

**DETERMINING BIOLOGICAL SIGNIFICANCE OF PUTATIVE EXONS
THROUGH THE APPLICATION OF ANNOTATION-AGNOSTIC METHODS
TO MEASURE POTENTIAL HALLMARKS OF FUNCTIONALITY**

By

Bianca Hoch

M.S. Bioinformatics (Hood College) 2022

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

In

BIOINFORMATICS

In the

GRADUATE SCHOOL

Of

HOOD COLLEGE

May 2022

Accepted:

DocuSigned by:
Conor Jenkins
15C65950742746C...

Conor Jenkins,
Committee Member

DocuSigned by:
Craig Laufer
0A64B5612BEE4DC...

Craig Laufer, Ph.D.
Committee Member

DocuSigned by:
Miranda M. Darby
CCFD0206D1064ED...

Miranda M. Darby, Ph.D.
Thesis Advisor
Director, Bioinformatics Program

STATEMENT OF USE AND COPYRIGHT WAIVER

I authorize Hood College to lend this thesis, or reproductions of it, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

DEDICATION

I would like to dedicate this study to students pursuing bioinformatics research opportunities, who will become the next generation of brilliant scientists and bring about a new wave of exciting discoveries. May your love for learning always guide you, and take you to great heights. I would also like to dedicate this work to the individuals who chose to donate their bodies to science. Their selfless contribution has advanced our understanding of the human form and all of its wonders.

ACKNOWLEDGEMENTS

First, I would like to thank my esteemed committee members, Dr. Miranda Darby of Hood College, Dr. Craig Laufer of Hood College, and Conor Jenkins. Dr. Darby is a fountain of strength and inspiration for me. She has served not only as my mentor in school, but as a mentor in life as well. Beyond being an excellent teacher, she is a wonderful mother, friend, and human being. I am blessed to have had the opportunity to learn so much from such an exceptional person.

I have had the pleasure of knowing Dr. Laufer since undergraduate school. He is well versed in microbiology, and the fermentation of fine wines and foods. Dr. Laufer is a shining example of a professor who teaches students how to think, and not just what to think. Some of my most challenging and rewarding experiences as a student occurred during his class. I will always be grateful for the opportunities he gave his students to be impressed with their own problem solving capabilities.

I am very lucky to call Conor Jenkins my colleague, my mentor, and my friend. He is a beacon of light in the realm of bioinformatics. He is unapologetically passionate about research, and has a habit of captivating the minds of others when he talks about his projects and ideas. Conor has set a new bar for how much drive and intellectual curiosity the human body can hold. I will always be grateful for the fact that I got to meet such a remarkable person at Hood College, and that he agreed to serve on my committee.

In addition to my committee members, I want to thank the Hood College Biology department. My undergraduate and graduate experience was phenomenal because of them. I am deeply appreciative for everything they have taught me, and for the lifelong memories and lessons they gave me.

Lastly, I would like to thank my mom and dad. They are my greatest support, and my guiding lights. When impostor syndrome or doubt crept in, they reassured me that I am and always will be the captain of my ship, and the master of my destiny. Because of my dad, I kept my nose to the grindstone. Thanks to my mom, I've hitched my wagon to the highest star. I am grateful that life has given me such wonderful people to learn from and love so dearly.

TABLE OF CONTENTS

TABLE OF CONTENTS	vi
LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	1
INTRODUCTION	2
Novel Splice Isoforms and Current Gene Annotations	2
Annotation Agnostic Approaches	4
Investigating Putative Exons and Splice Junctions	4
Characterization of Novel Isoforms in Brain Tissue	6
Potential Hallmarks of Functionality	8
Overall Expression	8
Detection Across Samples	10
Tissue Specificity	11
MATERIALS AND METHODS	14
Pipeline Workflow	14
Compilation of Query Regions and GTEx Tissue Sample Tissue Sample Groups	15
Selection of GTEx Brain Samples	16
Defining Genomic Regions for Snaptron Queries	18
Lift Over from hg19 to hg38 Genome Builds	19
Compilation of Canonical Junction Query Regions	22
Compilation of Flanking Exon-Exon Junction Query Regions	24
Generation of Snaptron Queries	25
Confirmation of Canonical Junction Presence in GTEx Brain Tissues	26
Validation of Putative Exons Presence in GTEx Brain Tissues	27
Characterization of Canonical and Flanking Junctions	28
Overall Expression and Prevalence Across Samples	29
Analysis of Canonical Junctions	30
Analysis of Flanking Junctions	32
Thresholds Used for Isoform Classification	37
Tissue specificity	37
RESULTS	40
Detection of Canonical Junctions	40
Validation of Putative Exons in Public RNA-seq Data	40

Characterization of Canonical and Putative Exon-Exon Junctions	43
Overall Expression and Detection Across Samples	43
Tissue Specificity	47
DISCUSSION	55
WORKS CITED	59
APPENDICES	62
Appendix 1. Novel Putative Exons with a Classification of Common Major	62
Appendix 2. Novel Putative Exons which are Tissue Specific (With Isoform Classification)	62
Appendix 3. Annotated Putative Exons which are Tissue Specific (With Isoform Classification)	65
Appendix 4. Canonical Junctions which are Tissue Specific (Without Isoform Classification)	67
Appendix 5. Canonical Junctions which are Tissue Specific (With Isoform Classification)	67
Appendix 6. Annotated Putative Exons which are Tissue Specific (Without Isoform Classification)	69
Appendix 7. Novel Putative Exons which are Tissue Specific (Without Isoform Classification)	71

LIST OF TABLES

Tables		Page
1	Summary Statistics for Canonical Junction Read Coverage.	43
2	Summary Statistics for Upstream Junction Coverage.	44
3	Summary Statistics for Downstream Junction Coverage.	44
4	Classification of Canonical Junctions.	46
5	Classification of Putative Exons.	46

LIST OF FIGURES

Figures		Page
1	Summary of the pipeline workflow which characterizes canonical and flanking splice junctions.	15
2	Summary of the number of GTEx samples which represent each tissue type.	18
3	Illustration of the exons and splice junctions on a positive and negative strand of DNA.	23
4	Illustration of putative exon-exon junction regions on both a positive and negative strand of DNA.	24
5	A scatterplot showing the distribution of canonical and flanking junctions across sample count and read coverage for anterior cingulate cortex samples.	36
6	Distribution of canonical junctions and validated putative exons across all GTEx brain tissue groups.	41
7	Distribution of validated putative exons across all brain tissue groups, categorized by annotation status.	42
8	Distribution of tissue specific and non-specific canonical junctions across the 13 brain tissues from GTEx.	48
9	Distribution of tissue specific canonical junctions across the brain tissue samples from GTEx.	49
10	Distribution of annotated tissue specific and non-specific putative exons across the 13 brain tissues from GTEx.	50
11	Distribution of unannotated tissue specific and non-specific putative exons across the 13 brain tissues from GTEx.	51
12	Distribution of tissue specific annotated putative exons across the brain tissue samples from GTEx.	52
13	Distribution of tissue specific putative exons across the brain tissue samples from GTEx.	52

ABSTRACT

The aim of this study is to infer the potential for biological function of putative exons which arise from repetitive element (REL) exonization events in human brain tissues. The data pipeline in this project utilizes existing annotation-agnostic methods and public RNA-seq data consortiums to determine the exact boundaries and levels of expression of these putative exons which may be unrepresented in current gene annotations. Biological function of putative exons is assessed by examining splicing events which could give rise to novel transcripts that contain these putative exons. These splicing events are assessed in regard to overall expression, prevalence across samples, and tissue specificity.

Documentation and source code is available at:

<https://github.com/biancabifx/PutativeExonFunctionPipeline>.

INTRODUCTION

Novel Splice Isoforms and Current Gene Annotations

Gene annotations offer insightful information about the structure and function of exons, introns, transposons, promoters, and other genomic elements. Structural annotation can be used to determine the physical location of genomic elements within the genome, in addition to identifying open reading frames which give rise to proteins. Functional annotation describes biological roles of genetic elements, including any interactions or pathways which they take part in. Projects such as GENCODE (Frankish et al., 2021), Ensembl (Cunningham et al., 2022), and RefSeq (O’Leary et al., 2016) provide researchers with open source access to their databases which contain vast quantities of curated gene annotations.

A practical use for gene annotations in bioinformatics is demonstrated through RNA sequencing (RNA-seq). RNA-seq is used to profile the transcriptome of organisms by detecting and quantifying the expression of different RNA transcripts within biological samples. One step in RNA-seq is to align sequence reads, which are inferred arrangements of base pairs, to reference genomes or transcriptomes. Reads are then classified into the following categories which pertain to structural annotation: exonic reads, junction reads, and sometimes poly(A) end-reads. Gene expression is determined by the total number of reads which overlap exons, introns, and junctions during the read alignment process.

Until recently, the human reference genome did not capture heterochromatic regions, which compose 8% of the genome. These gaps were brought on by limitations in the methods used for constructing the reference, resulting in unfinished repetitive and polymorphic regions. Advancement of sequencing technologies has led to the assembly of a complete sequence of the human genome, the T2T-CHM12 reference assembly (Nurk et al., 2022). This breakthrough, coupled with our growing understanding of the human genome, has fueled the discovery of novel transcript isoforms that are not represented in current genome annotations.

Some studies suggest that unannotated isoforms tend to have low levels of expression. In contrast, annotated transcript isoforms tend to be well expressed and are found across many samples. One study which investigated the extent of unannotated RNA-seq splice junctions across 21,504 Illumina-sequenced human samples observed that “18.6% of junctions that appeared in 1000 or more samples did not appear in annotation” (Nellore et al., 2016). This suggests that current annotations are incomplete, which creates potential implications for bioinformatics methods and studies which utilize gene annotations.

To create a more comprehensive expression profile for genes with lower levels of expression, information from both annotated and unannotated events should be taken into consideration. To accomplish this, researchers may consider incorporating annotation agnostic tools into their data analysis pipelines.

Annotation Agnostic Approaches

Completely characterizing complex transcript structures with short read sequencing technologies is particularly difficult to accomplish. This is why most statistical methods used in RNA-seq analysis depend on mapping reads to existing annotations in reference genomes when defining genomic regions of interest. Although these methods are capable of producing robust results, it must be kept in mind that there are limitations which arise when incomplete or incorrect annotations are utilized. Relying solely on annotations may create discrepancies in downstream modeling of read numbers (Leonardo Collado-Torres, 2015).

The sensitivity of RNA-Seq, and its capability of producing single-base resolution of transcripts, results in data that is ideal for annotation-agnostic tools and methods. Expressed regions can be identified by observing the read counts directly, which can lead to the discovery of novel transcribed regions and splice isoforms.

Investigating Putative Exons and Splice Junctions

The detection of novel exons using RNA-seq has brought awareness to unannotated transcription within humans and other organisms. For the purpose of this study, the terms “novel exon” and “putative exon” refer to transcribed regions that map to areas which are not annotated or fall within regions that are annotated as intronic or intergenic. Studies are suggesting that genomic regions which have been coined “junk DNA” are giving rise to new, potentially functional, transcript isoforms (Darby et al., 2016). To detect and characterize these novel exons and transcripts, more annotation-agnostic tools and

methods are becoming available for researchers to gain new insight into the transcriptomes of organisms.

For this project, the selection of annotation agnostic tools and methods for the data pipeline are justified by methods found in scientific literature. Recent literature shows that the examination of splicing patterns has become a useful tool for studying the transcriptomes of organisms. In regard to measuring transcriptomic diversity in RNA-seq datasets, studies have resorted to comparing exon-exon junctions detected in RNA-seq samples to exon-exon junctions in genome assemblies (Nellore et al., 2016). In lieu of using full transcripts for comparison, short RNA-seq reads that cross an exon-exon junction, known as intersecting junction reads, are deemed as more reliable (Steijger et al., 2013). This is because accurately calling unannotated transcript isoforms from short-read RNA-seq data is particularly difficult to accomplish.

Justification for investigating splicing events surrounding putative exons stems from another recent study which sought to find unannotated protein-coding transcripts. The methods combined unannotated transcription data with junction read data to link these regions to specific genes of interest (Zhang et al., 2020). Boundaries of putative exons were inferred by using intersecting junction reads. In addition, putative exons were characterized as potentially protein-coding if there were reads which indicated the presence of flanking junctions on either side of a putative exon. If the flanking junctions connected the putative exons to known protein coding exons, they were considered protein-coding.

This project aims to study the potential functional significance of putative exons by analyzing splice junction data from publicly available RNA-seq databases. To accomplish this, the pipeline utilizes Snaptron (Wilks et al., 2018) to comb through sequencing data. Snaptron is an open source search engine which queries over 146 million exon-exon splice junctions from over 70,000 RNA-seq samples. This tool was selected for its ability to query large repositories of RNA-seq data, and for the fact that queries can be tailored to find both canonical and flanking junctions. The term “canonical junctions” in this paper refers to canonical splice sites which are present in the most abundant splice isoforms. The mention of “flanking junctions” in this paper refers to junctions which flank the putative exons, connecting them to annotated exons that are upstream and downstream from the putative exon loci.

In addition to the ease of combing through large datasets, Snaptron also provides useful metadata pertaining to the junctions which satisfy the query parameters. This information includes splice junction annotation status, and summary statistics of read coverage which are examined later in the pipeline.

Characterization of Novel Isoforms in Brain Tissue

Studies show that the brain has an extensive alternative splicing program compared to other organs and tissues (Melé et al., 2015). These splicing events give rise to a large proportion of novel transcripts. The putative exon regions which are examined in this

project were compiled from a previous study by Darby, et al. (2016). Their aim was to investigate the expression of REL in the brain. They examined directional RNA-seq data from 59 orbitofrontal cortex (OFC) samples and found that 30,000 RE were consistently expressed across subjects. The results indicated that 14,055 RE were present in annotated exons, 7,288 in annotated introns, and 10,405 in intergenic regions.

Further analysis suggested that exonization of parts of introns and intergenic regions containing RE yields thousands of novel mRNA transcripts. At the time of their discovery, these transcripts were not represented in gene annotations. In addition to the RE that were spliced into coding regions of gene transcripts, a lower frequency of RE splicing also occurred in non-coding transcripts. Non-coding RNAs are not translated into proteins and are involved in cellular processes such gene expression regulation at the transcriptional and post-transcriptional level. This study aims to investigate the putative exons from the Darby, et al (2016) study that are identified as being intronic or intergenic, and are found to be spliced into coding regions of gene transcripts.

The data produced by their study, which will be examined in this pipeline, provides: genomic regions and coverage of REL-derived putative exons found in the OFC, genomic regions and coverage of downstream and upstream exons, identifiers of genes which the transcripts were mapped to, base pair sequence of putative exons, and information pertaining to putative exon reading frames. Due to the fact that these putative exons were found in the OFC, this study aims to find evidence of splice patterns resulting in the transcription of these putative exons in brain tissues. To accomplish this, the

pipeline utilizes Snaptron to search for flanking junctions within RNA-seq data from Genotype-Tissue Expression project (GTEx) brain tissue samples. GTEx hosts sequencing data collected from 54 non-diseased tissues across approximately 1000 individuals. Due to the fact that these tissues are non-diseased, using GTEx sequencing data mitigates the impact that disease could have on the expression of the splice junctions examined in this project.

Potential Hallmarks of Functionality

In addition to finding evidence for the incorporation of intronic and intergenic putative exons into mRNA transcripts, another aim of this study is to investigate the potential biological function of these putative exons. This project infers the potential biological function of the putative exons through the application of potential hallmarks of functionality. While some of the potential hallmarks of functionality examined in this study have been greatly discussed in literature, there is little consensus as to the utility of each hallmark. The data pipeline in this study is designed to evaluate putative exons based on their overall expression, detection across samples, and tissue specificity.

Overall Expression

Overall expression indicates the abundance of a transcript within a given sample. When determining the overall expression of gene transcripts, utilizing RNA-seq data is particularly advantageous. High-throughput sequencing data produced by RNA-seq paints a picture of the entire transcriptome in a quantitative manner. The number of

sequencing reads which align to genomic regions is indicative of the number of transcript fragments present within the sample.

To conserve a cell's energy, the synthesis of RNA is done in a highly regulated process. Synthesis of RNA does come at a cost; therefore, cellular energy is devoted towards the transcription of functional RNAs as opposed to non-functional RNAs that would prove to be an energetic burden (Palazzo & Lee, 2015). It is hypothesized that higher levels of expression may be a characteristic of gene isoforms which have inherent function. In the remainder of this paper, protein-coding gene isoforms which exhibit high levels of expression shall be referred to as "major isoforms". To grant the status of "functional" to only major isoforms would be inappropriate, given the breadth of expression levels in the transcriptome. Protein-coding isoforms which are comparatively less expressed than major isoforms could still be considered functional. In this paper, these isoforms shall be referred to as "minor isoforms."

It must be noted that transcripts cannot be deemed functional for the sole reason that they are expressed. Although inappropriate transcription of junk DNA is limited by heterochromatin structures, studies have found that heterochromatin can be loosened, allowing for RNA polymerases to transcribe these regions (Palazzo & Lee, 2015) (Nurk et al., 2022). Therefore, additional hallmarks of functionality that go beyond expression levels must be considered prior to classifying a transcript as functional.

Detection Across Samples

Despite our wide range of phenotypes, human beings are physiologically and genetically comparable to one another. With the aid of RNA sequencing technologies, researchers are now aware that humans share commonalities at the transcriptomic level as well. Studies have shown evidence of conserved transcriptional regulatory networks, splicing patterns, and shared transcript isoforms (Melé et al., 2015). In the remainder of this paper, the term “common isoform” is used to refer to these widely prevalent isoforms.

In addition to common isoforms, there may be instances where rare transcript isoforms are observed within a small subset of individuals. Organisms which produce rare isoforms can have exceptional phenotypes. Rare isoforms can be the result of genomic variants and disease mutations that result in alternative splicing patterns. The transcript isoforms that arise from these events can be non-functional or have an alternative function compared to those which are ubiquitously expressed across a population.

In regard to the biological function of the putative exons examined in this study, this study aims to investigate how prevalent the junctions which flank these putative exons are across different individuals, and to use this metric as a way to characterize their functional significance. If their flanking junctions are detected in a majority of individuals, it could indicate that the putative exon is being transcribed into a conserved functional isoform, or common isoform. If the flanking junctions are detected in a minority of individuals, the putative exons could be transcribed into rare isoforms with altered functionality.

With this being said, it must be stated that distinguishing functional isoforms from those that are non-functional cannot be done by solely measuring prevalence across samples. For example, an isoform which is found to be in a small subset of individuals could very well be the result of a nonsense variant. In contrast, it could also be the case that an isoform which appears abundant is actually the result of a neutral non-coding RNA locus that is close to a genomic region undergoing positive selection (Palazzo & Lee, 2015). Therefore, prevalence across samples alone is not conclusive evidence for the presence or absence of function and should be viewed in conjunction with other characteristics.

Tissue Specificity

A singular organism is capable of producing a variety of tissues which have distinct functions. This functional diversity can largely be attributed to differential expression of genes and alternative splicing. Studies have shown strong evidence that gene expression varies between tissues, and that each tissue has their own unique transcriptome (Melé et al., 2015). It has been estimated that thousands of differentially expressed genes contribute to the diverse transcriptomes of tissues. The post-transcriptional process of alternative splicing adds an additional layer of diversity by allowing a single gene to produce different mRNA isoforms that may code for a diverse set of protein isoforms. Although the mRNA isoforms of the same gene may have similar sequences, they can have significantly different cellular function.

Many mRNA isoforms can be considered tissue and condition specific. Their formation is attributed to various factors such as environmental stress or the activation of biological pathways (immune responses, etc.). In an effort to understand the functional implications of mRNA isoforms, some studies have analyzed the presence or absence of tissue specificity, suggesting that tissue specificity is a proxy for biological function (Kandoi & Dickerson, 2019). Cross-tissue comparison of RNA-seq data may reveal that certain isoforms are more abundant in some tissues than others. If mRNA isoforms are not exclusively present in a singular tissue, it may be that they are found across different tissues at varying transcription levels.

This study aims to characterize the potential for biological function of putative exons by determining whether they are tissue specific or not. To accomplish this, the pipeline examines data from 13 different brain tissue regions present in the GTEx RNA-seq data repository. The brain tissues represented in GTEx include: anterior cingulate cortex, amygdala, caudate, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens, putamen, spinal cord, and substantia nigra. The pipeline in this study utilizes Snaptron to search for flanking junctions in each of these brain tissues, and then quantifies the number of times a splicing event occurs which could incorporate a putative exon into a final transcript. For this study, a splicing event consists of a pair of flanking junctions which span the introns on either side of a putative exon.

If a putative exon is found to only be spliced into transcripts for one specific tissue, it is presumed that the putative exon is present in a tissue-specific transcript isoform.

Therefore, it may be inferred that the function of the putative exon is unique to that tissue. On the other hand, if a putative exon is spliced into transcripts across multiple tissue types, it could be that the putative exon is present in a transcript that is involved in cellular functions which are not specific to a certain tissue type.

With this being said, it must be noted that there is no clear consensus as to whether tissue specificity can serve as a hallmark of functionality, according to existing literature. In order to address whether tissue specificity can be used as a hallmark of functionality, this study examines canonical isoforms with known function in the brain and measures how frequently they are tissue specific. This data is intended to serve as a comparison point for evaluating the putative exons examined in this study.

MATERIALS AND METHODS

Pipeline Workflow

This project presents a pipeline which: (1) Compiles a list of canonical junctions overlapping putative exons in public RNA-seq data; (2) Searches for canonical junctions within public RNA-seq data using Snaptron; (3) Validates existence of putative exons in public RNA-seq data by searching for exon-exon splice junctions which flank them; (4) Infers the potential for biological function of both canonical and putative exon splice junctions by utilizing potential hallmarks of functionality. For a first pass analysis, the pipeline is applied to canonical junctions which overlap the putative exon regions detected in human OFC samples by Darby et al., 2016. These junctions are assessed for overall expression, detection across samples, and tissue specificity. After analyzing the canonical junctions, the flanking junctions of the putative exons are evaluated in the same manner. Lastly, the results between canonical and flanking junctions are compared to determine any negative or positive correlations in coverage, sample count, and tissue specificity. Figure 1 provides an illustration of the pipeline workflow.

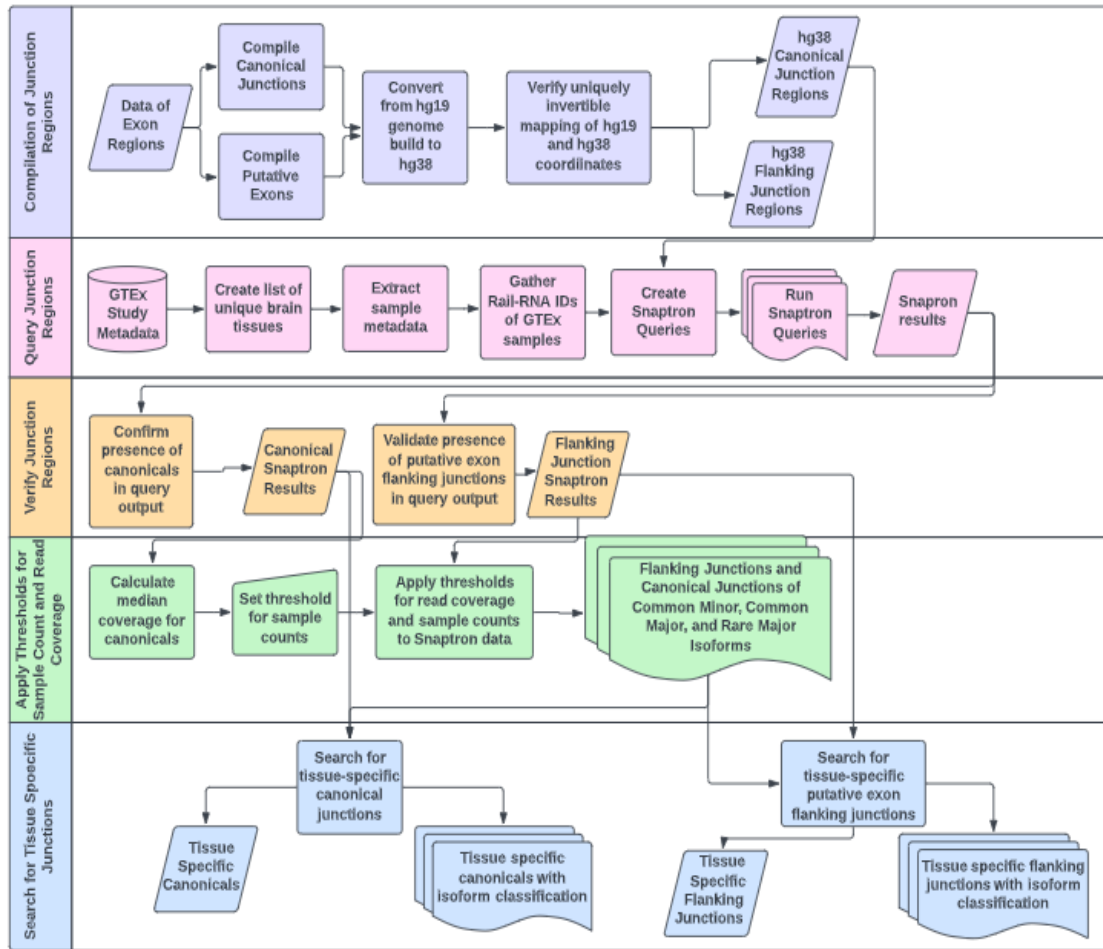


Figure 1: Summary of the pipeline workflow which characterizes canonical and flanking splice junctions. This diagram represents the flow of the original data through all transformations and analyses used in this pipeline.

Compilation of Query Regions and GTEX Tissue Sample Tissue Sample Groups

Snaptron is a search engine designed to filter through repositories of RNA-seq data to find splice junctions which match search criteria defined by the user. For this study, we will be searching for two sets of junctions, canonical junctions and flanking junctions, within GTEX brain tissue samples. To accomplish this, the Snaptron queries used in the pipeline require a list of unique rail-RNA IDs which represent the samples we will be

searching against. In addition to these IDs, the queries require a list of genomic regions of interest. These genomic regions specify where the putative exons' flanking junctions are located, as well as the canonical junctions which span the intron between the exons directly upstream and downstream from the putative exon.

Selection of GTEx Brain Samples

Snaptron provides lists of samples that are uniquely identified by Rail IDs, sample run IDs, and the tissues they were taken from. Rail IDs are assigned to samples which were analyzed in the Snaptron project using Rail-RNA, an intron-aware RNA-seq analysis for splicing and coverage. The Rail IDs are mapped to the run IDs, which are unique identifiers that represent run accessions for sequencing data. Many of the samples in Snaptron are from human brain tissues, however, the specific brain regions are not listed in the sample data. This study aims to compare coverage, tissue specificity, and prevalence of splice junctions across different brain regions. Therefore, the Snaptron queries need to be tailored to search for splice junctions in specific brain regions, rather than the brain as a whole.

To accomplish this, the start of the pipeline utilizes the Bioconductor package 'recount' (Collado-Torres et al., 2017). This package allows users to access data for over 70,000 publicly available RNA-seq samples, which were compiled and made available through the recount2 resource. The samples are grouped by projects which originate from SRA, GTEx, and the Cancer Genome Atlas (TCGA). For this project, 'recount' is utilized to access sample metadata from GTEx.

First, ‘recount’ is used to determine the number of unique brain tissues available through GTEx. In total, 13 different brain tissues are represented in the GTEx consortium, with a minimum sample size of 71 representing each tissue (Figure 2). After the unique brain tissues are listed, ‘recount’ is used to filter the GTEx sample metadata by tissue type. Then the run IDs of all samples representing each brain tissue are stored in a data table. These IDs are necessary for pairing the GTEx brain tissue regions to the brain samples present in Snaptron.

Once the run IDs from GTEx are gathered, they are compared to the run IDs available through Snaptron. The Snaptron run IDs are imported from a file “samples.rid2run_acc.tsv” that is available through the project’s data repository (<http://snaptron.cs.jhu.edu/data/gtex>). By joining the GTEx and Snaptron Rail ID dataframes based on shared run IDs, the Snaptron samples are paired with their respective GTEx brain tissue regions. This information is then stored in a series of 13 files, each containing the Rail IDs of samples corresponding to each of the unique brain tissue. These files are referenced at a later point in the pipeline when the Snaptron queries are compiled.

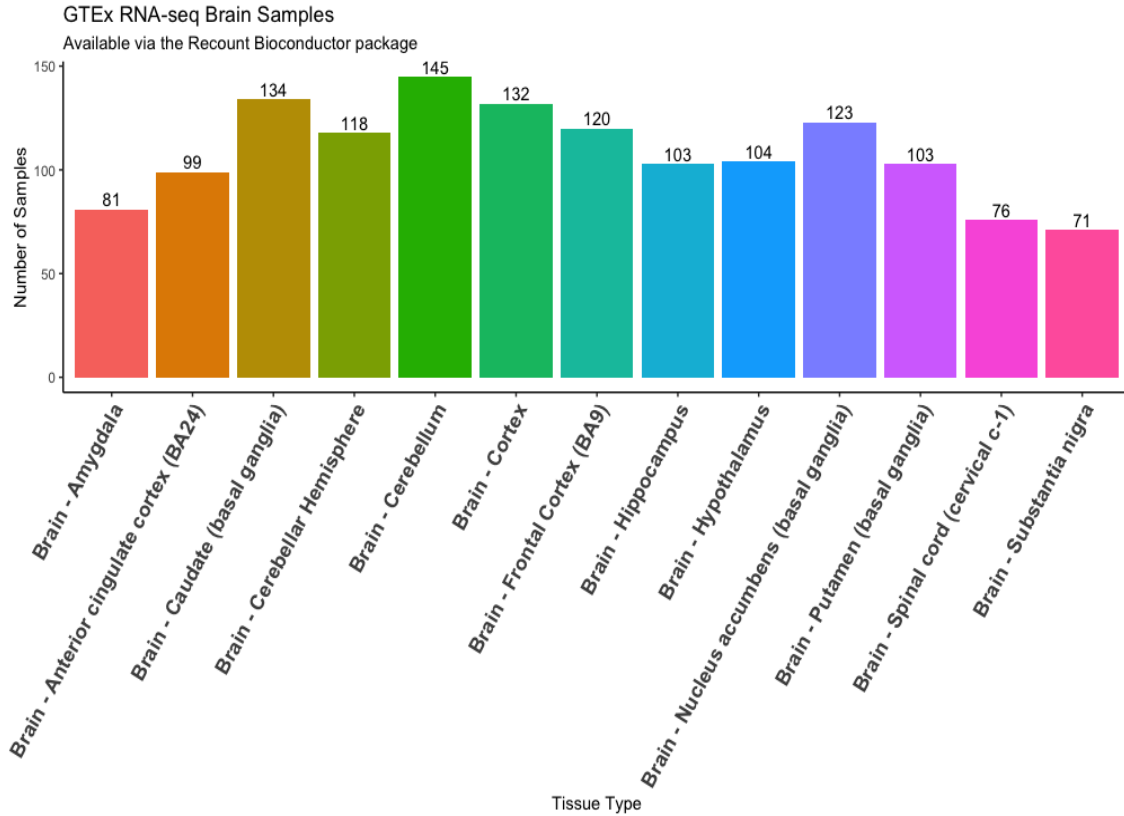


Figure 2: Summary of the number of GTEx samples which represent each tissue type. Snaptron is used to search for canonical and flanking junctions in the sequencing data produced by each of these samples.

Defining Genomic Regions for Snaptron Queries

The putative exons analyzed in this pipeline were identified in the Darby et al., 2016 study. The data are presented in a table consisting of 3,906 sets of genomic regions.

These regions represent locations where RE are spliced into coding regions of gene transcripts. For every putative exon, the coordinates for the exons directly upstream and downstream from the putative exon are also provided. In addition to these coordinates, the chromosome number, strand information, and gene identifier are also available in the data table. With this information, the pipeline aims to define the coordinates for the canonical junctions which span the intron between the exons upstream and downstream

from the putative exons as well as the junctions which flank the putative exons. These junction coordinates are used to generate the Snaptron queries at a later point in the pipeline.

Lift Over from hg19 to hg38 Genome Builds

To begin the process of compiling the junction coordinates for the Snaptron queries, the genomic regions provided by the Darby et al., 2016 study are converted from the hg19 genome build to the hg38 build. Genome builds represent different versions of the human genome reference that have been published since the original draft of the Human Genome Project publication. The Snaptron compilation for GTEx uses the hg38 genome build, whereas the Darby et al., 2016 study uses the hg19 build. In order to query the hg19 genomic coordinates against the GTEx data using Snaptron, they must be converted to the hg38 build. To accomplish this, the coordinates need to be loaded into data containers that facilitate the conversion of genomic coordinates between two genome builds.

The GenomicRanges package (Lawrence et al., 2013) is utilized for the first step of converting the regions. The tools in this package are used to take the genomic regions from the Darby et al., 2016 study and store them in GRanges objects. These objects are data containers which store genomic coordinates, chromosome numbers, strand information, and metadata pertaining to the genomic regions. Prior to the creation of the GRanges objects, each row from the original data table is given a unique identifier. This step ensures that the original putative, upstream, and downstream exon combinations can be reassembled once the conversion is complete.

GenomicRanges is used to create five GRanges objects containing the following information: putative exon coordinates, upstream exon coordinates on the positive strand, downstream exon coordinates on the positive strand, upstream exon coordinates on the negative strand, and downstream exon coordinates on the negative strand. All GRanges objects are passed in metadata containing the exon's original hg19 coordinates, and the unique identifiers assigned to them in the original data table. This metadata is necessary for tracking the exons as they are converted from one genome build to another. Once the GRanges objects are compiled, the data is ready to be converted from the hg19 build to hg38.

The rtracklayer package is used to lift the genomic coordinates from one build to the other. The tools from the package which are used are 'import.chain' and 'liftOver'. The 'import.chain' tool is used to import the 'hg19ToHg38.over.chain' chain file, which can be downloaded from a University of California, Santa Cruz (UCSC) repository (<https://hgdownload.soe.ucsc.edu/gbdb/hg19/liftOver/>). A chain file contains information about genome regions and is used to translate features from one genome build to another. This chain file, in addition to the GRanges objects, are passed in as input to the 'liftOver' tool, which executes the conversion from hg19 to hg38. The output from the tool is a new set of five GRanges objects containing the hg38 coordinates of the exons, and the metadata from the original GRanges objects.

The tools from the rtracklayer package are used a second time to confirm the coordinates have uniquely invertible mappings between the hg19 and hg38 builds. When lifting genomic coordinates from one genome build to another, there is a possibility that the original regions map to multiple regions in the new build. This can occur when coordinates overlap segmental duplications or contiguous sequences that differ between hg19 and hg38. To confirm that the coordinates are uniquely invertible between builds, the rtracklayer package is used to lift the hg38 coordinates back to hg19 to see if the original hg19 regions from the Darby et al. (2016) study are returned.

For the confirmation lift from hg38 back to hg19, the 'import.chain' function is used to import the 'hg38ToHg19.over.chain' chain file, which can be downloaded from a UCSC repository (<http://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/>). The input for the 'liftOver' tool consists of this chain file and the five GRanges objects containing the hg38 coordinates for the putative exons and upstream/downstream exons. The output of the tool is a new set of five GRanges objects containing the hg19 coordinates. Once the hg19 and hg38 GRanges objects are compiled, they are compared to see which of the original hg19 coordinates are returned. The hg38 coordinates of exons which are confirmed to be unilaterally mapped are passed onto the next step in the pipeline, which compiles the coordinates of the canonical and flanking junctions from the exon coordinates.

To define the junction regions, the GRanges objects containing the hg38 coordinates for the putative, upstream, and downstream exons are converted into data frames. The data

frames are then joined based on the unique identifiers assigned to the exons prior to their separation into different GRanges objects. Joining the data frames produces a single data frame which has a row for each combination of putative, upstream, and downstream exons.

Compilation of Canonical Junction Query Regions

For this project, the Snaptron queries are designed to search for splice junctions within GTEx tissue data. In preparation for creating the queries, a list of genomic coordinates for each canonical junction is compiled. Canonical junctions, or canonical splice sites, are the most common splicing events that overlay the majority of introns. For the purpose of this study, the data pipeline prioritizes canonical junctions which overlap putative exons of interest and share their boundaries with two flanking exons that are on either side of the putative exon (Figure 3).

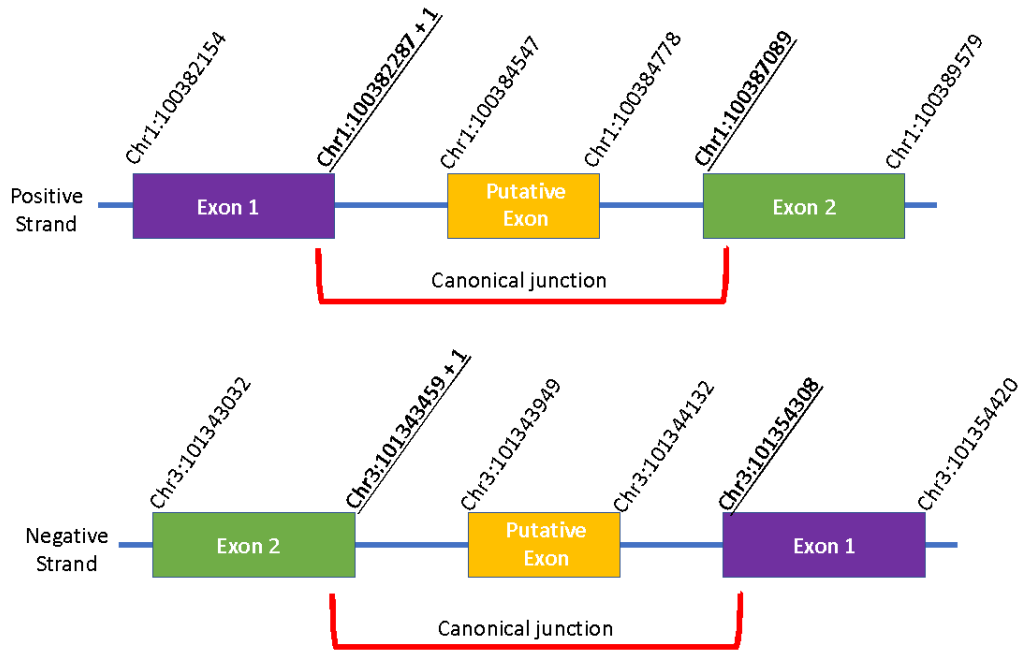


Figure 3: Illustration of the exons and splice junctions on a positive and negative strand of DNA. The canonical junctions are represented by the red brackets, which span the intron between the exons upstream and downstream of the putative exon.

The coordinates of the upstream and downstream exons are used to define the genomic coordinates of the canonical junctions. The starting position is the end of the upstream exon plus one base pair, and the end coordinate is the start of the downstream exon. A base pair is added to the starting coordinate for every junction to account for the 1-based coordinate system which is utilized in Snaptron. The coordinates for the canonical junctions are stored in a table which is referenced during the creation of the Snaptron queries.

Compilation of Flanking Exon-Exon Junction Query Regions

In addition to the canonical junctions, this study aims to investigate the junctions flanking the putative exons, which are referred to as “flanking junctions”. Flanking junctions are exon-exon junctions which span the introns between the putative exons and the exons that are directly upstream and downstream of the putative. Flanking junction pairs are composed of both an upstream and downstream junction (Figure 4).

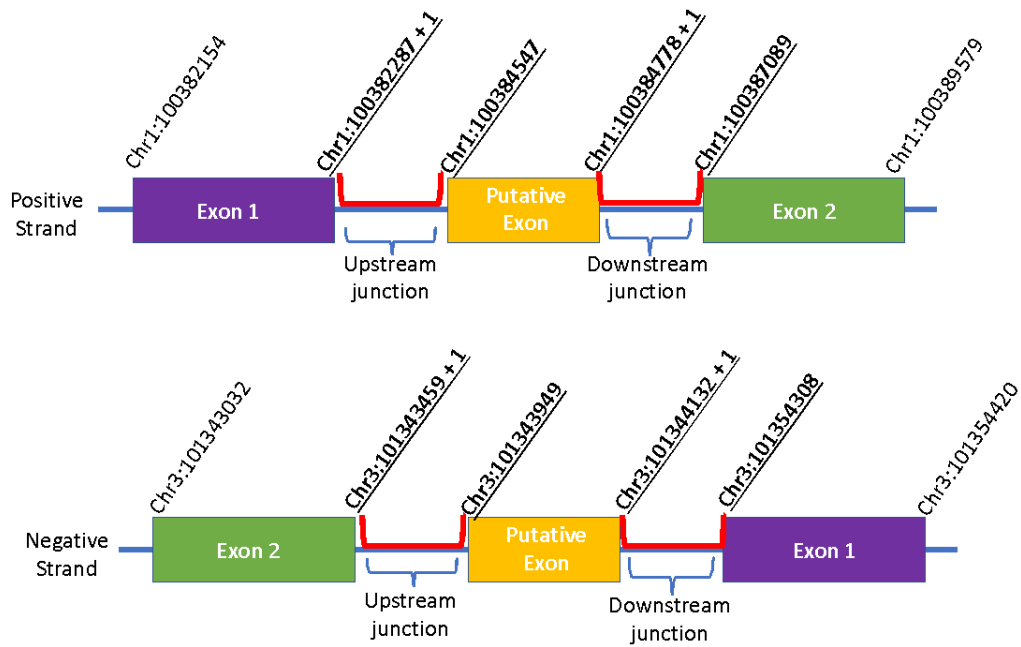


Figure 4: Illustration of putative exon-exon junction regions on both a positive and negative strand of DNA. The exon-exon junctions are represented by the red brackets, which span the introns between the putative exon and the upstream and downstream exons.

For upstream junctions, the start coordinate is the end of the upstream exon plus one base pair and the end coordinate is the start of the putative exon. For downstream junctions, the start coordinate is the end of the putative exon plus one base pair and the end

coordinate is the start of the downstream exon. A base pair is added to the starting coordinate for every junction to account for the 1-based coordinate system which is utilized in Snaptron. The coordinates for the flanking junctions are stored in a table which is referenced at a later point in the pipeline.

Generation of Snaptron Queries

For this project, Snaptron is used to query junctions that either match or fall within the bounds of the canonical junction coordinates defined earlier in the pipeline. It is anticipated that the flanking junctions which fall within the canonical junctions will be returned in the query results.

To create the Snaptron queries, the pipeline utilizes a script which iterates over the data frame containing the canonical junction regions, and the tables which contain the Rail-RNA IDs of the 13 GTEx brain tissue sample groups. Each canonical junction is entered into 13 queries, with each query searching for a canonical junction within a specific brain tissue. Each query is tailored to have the following parameters: the genomic coordinates of a canonical junction, a parameter called “contains” which tells the search engine to return junctions whose start and end coordinates fall within the boundaries of the query region, the database that is being queried (GTEx), a list of sample Rail IDs specific to one tissue type, and the strand information.

Once the Snaptron queries are written, the script stores the queries in a series of 13 shell scripts, one for each brain tissue. To execute the queries, the shell scripts are run using

the terminal. Each query produces a table which contains junctions from the GTEX data that match the search criteria. Each row of the table corresponds to a unique junction which is accompanied by descriptive metadata. This metadata includes annotation status, summary statistics for read coverage, and the total number of samples which have one or more reads covering the junction. These variables are needed at a later point in the pipeline, when the hallmarks of functionality are applied to the canonical and flanking junctions.

Confirmation of Canonical Junction Presence in GTEX Brain Tissues

Following the execution of the Snaptron queries, the output is filtered to verify which of the canonical and flanking junctions defined earlier in the pipeline are present in the GTEX brain tissues. To verify the presence of the canonicals, a series of scripts combine the results of all Snaptron queries specific to one brain region into a single GRanges object. The script also imports the data frame containing all the canonical junction coordinates. This data frame is used to create another GRanges object. Both GRanges objects are then compared to see which Snaptron junctions match the genomic coordinates of the canonical junctions.

The ‘findOverlaps’ tool in the IRanges package (Lawrence et al., 2013) is used to find junctions in the Snaptron query output that are exact matches to the canonical and flanking junctions defined at the beginning of the pipeline. The input is two GRanges objects, one being the ‘query’ object and the other is the ‘subject’ object, and two parameters which specify the type of overlap which is applied to the GRanges objects and

whether or not the regions need to be on the same strand. For this study, the type of overlap is set to 'equal' to find exact matches between the regions in the GRanges objects. In addition, the strand information has to be the same for the two regions. The output is a hits object, which specifies which regions between the two GRanges objects are exact matches and located on the same strand. The hits object is used to extract the rows from the Snaptron tables which contain junctions that share the same genomic coordinates as the canonicals. These results are stored in a series of .csv files and are imported at later points in the pipeline so that the canonical junctions can be evaluated for their read coverage, prevalence across different samples of the same tissue, and tissue specificity.

Validation of Putative Exons Presence in GTEx Brain Tissues

The second round of scripts using the tools from the 'IRanges' package verifies which of the junctions flanking the putative exons defined earlier in this pipeline are returned by the Snaptron queries. The 'findOverlaps' tool is used again, in a manner which is similar to the verification of the canonical junction regions aside from a few minor differences. The script imports a data frame containing all of the putative exons and flanking junctions mapped to the hg38 genome build. This data frame is used to create two GRanges objects for storing the upstream and downstream junctions separately. The script then creates a third GRanges object which contains the combined output of the Snaptron queries specific to one tissue.

The ‘findOverlaps’ tool is then used to compare the GRanges objects of the upstream and downstream exons to the GRanges object containing the splice junctions returned by the Snaptron queries. Like the canonicals, the desired overlap type is set to ‘equal’ to find exact matches between the GRanges objects. The strand information has to be the same as well. The resulting hits objects are used to extract the junctions from the Snaptron data which match the upstream and downstream junctions defined earlier in the pipeline. The result of this is two data frames of Snaptron data, one for the upstream junctions and one for the downstream. The upstream and downstream junction data tables are then joined into a single data frame based on the shared putative exon region which they flank. Putative exons present in this data frame are considered validated. It is presumed that putative exons which are incorporated into final transcripts have both upstream and downstream flanking junctions present in the brain tissue data.

Characterization of Canonical and Flanking Junctions

This study aims to assess the following functional hallmarks of the canonical and flanking junctions: overall expression, prevalence across samples, and tissue specificity. These hallmarks are used to infer the biological function of the transcript isoforms these junctions are found in. By doing so, it is anticipated that the functional significance of the junctions will provide some insight on the putative exons themselves. In addition to the functional hallmarks, the annotation status of the splice junctions are utilized to discern whether or not the putative exons could be spliced into a novel isoform.

Overall Expression and Prevalence Across Samples

Both canonical and flanking junctions found in the GTEx brain tissues are passed through a series of three functions which categorize splice junctions as belonging to common minor, common major, or rare major isoforms. These functions iterate through junction records in the data tables produced by the Snaptron query output and categorize them based on their overall expression and prevalence across samples from each tissue group. Expression levels are used to distinguish junctions which could be present in major isoforms from those which may be in minor isoforms. Major isoforms exhibit high levels of expression, whereas minor isoforms are expressed at lower levels. Sample counts are used to separate junctions with the potential for being present in common isoforms from those which could be incorporated into rare isoforms. Common isoforms are considered to be present in a high percentage of samples from a specific brain tissue, whereas rare isoforms are found in a small minority of samples.

As a first pass analysis, the data tables containing Snaptron data for the canonical junctions are entered into the functions first. These functions are applied to the canonicals first in the pipeline so that their results can be compared to the junctions flanking the putative exons from the Darby et al., 2016 study. The expression levels and sample counts of the canonicals are used to set thresholds which the functions apply to the flanking junctions. The analysis of canonical junctions followed by that of the flanking junctions is repeated for each set of Snaptron results for the 13 GTEx brain tissues.

Analysis of Canonical Junctions

At this point in the pipeline, three functions are used to categorize canonical junctions as belonging to common minor, common major, or rare major isoforms. The functions categorize the canonical junctions by evaluating them for two of the three functional hallmarks utilized in this study: overall expression, and detection across samples.

Existing literature describes canonical junctions, or canonical splice sites, as being the most abundant form of splice junction (Burset et al., 2000). With this characteristic in mind, these functions are used to verify whether or not the canonical junction regions defined earlier in the pipeline align with the description of what a canonical junction is. It is anticipated that a majority of these junctions exhibit high levels of expression and are found across a majority of brain tissue samples. Therefore, it is expected that a majority of the canonical junctions are classified as belonging to common major isoforms.

For the analysis of the canonical junctions, the three functions require the following values as input: 1) a Snaptron data frame which contains the canonical junctions found in samples from one brain tissue, 2) the total number of samples present in the GTEx brain tissue group (the length of the Rail-RNA ID list), 3) a percentile value which tells the function to return junctions in the Snaptron data frame which are found in 'n' number of samples, 4) the GTEx brain tissue which the junctions were detected in. To distinguish major isoforms from minor isoforms, the functions set a threshold for read coverage. Major isoforms fall above the threshold, while minor isoforms fall below it. The value used for the threshold is the median coverage of all the canonical junctions in a specific tissue. To calculate this value, the functions take the average coverage values

(‘coverage_avg’) of all junctions in the data frame and calculate the median coverage. The median coverage is used for the threshold as opposed to the average coverage to negate the effects of outliers present in the data. These outliers are canonical junctions which have thousands of reads.

To classify the canonical junctions as belonging to either rare or common isoforms, the functions evaluate how many samples the junctions are detected in for a specific tissue. The functions can be tailored to specify the cut off for how many samples a junction needs to be present in to be considered part of common or rare isoform. The cutoff is entered as a percentage, which is applied to the total number of samples present in the GTEx brain tissue group. The result of this calculation is the threshold which classifies junctions as belonging to common or rare isoforms. The functions compare the number of samples each Snaptron junction is detected in (‘samples_count’) to the threshold.

Each of the three functions applies the thresholds differently to the Snaptron data. The function which classifies canonical junctions as belonging to common major isoforms filters the Snaptron data frame and returns canonical junctions which meet the following criteria: 1) have an average coverage that is equal to or higher than the median coverage of all detected canonicals found in one tissue type, 2) Are present in ‘n’ percent of samples or more. The function which classifies canonical junctions as belonging to common minor isoforms filters the Snaptron data frame and returns canonical junctions which meet the following criteria: 1) Have an average coverage that is equal to or lower than the median coverage of all detected canonicals found in one tissue type, 2) Are

present in 'n' percent of samples or more. The function which classifies canonical junctions as belonging to rare major isoforms filters the Snaptron data frame and returns canonical junctions which meet the following criteria: 1) have an average coverage that is equal to or higher than the median coverage of all detected canonicals found in one tissue type, 2) Are present in 'n' percent of samples or less.

The output of each function is a csv file containing the canonical junctions which are classified as belonging to a common minor isoform, a common major isoform, or a rare major isoform. The data frames contain all of the original columns from the Snaptron tables, in addition to the following new columns which are created by the functions: 1) the tissue type which produced the data, 2) the isoform classification (common major, common minor, rare major), 3) the median coverage of the canonicals used for the coverage threshold, 4) the percentage used to calculate the number of samples the junctions needed to be present in, 5) the number of samples used as the threshold, 6) the total number of samples used in the Snaptron queries for a specific brain tissue, 7) the junction type (canonical).

Analysis of Flanking Junctions

Following the analysis of the canonical junctions, the flanking junctions are examined using the same three functions, but with slightly different methods. These functions are designed to infer the biological function of the putative exons by evaluating the flanking junctions. The flanking junctions are assessed for two of the three functional hallmarks examined in this study: overall expression, and detection across samples. It is anticipated

that a majority of the flanking junctions exhibit low levels of expression, since the putative exons identified in the Darby et al. 2016 study were found to be expressed at lower levels compared to the surrounding exons. Therefore, it is expected that a majority of flanking junctions, and consequently the putative exons, are classified as belonging to minor isoforms.

For the analysis of the flanking junctions, the three functions require the following values as input: 1) a Snaptron data frame which contains the canonical junctions found in one tissue, 2) a Snaptron data frame which contains the flanking junctions found in one tissue, 3) the number of samples used in the Snaptron query (the length of the Rail-RNA ID list) 4) a percentile value which tells the function to return junctions in the Snaptron data frame which are found in 'n' number of samples, 5) the GTEx brain tissue which the junctions were detected in. To distinguish major isoforms from minor isoforms, the functions set a threshold for read coverage. The value used for the threshold is the median coverage of the canonical junctions found in the same tissue as the flanking junctions. To calculate this value, the function takes the average coverage values ('coverage_avg') of all junctions in the canonical junction data frame and calculates the median coverage. The median coverage of the canonicals is used as a threshold to see how the read coverage of flanking junctions compares to that of the canonicals. Major isoforms fall above the threshold, while minor isoforms fall below it.

To classify the flanking junctions as belonging to either rare or common isoforms, the functions evaluate how many samples of a specific tissue type the junctions are detected

in. The functions allow the user to specify the cut off for how many samples the junctions need to be present in to be considered part of common or rare isoform. The cutoff is entered as a percentage, which is applied to the total number of samples entered in the Snaptron query. The result of this calculation is the threshold which classifies flanking junctions as belonging to common or rare isoforms. The function compares the number of samples each Snaptron junction is detected in ('samples_count') to the threshold.

These functions also incorporate an additional step which filters flanking junction pairs by their annotation status (annotated vs. unannotated). This value is present in the 'annotated' column of the Snaptron data frame. The function which classifies flanking junctions as belonging to common major isoforms filters the Snaptron data frame containing the flanking junctions and returns those which meet the following criteria: 1) have an average coverage that is equal to or higher than the median coverage of all detected canonicals found the same tissue type, 2) are present in 'n' percent of samples or more. The function which classifies flanking junctions as belonging to common minor isoforms filters the Snaptron data frame and returns junctions which: 1) have an average coverage that is equal to or lower than the median coverage of all detected canonicals found in the same tissue type, 2) are present in 'n' percent of samples or more. The function which classifies flanking junctions as belonging to rare major isoforms filters the Snaptron data frame and returns junctions which: 1) have an average coverage that is equal to or higher than the median coverage of all detected canonicals found in the same tissue type, 2) are present in 'n' percent of samples or less.

The output of each function is a series of csv files containing the putative exons whose flanking junctions meet the thresholds for read coverage and sample counts. A total of six csvs are created, storing putative exons and flanking junctions which are classified as belonging to novel or annotated common minor isoforms, common major isoforms, or rare major isoforms. The data frames in these files contain all of the original columns from the Snaptron tables for both the upstream and downstream flanking junctions. In addition to these columns, the following new columns which are created by the functions are also present: 1) the tissue type specified by the user, 2) the putative exon which the junctions flank, 3) the isoform classification (common major, common minor, rare major, 4) The median coverage of the canonicals used for the coverage threshold, 5) the percentage used to calculate the number of samples the junctions needed to be present in, 6) the number of samples used as the threshold, 7) the total number of samples used in the Snaptron queries for a specific brain tissue, 8) the junction type (flanking junction).

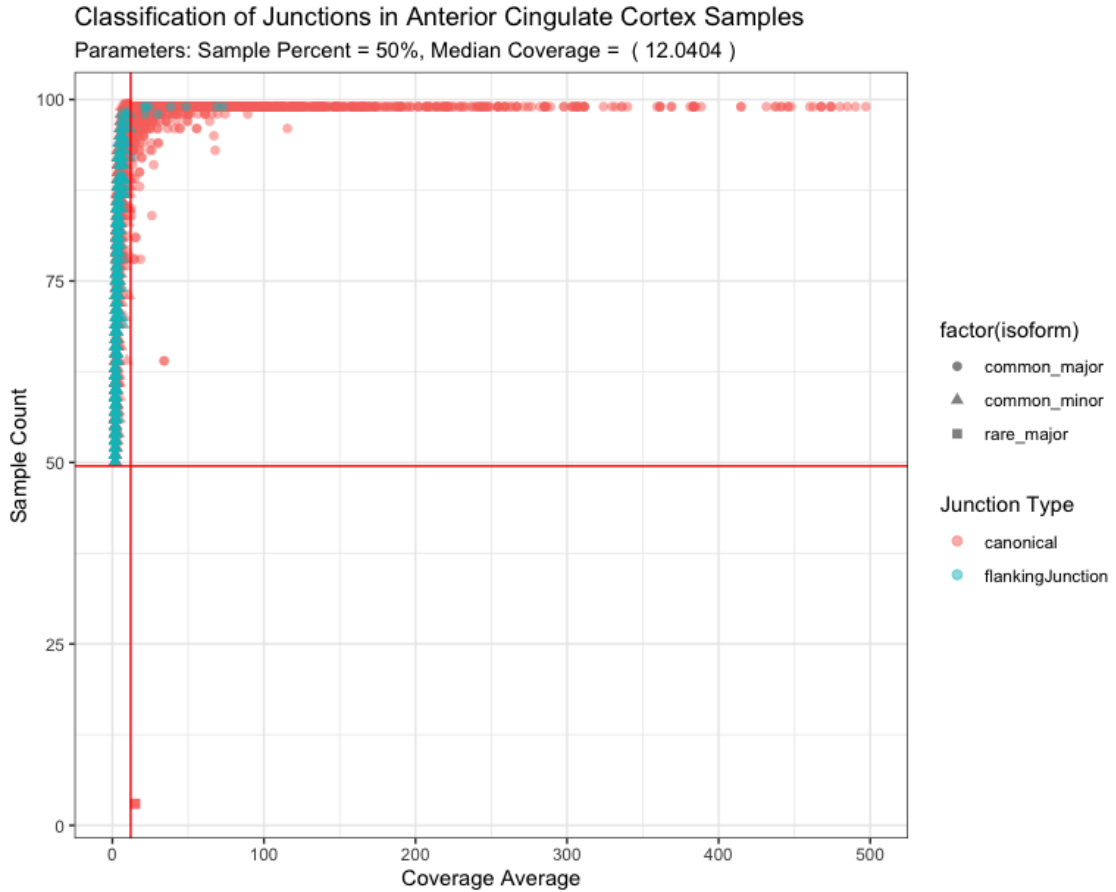


Figure 5. A scatterplot showing the distribution of canonical and flanking junctions across sample count and read coverage for anterior cingulate cortex samples. The red lines are indicative of thresholds for sample count and read coverage. The sample count is set as a percentage, and the coverage threshold is set by the median coverage for canonical junctions. The quadrants set by the thresholds represent different isoform groupings. The shape of the points indicates the isoform group assigned to the junctions. The color indicates whether the point represents a canonical or flanking junction.

Thresholds Used for Isoform Classification

For a first pass analysis, the following parameters for common minor isoform classification are used for both the canonical and flanking junctions: detection in at least 50% of samples and coverage averages less than or equal to the medium coverage calculated for all canonical junctions found in the same tissue. The following thresholds are used for common major isoform classification: detection in at least 50% of samples and coverage averages greater than or equal to the medium coverage calculated for all canonical junctions found in the same tissue. Lastly, the following parameters are used for rare major isoform classification: detection in less than 50% of samples and coverage averages greater than or equal to the medium coverage calculated for canonical junctions found in the same tissue.

Tissue specificity

Both the canonical junctions and putative exons found in the GTEx brain tissues are entered into a script which evaluates their tissue specificity, one of the three functional hallmarks assessed in this study. This script is designed to identify tissue specific canonicals and putative exons which are present in only one of the 13 selected brain tissues. It is postulated that tissue specific canonicals and putative exons could be incorporated into transcript isoforms which are exclusive to a certain brain tissue.

Four sets of data tables are processed by the script. The first two sets of data contain Snaptron results for all detected canonical junctions and validated putative exons. These tables contain canonical junctions and putative exons which have not been classified as

belonging to common major, common minor, or rare major isoforms. Data in these tables are analyzed with the intention of identifying any tissue specific canonical junctions or putative exons whose flanking junctions may have a combination of low coverage and sample counts. Splice junctions with low values for these two variables may be filtered out when sample count and read coverage thresholds are applied to the data during the evaluation of the other two hallmarks in this project. The last two data tables evaluated for tissue specificity contain canonical junctions and putative exons which have been classified as belonging to a common major, common minor, or rare major isoform. These tables are analyzed separately from the previous two tables to see if any of the canonicals or putative exons which received isoform classifications are also tissue specific.

To begin the analysis, the script imports all of the tables containing the Snaptron data for one of the previously mentioned data sets and combines them into a single table. It then compiles a list of all the unique regions present in the table. For canonical junctions, the script compiles a list of unique canonical junction regions. For tables containing validated putative exons and their corresponding flanking junction pairs, a list containing unique putative exons is assembled. The script also creates a list of all the unique brain tissues from the GTEx project. The lists and data table are then passed into a function which iterates over data table entries and the lists, and tallies how many tissues each canonical junction or putative exon is found in.

The function produces a table which contains the following data: 1) the unique genomic region (canonical junction or putative exon), 2) the presence (denoted as “1”) or absence

(denoted as “0”) of the region in each of the 13 different brain tissues, 3) the sum of how many tissues each region was found in. This table is used to identify tissue specific splice junctions and putative exons, which are only found in one tissue. In addition, the table is used to generate a series of histograms for visualizing the distribution of tissue specific canonical junctions and putative exons across the different brain tissues.

RESULTS

Detection of Canonical Junctions

A total of 3,367 unique canonical junction regions were assembled using the data provided by the Darby et al., 2016 study. Querying these regions using Snaptron indicated that 3,313 junctions present in the GTEx brain data have coordinates which match those of the canonical junction coordinates provided to the search engine. All junctions included in the output of the queries have at least one read covering them.

The data present in the Snaptron output provides evidence that the reconstruction of canonical splice sites from the Darby et al., 2016 study was successful. Canonical splice sites are the most common form of splicing and correspond with a majority of introns. In contrast, non-canonical splice junctions are rarer and are more likely to be tissue specific (Sibley et al., 2016). The Snaptron results indicate that the majority of canonical junctions defined in this study are detected across all selected brain tissues. Figure 6 suggests these junctions could be expressed ubiquitously in the brain, without corresponding to a particular region or tissue type.

Validation of Putative Exons in Public RNA-seq Data

The Darby et al., 2016 study produced 3,906 unique putative exon regions. Some of these exons occur in multiple splicing patterns, with a different set of flanking splice junctions at each occurrence. The data present a total of 4,277 unique splicing patterns. Lifting the putative exon regions from the hg19 genome build to hg38 generates a total of 3,871 unique putative exons and splicing patterns which are uniquely mapped between the two

builds. Querying the GTEx brain tissue data with Snaptron returns exact matches for the splice junctions presented in these patterns. A total of 3,723 putative exons have at least one pair of flanking splice junctions present in the GTEx brain tissue data which are overlapped by at least one read. These putative exons are considered validated, since there is evidence of at least one splice pattern that could include them in a final transcript.

The data presented in Figure 6 indicate that all brain tissues express at least two-thirds of validated putative exons. Much like the detection of canonical junctions, it is plausible that a majority of the putative exons are expressed across brain tissues. None of the tissue types appear to have a significantly higher or lower number of splicing events.

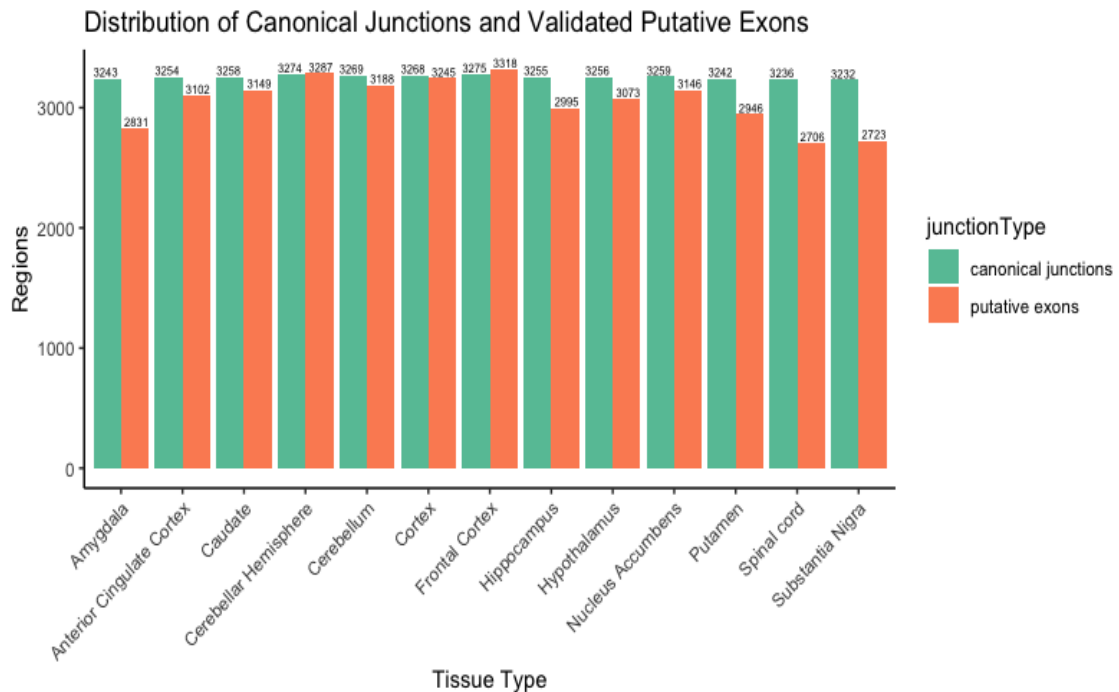


Figure 6. Distribution of canonical junctions and validated putative exons across all GTEx brain tissue groups. The data returned by the Snaptron queries provided evidence that these regions are present in the sequencing data of the GTEx samples.

Of the validated putative exons, 2,677 are classified as belonging to novel isoforms because they are flanked by at least one unannotated splice junction, indicating that they could be spliced into potentially novel transcripts. A total of 1,046 putative exons are categorized as belonging to annotated isoforms because they are flanked by annotated splice junctions. The presence of these annotated putative exons in the data suggests that some of the original putative exons from the Darby et al., 2016 study have been confirmed to exist since their discovery. Figure 7 illustrates that a majority of the validated putative exons for each tissue type are classified as novel.

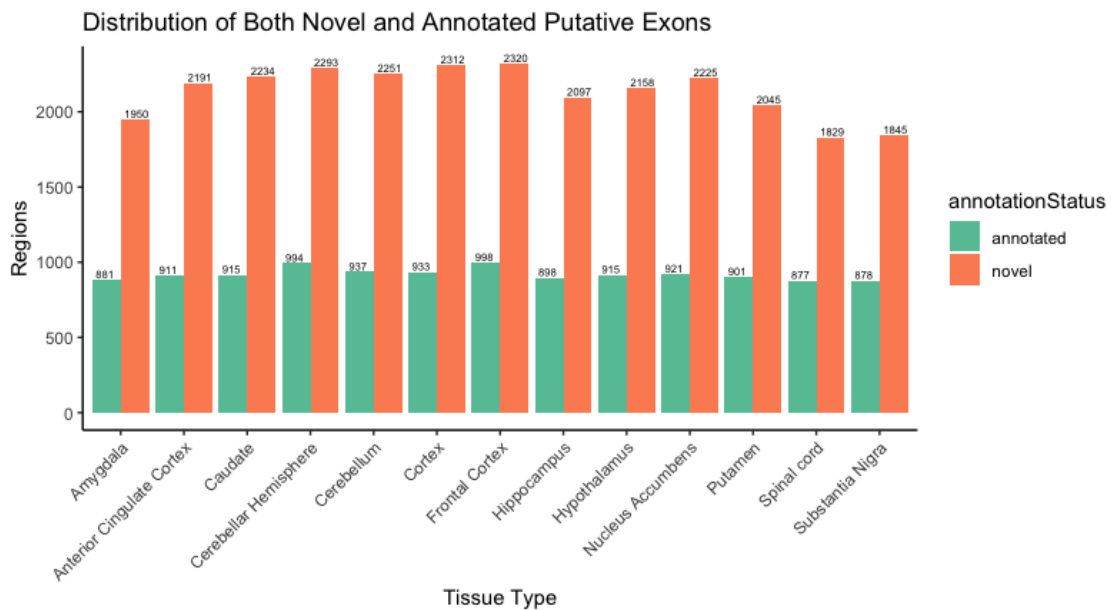


Figure 7. Distribution of validated putative exons across all brain tissue groups, categorized by annotation status. Annotated putative exons are flanked by two annotated splice junctions. Novel putative exons are flanked at least one annotated splice junction.

Characterization of Canonical and Putative Exon-Exon Junctions

Overall Expression and Detection Across Samples

Summary statistics for the Snaptron output show that canonical junctions have much higher read coverage values on average compared to junctions which flank putative exons. This is true for all tissue types, suggesting that most of the canonical junctions are present in major isoforms and are well expressed in brain tissues (Table 1). In contrast, a majority of flanking junctions appear to be expressed at much lower levels, suggesting that most of the putative exons may be incorporated into minor isoforms (Tables 2 and 3).

Table 1. Summary Statistics for Canonical Junction Read Coverage. These summary statistics correspond to the read coverage for the canonical junctions returned by the Snaptron queries. Summary statistics are presented for canonical junctions found in each of the 13 tissues.						
Tissue Type	Minimum	1 st Quartile	Median	Mean	3 rd Quartile	Maximum
Anterior Cingulate Cortex	1.00	3.868	12.040	47.147	35.758	11089.758
Amygdala	1.00	3.531	9.814	41.087	27.506	20979.852
Caudate	1.00	3.798	11.408	40.465	31.985	8538.843
Cerebellar Hemisphere	1.00	5.478	17.323	53.726	48.271	6621.237
Cerebellum	1.00	3.294	11.193	34.104	31.655	2324.586
Cortex	1.00	3.697	11.890	43.747	35.105	7349.136
Frontal Cortex	1.00	5.425	17.606	64.422	51.192	10455.192
Hippocampus	1.00	3.545	9.772	41.941	27.650	26021.117
Hypothalamus	1.00	4.342	12.067	44.614	34.729	25266.087
Nucleus Accumbens	1.00	4.039	12.065	43.103	34.217	9055.325
Putamen	1.00	3.131	9.238	38.616	26.869	17037.233
Spinal Cord	1.00	3.63	10.52	72.30	29.86	123050.58
Substantia Nigra	1.00	3.76	10.21	51.53	29.59	55190.24
All Tissues	1.00	3.89	11.75	47.44	33.94	123050.58

Table 2. Summary Statistics for Upstream Junction Coverage. These summary statistics correspond to the read coverage for the upstream flanking junctions returned by the Snaptron queries. Summary statistics are presented for upstream junctions found in each of the 13 tissues.

Tissue Type	Minimum	1 st Quartile	Median	Mean	3 rd Quartile	Maximum
Anterior Cingulate Cortex	1.000	1.056	1.333	1.803	1.667	304.727
Amygdala	1.000	1.000	1.375	1.792	1.727	185.700
Caudate	1.000	1.111	1.326	1.703	1.636	134.858
Cerebellar Hemisphere	1.000	1.250	1.600	2.404	2.211	283.042
Cerebellum	1.000	1.111	1.333	1.909	1.667	281.924
Cortex	1.000	1.111	1.316	1.771	1.625	329.603
Frontal Cortex	1.000	1.200	1.519	2.177	2.024	311.392
Hippocampus	1.000	1.000	1.333	1.755	1.682	193.039
Hypothalamus	1.000	1.111	1.360	1.774	1.721	161.962
Nucleus Accumbens	1.000	1.111	1.350	1.782	1.700	171.447
Putamen	1.000	1.000	1.294	1.671	1.600	119.796
Spinal Cord	1.000	1.000	1.400	1.864	1.795	81.658
Substantia Nigra	1.000	1.000	1.400	1.833	1.833	97.831
All Tissues	1.000	1.111	1.368	1.870	1.760	329.603

Table 3. Summary Statistics for Downstream Junction Coverage. These summary statistics correspond to the read coverage for the downstream flanking junctions returned by the Snaptron queries. Summary statistics are presented for the downstream junctions found in each of the 13 tissues.

Tissue Type	Minimum	1 st Quartile	Median	Mean	3 rd Quartile	Maximum
Anterior Cingulate Cortex	1.000	1.077	1.333	1.734	1.713	73.152
Amygdala	1.000	1.067	1.385	1.727	1.785	1.785
Caudate	1.000	1.100	1.333	1.707	1.658	63.799
Cerebellar Hemisphere	1.000	1.259	1.614	2.525	2.283	77.678
Cerebellum	1.000	1.111	1.328	1.955	1.684	63.048
Cortex	1.000	1.111	1.333	1.742	1.643	73.265
Frontal Cortex	1.000	1.231	1.567	2.167	2.116	105.642
Hippocampus	1.000	1.083	1.333	1.710	1.700	61.767
Hypothalamus	1.000	1.125	1.386	1.755	1.750	76.510
Nucleus Accumbens	1.000	1.125	1.349	1.755	1.721	67.081
Putamen	1.000	1.000	1.286	1.651	1.618	63.466
Spinal Cord	1.000	1.071	1.375	1.887	1.844	78.553
Substantia Nigra	1.000	1.000	1.400	1.816	1.844	75.873
All Tissues	1.000	1.111	1.375	1.863	1.800	105.642

As a first pass analysis for examining the overall expression of the splice junctions, the pipeline in this study categorizes junctions as belonging to major and minor isoforms. In order to categorize the junctions, the median coverage for canonical junctions found in each of the 13 brain tissues is set as the threshold for read coverage (Table 1). Canonical and flanking junctions are classified as belonging to major isoforms when their average read coverages are higher than the threshold. Junctions which have average coverages that fall below the threshold are categorized as being present in minor isoforms. For a putative exon to be considered part of a major or minor isoform, they must be flanked by junctions which both fall into the same isoform category.

To investigate the prevalence of splice junctions across samples from each brain tissue, the pipeline categorizes junctions as belonging to common or rare isoforms. To categorize the junctions into these two groups, the threshold is set at 50% of the total samples in each brain tissue group. A junction which is found in less than 50% of samples is considered part of a rare isoform, whereas a junction found in more than 50% of samples is considered part of a common. For a putative exon to be considered part of a common or rare isoform, they must be flanked by junctions which both fall into the same isoform category.

After applying these thresholds, a majority of canonical junctions for all tissue types are classified as common major isoforms (Table 4). In contrast, a majority of putative exons are flanked by junctions which are categorized as belonging to common minor isoforms (Table 5). This is to be expected, given the high levels of expression seen in the Snaptron

output for canonical junctions, and the lower levels of expression for the flanking junctions.

Table 4. Classification of Canonical Junctions. This table summarizes how many canonical junctions are classified as belonging to common major, common minor, and rare major isoforms in each of the 13 different tissues.			
Tissue	Common Major	Common Minor	Rare Major
Anterior Cingulate Cortex	1872	1428	4
Amygdala	1866	1333	6
Caudate	1877	1405	1
Cerebellar Hemisphere	1855	1369	33
Cerebellum	1886	1350	0
Cortex	1884	1455	0
Frontal Cortex	1854	1473	35
Hippocampus	1874	1346	4
Hypothalamus	1876	1408	0
Nucleus Accumbens	1876	1399	1
Putamen	1870	1315	0
Spinal Cord	1865	1335	0
Substantia Nigra	1865	1316	0

Table 5. Classification of Putative Exons. This table summarizes how many putative exons are classified as belonging to common major, common minor, and rare major isoforms in each of the 13 different tissues.			
Tissue	Common Major	Common Minor	Rare Major
Anterior Cingulate Cortex	5	187	0
Amygdala	0	148	0
Caudate	0	195	0
Cerebellar Hemisphere	7	394	0
Cerebellum	10	335	0
Cortex	5	243	0
Frontal Cortex	0	236	0
Hippocampus	0	146	0
Hypothalamus	0	180	0
Nucleus Accumbens	0	208	0
Putamen	6	158	0
Spinal Cord	8	207	0
Substantia Nigra	7	154	0

However, there are 48 instances where putative exons are flanked by junctions with average read coverages above the threshold and are categorized as common major isoforms. There are ten instances of these putative exons being flanked by at least one unannotated junction, suggesting that they could be part of novel common major isoforms (see Appendix 1).

Tissue Specificity

As a first pass analysis for examining the tissue specificity of splice junctions, the pipeline in this study quantifies how many canonical junctions are tissue specific. The canonical junction data produced by Snaptron is evaluated twice for tissue specificity: once before the application of the thresholds for read coverage and sample count to the data, and a second time following the application of the thresholds.

Examining the tissue specificity of the canonicals prior to evaluating them for the other hallmarks reveals that a majority seem to be expressed in more than one brain tissue (Figure 8). Therefore, it can be speculated that the canonical junctions defined in this study are representative of splicing events which are not specific to any of the 13 brain tissues. In total there are 12 canonical junctions which appear to be tissue-specific (see Appendix 4). These tissue specific canonical junctions have low read coverage and sample counts, implying that they occur rarely and are expressed at a lower rate, so the apparent tissue specificity may simply indicate that the expression levels are near our detection threshold.

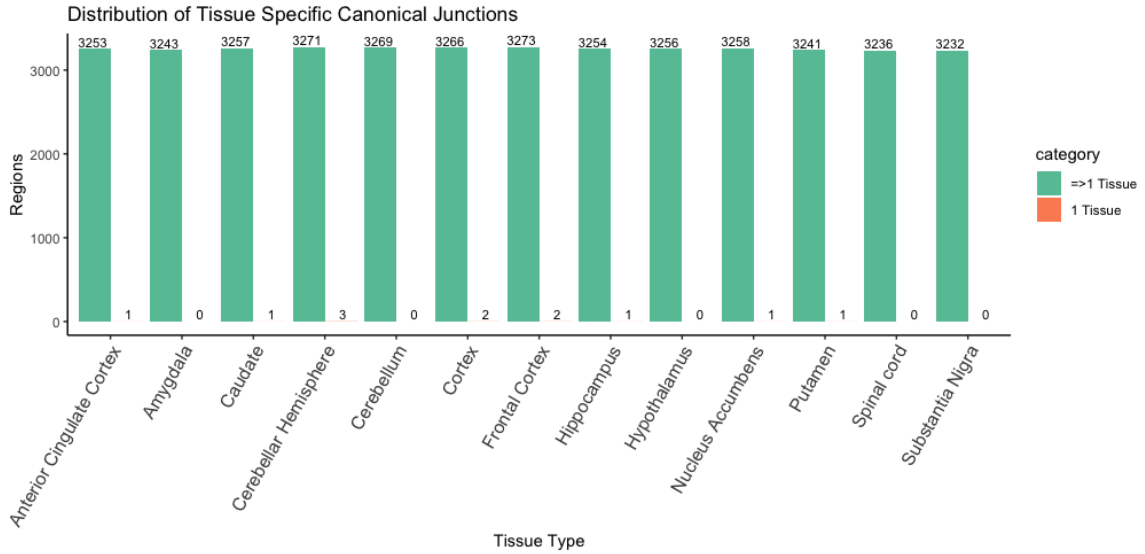


Figure 8. Distribution of tissue specific and non-specific canonical junctions across the 13 brain tissues from GTEx. The data used for this graph include all detected canonical junctions prior to the application of thresholds for read coverage and detection across samples.

Following the application of sample count and read coverage thresholds to the data, the number of tissue specific canonicals increases to 46 (Appendix 5). A total of 41 tissue specific canonical junctions are associated with common minor isoforms, and 5 are classified as rare major isoforms (Figure 9). This increase in tissue specific canonical junctions following the application of the thresholds suggests that these junctions may have been found in multiple brain tissues in the original data, prior to applying the thresholds. However, the instances of the junctions occurring in other tissues may have been removed due to having read coverages and sample counts lower than what the thresholds allow for isoform classification.

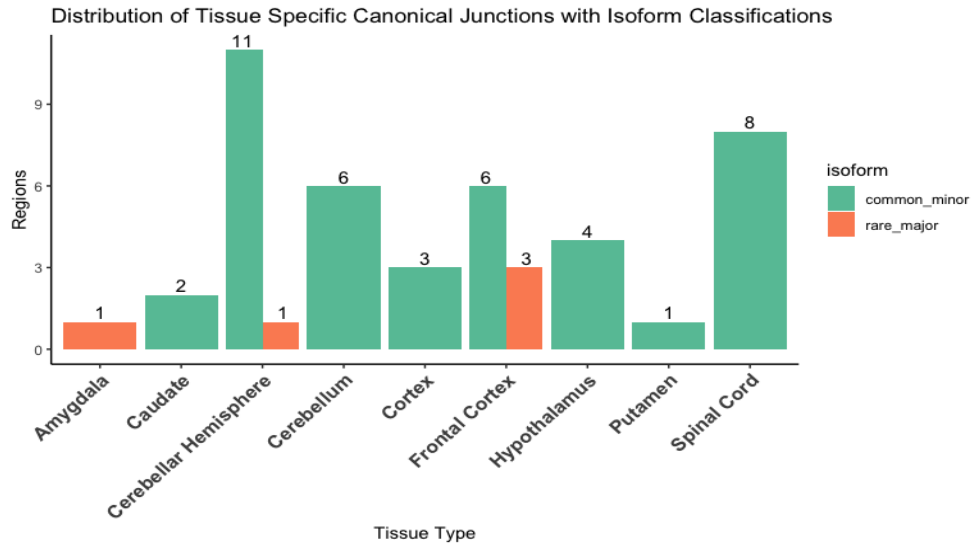


Figure 9. Distribution of tissue specific canonical junctions across the brain tissue samples from GTEx. The data used for this graph include canonical junctions which satisfied the threshold requirements for read coverage and detection across samples.

Whether these canonicals should be considered tissue specific is dependent upon including or excluding the other occurrences that were present in the original data. It can be argued that only the canonical junction results which have read coverages and sample counts that meet at least one of the thresholds should be considered, as they are more representative of canonical splicing events. However, seeing as the thresholds can be lowered so that they are less stringent, it may be appropriate to account for the other canonical junction results as well. In summation, the results suggest that a majority of canonical junctions are not tissue specific. For those that appear to be tissue specific, additional research is needed to test their tissue specificity and to see if they are actually representative of canonical junctions.

Repeating the same analysis for the putative exons yields similar results. Examination of the tissue specificity of putative exons, prior to evaluating their flanking junctions for the other hallmarks, shows that a majority appear to be spliced into transcripts that are found in more than one brain tissue. This observation is applicable to putative exons with a status of annotated (See Figure 10) and novel (See Figure 11). For putative exons which are flanked by two annotated junctions, 25 are found to be tissue specific (see Appendix 6). For putative exons flanked by an unannotated flanking junction, 81 are found to be tissue specific (see Appendix 7). These results propose that, regardless of annotation status, the putative exons in this study are mostly spliced into transcripts which are not specific to any of the 13 brain tissues.

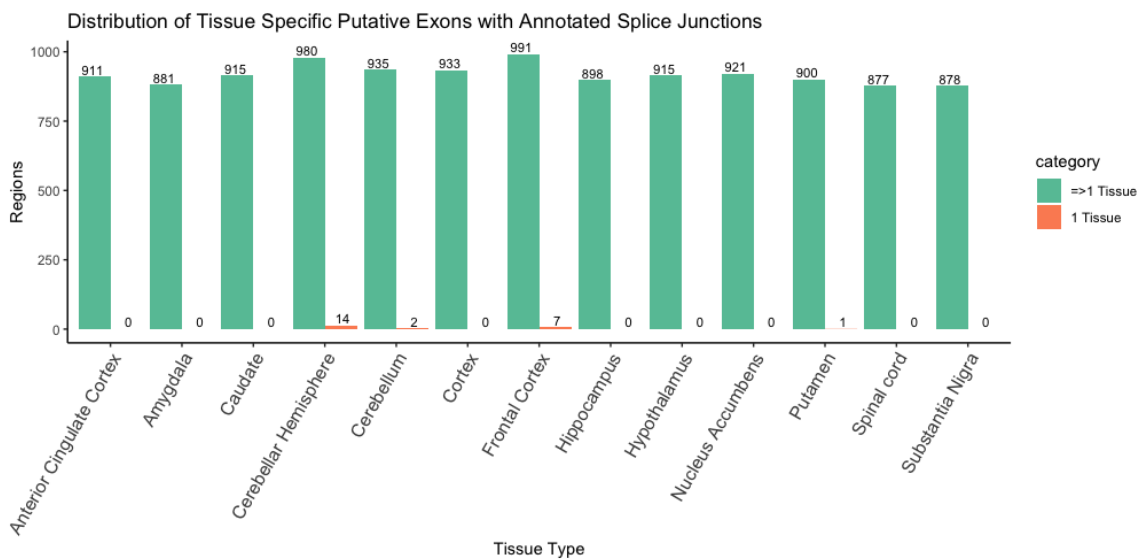


Figure 10. Distribution of annotated tissue specific and non-specific putative exons across the 13 brain tissues from GTEx. The data used for this graph include all putative exons prior to the application of thresholds for read coverage and detection across samples. These putative exons are flanked by two annotated junctions.

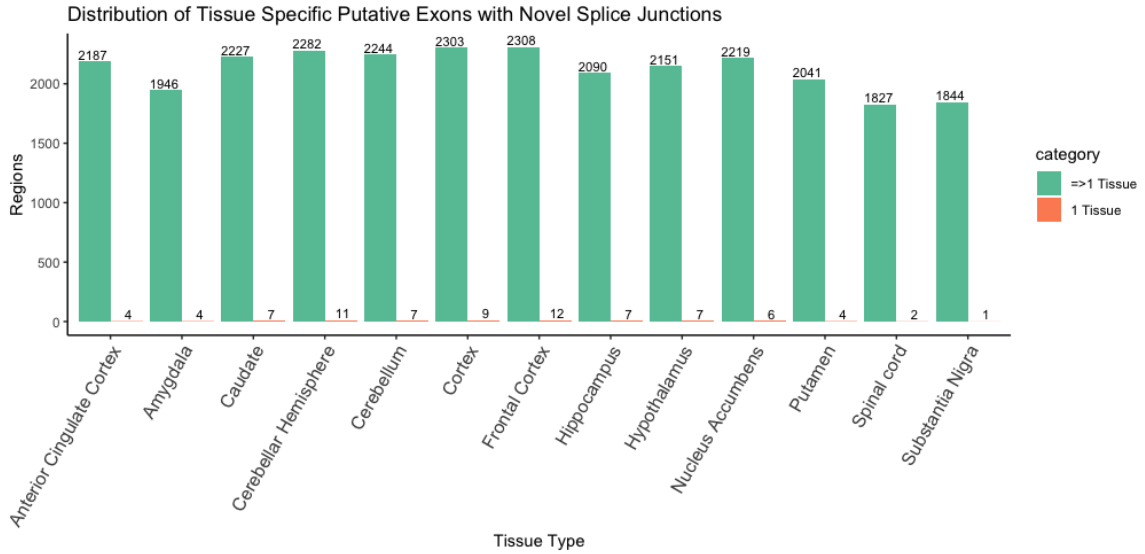


Figure 11. Distribution of unannotated tissue specific and non-specific putative exons across the 13 brain tissues from GTEx. The data used for this graph includes all novel putative exons prior to the application of thresholds for read coverage and detection across samples. These putative exons are flanked by at least one unannotated junction.

After applying the sample count and read coverage thresholds to the data, the number of tissue specific annotated putative exons increases to 38 (see Appendix 3). For novel putative exons, a total of 68 are found to be tissue specific (see Appendix 2). All putative exons found to be tissue specific are associated with common minor isoforms (Figures 12 and 13).

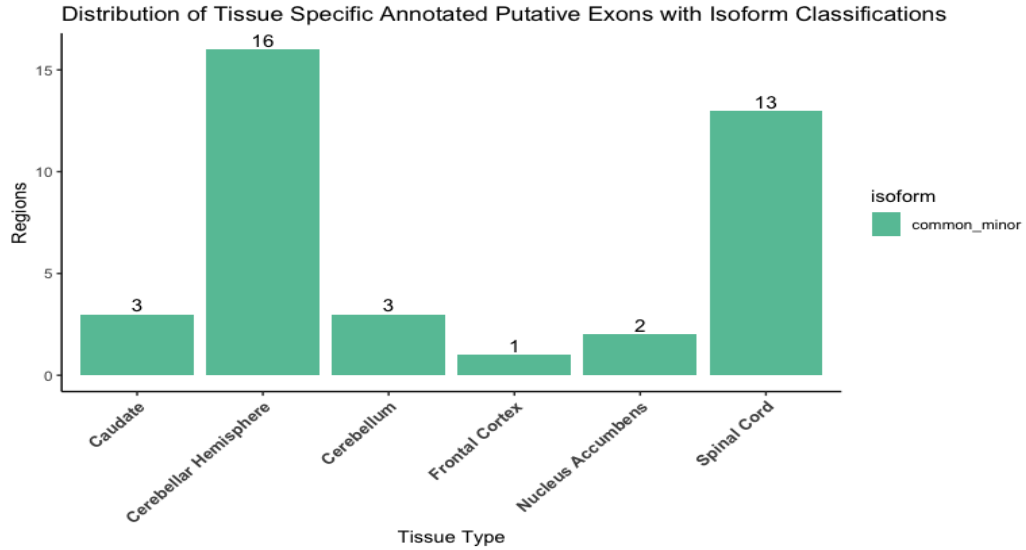


Figure 12. Distribution of tissue specific annotated putative exons across the brain tissue samples from GTEx. The data used for this graph include putative exons whose flanking junctions satisfied the threshold requirements for read coverage and detection across samples. These putative exons are flanked by two annotated junctions.

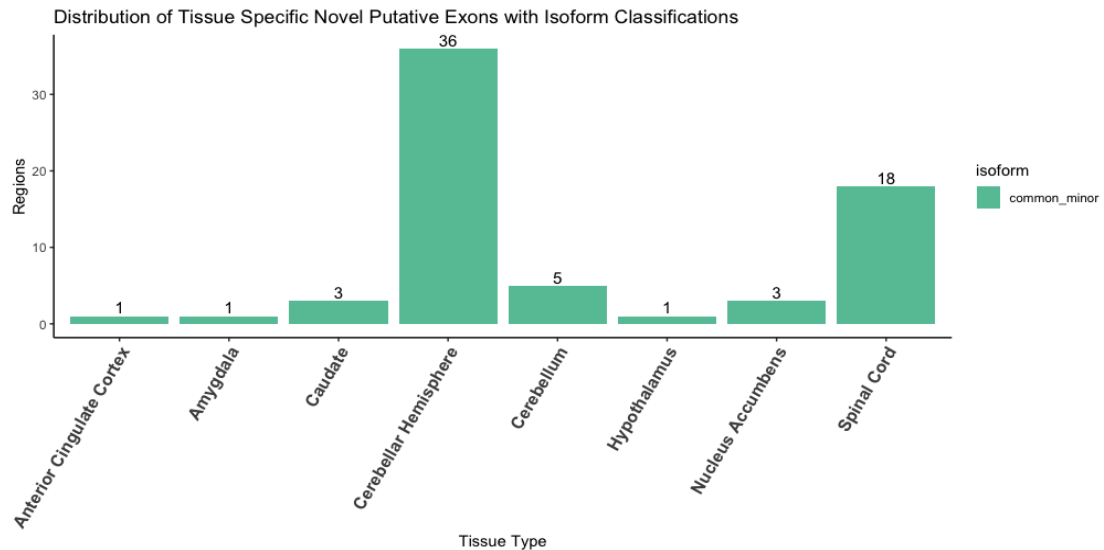


Figure 13. Distribution of tissue specific putative exons across the brain tissue samples from GTEx. The data used for this graph include putative exons whose flanking junctions satisfied the threshold requirements for read coverage and detection across samples. In addition, at least one of the junctions flanking the putative exon are unannotated.

Comparison of these tissue specific putative exons to those found in the original data, prior to application of the thresholds, shows that the putative exon regions are not the same. It is possible that the sample count and read coverage thresholds are influencing the tissue specificity results of the putative exons in a manner similar to the canonicals. The putative exons identified as common isoforms may have been detected in multiple brain tissues in the original data, prior to the application of the thresholds. However, these were potentially omitted from this dataset due to having read coverage or sample count values lower than what is allowed by the thresholds.

Much like the canonicals, assigning tissue specificity to the putative exons is dependent upon including or excluding the other occurrences found in the original data. It can be argued only putative exons whose flanking junctions satisfied the threshold requirements should be considered for tissue specificity, since they are more robust in regard to overall expression and detection across samples. When looking at the tissue specific putative exons from the original data, it is evident that most of the annotated (see Appendix 6) and novel putative exons (see Appendix 7) are flanked by at least one junction which has relatively low sample count or read coverage. The tissue specific putative exons associated with common minor isoforms have much higher levels of read coverage and sample counts in comparison. However, since the characteristics of the putative exons are not as well defined as those of the canonicals, read coverage and sample counts cannot necessarily be used as definitive reasons to exclude putative exon data from being evaluated for tissue specificity. Additional research which goes beyond the scope of this

pipeline is needed to test their tissue specificity. These results suggest that a majority of putative exons are incorporated into transcripts which are not tissue specific.

It is possible that the lack of tissue-specific canonical and putative exon regions identified in this project is due to the fact that this study utilizes sequencing data from a single organ, the brain. Overlap of expression between the 13 brain tissues is to be expected, provided that the majority of cell types found in these tissues are composed of neurons and glial cells. This overlap may not be found if other more distinct tissues were compared to brain tissue, such as skin or blood. Broadening the number of tissues which are examined in this study would make the criteria for tissue-specificity less stringent. Doing so may provide further insight into how many canonical splice junctions and putative exons are exclusive to the brain, and if they are expressed in any other organs or tissues.

DISCUSSION

This pipeline is designed to investigate the biological significance of putative exons derived from REL exonization events in human brain tissues by utilizing publicly available tools and resources. Putative exon function is assessed using an annotated agnostic approach which evaluates the splicing events that could potentially incorporate the exons into final transcripts. These splicing events, or flanking junctions, are evaluated using the following hallmarks of functionality: overall expression, prevalence across samples, and tissue specificity. This multipronged approach provides valuable insight as to how these hallmarks can be utilized to prioritize genomic features such as exons and splice junctions for future study. In addition, this pipeline is intended to supplement ongoing studies which are geared towards the investigation of novel exons and splicing events. This goal in particular has been achieved through the analysis of the putative exons provided by the Darby et al., 2016 study.

The methods utilized in this study provide independent verification of some Darby et al., 2016 study findings regarding one of the functional hallmarks: overall expression. Their analysis of putative exons in the OFC found that a majority of putative exonization events were expressed at lower levels compared to the surrounding exons. Therefore, most of the putative exons were determined to give rise to minor splice variants. This study corroborates their observations, as a majority of flanking junctions surrounding the putative exons were classified as minor isoforms when their read coverage was compared to that of the canonical junctions. This independent verification suggests that analyzing

transcription peaks occurring in introns is a valid method for identifying potential exons, and that these exons can be verified using splice junctions.

Evaluation of the canonical junctions and flanking junctions for sample prevalence indicated that a majority were found to be associated with common isoforms. Both sets of junctions are found in more than 50% of samples within each brain tissue group. Very few canonical junctions are found in less than 50% of samples, which suggests that a majority of the canonical junctions defined in this pipeline are representative of actual canonical junctions which are the most common form of splicing. None of the putative exons are found to be in rare major isoforms. This suggests that the putative exons may have an inherent biological function that is needed for human brain function at a certain baseline. However, additional research is needed to confirm the specific function of these putative exons. Future analyses will be needed to determine if the putative exons are transcribed into functional proteins, and if those proteins can be identified in brain proteomics data.

Assessing the tissue specificity of canonical junctions and putative exons revealed that a majority of these regions are potentially incorporated into transcripts across multiple brain tissues. With the understanding that canonical junctions tend to be incorporated into functional isoforms, it can be deduced that lack of tissue specificity does not necessarily equate to lack of function. Therefore, the putative exons in this study whose flanking junctions are found to not be tissue specific may still be considered functional. This could

suggest that the canonical junctions and putative exons investigated in this study play a biological role that is not restricted to a specific brain tissue.

Very few regions are found to be part of tissue specific isoforms, and there is a possibility that they may not be truly tissue specific. It must be noted that the process of evaluating tissue specificity in this study raised some uncertainties in defining what data is suitable for tissue specificity evaluation. The Snaptron results for canonical junctions and putative exons were evaluated twice, once before and once after the application of thresholds used for the assessment of the other two functional hallmarks. Each round of analysis yielded different results, suggesting that the thresholds were impacting which regions were defined as tissue specific. Additional investigation is needed to determine which of these regions is truly tissue specific.

In addition to tissue specificity, application of the other functional hallmarks to the putative exons provided insight as to which exons should be prioritized for future study. This is particularly true for those which are flanked by unannotated junctions and are classified as belonging to potentially novel common major isoforms, with exceptional read coverage and sample counts (See Appendix 1). It would be interesting to see if these specific isoforms have a unique function in the tissues they are expressed in.

This pipeline has provided valuable guidance as to how publicly available tools and resources can be used to study putative exons and splicing events at the level of the transcriptome. The methods of this pipeline can be applied to other RNA-seq studies in a

similar fashion and can be used to prioritize other transcriptomic elements for future study. In regard to future research, it would be beneficial to investigate the functional significance of these putative exons in disease states, and their impact on the translation of proteins. Looking at the final proteins which result from transcripts that include these putative exons would also provide valuable insight into any cellular functions or biological pathways that may be influenced by the proteins. These topics are being investigated by Mr. Conor Jenkins, and Ms. Alyssa Klein, who are collaborating with myself and Dr. Darby to research the functional impact of novel exons.

WORKS CITED

- Burset, M., Seledtsov, I. A., & Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, 28(21), 4364–4375.
- Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., & Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*, 35(4), 319–321.
<https://doi.org/10.1038/nbt.3838>
- Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., ... Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, 50(D1), D988–D995. <https://doi.org/10.1093/nar/gkab1049>
- Darby, M. M., Leek, J. T., Langmead, B., Yolken, R. H., & Sabunciyan, S. (2016). Widespread splicing of repetitive element loci into coding regions of gene transcripts. *Human Molecular Genetics*, 25(22), 4962–4982.
<https://doi.org/10.1093/hmg/ddw321>
- Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., ... Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, 49(D1), D916–D923. <https://doi.org/10.1093/nar/gkaa1087>
- Kandoi, G., & Dickerson, J. A. (2019). Tissue-specific mouse mRNA isoform networks.

Scientific Reports, 9(1), 13949. <https://doi.org/10.1038/s41598-019-50119-x>

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 9(8), e1003118.

<https://doi.org/10.1371/journal.pcbi.1003118>

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., ... Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, 348(6235), 660–665. <https://doi.org/10.1126/science.aaa0355>

Nellore, A., Jaffe, A. E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips III, R. A., Karbhari, N., Hansen, K. D., Langmead, B., & Leek, J. T. (2016). Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology*, 17(1), 266. <https://doi.org/10.1186/s13059-016-1118-6>

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizkadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., ... Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44–53.

<https://doi.org/10.1126/science.abj6987>

O’Leary, N. A., Wright, M. W., Brister, J. R., Ciuffo, S., Haddad, D., McVeigh, R., Rajput,

- B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733-745. <https://doi.org/10.1093/nar/gkv1189>
- Palazzo, A. F., & Lee, E. S. (2015). Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics*, 6. <https://www.frontiersin.org/article/10.3389/fgene.2015.00002>
- Sibley, C. R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews Genetics*, 17(7), 407–421. <https://doi.org/10.1038/nrg.2016.46>
- Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., & Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12), 1177–1184. <https://doi.org/10.1038/nmeth.2714>
- Wilks, C., Gaddipati, P., Nellore, A., & Langmead, B. (2018). Snaptron: Querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics (Oxford, England)*, 34(1), 114–116. <https://doi.org/10.1093/bioinformatics/btx547>
- Zhang, D., Guelfi, S., Garcia-Ruiz, S., Costa, B., Reynolds, R. H., D'Sa, K., Liu, W., Courtin, T., Peterson, A., Jaffe, A. E., Hardy, J., Botía, J. A., Collado-Torres, L., & Ryten, M. (2020). Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances*, 6(24), eaay8299. <https://doi.org/10.1126/sciadv.aay8299>

APPENDICES

Appendix 1. Novel Putative Exons with a Classification of Common Major

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Sample Threshold	Classification
chr18:79904184-79930227	chr18:79930227-79930311	chr18:79930312-79933968	Anterior Cingulate Cortex	=> 50%	common_major
chr16:29817293-29819011	chr16:29819011-29819202	chr16:29819203-29819554	Cerebellar Hemisphere	=> 50%	common_major
chr16:29817293-29819011	chr16:29819011-29819202	chr16:29819203-29819554	Cerebellum	=> 50%	common_major
chr19:3661084-3661372	chr19:3661372-3661526	chr19:3661527-3661870	Cerebellum	=> 50%	common_major
chr18:79904184-79930227	chr18:79930227-79930311	chr18:79930312-79933968	Cortex	=> 50%	common_major
chr12:54580393-54580703	chr12:54580703-54580778	chr12:54580779-54580943	Putamen	=> 50%	common_major
chr9:98005154-98007501	chr9:98007501-98007702	chr9:98007703-98011270	Spinal Cord	=> 50%	common_major
chr17:69079266-69079957	chr17:69079957-69080027	chr17:69080028-69081065	Spinal Cord	=> 50%	common_major
chr17:69079266-69079957	chr17:69079957-69080137	chr17:69080138-69081065	Spinal Cord	=> 50%	common_major
chr9:98005154-98007501	chr9:98007501-98007702	chr9:98007703-98011270	Substantia nigra	=> 50%	common_major

Appendix 2. Novel Putative Exons which are Tissue Specific (With Isoform Classification)

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr22:41952422-41956104	chr22:41956104-41956152	chr22:41956153-41957671	Anterior Cingulate Cortex	common_minor	1.352941	4.39759	51	83
chr8:141190418-141191908	chr8:141191908-141191990	chr8:141191990-141192330	Amygdala	common_minor	1.738095	2.586207	42	58
chr11:12021511-120220457	chr11:120220457-120220534	chr11:120220535-120225660	Caudate	common_minor	1.597015	2.55914	67	93
chr2:26176485-26177132	chr2:26177132-26177214	chr2:26177215-26182966	Caudate	common_minor	1.767123	1.975	73	80
chrX:108677634-108680228	chrX:108680228-108680295	chrX:108680296-108680678	Caudate	common_minor	1.960526	2.044776	76	67

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr11:108723463-108767936	chr11:108767936-108768001	chr11:108768002-108838445	Cerebellar Hemisphere	common_misor	1.852459	2.166667	61	72
chr13:99318070-99329141	chr13:99329141-99329200	chr13:99329201-99340434	Cerebellar Hemisphere	common_misor	2.45	2.101695	60	59
chr19:54126618-54127664	chr19:54127664-54127740	chr19:54127741-54128072	Cerebellar Hemisphere	common_misor	2.358209	2.183333	67	60
chr19:55308729-55310292	chr19:55310292-55310498	chr19:55310499-55311910	Cerebellar Hemisphere	common_misor	3.293478	1.985294	92	68
chr2:171707378-171710727	chr2:171710727-171710845	chr2:171710846-171712766	Cerebellar Hemisphere	common_misor	1.915254	3.65625	59	96
chr22:24302385-24311993	chr22:24311993-24312131	chr22:24312132-24313312	Cerebellar Hemisphere	common_misor	2.081967	2.978495	61	93
chr3:9750424-9750664	chr3:9750664-9750755	chr3:9750756-9750944	Cerebellar Hemisphere	common_misor	1.580645	1.633333	62	60
chr4:11225798-112258332	chr4:112258332-112258479	chr4:112258480-112260765	Cerebellar Hemisphere	common_misor	2.698795	3.306818	83	88
chr4:165379529-165401072	chr4:165401072-165401191	chr4:165401192-165464389	Cerebellar Hemisphere	common_misor	2.932584	2.344262	89	61
chr5:179833783-179835729	chr5:179835729-179835833	chr5:179835834-179836435	Cerebellar Hemisphere	common_misor	1.677966	5.442478	59	113
chr6:42688387-42688989	chr6:42688989-42689143	chr6:42689144-42689568	Cerebellar Hemisphere	common_misor	2.078125	3.389474	64	95
chr7:87628946-87635162	chr7:87635162-87635211	chr7:87635212-87650821	Cerebellar Hemisphere	common_misor	1.921053	1.85	76	60
chr8:84183579-84237609	chr8:84237609-84237715	chr8:84237716-84529298	Cerebellar Hemisphere	common_misor	2.016949	4.414894	59	94
chr1:27534917-27545255	chr1:27545255-27545321	chr1:27545322-27547260	Cerebellar Hemisphere	common_misor	2.056338	2.083333	71	60
chr1:45058760-45059415	chr1:45059415-45059459	chr1:45059460-45060098	Cerebellar Hemisphere	common_misor	2.707692	2.369231	65	65
chr11:118168752-118170318	chr11:118170318-118170422	chr11:118170423-118176361	Cerebellar Hemisphere	common_misor	2.014493	1.552239	69	67
chr11:6396216-6397691	chr11:6397691-6397805	chr11:6397806-6400988	Cerebellar Hemisphere	common_misor	14.294643	2.608696	112	69
chr11:6396216-6397691	chr11:6397691-6397808	chr11:6397809-6400988	Cerebellar Hemisphere	common_misor	14.294643	2.142857	112	70
chr11:9021956-9022411	chr11:9022411-9022756	chr11:9022757-9025701	Cerebellar Hemisphere	common_misor	2.482353	1.661765	85	68
chr12:101753540-10174023	chr12:10174023-101754085	chr12:101754086-101757211	Cerebellar Hemisphere	common_misor	4.632075	2.40625	106	64
chr15:101686063-101693247	chr15:101693247-101693272	chr15:101693273-101701085	Cerebellar Hemisphere	common_misor	3.166667	2.444444	66	63
chr16:1511151-1516360	chr16:1516360-1516411	chr16:1516412-1518215	Cerebellar Hemisphere	common_misor	1.548387	2.179104	62	67
chr16:21269892-21275194	chr16:21275194-21275260	chr16:21275261-21275531	Cerebellar Hemisphere	common_misor	1.825397	2.421875	63	64
chr16:66803119-66805513	chr16:66805513-66805628	chr16:66805629-66805776	Cerebellar Hemisphere	common_misor	2.276923	6.232143	65	112

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr16:66803119-66805513	chr16:66805513-66805634	chr16:66805635-66805776	Cerebellar Hemisphere	common_menor	2.276923	2.059524	65	84
chr17:51931133-52070408	chr17:52070408-52070523	chr17:52070524-52072318	Cerebellar Hemisphere	common_menor	2.183099	2.449275	71	69
chr17:5474790-5477843	chr17:5477843-5477999	chr17:5478000-5480053	Cerebellar Hemisphere	common_menor	1.634921	5.517857	63	112
chr19:13145396-13145523	chr19:13145523-13145587	chr19:13145588-13149028	Cerebellar Hemisphere	common_menor	1.887324	1.820896	71	67
chr2:240529510-240532025	chr2:240532025-240532135	chr2:240532136-240552913	Cerebellar Hemisphere	common_menor	2.169492	1.974026	59	77
chr2:27450097-27453176	chr2:27453177-27453270	chr2:27453271-27453383	Cerebellar Hemisphere	common_menor	1.866667	5.645455	60	110
chr2:61533904-61534557	chr2:61534557-61534654	chr2:61534655-61537561	Cerebellar Hemisphere	common_menor	5.933962	1.79661	106	59
chr3:197522782-197527886	chr3:197527886-197527987	chr3:197527988-197532411	Cerebellar Hemisphere	common_menor	14.40367	1.854839	109	62
chr6:151405837-151408642	chr6:151408642-151408900	chr6:151408901-151417278	Cerebellar Hemisphere	common_menor	2.419355	2.283333	62	60
chr7:130192824-130197414	chr7:130197414-130197452	chr7:130197453-130201849	Cerebellar Hemisphere	common_menor	2.378378	2.106061	74	66
chr7:92201424-92210935	chr7:92210935-92211004	chr7:92211005-92213194	Cerebellar Hemisphere	common_menor	2.605634	1.983871	71	62
chrX:47607199-4761356	chrX:4761356-47614421	chrX:47614422-47619351	Cerebellar Hemisphere	common_menor	4.090909	1.8	99	60
chr19:17568714-17571865	chr19:17571865-17571991	chr19:17571992-17572482	Cerebellum	common_menor	1.6	4.927536	80	138
chr3:121282338-121306258	chr3:121306258-121306329	chr3:121306330-121318474	Cerebellum	common_menor	1.75	1.623377	80	77
chr4:500602-501740	chr4:501740-501925	chr4:501926-505717	Cerebellum	common_menor	2.076087	1.648649	92	74
chr9:32436398-32438316	chr9:32438316-32438527	chr9:32438528-32440464	Cerebellum	common_menor	2.153061	1.545455	98	77
chr6:99476231-99479765	chr6:99479765-99479902	chr6:99479903-99482752	Cerebellum	common_menor	2.081395	2.181818	86	77
chr11:112228674-112229441	chr11:112229441-112229486	chr11:112229487-112230207	Hypothalamus	common_menor	1.857143	5.234694	56	98
chr19:34396658-34397167	chr19:34397167-34397397	chr19:34397398-34399206	Nucleus accumbens	common_menor	2.080645	5.157895	62	95
chr3:15946856-16137390	chr3:16137390-16137498	chr3:16137499-16195759	Nucleus accumbens	common_menor	2.670213	2.666667	94	69
chr2:208187062-208187901	chr2:208187901-208188037	chr2:208188038-208189952	Nucleus accumbens	common_menor	6.417476	2.478873	103	71
chr10:119751157-119757894	chr10:119757894-119757977	chr10:119757978-119781634	Spinal Cord	common_menor	2.044444	2.1	45	50
chr14:65054690-65059393	chr14:65059393-65059442	chr14:65059443-65061180	Spinal Cord	common_menor	3.921875	5.014925	64	67
chr14:76763131-76763597	chr14:76763597-76763643	chr14:76763644-76769962	Spinal Cord	common_menor	1.702128	2.404255	47	47

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr15:68695257-68704432	chr15:68704432-68704492	chr15:68704493-68710731	Spinal Cord	common_menor	2.777778	3.12766	45	47
chr17:42663504-42664653	chr17:42664653-42664751	chr17:42664752-42665475	Spinal Cord	common_menor	2.585366	2.190476	41	42
chr2:219280688-219281015	chr2:219281015-219281086	chr2:219281087-219281717	Spinal Cord	common_menor	1.820513	1.926829	39	41
chr3:170708195-170794427	chr3:170794427-170794502	chr3:170794503-170859982	Spinal Cord	common_menor	1.974359	2.369565	39	46
chr4:38689880-38694473	chr4:38694473-38694552	chr4:38694553-38694745	Spinal Cord	common_menor	2.309524	2.767857	42	56
chr7:36427029-36434332	chr7:36434332-36434386	chr7:36434387-36439203	Spinal Cord	common_menor	2.177778	2.266667	45	45
chr8:78715329-78716921	chr8:78716921-78717006	chr8:78717007-78717327	Spinal Cord	common_menor	2.75	3.38	48	50
chr9:100240241-100241474	chr9:100241474-100241556	chr9:100241557-100242569	Spinal Cord	common_menor	1.657895	3.968254	38	63
chr9:128419279-128419540	chr9:128419540-128419625	chr9:128419626-128420186	Spinal Cord	common_menor	2.068182	6.4	44	70
chr1:109355470-109366779	chr1:109366779-109366904	chr1:109366905-109367407	Spinal Cord	common_menor	2.219512	2.704545	41	44
chr1:205663619-205663719	chr1:205663719-205663856	chr1:205663857-205664484	Spinal Cord	common_menor	3.304348	3	46	41
chr12:89425380-89425869	chr12:89425869-89425986	chr12:89425987-89459637	Spinal Cord	common_menor	3.383333	2.095238	60	42
chr18:47870565-47873896	chr18:47873896-47873955	chr18:47873956-47896520	Spinal Cord	common_menor	1.857143	1.875	49	40
chr4:6372701-6372902	chr4:6372902-6372959	chr4:6372960-6375818	Spinal Cord	common_menor	1.454545	1.560976	44	41
chr9:112694234-112697438	chr9:112697438-112697582	chr9:112697583-112716460	Spinal Cord	common_menor	2.066667	1.605263	45	38

Appendix 3. Annotated Putative Exons which are Tissue Specific (With Isoform Classification)

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr11:62882067-62882529	chr11:62882529-62882618	chr11:62882619-62882907	Caudate	common_menor	2.168831	9.061069	77	131
chr11:62885609-62886436	chr11:62886436-62886568	chr11:62886569-62888134	Caudate	common_menor	1.647059	3.009174	68	109
chr1:154970200-154973292	chr1:154973292-154973508	chr1:154973509-154974247	Caudate	common_menor	1.614286	2.380952	70	84
chr12:119668404-119671494	chr12:119671494-119671668	chr12:119671669-119672300	Cerebellar Hemisphere	common_menor	1.850746	2.040541	67	74
chr12:95217887-95224179	chr12:95224179-95224236	chr12:95224237-95251939	Cerebellar Hemisphere	common_menor	2.1	2.305085	60	59
chr17:32207030-32207511	chr17:32207511-32207563	chr17:32207564-32208106	Cerebellar Hemisphere	common_menor	2.606557	2.47561	61	82

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr2:101844 312-1018453 92	chr2:10184 5392-10184 5602	chr2:1018456 03-10185597 6	Cerebellar Hemisphere	common_mi nor	2.298701	2.457627	77	59
chr2:287769 83-28778350	chr2:28778 350-287785 89	chr2:2877859 0-28778808	Cerebellar Hemisphere	common_mi nor	3.823529	2.983871	68	62
chr20:46126 702-4612720 6	chr20:4612 7206-46127 289	chr20:461272 90-46128137	Cerebellar Hemisphere	common_mi nor	2.213333	2.146667	75	75
chr22:41205 387-4120600 6	chr22:4120 6006-41206 125	chr22:412061 26-41209695	Cerebellar Hemisphere	common_mi nor	1.9	2.56338	60	71
chr3:632787 43-63289374	chr3:63289 374-632894 99	chr3:6328950 0-63480831	Cerebellar Hemisphere	common_mi nor	3.234568	5.070707	81	99
chr5:175809 069-1758093 32	chr5:17580 9332-17580 9392	chr5:1758093 93-17587865 1	Cerebellar Hemisphere	common_mi nor	9.913793	2.468354	116	79
chr6:831946 53-83195194	chr6:83195 194-831952 71	chr6:8319527 2-83195594	Cerebellar Hemisphere	common_mi nor	2.016949	4.5	59	102
chrX:484830 28-48483240	chrX:48483 240-484833 41	chrX:4848334 2-48485735	Cerebellar Hemisphere	common_mi nor	1.746032	2.417722	63	79
chr11:11524 0421-115445 760	chr11:11544 5760-11544 5836	chr11:115445 837-1155042 70	Cerebellar Hemisphere	common_mi nor	2.779412	2.338983	68	59
chr17:42961 352-4296217 4	chr17:4296 2174-42962 295	chr17:429622 96-42964105	Cerebellar Hemisphere	common_mi nor	3.264706	2.090909	102	66
chr2:202207 322-2022078 67	chr2:20220 7867-20220 8002	chr2:2022080 03-20221073 4	Cerebellar Hemisphere	common_mi nor	2.520548	2.424242	73	66
chr5:115832 665-1158335 11	chr5:115833 511-115833 598	chr5:1158335 99-11583762 7	Cerebellar Hemisphere	common_mi nor	4.14	2.242857	100	70
chr8:119760 739-1197616 07	chr8:119761 607-119761 772	chr8:1197617 73-11976241 4	Cerebellar Hemisphere	common_mi nor	2.395349	1.876923	86	65
chr10:12085 2636-120856 081	chr10:1208 56081-1208 56130	chr10:120856 131-1208586 42	Cerebellum	common_mi nor	1.592105	3.71875	76	128
chr3:138471 027-1384716 74	chr3:13847 1674-13847 1748	chr3:1384717 49-13847236 2	Cerebellum	common_mi nor	1.938272	3.055556	81	108
chr10:68470 164-6847048 8	chr10:6847 0488-68470 612	chr10:684706 13-68471790	Cerebellum	common_mi nor	2.218182	1.61039	110	77
chr12:76060 280-7606100 9	chr12:7606 1009-76061 087	chr12:760610 88-76067370	Frontal cortex	common_mi nor	4.630137	3.564516	73	62
chr18:79943 463-7994951 8	chr18:7994 9518-79949 655	chr18:799496 56-79950723	Nucleus accumbens	common_mi nor	11.218487	1.769231	119	65
chr19:50662 683-5066309 6	chr19:5066 3096-50663 169	chr19:506631 70-50666191	Nucleus accumbens	common_mi nor	8.378378	1.714286	111	63
chr1:205228 772-2052407 31	chr1:20524 0731-20524 0920	chr1:2052409 21-20524150 4	Spinal Cord	common_mi nor	2.090909	2.521739	44	46
chr18:41970 457-4198494 4	chr18:4198 4944-41985 107	chr18:419851 08-41987811	Spinal Cord	common_mi nor	1.921053	2.131579	38	38
chr2:111720 46-11172750	chr2:111727 50-1117282 4	chr2:1117282 5-11174967	Spinal Cord	common_mi nor	5.597015	2.767442	67	43
chr2:993375 90-99354036	chr2:99354 036-993540 73	chr2:9935407 4-99360235	Spinal Cord	common_mi nor	2.133333	2	45	40
chr20:34516 462-3451941 9	chr20:3451 9419-34519 498	chr20:345194 99-34526271	Spinal Cord	common_mi nor	7.915493	2.021277	71	47

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Isoform	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr3:184870806-184882348	chr3:184882348-184882433	chr3:184882434-184886109	Spinal Cord	common_minor	1.777778	2.075	45	40
chr8:47284095-47285632	chr8:47285632-47285791	chr8:47285792-47291032	Spinal Cord	common_minor	2.74	1.704545	50	44
chr10:125881072-125885364	chr10:125885364-125885676	chr10:125885677-125896217	Spinal Cord	common_minor	3.102041	2.1	49	40
chr10:32901636-32906464	chr10:32906464-32906580	chr10:32906581-32908367	Spinal Cord	common_minor	2.705882	2.5	51	50
chr19:19150765-19157215	chr19:19157215-19157283	chr19:19157284-19180685	Spinal Cord	common_minor	2.088889	1.8	45	40
chr3:118930276-118981790	chr3:118981790-118981940	chr3:118981941-119034530	Spinal Cord	common_minor	1.975	2.190476	40	42
chr3:195579496-195582616	chr3:195582616-195582739	chr3:195582740-195583877	Spinal Cord	common_minor	3.625	2.363636	56	44
chr7:106091940-106092547	chr7:106092547-106092669	chr7:106092670-106092948	Spinal Cord	common_minor	4.442308	7.015152	52	66

Appendix 4. Canonical Junctions which are Tissue Specific (Without Isoform Classification)

Canonical Junction	Tissue	Average Coverage	Sample Count
chr14:59596788-59603060	Anterior Cingulate Cortex	1	1
chr19:37089067-37155864	Caudate	1	1
chr16:10923326-10936632	Cerebellar Hemisphere	1	1
chr17:16773642-16786947	Cerebellar Hemisphere	4	1
chr10:95037310-95040921	Cerebellar Hemisphere	2	1
chr14:21501048-21503172	Cortex	1	1
chr15:78274998-78283395	Cortex	1	1
chr2:24190554-24216099	Frontal Cortex	1	1
chrY:5581799-5737271	Frontal Cortex	5	1
chr10:73208705-73243842	Hippocampus	1	1
chr12:71854854-71861404	Nucleus Accumbens	1	1
chr11:74906450-74907144	Putamen	1	1

Appendix 5. Canonical Junctions which are Tissue Specific (With Isoform Classification)

Canonical Junction	Tissue	Sample Count	Coverage Average	Isoform
chr8:95800656-95810018	Caudate	74	2.283784	common_minor
chr2:55180713-55206269	Caudate	69	1.811594	common_minor

Canonical Junction	Tissue	Sample Count	Coverage Average	Isoform
chr10:110569014-110575335	Cerebellar Hemisphere	65	3.046154	common_minor
chr12:99215-137555	Cerebellar Hemisphere	76	1.644737	common_minor
chr16:89321228-89323098	Cerebellar Hemisphere	64	1.765625	common_minor
chr19:44072292-44076321	Cerebellar Hemisphere	66	2.045455	common_minor
chr3:40506377-40511469	Cerebellar Hemisphere	62	2.758065	common_minor
chr5:180550118-180564488	Cerebellar Hemisphere	69	2.072464	common_minor
chr8:143636416-143649897	Cerebellar Hemisphere	63	1.746032	common_minor
chr11:31463306-31473723	Cerebellar Hemisphere	64	2.171875	common_minor
chr11:96384762-96387529	Cerebellar Hemisphere	76	2.868421	common_minor
chr4:6378573-6472159	Cerebellar Hemisphere	63	2.206349	common_minor
chr8:81676266-81679125	Cerebellar Hemisphere	72	2.180556	common_minor
chr11:103129006-103133554	Cerebellum	77	2.428571	common_minor
chr15:83943083-83970483	Cerebellum	83	1.53012	common_minor
chr20:59970187-59971961	Cerebellum	74	1.824324	common_minor
chr12:63608676-63617303	Cerebellum	87	1.862069	common_minor
chr4:112606092-112609378	Cerebellum	93	2.215054	common_minor
chr4:112623877-112631929	Cerebellum	90	2.4	common_minor
chr10:115215881-115265192	Cortex	76	3.052632	common_minor
chr4:176096586-176111574	Cortex	70	2	common_minor
chr4:37439391-37443284	Cortex	74	2.527027	common_minor
chr1:196342229-196373139	Frontal Cortex	66	2.515152	common_minor
chr1:217491751-217498355	Frontal Cortex	60	2.95	common_minor
chr13:50713241-50843187	Frontal Cortex	61	2.42623	common_minor

Canonical Junction	Tissue	Sample Count	Coverage Average	Isoform
chr15:30362812-30367417	Frontal Cortex	62	4.612903	common_minor
chr3:155596379-155675996	Frontal Cortex	64	3.71875	common_minor
chr4:163854170-164111587	Frontal Cortex	62	2.225806	common_minor
chr10:122888724-122897792	Hypothalamus	52	1.846154	common_minor
chr22:43687582-43711474	Hypothalamus	53	2.264151	common_minor
chr5:110642372-110667227	Hypothalamus	52	3.442308	common_minor
chr7:102926375-102931880	Hypothalamus	54	1.981481	common_minor
chr20:59322453-59324330	Putamen	52	2.230769	common_minor
chr16:249659-254380	Spinal Cord	45	1.777778	common_minor
chr18:62349938-62354390	Spinal Cord	52	2.923077	common_minor
chr6:36948405-36954505	Spinal Cord	44	2.5	common_minor
chr12:111938211-111991757	Spinal Cord	40	1.8	common_minor
chr3:4068746-4376329	Spinal Cord	39	2.307692	common_minor
chr4:102991713-103019598	Spinal Cord	48	2.145833	common_minor
chr5:65542849-65551563	Spinal Cord	39	2.076923	common_minor
chr5:71011375-71012347	Spinal Cord	46	3.934783	common_minor
chr15:76932467-76943757	Amygdala	1	10	rare_major
chr16:28353062-28362702	Cerebellar Hemisphere	3	25.333333	rare_major
chr1:156044224-156048500	Frontal Cortex	3	32	rare_major
chr2:128268871-128318036	Frontal Cortex	4	26.25	rare_major
chr22:23958702-23960293	Frontal Cortex	12	21.833333	rare_major

Appendix 6. Annotated Putative Exons which are Tissue Specific (Without Isoform Classification)

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr1:40763047-40763362	chr1:40763362-40763489	chr1:40763490-40766595	Cerebellar Hemisphere	1.333333	1	6	1

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr12:565220 90-56539685	chr12:56539 685-565397 64	chr12:56539765 -56562416	Cerebellar Hemisphere	1	5	2	1
chr14:103428 441-1034414 69	chr14:10344 1469-10344 1563	chr14:10344156 4-103448918	Cerebellar Hemisphere	2	1.625	1	8
chr16:120274 45-12029612	chr16:12029 612-120297 34	chr16:12029735 -12042896	Cerebellar Hemisphere	1.357143	2	14	1
chr16:294619 51-29462692	chr16:29462 692-294627 85	chr16:29462786 -29463344	Cerebellar Hemisphere	1.5	1	4	3
chr3:1425965 10-14262781 5	chr3:142627 815-142627 908	chr3:142627909 -142664201	Cerebellar Hemisphere	1.75	4	4	1
chr4:9465464 1-94656817	chr4:946568 17-9465691 1	chr4:94656912- 94657426	Cerebellar Hemisphere	1	2.833333	1	6
chr7:1492398 03-14924289 4	chr7:149242 894-149242 985	chr7:149242986 -149250158	Cerebellar Hemisphere	2	1.272727	2	11
chr1:7774560 0-77746398	chr1:777463 98-7774655 6	chr1:77746557- 77759642	Cerebellar Hemisphere	1	1.6	1	5
chr11:832659 80-83266866	chr11:83266 866-832670 38	chr11:83267039 -83273646	Cerebellar Hemisphere	1.52381	2	42	2
chr16:151145 72-15123435	chr16:15123 435-151235 67	chr16:15123568 -15125690	Cerebellar Hemisphere	5	7.416667	1	12
chr2:3834351 3-38358442	chr2:383584 42-3835868 8	chr2:38358689- 38377142	Cerebellar Hemisphere	2	1.571429	2	7
chr2:9751594 6-97521995	chr2:975219 95-9752207 2	chr2:97522073- 97523325	Cerebellar Hemisphere	1	2.333333	2	3
chr3:5274690 5-52747329	chr3:527473 29-5274745 1	chr3:52747452- 52749691	Cerebellar Hemisphere	2.373333	2	75	2
chr2:1197603 85-11979115 4	chr2:119791 154-119791 272	chr2:119791273 -119809836	Cerebellum	1	1	4	1
chr3:1138762 61-11388079 9	chr3:113880 799-113880 851	chr3:113880852 -113882751	Cerebellum	1	1	1	1
chr2:1084524 67-10846258 7	chr2:108462 587-108462 685	chr2:108462686 -108468979	Frontal Cortex	1	1	1	1
chr6:3793000 0-38003690	chr6:380036 90-3800378 3	chr6:38003784- 38061592	Frontal Cortex	1.190476	2	21	2
chr12:543717 16-54372893	chr12:54372 893-543730 31	chr12:54373032 -54373972	Frontal Cortex	1	1.565217	1	46
chr14:927328 16-92741704	chr14:92741 704-927418 59	chr14:92741860 -92748488	Frontal Cortex	1	1	5	1
chr3:3254596 6-32566612	chr3:325666 12-3256673 5	chr3:32566736- 32570345	Frontal Cortex	2.5	1.25	2	4
chr5:1763557 50-17635717 5	chr5:176357 175-176357 273	chr5:176357274 -176361601	Frontal Cortex	4	2	1	1

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr8:8000318 1-80028694	chr8:800286 94-8002879 6	chr8:80028797- 80030053	Frontal Cortex	3.13253	11	83	1
chr12:969201 26-96924832	chr12:96924 832-969249 42	chr12:96924943 -96936610	Putamen	1.071429	1	14	1

Appendix 7. Novel Putative Exons which are Tissue Specific (Without Isoform

Classification)

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr1:6998045 4-69983489	chr1:699834 89-6998361 4	chr1:6998361 5-69986248	Anterior Cingulate Cortex	1.111111	1	9	2
chr7:1075671 87-10756781 3	chr7:107567 813-107567 865	chr7:1075678 66-107571144	Anterior Cingulate Cortex	1	1	1	3
chr3:3174812 1-31797754	chr3:317977 54-3179784 1	chr3:3179784 2-31876432	Anterior Cingulate Cortex	1	1	2	1
chr4:4608405 4-46091719	chr4:460917 19-4609187 4	chr4:4609187 5-46097200	Anterior Cingulate Cortex	1	1	1	2
chr6:1465497 42-14655267 0	chr6:146552 670-146552 933	chr6:1465529 34-146554455	Amygdala	1	1.333333	1	3
chr1:3571632 7-35716500	chr1:357165 00-3571660 6	chr1:3571660 7-35718706	Amygdala	1.090909	1	11	1
chr20:397440 1-3989584	chr20:39895 84-3989853	chr20:398985 4-4015436	Amygdala	1.285714	1	14	2
chr4:7809764 -7815154	chr4:781515 4-7815243	chr4:7815244- 7816017	Amygdala	2	1	1	1
chr17:639990 75-63999315	chr17:63999 315-639993 48	chr17:639993 49-64001083	Caudate	1	1.166667	1	12
chr2:1356803 25-13569553 7	chr2:135695 537-135695 906	chr2:1356959 07-135709432	Caudate	1	2.152941	1	85
chr3:4050637 7-40507842	chr3:405078 42-4050789 0	chr3:4050789 1-40515859	Caudate	1.4	1	5	1
chr5:1041740 5-10420134	chr5:104201 34-1042034 1	chr5:1042034 2-10423734	Caudate	1	1	1	2
chr1:1615449 59-16154641 2	chr1:161546 412-161546 487	chr1:1615464 88-161548420	Caudate	1	1.333333	1	3
chr11:737308 24-73746644	chr11:73746 644-737467 53	chr11:737467 54-73760565	Caudate	1	1	1	1
chr16:809818 78-80991753	chr16:80991 753-809919 10	chr16:809919 11-80997313	Caudate	1.375	1	40	1
chr12:101877 921-1018900 99	chr12:10189 0099-10189 0221	chr12:101890 222-10189786 2	Cerebellar Hemisphere	1.428571	2.5	7	2

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr17:47579512-47581107	chr17:47581107-47581172	chr17:47581173-47582741	Cerebellar Hemisphere	3	3.2	1	5
chr19:42651182-43101560	chr19:43101560-43101700	chr19:43101701-43415880	Cerebellar Hemisphere	1	1	2	1
chr2:97233872-97240394	chr2:97240394-97240637	chr2:97240638-97241265	Cerebellar Hemisphere	1	2.833333	1	6
chr20:33848643-33849794	chr20:33849794-33849918	chr20:33849919-33850951	Cerebellar Hemisphere	1	1.5	1	4
chr7:66769026-66770096	chr7:66770096-66770130	chr7:66770131-66771882	Cerebellar Hemisphere	2	1	4	1
chr7:93239951-93240715	chr7:93240715-93240818	chr7:93240819-93246101	Cerebellar Hemisphere	1	1	2	1
chr15:43673929-43675271	chr15:43675271-43675402	chr15:43675403-43675935	Cerebellar Hemisphere	1	5.5	2	2
chr16:11703751-11721313	chr16:11721313-11721435	chr16:11721436-11721576	Cerebellar Hemisphere	1	1.307692	3	52
chr2:97515946-97516559	chr2:97516559-97516802	chr2:97516803-97523325	Cerebellar Hemisphere	7.5	1	2	1
chr3:74499786-74501049	chr3:74501049-74501502	chr3:74501503-74521057	Cerebellar Hemisphere	2	1.5	1	2
chr12:99215-119609	chr12:119609-119746	chr12:119747-137555	Cerebellum	1.333333	2	9	1
chr2:119760385-119802708	chr2:119802708-119802810	chr2:119802811-119809836	Cerebellum	1	1	4	1
chr18:33883425-33933051	chr18:33933051-33933083	chr18:33933084-33943064	Cerebellum	2	1.5	1	2
chr2:47521809-47540485	chr2:47540485-47540588	chr2:47540589-47569940	Cerebellum	1	1.594595	1	74
chr4:105449404-105453186	chr4:105453186-105453237	chr4:105453238-105453597	Cerebellum	1	1	2	1
chr6:108054588-108062203	chr6:108062203-108062362	chr6:108062363-108064184	Cerebellum	1	1	1	2
chr7:35132370-35139820	chr7:35139820-35139894	chr7:35139895-35141032	Cerebellum	1	1.142857	1	7
chr12:80271811-80275711	chr12:80275711-80275873	chr12:80275874-80278167	Cortex	1	1.4	6	5
chr18:54884437-54887832	chr18:54887832-54887881	chr18:54887882-54887994	Cortex	1	4	2	1
chr20:50830381-50832734	chr20:50832734-50832817	chr20:50832818-50841765	Cortex	1	1.230769	1	13
chr3:23889265-23898008	chr3:23898008-23898057	chr3:23898058-23900846	Cortex	1	1	1	4

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr4:128122189-128177210	chr4:128177210-128177266	chr4:128177267-128178430	Cortex	1	1.222222	1	9
chr1:216677492-216905247	chr1:216905247-216905683	chr1:216905684-216939581	Cortex	1	1.2	1	5
chr11:2299708-2347545	chr11:2347545-2347796	chr11:2347797-2377585	Cortex	1	1.32	1	25
chr11:30536176-30570081	chr11:30570081-30570252	chr11:30570253-30580245	Cortex	1	1.461538	1	13
chr9:96972949-97003057	chr9:97003057-97003124	chr9:97003125-97013258	Cortex	2	1	2	1
chr2:190408504-190425870	chr2:190425870-190425951	chr2:190425952-190435976	Frontal Cortex	3	1.333333	1	6
chr2:208428304-208429983	chr2:208429983-208430080	chr2:208430081-208437536	Frontal Cortex	1.666667	1	3	1
chr21:46459223-46481469	chr21:46481469-46481498	chr21:46481499-46484756	Frontal Cortex	1	1.3125	1	32
chr3:111147485-111185165	chr3:111185165-111185309	chr3:111185310-111192350	Frontal Cortex	1	1	1	1
chr3:184920199-184922437	chr3:184922437-184922483	chr3:184922484-184924861	Frontal Cortex	1	1.166667	1	12
chr4:47406927-47409272	chr4:47409272-47409325	chr4:47409326-47425673	Frontal Cortex	2.571429	1	42	1
chr5:162142317-162148299	chr5:162148299-162148381	chr5:162148382-162149107	Frontal Cortex	3.5	6	2	1
chr7:73190138-73190759	chr7:73190759-73190843	chr7:73190844-73191370	Frontal Cortex	2	4.5	1	2
chr10:73208705-73217889	chr10:73217889-73218052	chr10:73218053-73243842	Frontal Cortex	2	2.660714	1	56
chr6:105278439-105280995	chr6:105280995-105281092	chr6:105281093-105281745	Frontal Cortex	1	1	2	1
chr7:124863641-124865128	chr7:124865128-124865155	chr7:124865156-124870910	Frontal Cortex	1.25	1	4	1
chr9:96551873-96559029	chr9:96559029-96559067	chr9:96559068-96562692	Frontal Cortex	1	1	1	1
chr11:120952965-120956649	chr11:120956649-120956697	chr11:120956698-120956779	Hippocampus	1	1.5	1	2
chr15:42327443-42328035	chr15:42328035-42328119	chr15:42328120-42329305	Hippocampus	1	1.7	1	40
chr15:76932467-76938725	chr15:76938725-76938815	chr15:76938816-76943757	Hippocampus	1	1.136364	1	22
chr1:53889464-53898347	chr1:53898347-53898403	chr1:53898404-53923903	Hippocampus	2	1.5	1	4

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr10:3292610-32926861	chr10:32926861-32926905	chr10:32926906-32928093	Hippocampus	1.5	1	2	1
chr2:10135171-101351379	chr2:101351379-101351418	chr2:101351419-101355252	Hippocampus	2	1.608696	1	23
chr2:10135171-101351379	chr2:101351379-101351457	chr2:101351458-101355252	Hippocampus	2	1.5	1	4
chr17:5074604-50748339	chr17:50748339-50748416	chr17:50748417-50750500	Hypothalamus	1	3.487805	1	82
chr3:119751099-119758175	chr3:119758175-119758271	chr3:119758272-119765051	Hypothalamus	1	1.2	1	5
chr5:81987679-82019420	chr5:82019420-82019461	chr5:82019462-82178489	Hypothalamus	1.076923	1	13	1
chr1:217491751-217492671	chr1:217492671-217492693	chr1:217492694-217498355	Hypothalamus	1	1.1	1	10
chr12:21275425-21276604	chr12:21276604-21276729	chr12:21276730-21292163	Hypothalamus	1	1	1	1
chr16:2860860-28609613	chr16:28609613-28609748	chr16:28609749-28609936	Hypothalamus	1	1	2	1
chr16:56386023-56388368	chr16:56388368-56388399	chr16:56388400-56389184	Hypothalamus	1.388889	1	18	1
chr12:64490120-64494752	chr12:64494752-64494872	chr12:64494873-64495482	Nucleus Accumbens	1	2	3	1
chr17:76906188-76913668	chr17:76913668-76913784	chr17:76913785-76932644	Nucleus Accumbens	1	1	1	1
chr2:23835320-238353670	chr2:238353670-238353714	chr2:238353715-238356003	Nucleus Accumbens	1	1	1	1
chr2:38004217-38012221	chr2:38012221-38012362	chr2:38012363-38017185	Nucleus Accumbens	1.25	1	8	1
chr16:61655722-61674563	chr16:61674563-61674631	chr16:61674632-61713840	Nucleus Accumbens	2	1	1	1
chr4:75982038-75989408	chr4:75989408-75989472	chr4:75989473-75990751	Nucleus Accumbens	1	1	1	1
chr17:81057968-81062801	chr17:81062801-81062899	chr17:81062900-81084831	Putamen	1.5	1.533333	2	30
chr19:56141358-56155760	chr19:56155760-56155973	chr19:56155974-56158493	Putamen	1	1.949153	1	59
chr13:95633322-95634639	chr13:95634639-95634724	chr13:95634725-95641294	Putamen	1	1	1	1
chr9:111369174-111370242	chr9:111370242-111370348	chr9:111370349-111370434	Putamen	2.5	1	2	1
chr11:66292089-66293709	chr11:66293709-66293823	chr11:66293824-66294321	Spinalcord	1	1	1	5

Upstream Junction	Putative Exon	Downstream Junction	Tissue	Upstream Coverage Avg.	Downstream Coverage Avg.	Upstream Sample Count	Downstream Sample Count
chr4:1223182 71-12232312 0	chr4:122323 120-122323 183	chr4:1223231 84-122324426	Spinalcord	1	1	1	1
chr22:239587 02-23958895	chr22:23958 895-239589 63	chr22:239589 64-23960293	Substantia Nigra	3	1.944444	1	18