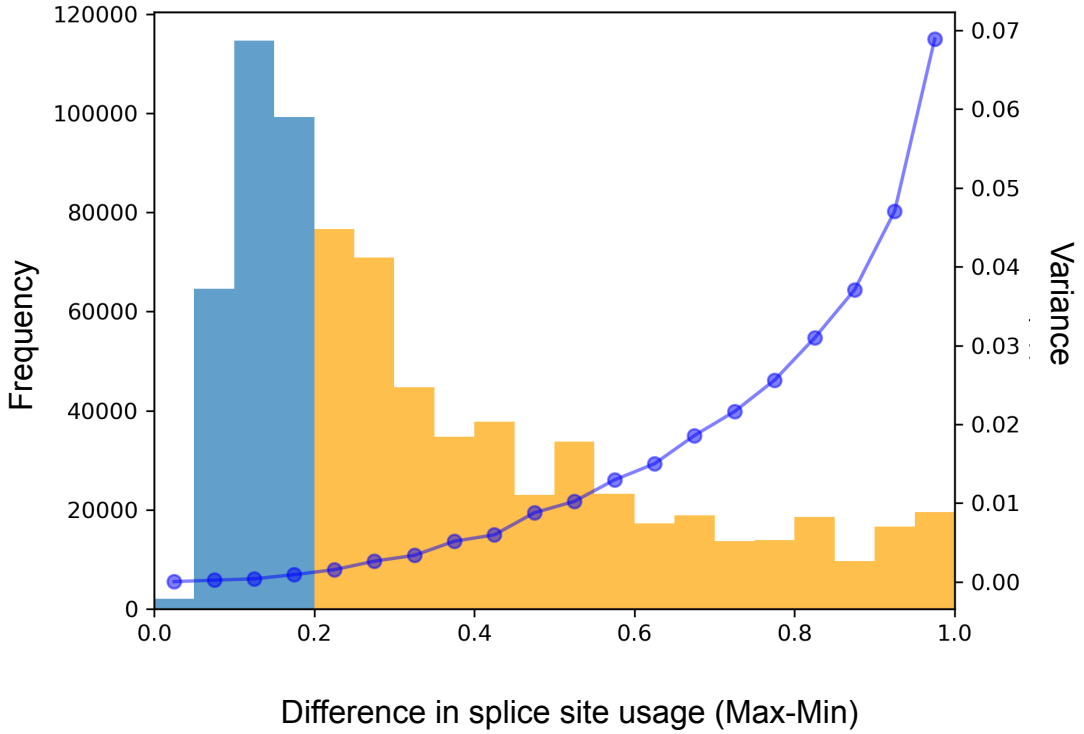


Figure 3.2: Distribution of splice site variation. The figure illustrates the distribution of splice site variation in humans. The sites are categorised based on the maximum difference in their usage, with the yellow region indicating sites that exhibit over a 20% difference between extreme

individuals. The blue dots represent the average variance in splice site usage across all individuals within each bin.

Since it was not feasible to calculate the heritability of these sites, due to the lack of replicates in the GTEx dataset, we included all sites in the upper quartile for variance (variance > 0.0072) in SSE among individuals for the GWAS analysis, which resulted in a total of 132,569 sites selected for GWAS.

3.3 Most of the genetically controlled variation in splicing is *"cis"*.

Out of the total 132,569 GWAS done, we obtained clean hits for 12,640 sites (~9.5%). Initially, we checked how many associations were on the same chromosome and found that a total of 12,239 sites did map to the same chromosome. To look further, we plotted splice-site positions against the highest associated variants, which revealed that a good proportion of the highest associated genomic variants were located near splice sites. We classified associations within 1Mb from splice sites as "*cis*" and the rest as "*trans*". Of the 12,640 significant associations identified by the SpliSER-GWAS SNP-calling pipeline, 11,281 were "*cis*" associations, with the rest being trans. "*cis*" associations represented around 91% of all associations.

To visualise it, we plotted the genomic distribution of splice sites against the top associated SNP shown in Figure 3.3. The plot showed a clear diagonal line, indicating that most genotype-dependent splicing variation is mediated via "*cis*" rather than "*trans*" genetic variation in humans. While some "*trans*" regulatory variation was identified, we did not detect any major vertical lines in the scatter plot, suggesting there were no major trans clustering or hot spots detected. This suggests that most associations deciphered by SpliSER-GWAS are "*cis*" in nature.

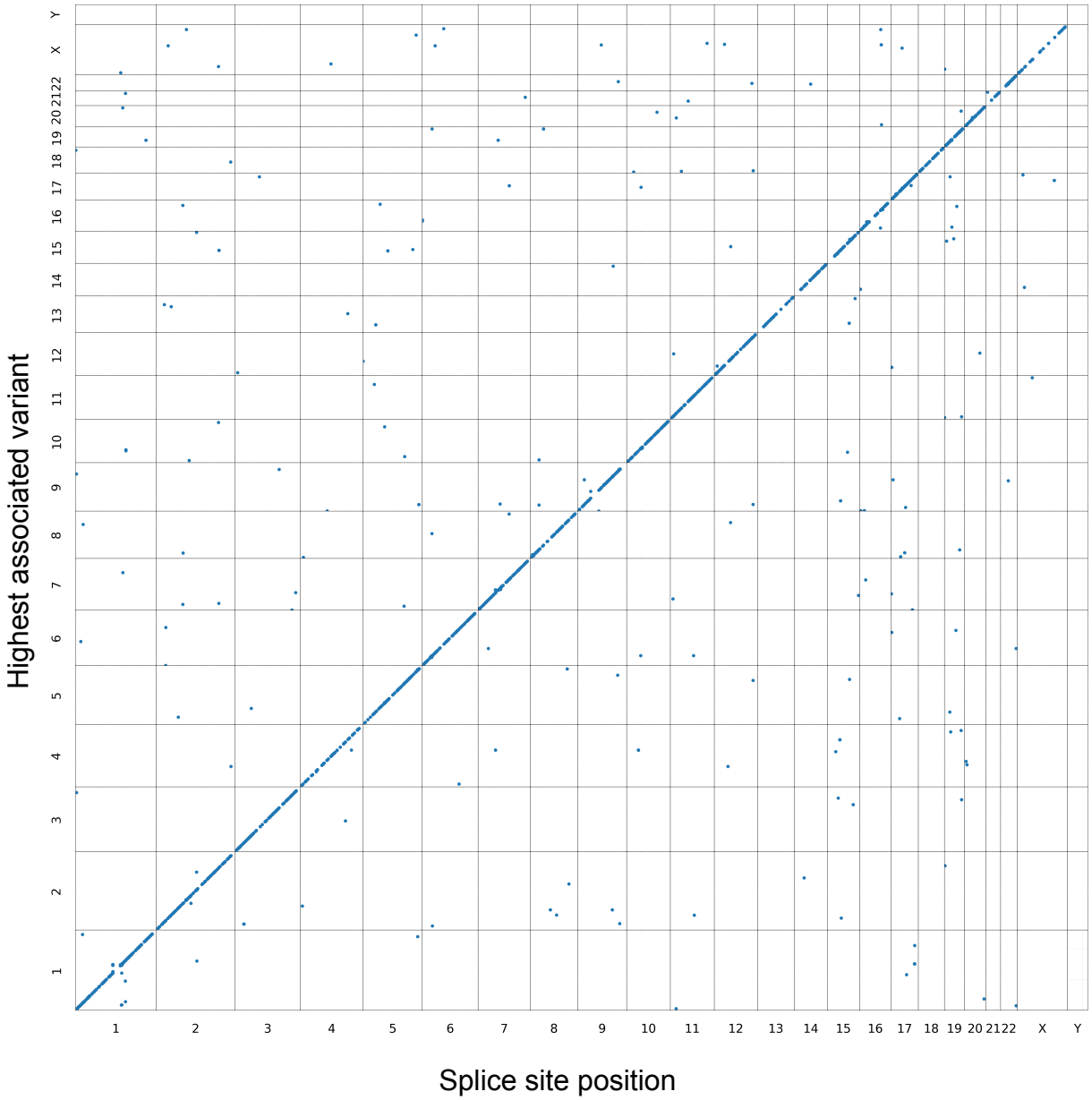


Figure 3.3: Splicing variability is mostly "*cis*" regulated. Scatter plot of splice-site positions vs their highest associated SNPs in the human testis across all the chromosomes.

3.4 Nucleotide variation around splice-sites modulates splice-site choice.

To evaluate features of splicing variation, we looked at the distances between the highest and closest associated SNPs to the respective splice site for all sites. This analysis revealed that the majority of associations map near the splice site, as shown in Figure 3.4. A good proportion of the highest associated SNPs (991/12,640) fell within +/-50bp of the splice site.

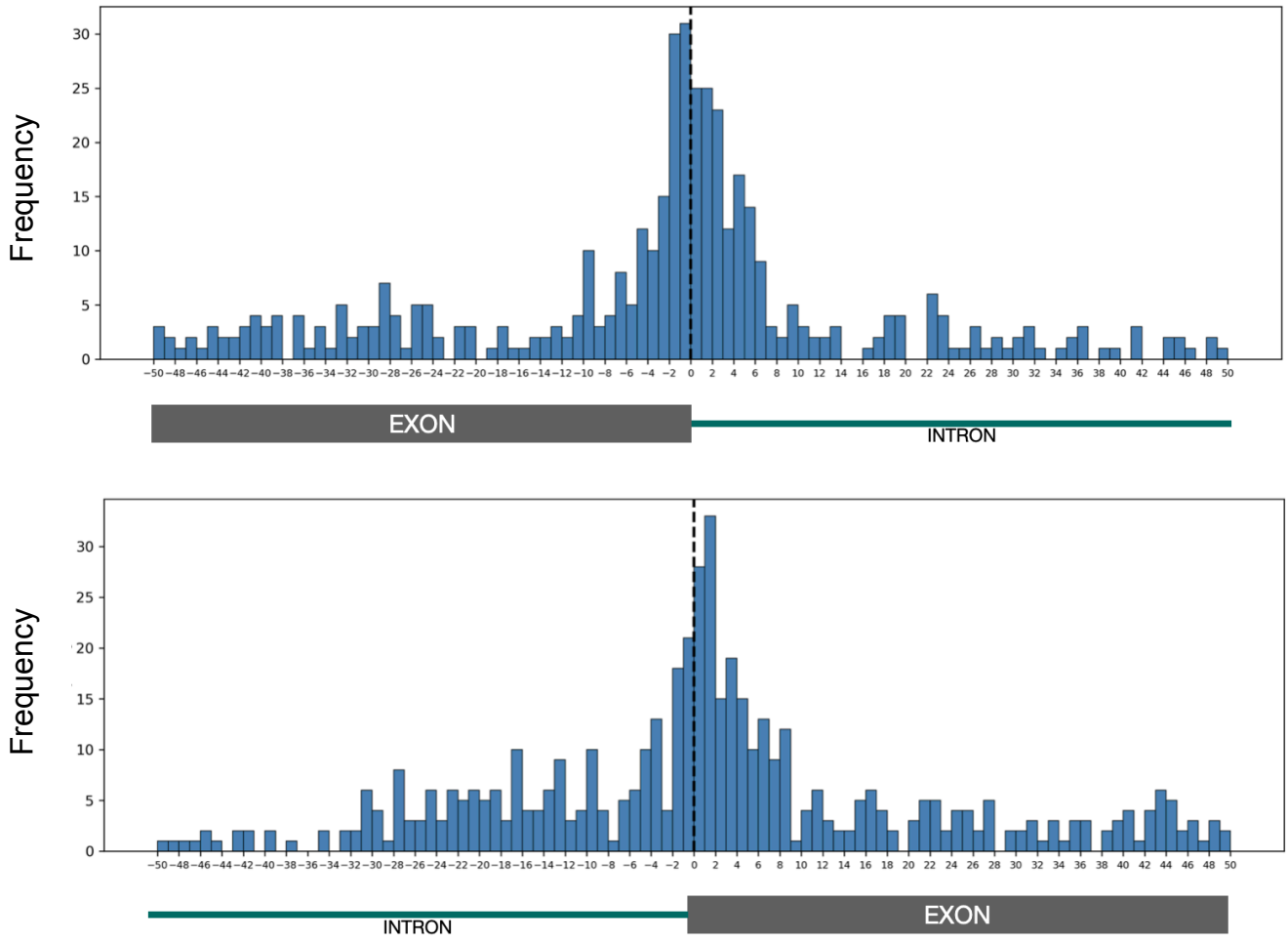


Figure 3.4: Genetic variation that impacts splice-site choice frequently occurs near the splice-site. The distribution of the distances of the highest associated SNPs identified in SpliSER-GWAS for donors (top) and acceptors (bottom). The nucleotide “G” of GT/AG is plotted as position 0.

3.5 Highest associated SNPs are linked with human diseases

Many diseases, including various cancers, are linked to differential splicing. One of the major aims of this project was to connect genetic variation, RNA splicing, and human diseases. Given that many disease phenotypes are linked to splicing changes, understanding the role of differential splicing in diseases is important. While efforts like SpliceAI and SpliceVault have linked splicing variants with diseases, a genome-wide approach is lacking. SpliSER-GWAS aims to bridge this gap by linking genome-wide variants to splicing changes.

To evaluate the potential impact on human traits, particularly diseases, we analysed whether any of our identified SNPs overlap with SNPs that have been previously associated with diseases using the GWAS Catalog from NHGRI (National Human Genome Research Institute) highlights SNPs associated with diseases. It is important to note that even if these SNPs are catalogued as disease/trait-associated SNPs, whether they are really causal SNPs and even if they are causal, the underlying mechanisms for some of the SNP-disease associations is not clear.

Comparing the top 12,640 associations from SpliSER-GWAS with known disease SNPs can reveal overlapping SNPs associated with both splicing changes and disease phenotypes. If a variant identified by SpliSER-GWAS is also significant in disease GWAS studies, it suggests a possible connection between the disease and splicing machinery. SpliSER-GWAS identified 4,665 unique SNPs associated with splicing variation, and of these, 379 SNPs (8.1%) were found to be associated with certain human traits or diseases. Thus this approach sheds light on the intricate relationship between genetic variation, RNA splicing, and human diseases, providing a valuable resource for further research into the molecular mechanisms underlying complex diseases and could pave the way for developing targeted therapeutics that modulate RNA splicing, offering new avenues for treating a wide range of diseases linked to splicing dysregulation.

Results - Section - B - Tissue specificity

3.6 Splice-sites can differ in their usage between tissues

Alternative splicing plays a crucial role in producing tissue-specific mRNA isoforms and, thus, tissue-specific proteins (Wen *et al.*, 2010; Wang *et al.*, 2009). It is regulated by various RNA-binding factors, which impact the spliceosome assembly and, thus, the usage of a splice site. This leads to differential usage of several splice sites in different tissues, leading to tissue-specific alternative splicing events, producing different isoforms and, thus, different proteins across tissues. Tissue-specific splicing is a critical process, yet it is not fully understood. Key questions remain unanswered, such as what are the factors that govern tissue-specific splicing (the specific RNA-binding factors involved) and the genomic determinants of this process.

Using the SSE value, we identified splice sites with differing usage across tissues (heart atrial and testis tissue). After obtaining SSE values for all sites in heart atrial and testis tissue, we focused on sites used in one tissue but not the other. This analysis identified 1537 sites corresponding to 697 genes used in the testis and not in the heart, and 1910 sites representing 567 genes used in the heart but not in the testis alongside 247,738 common sites utilised roughly equally in both tissues. An IGV view, for example sites that are used in a tissue-specific way, is depicted in Figure 3.5.

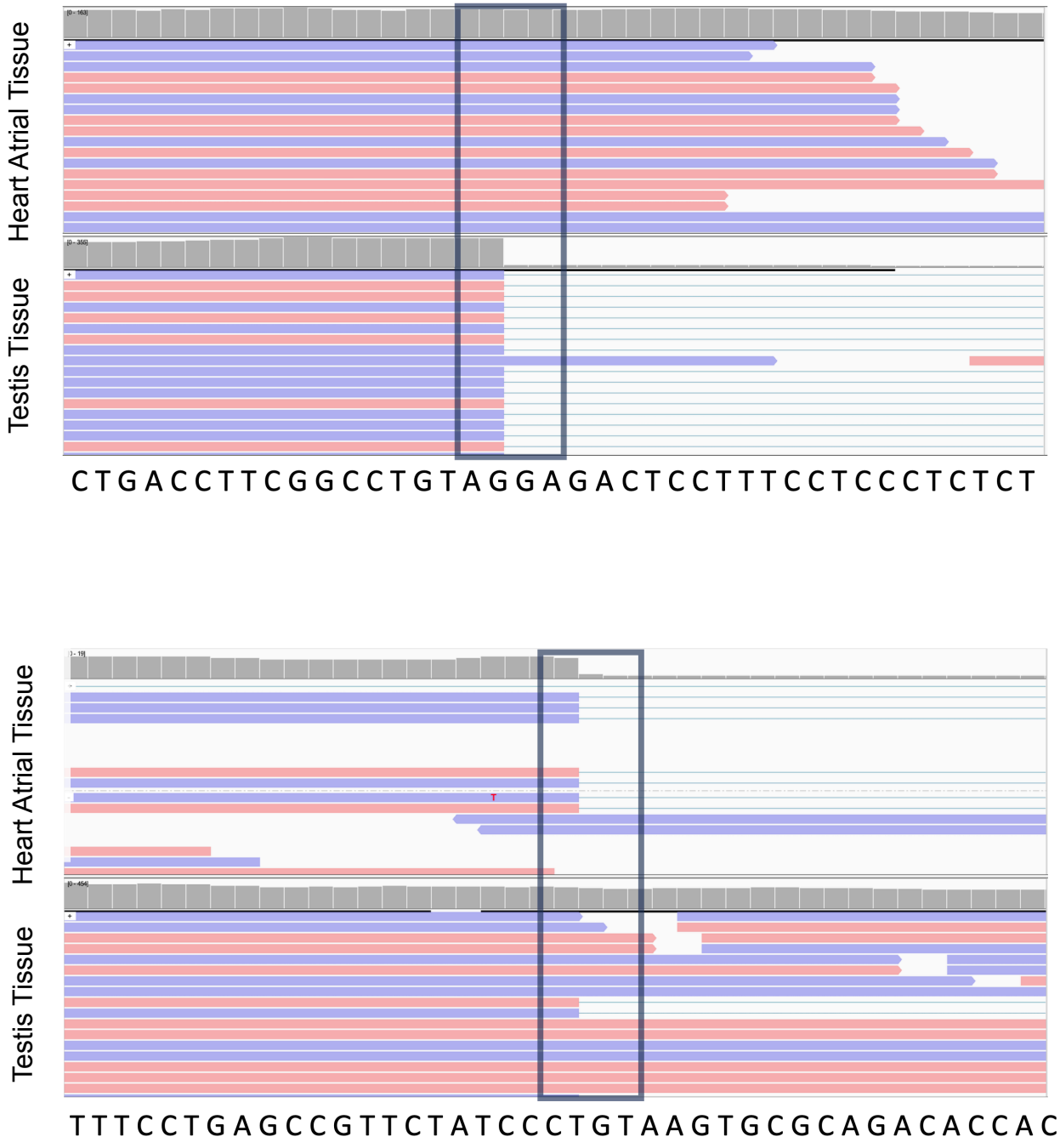
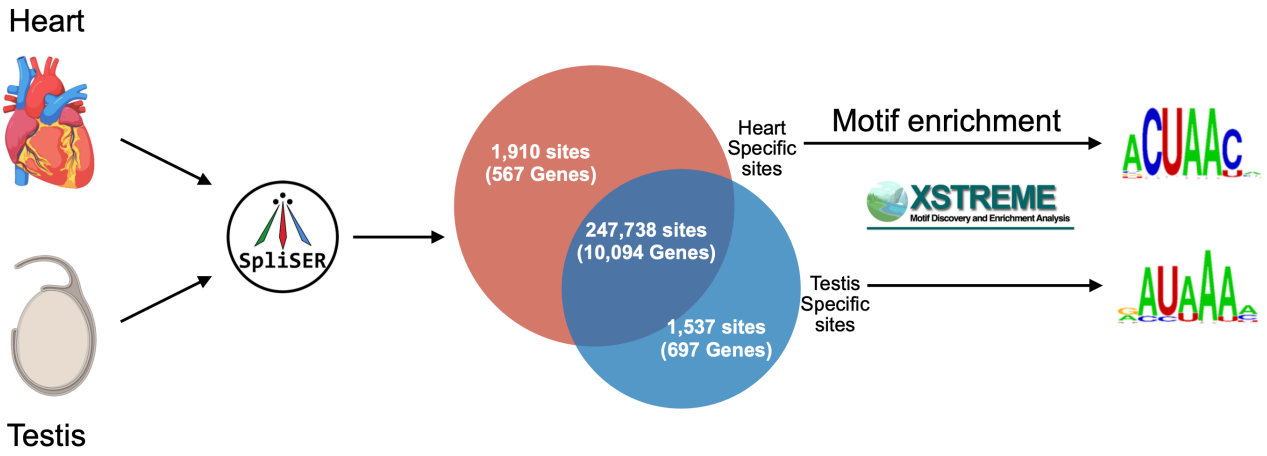


Figure 3.5: IGV view of sites used in a tissue-specific way. (a) showcases a site located at position Chr7:106112209 (located on a negative strand) used in testis tissue but not in the heart, while (b) displays a site at position Chr1:100499263 utilised in the heart but not in the testis. The splice sites are boxed.

3.7 Motif enrichment around tissue-specific splice sites gives motifs



(Heart, Testis image source: Biorender)

Figure 3.6: Motif enrichment analysis for tissue-specific splice sites. SpliSER was run on both testis and heart atrial tissue, and the splice sites, which are used in a tissue-specific way, were noted. A motif enrichment analysis was done for those sites using XSTREME.

Once we extracted the sites that are utilised in a tissue-specific way, we did a motif enrichment in a 600-base pair window centred at the splice site to look for sequence motifs that are enriched among those sites (see pipeline in Figure 3.6) and identified several motifs. Many of them are known to be associated with known RNA binding factors involved in splicing. See examples in Tables 2 & 3. One of the examples is the motif "GAUAAAA" (Figure 3.7), which was discovered through motif enrichment around splice site donors used in the testis but not in the heart. The known binding factor, KHDRBS2, is predicted to have a role in mRNA splicing (Bult and Sternberg, 2023). To assess whether the factor is expressed in a tissue-specific manner, we analysed the human protein atlas. We found that KHDRBS2 is expressed in testis but not in heart consistent with the enrichment of its binding site.

Conversely, we found the "ACUAACA" (Figure 3.7) motif as one of the examples for enrichment in heart-specific sites, with QKI (KH domain containing RNA binding) being the known binding factor for this motif. Again, the QKI is known to be present at higher levels

in the heart compared to the testis (HPA), consistent with its role in mRNA splicing in the heart (Chen *et al.*, 2021).

(Expression source: Human Protein Atlas)

Figure 3.7: Sequence logo for "GAUAAAA" motif, which is known to bind to KHDRBS2 (Left) and Sequence logo for "ACUAACA" motif, which is known to bind to QKI (Right). The binding factor's expression is depicted using the blue and red bars for the testis and heart, respectively.

3.8 Mutations in the motif region affect the SSE of the corresponding splice site

To validate if these motifs play a role in tissue-specific splicing, we looked at individuals with mutations in these motif regions to see if the SSE value for the corresponding splice site had changed. It was rare to find individuals with mutations in the motif region, as we had data from normal individuals and mutations in these regions can affect splicing patterns and thus affect an individual. However, we did find some cases where individuals had mutations in the motif regions. For example, in the case of motif KHDRBS2, which was found enriched in testis-specific splice sites, we found that the average SSE for the corresponding site was 0.579 for individuals with the normal motif. However, for the individual with a GATAAAA to G**G**TAAAA mutation in the motif, the SSE value dropped to 0.236 for the corresponding site (Figure 3.8). This analysis gives us more confidence that the identified motifs play a role in tissue-specific splicing. Since some of the detected motifs were not previously known to bind to a binding factor, we can further investigate those motifs to see if mutations in those motifs make a difference. We can then conduct experiments to identify a binding factor that binds to that motif. This method can help us detect novel tissue-specific splicing factors.

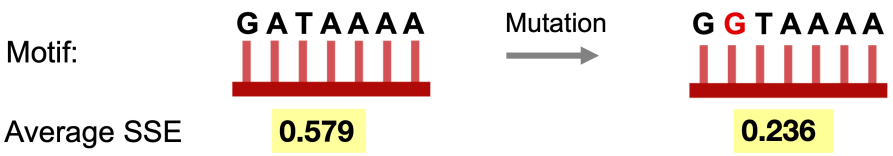


Figure 3.8: Mutation in the motif region affects the SSE of the corresponding splice-site. Illustrates the mutation in the motif and its effect on the SSE of the corresponding splice-site. The SSE values are highlighted in yellow.

3.9 Can genetic variation affect tissue-specific splicing?

We have shown that genetic variation can affect splicing. But then we asked, can genetic variation affect splicing in a tissue-specific way? To answer that question, we looked for sites where genetic variation affects splicing in one tissue but not the other. For that, we searched for sites where variation in the SSE was mapped in one tissue but not the other by SpliSER-GWAS.

We found 3104 sites that mapped in the testis but did not map in heart atrial tissue and 2036 sites that mapped in the heart but not in testis tissue. In other words, the nucleotide change at the highest associated SNP makes a difference in the SSE for the corresponding site in one tissue but not the other. One of the examples is shown in Figure 3.9 (A), where the SSE value for the splice site drops to around 0.25 for individuals with 'A' instead of 'G' at the highest associated SNP in testis, but it makes no difference for the SSE in heart. In fact, there was no significantly associated allele for the splice site in the case of the heart. Therefore, we see a clear mapping and, thus, a clean peak in the GWAS for testis tissue but not the heart, as shown in Figure 3.9 (C).

We speculated that a splicing factor could be binding on or near the highest associated SNP, potentially influencing the usage of the corresponding splice site. This factor might be present in a tissue-specific manner, meaning that a change in the nucleotide could affect its binding affinity in one tissue but not in another. To test this hypothesis, we tried to do a motif search around the highest associated SNP and see if we found any meaningful motifs. For the motif search, we identified 44 donor sites where the variation in the SSE

was mapped in the testis but not in the heart. Conversely, we found 36 donor sites where the variation in the SSE was clearly mapped in the heart but not in the testis.

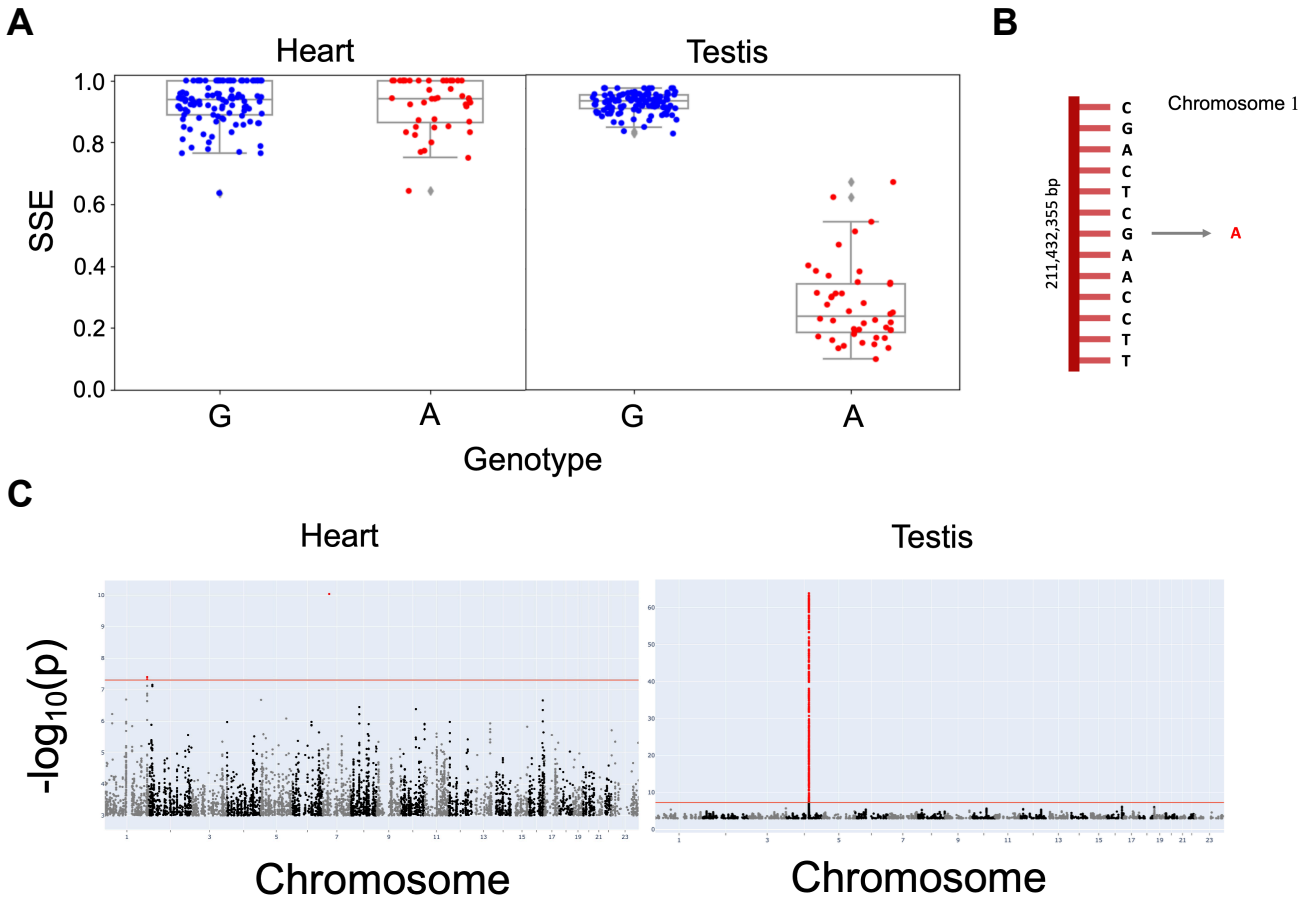


Figure 3.9: (A) depicts the SSE for individuals with the 'G' (blue) and 'A' (red) nucleotides at the highest associated SNP in heart and testis. A noticeable change in SSE is observed in the testis but not in the heart. (B) displays the position of the highest associated SNP. (C) illustrates the corresponding GWAS Manhattan plots, revealing a clean mapping in the case of testis but not in the heart. The Bonferroni threshold is indicated by a red line

3.10 Motif search around the highest associated SNP gives us relevant motifs

After identifying all sites, we conducted a motif search within a small window surrounding the highest associated SNP. We found several motifs in that region, many of which were linked to known binding factors. Upon examining the expression levels of these factors, we noted significant differences between the two tissues for many of them.

For example, when we focused on a motif associated with sites that mapped in testis but not to heart, we identified "UAGGUAG" (Figure 3.10) as the motif known to bind to the factor DAZAP1, which is a known splicing activator (Choudhury *et al.*, 2014). Interestingly, the protein expression of DAZAP1 was higher in the testis compared to the heart (Source: HPA). Conversely, the motif "AAGAA" was found at the highest associated SNP for sites that are mapped in the heart but not in the testis. This Motif is known to be bound by the TRA2A splicing factor (Tacke and Manley, 1999). Here also, we found expression for TRA2A to be higher in the heart compared to the testis (source: HPA). These findings suggest that these factors play a crucial role in driving tissue-specific splicing patterns and thus provide evidence that this approach can help us understand how genomic variation affects splicing in a tissue-specific way.

(Expression source: Human Protein Atlas)

Figure 3.10: Sequence logo for "AAGA" motif, which is known to bind to TRA2A (Left) and Sequence logo for "UAGGUAG" motif, which is known to bind to DAZAP1 (Right). The binding factor's expression is depicted using the blue and red bars for the testis and heart, respectively.

Results - Section - C - Splicing code

3.11 Deciphering the "splicing code"

From the SpliSER-GWAS analysis on the testis, we get many of the highest associations to fall near the splice sites. As we have many associations at each position relative to the splice site, we can identify nucleotides for each position relative to the splice site that enhance or repress splicing (Figure 3.11). For instance, in cases where the top SNP is at the splice-site, we observed that the 'G' nucleotide enhances splicing, while for top associated SNPs at position +1 relative to the splice site, a 'T' nucleotide promotes splicing, which aligns with the 'GT' motif commonly found at splice sites. The fact that these were identified solely through our GWAS analysis highlights the power of this approach. Similarly, we can go beyond and look for splice-promoting nucleotides for a larger window around the splice site. By deciphering the splice-promoting bases at each position near the splice site, we can potentially design an "ideal" or "best" intron by combining splice-promoting nucleotides for each position. Similarly, we can decipher the sequence for a "worst" intron by combining splice-reducing nucleotides for each position.

EXON — INTRON

Position	-2	-1	0	1	2	3	4	5
A	**14**	3	1	0	**17**	**10**	1	1
C	11	3	2	2	0	2	0	2
G	2	**24**	**17**	1	3	0	**15**	3
T	0	0	2	**21**	1	0	1	**8**
Total	27	30	22	24	21	12	17	14
Best	**A**	**G**	**G**	**T**	**A**	**A**	**G**	**T**

INTRON — EXON

-5	-4	-3	-2	-1	0	1	2	Position
5	1	1	2	**15**	0	4	1	A
0	**5**	0	**10**	0	6	**12**	2	C
0	0	0	0	4	**20**	**12**	3	G
4	4	**3**	6	1	1	2	**8**	T
27	30	22	24	21	12	17	14	Total
A	**C**	**T**	**C**	**A**	**G**	**C/G**	**T**	Best

Figure 3.11: The distribution of alleles that promote splicing for positions near the splice-site donors (-2 to +5) (left) and acceptors (-5 to +2) (right). The most common nucleotide that enhances splicing is identified for each position.

To expand this analysis, we aggregated data from Arabidopsis, Drosophila, and human atrial tissue previously generated in our lab to increase statistical power. By doing so, we can determine the nucleotide at each position relative to the splice site that enhances splicing. From this, we can design a splice-promoting intron, which we can refer to as the

"best intron", and a splice-repressing intron, which we can refer to as the "worst intron". James, a PhD student in our lab, has experimentally validated the functionality of both the best and worst introns. He demonstrated perfect splicing in the case of the "best intron" and no splicing in the case of the "worst intron". Both introns contain the canonical 'GT/AG' at the splice site and the branch point.

3.12 Hexamer is the primary determinant of splice-site choice

In a previous analysis done in our lab, we analysed the relationship between splice site strength (SSE) and their surrounding sequences to understand the "splicing code". We explored various sets of sequences around splice sites (Figure 3.12) to determine which best explains splice site selection, using different strategies to identify the optimal sequence length and distance from the splice site. Our results suggested that the hexamer sequence ($GT[N]_4$ & $[N]_4AG$) around the splice sites best explains the variation in the splice site strength (SSE). To validate this idea further, I used two different approaches.

Figure 3.12 The illustration of various sets of potential sequences around the splice-site that could best explain splice site selection.

We grouped all possible splice-sites into distinct *k*-mer groups (e.g., GTNNNN with 256 sequence combinations) and calculated the average strength for each group by taking the average of SSE values for the sites in a group. We created a *k*-mer rank based on the average SSE for each group. The first approach involved testing how well the rankings explained splice-site choice in three species. This score is shown in Figure 3.13 for Arabidopsis, Drosophila and Humans. We found that, for donors, $GT[N]_4$ hexamers explain most of the splice-site choices. With acceptors, we found both hexamers $[N]_4AG$ and

$[N]_5$AG had comparable scores, though the hexamers, on average, slightly outperformed the septamers.

We found that hexamer ranking, based on their average SSE rank, could explain most splice-site choices. For humans, out of 182,512 splice donor sites with GT, the hexamer ranking explained around 73% of the detected splice sites. Similarly, for the acceptor sites with AG, hexamer ranking explained the choice for around 61% of the sites. Similar trends were observed in Arabidopsis and Drosophila, with hexamer ranking explaining 58% and 81% of splice site choices, respectively. For splice acceptor sites, hexamer ranking accounted for around 65% and 72% in Arabidopsis and Drosophila, respectively. We then investigated whether the ranking of hexamers can explain splice-site choice across different species. We analysed RNA-seq data from 20 eukaryotic species to establish hexamer ranks based on their strength and frequency. We computed hexamer rankings for

k-mers	Percentage of splice site choices explained			
	Arabidopsis	Drosophila	Humans	Average
Donors				
GT	0.66	0.76	0.87	0.76
GTN	28.56	15.02	26.73	23.43
GTNN	38.45	48.73	52.42	46.53
GTNNN	56.06	77.47	69.87	67.80
GTNNNN	**58.26**	**81.41**	**72.86**	**70.84**
GTNNNNN	48.95	65.84	69.05	61.28
GTNNNNNN	24.60	40.90	53.24	39.58
Acceptors				
AG	0	1.93	0	0.64
NAG	30.24	28.46	19.49	65.19
NNAG	48.95	46.04	32.14	42.37
NNNAG	60.98	64.05	46.46	57.16
NNNNAG	**65.73**	**71.02**	**55.25**	**64.00**
NNNNNAG	65.71	65.03	**57.72**	62.00
NNNNNNAG	53.92	43.29	51.10	49.43

Figure 3.13: It illustrates how the ranking of hexamers can explain the selection of splice sites. The figure shows the percentage of splice site selections explained by various intronic *k*-mers in three species, along with their average. The best *k*-mers for donor and acceptor sites are highlighted in blue.

each species and evaluated the proportion of splice site choices that these rankings could explain (Table 4). Remarkably, hexamer rankings based on splice site strength accounted for most splice site choices (approximately 60-85%) across diverse species. Overall, hexamer-ranking provided a robust explanation for most splice site choices across eukaryotic transcriptomes, suggesting a fundamental framework underlying splicing regulation.

3.13 Meta-analysis to test hexamers as determinants of splice-site choice

We utilised data from Rosenberg et al.'s study to examine whether hexamer ranking is the critical factor in splice-site choice. We conducted a meta-analysis focusing on hexamer ranking (Rosenberg et al., 2015). In their study, Rosenberg et al. created numerous mini-gene constructs with embedded random sequences capable of forming potential novel splice sites. They then sequenced the DNA and RNA to analyse splice-site usage. Using SpliSER on this RNA-seq data, we quantified the splice-site usage for all splice sites.

We identified a total of 466,000 potential competing donor sites that could be evaluated based on hexamer rankings. For each combination, we examined the hexamers surrounding the splice sites and determined how many of the winning splice sites could be explained by the hexamer rankings. Our analysis of approximately 4000 unique competing pairs indicated that hexamer ranking could account for 68% of splice site choices (Fig 3.14). This percentage increased to 91% when considering constructs where the differences in hexamer strengths were greater than 25% (according to the ranking), highlighting the significant role of hexamers as primary determinants of splice site choice.

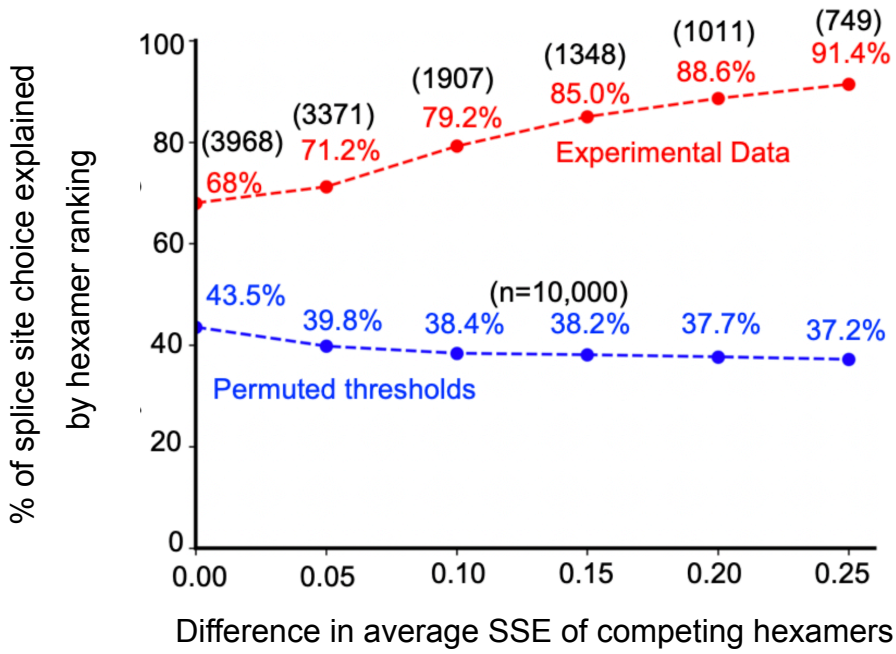


Figure 3.14: Hexamers determine splice-site selection in an experiment by comparing competing hexamers in minigene constructs from Rosenberg et al. (2015). The blue line indicates thresholds derived from 10,000 permutations, while the red line represents the actual experimental data, showing different combinations of hexamers. The number of unique competing pairs tested is displayed above the percentages.

Chapter 4

Discussion

There is extensive, genetically determined variation in splicing in the human testis that can be mapped to the specific nucleotide variation by combining empirically quantified estimates of splice-site strength with GWAS. A major debate in splicing revolves around whether the genomic determinants influencing splicing variation are primarily cis or trans. The literature presents conflicting views, with some groups arguing for predominant cis effects (Garrido-Martín *et al.*, 2021) and others for predominant trans effects (Khokhar *et al.*, 2019); however, many of them have insufficient data. Our genome-wide approach offers a unique opportunity to answer this question. We show that the majority of splicing variation in the human testis is driven by "cis" rather than "trans" regulatory changes. While some trans-acting factors are essential for splicing (Lee and Rio, 2015), natural variation in trans effects can occur through two main mechanisms. One involves sequence changes in trans-acting factors, affecting multiple genes, while the other involves changes in potential binding sites, leading to specific changes in individual genes or splice sites. Our findings suggest that cis-regulatory sequence variation is favoured over trans-regulatory sequence variation, as we found no major trans-regulatory hotspots. The trans-regulatory sequence variation can affect multiple genes, while the cis-regulatory changes enable targeted changes to individual genes or splice sites, which can be more evolutionarily advantageous than widespread effects on multiple genes (Singh and Ahi, 2022).

GWAS studies on various traits/genetic diseases have identified important genetic regions and variants (Abdellaoui et al., 2023). Despite this progress, understanding how these variants function remains challenging (Lappalainen and MacArthur, 2021). SpliSER-GWAS analysis has shown promise in identifying specific SNPs that affect splicing, which in many cases are causal polymorphisms that drive a change in splicing. As we show, around 8% of the highest associated SNP are found to be associated with human traits/diseases.

Splicing serves a crucial function in conferring tissue specificity, evident in the diverse splicing patterns observed across various tissues. However, investigating tissue-specific

splicing encounters a significant hurdle: the predominant adoption of an event-based approach. This method, focused on splicing events, complicates our understanding of the underlying biological mechanisms. We show the effectiveness of the SpliSER-based approach in identifying motifs and potential regulators of tissue-specific splicing. Since we found numerous motifs enriched in tissue-specific sites, many of which are unassociated with any RNA-binding factors, we can now explore these motifs to uncover previously unknown regulators of tissue-specific splicing. Tissue-specific splicing is regulated by a combination of splicing factors that are specific to tissues as well as those that are expressed ubiquitously (Wang *et al.*, 2008). Our result agrees with this notion, where we find factors that are expressed in a tissue-specific way driving tissue-specific splicing.

We demonstrated that the hexamers ($GT[N]_4$ and $[N]_4AG$) surrounding splice donor and acceptor sites play a pivotal role in splice site selection. Through our analysis, we have established a ranking of hexamers based on the average SSE, showing that the majority of splice site choices across diverse species can be explained by this ranking. These results indicate that the hexamer ranking forms a primary part of the fundamental logical framework that governs splice site selection in eukaryotic organisms. The exact mechanism behind the hexamer is currently unclear, but one possibility is that it involves differential pairing with snRNPs. For instance, variability in the donor site sequence could affect how U1 SnRNA binds, potentially leading to different binding kinetics of U1 SnRNA and its associated proteins. This could impact splice site choice (Rogalska et al., 2023). A similar scenario could explain variance at acceptor sites. Further experiments are needed to confirm these ideas, but it's possible that sequence variation-driven changes could directly or indirectly affect RNA-SnRNP interactions. These differential interactions might be the underlying mechanism through which hexamers influence splice site choices.

Data Availability

A part of this work is written up and it is part of a paper that has been submitted and is expected to appear in BioRxives in the next couple of days. This thesis presented the summary of results and the raw data is available at Monash University, Australia.

Publication Reference.

Dent, CI*., Produce, S*., Balakrishnan, A*., Georges, J$, **Chhabra, A$,** Mukherjee, S., Coutts, J., Gitonobel, M., Sarwade, RD., Rosenbluh, J., D'Amato, M., Das, PP., Guo, Y-L., Fournier-Level, A., Burke, R., Sureshkumar, S., Powell, D and Balasubramanian, S (2024) A basic framework governing splice-site choice in eukaryotes, BioRxives (to appear soon)

Tables

Tissue	JSC
Heart - Atrial Appendage	1
Heart - Left Ventricle	0.92
Artery - Coronary	0.9137
Esophagus - Gastroesophageal Jun	0.9089
Esophagus - Muscularis	0.9087
Artery - Aorta	0.9028
Colon - Sigmoid	0.9002
Artery - Tibial	0.8962
Adipose - Visceral (Omentum)	0.8962
Adipose - Subcutaneous	0.8914
Bladder	0.8878
Adrenal Gland	0.8859
Breast - Mammary Tissue	0.8853
Stomach	0.8837
Uterus	0.8816
Vagina	0.8753
Colon - Transverse	0.8741
Nerve - Tibial	0.8718
Kidney - Cortex	0.8705
Thyroid	0.8691
Lung	0.8684
Ovary	0.8653
Prostate	0.8653
Esophagus - Mucosa	0.8649
Fallopian Tube	0.8644
Cervix - Ectocervix	0.861
Muscle - Skeletal	0.8608
Minor Salivary Gland	0.8583
Skin - Sun Exposed (Lower leg)	0.8576
Cervix - Endocervix	0.8551
Pancreas	0.8547
Small Intestine - Terminal Ileum	0.8537
Skin - Not Sun Exposed (Suprapub	0.8536
Brain - Spinal cord (cervical c-1)	0.851
Brain - Substantia nigra	0.8452
Pituitary	0.8421
Spleen	0.8413
Brain - Hippocampus	0.8395
Liver	0.8389
Brain - Hypothalamus	0.8382
Brain - Amygdala	0.8356
Brain - Caudate (basal ganglia)	0.8346
Brain - Putamen (basal ganglia)	0.834
Cells - Cultured fibroblasts	0.8335
Brain - Cortex	0.8321
Brain - Nucleus accumbens (basal	0.828
Brain - Anterior cingulate cortex (BA	0.8276
Brain - Frontal Cortex (BA9)	0.8246
Brain - Cerebellar Hemisphere	0.8157
Brain - Cerebellum	0.8149
Kidney - Medulla	0.8128
Cells - EBV-transformed lymphocyt	0.7945
Whole Blood	0.7673
Testis	**0.6776**

Table 1: List of Jaccard Similarity coefficient for all tissues with heart atrial tissue for gene expression. They are arranged in descending order.

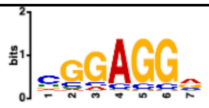
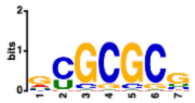
Motif	E-Value	Binding Factor	Similar Binding Factor
	3.97e-002	LIN28A (RNCMPT00162)	-
	3.56e-004	RBM8A (RNCMPT00056)	-
	2.28e-003	-	-
	6.95e-003	SRSF1 (RNCMPT00109)	-
	1.46e-002	SRSF9 (RNCMPT00067)	-
	1.16e-002	QKI (RNCMPT00047)	
	2.40e-005	-	-
	9.15e-003	-	HNRNPH2 (RNCMPT00160)
	3.00e-011	-	PCBP1 (RNCMPT00186) PCBP2 (RNCMPT00044) HNRNPK (RNCMPT00026)

Table 2: List of motifs found to be significantly enriched in sites that are used in heart but not in testis. The E-value for each motif is determined through the Fisher exact test.

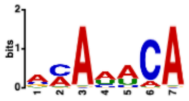
Motif	E-Value	Binding Factor	Similar Binding Factor
	1.36e-002	-	-
	4.94e-004	IGF2BP2 (RNCMPT00033)	-
	2.28e-003	SART3 (RNCMPT00064) PABPC1 (RNCMPT00155)	-
	8.36e-003	KHDRBS2 (RNCMPT00185)	-
	2.38e-002	-	-
	1.14e-001	-	RBM6 (RNCMPT00170)
	3.24e-002	HNRNPL (RNCMPT00027)	-
	2.38e-002	-	-
	7.00e-015	-	PCBP1 (RNCMPT00186) PCBP2 (RNCMPT00044) HNRNPK (RNCMPT00026)
	8.15e-001	-	

Table 3: List of motifs found to be significantly enriched in sites that are used in testis but not in heart. The E-value for each motif is determined through the Fisher exact test.

Number	Species	Donor/ Acceptor	Total Splice sites	Hexamer Rank 1 Sites	Percentage of sites explained by hexamer ranks
1	Human	Donors	182512	139923	76.67
2	Human	Acceptors	169860	103513	60.94
3	Drosophila	Donors	41983	35430	84.39
4	Drosophila	Acceptors	39399	29100	73.86
5	Arabidopsis	Donors	92246	55854	60.55
6	Arabidopsis	Acceptors	85663	60120	70.18
7	Rice	Donors	64898	37759	58.18
8	Rice	Acceptors	62030	43793	70.60
9	C elegans	Donors	1709	726	42.48
10	C elegans	Acceptors	1739	1444	83.04
11	Chicken	Donors	83736	64999	77.62
12	Chichen	Acceptors	80648	51581	63.96
13	Chimp	Donors	117165	92195	78.69
14	Chimp	Acceptors	114214	70227	61.49
15	Cobra	Donors	47319	34895	73.74
16	Cobra	Acceptors	47447	29579	62.34
17	Corn	Donors	46225	32147	69.54
18	Corn	Acceptors	47461	26163	55.13
19	Opium	Donors	18064	5102	28.24
20	Opium	Acceptors	18306	11404	62.30
21	Pig	Donors	14520	6723	46.30
22	Pig	Acceptors	17312	6802	39.29
23	Potato	Donors	47503	26248	55.26
24	Potato	Acceptors	47151	31856	67.56
25	Sorghum	Donors	76514	46476	60.74
26	Sorghum	Acceptors	72794	50963	70.01
27	Tomato	Donors	38109	18849	49.46
28	Tomato	Acceptors	36629	25480	69.56
29	Xenopus	Donors	122954	96211	78.25
30	Xenopus	Acceptors	119674	78537	65.63
31	ZebraFish	Donors	117839	76586	64.99
32	ZebraFish	Acceptors	113729	67856	59.66
33	Canola	Donors	173671	112532	64.80
34	Canola	Acceptors	168594	117280	69.56
35	Tegu	Donors	101340	80184	79.12
36	Tegu	Acceptors	97615	61826	63.34
37	Hydra	Donors	73004	30851	42.26
38	Hydra	Acceptors	71118	51566	58.12
39	Chara	Donors	4642	2675	57.63
40	Chara	Acceptors	4716	2900	61.49

Table 4: Percentage of splice site choice explained by hexamer ranking for 20 species for both acceptors and donor splice sites.

Bibliography

Abdellaoui, A, Yengo, L, Verweij, KJH, and Visscher, PM (2023). 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet 110, 179–194.

Bult, CJ, and Sternberg, PW (2023). The alliance of genome resources: transforming comparative genomics. Mamm Genome 34, 531–544.

Castle, JC, Zhang, C, Shah, JK, Kulkarni, AV, Kalsotra, A, Cooper, TA, and Johnson, JM (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40, 1416–1425.

Chen, X, Liu, Y, Xu, C, Ba, L, Liu, Z, Li, X, Huang, J, Simpson, E, Gao, H, Cao, D, et al. (2021). QKI is a critical pre-mRNA alternative splicing regulator of cardiac myofibrillogenesis and contractile function. Nat Commun 12.

Choudhury, R, Roy, SG, Tsai, YS, Tripathy, A, Graves, LM, and Wang, Z (2014). The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration. Nat Commun 5, 3078.

Clancy, s. (2008) rna splicing: introns, exons and spliceosome. nature education 1(1):31.

Cotto, KC, Feng, Y-Y, Ramu, A, Richters, M, Freshour, SL, Skidmore, ZL, Xia, H, McMichael, JF, Kunisaki, J, Campbell, KM, et al. (2023). Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. Nat Commun 14, 1589.

Dent, CI, Singh, S, Mukherjee, S, Mishra, S, Sarwade, RD, Shamaya, N, Loo, KP, Harrison, P, Sureshkumar, S, Powell, D, et al. (2021). Quantifying splice-site usage: a simple yet powerful approach to analyze splicing. NAR Genomics Bioinforma 3, lqab041.

Dwivedi, SL, Quiroz, LF, Reddy, ASN, Spillane, C, and Ortiz, R (2023). Alternative Splicing Variation: Accessing and Exploiting in Crop Improvement Programs. Int J Mol Sci 24, 15205.

Early, P, Rogers, J, Davis, M, Calame, K, Bond, M, Wall, R, and Hood, L (1980). Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell 20, 313–319.

Fica, SM, and Nagai, K (2017). Cryo-EM snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. Nat Struct Mol Biol 24, 791–799.

Fu, X-D, and Ares, M (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet 15, 689–701.

Garrido-Martín, D, Borsari, B, Calvo, M, Reverter, F, and Guigó, R (2021). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nat Commun 12, 727.

Giudice, G, Sánchez-Cabo, F, Torroja, C, and Lara-Pezzi, E (2016). ATtRACT-a database of RNA-binding proteins and associated motifs. Database J Biol Databases Curation 2016, baw035.

Grant, C, and Bailey, T (2021). XSTREME: Comprehensive motif analysis of biological sequence datasets.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330.

Haller, T, Tasa, T, and Metspalu, A (2019). Manhattan Harvester and Cropper: a system for GWAS peak detection. BMC Bioinformatics 20, 22.

Hocine, S, Singer, RH, and Grünwald, D (2010). RNA processing and export. Cold Spring Harb Perspect Biol 2, a000752.

Khokhar, W, Hassan, MA, Reddy, ASN, Chaudhary, S, Jabre, I, Byrne, LJ, and Syed, NH (2019). Genome-Wide Identification of Splicing Quantitative Trait Loci (sQTLs) in Diverse Ecotypes of Arabidopsis thaliana. Front Plant Sci 10.

Lappalainen, T, and MacArthur, DG (2021). From variant to function in human disease genetics. Science 373, 1464–1468.

Lee, Y, and Rio, DC (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem 84, 291–323.

Lukacsovich, D, Winterer, J, Que, L, Luo, W, Lukacsovich, T, and Földy, C (2019). Single-Cell RNA-Seq Reveals Developmental Origins and Ontogenetic Stability of Neurexin Alternative Splicing Profiles. Cell Rep 27, 3752-3759.e4.

Manning, KS, and Cooper, TA (2017). The roles of RNA processing in translating genotype to phenotype. Nat Rev Mol Cell Biol 18, 102–114.

Martin Anduaga, A, Evantal, N, Patop, IL, Bartok, O, Weiss, R, and Kadener, S Thermosensitive alternative splicing senses and mediates temperature adaptation in Drosophila. eLife 8, e44642.

Matera, AG, and Wang, Z (2014). A day in the life of the spliceosome. Nat Rev Mol Cell Biol 15, 108–121.

Moschall, R, Rass, M, Rossbach, O, Lehmann, G, Kullmann, L, Eichner, N, Strauss, D, Meister, G, Schneuwly, S, Krahn, MP, et al. (2019). Drosophila Sister-of-Sex-lethal reinforces a male-specific gene expression pattern by controlling Sex-lethal alternative splicing. Nucleic Acids Res 47, 2276–2288.

Nilsen, TW, and Graveley, BR (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463.

Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, MAR, Bender, D, Maller, J, Sklar, P, de Bakker, PIW, Daly, MJ, et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81, 559–575.

Rice, GR, Barmina, O, Luecke, D, Hu, K, Arbeitman, M, and Kopp, A (2019). Modular tissue-specific regulation of doublesex underpins sexually dimorphic development in Drosophila. Dev Camb Engl 146, dev178285.

Rogalska, ME, Vivori, C, and Valcárcel, J (2023). Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. Nat Rev Genet 24, 251–269.

Rosenberg, AB, Patwardhan, RP, Shendure, J, and Seelig, G (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell 163, 698–711.

Singh, P, and Ahi, EP (2022). The importance of alternative splicing in adaptive evolution. Mol Ecol 31, 1928–1938.

Su, C-H, D, D, and Tarn, W-Y (2018). Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci 5, 12.

Tacke, R, and Manley, JL (1999). Functions of SR and Tra2 proteins in pre-mRNA splicing regulation. Proc Soc Exp Biol Med Soc Exp Biol Med N Y N 220, 59–63.

Uffelmann, E, Huang, QQ, Munung, NS, de Vries, J, Okada, Y, Martin, AR, Martin, HC, Lappalainen, T, and Posthuma, D (2021). Genome-wide association studies. Nat Rev Methods Primer 1, 1–21.

Visscher, PM, Brown, MA, McCarthy, MI, and Yang, J (2012). Five Years of GWAS Discovery. Am J Hum Genet 90, 7–24.

Wang, ET, Sandberg, R, Luo, S, Khrebtukova, I, Zhang, L, Mayr, C, Kingsmore, SF, Schroth, GP, and Burge, CB (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476.

Wang, X, Wang, K, Radovich, M, Wang, Y, Wang, G, Feng, W, Sanford, JR, and Liu, Y (2009). Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. BMC Genomics 10, S4.

Wen, J, Chiba, A, and Cai, X (2010). Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. Nucleic Acids Res 38, 7895–7907.

Zhou, X, and Stephens, M (2012). Genome-wide Efficient Mixed Model Analysis for Association Studies. Nat Genet 44, 821–824.