# Alternative Splicing Analysis of Tripeptidyl Peptidase 1 Mutations in CLN2 Batten Disease Models

Marlene Goetz
ScB Computational Biology Honors Thesis, April 2021
Advisors: Eric Morrow, MD, PhD and Ece Uzun, PhD
Second Reader: Sorin Istrail, PhD

**Abstract**

CLN2 Batten disease is an inherited, recessive, neurodegenerative disorder with onset of clinical symptoms typically during the late infantile stage of childhood development between ages 2 and 4 years old. As one of the 13 Batten disease subtypes, CLN2 is caused by mutations in the TPP1 gene at chromosome 11p15. One of the most common variants observed in CLN2 disease is an intronic splice acceptor mutation (c.509-1G>C). This TPP1 splice-site mutation when induced in human embryonic stem cells (ESCs) may produce new splicing variants, but the splice variants in the mutant is not yet determined. Also, the extent to which there is splicing of the major isoform in the presence of the mutation is unknown. This study aims to use RNA-Sequencing replicates from both control ES cells and TPP1 mutant ES differentiated neurons to characterize the TPP1 splice-site mutation behavior. By comparing and assessing the output of differential splicing programs including rMATS, Leafcutter, and ASGAL, this study intends to identify the extent of normal splicing and alternative splicing patterns on the TPP1 gene. This work will contribute to a greater understanding of the biological progression of CLN2 Batten disease and support the ability to develop therapeutics with mechanisms that target identified alternative splicing pathways and/or enhance normal splicing.

# TABLE OF CONTENTS

# INTRODUCTION

## Batten Disease

Batten disease, also known as Spielmeyer-Vogt-Sjogren-Batten disease, is the common name for a heterogenous group of inherited, recessive, neurodegenerative disorders called neuronal ceroid lipofuscinoses (NCLs). NCLs are lysosomal storage diseases (LSDs), and are characterized pathologically by glial reactivity, neuronal loss, and accumulation of autofluorescent ceroid lipopigment in neurons and other cell types [1-3]. The accumulating material in NCLs is not a disease-specific substrate and the main storage materials found to accumulate are the subunit c of mitochondrial ATP synthase or sphingolipid activator proteins A and D [4-6]. Subtypes of batten disease vary in the gene or protein affected, the age of onset of symptoms, the specific neurological phenotypes, and the rate of disease progression. Intracellular localization and function of the defective NCL proteins are also different in the subtypes: four NCL types are caused by defects in lysosomal enzymes (CLN1, CLN2, CLN10, CLN13), other NCLs are caused by defects in transmembrane proteins (CLN3, CLN6, CLN7, CLN8) [6]. Though precise function of all NCL proteins remains largely unknown, the similarity in clinical phenotypes indicates the potential presence of their shared or convergent biological pathways [7].

NCLs occur around 1 in 12,500 live births [8]. Most forms of NCLs begin during childhood and children appear to develop normally before onset of symptoms. Clinical symptoms of Batten disease often begin with loss of vision or epileptic seizures, and grow to include intellectual and motor deterioration, loss of ability to walk or talk, cognitive decline, and eventual premature death [9-11]. Children with all subtypes of Batten disease have significantly shortened life expectancy. Batten disease had originally been characterized by the age of onset of these symptoms (congenial, infantile, late infantile, juvenile or adult), but the new NCL nomenclature classifies both the defective gene and the age of onset of symptoms [12].

There are currently 13 human genes that have been identified to be linked with NCL disorders [13, 14]. Most NCL genes have a typical disease phenotype associated with complete loss of function and

are inherited in a recessive manner [10]. Over 430 mutations underlying NCLs have been identified, but the function of the causative gene has, in most cases, not been fully detailed. All mutations in NCL genes can be found in the NCL mutation database (https://www.ucl.ac.uk/ncl-disease/).

There are two main therapeutic strategies for Batten disease. The first strategy is gene replacement therapy, in which viral vectors introduce functional copies of the mutated gene stably into patients via a viral vector. The patients' cells are then able to synthesize their own enzyme that can replace the mutated enzyme. The second strategy is enzyme replacement therapy, in which enzymes that are deficient due to mutation are delivered directly to patients. Since Batten disease is defined as a classic lysosomal storage disease, intravenous injection of the deficient lysosomal enzyme in this way can potentially reduce accumulation of storage material and improve clinical measures [1, 15, 16].

## CLN2 Batten Disease Subtype

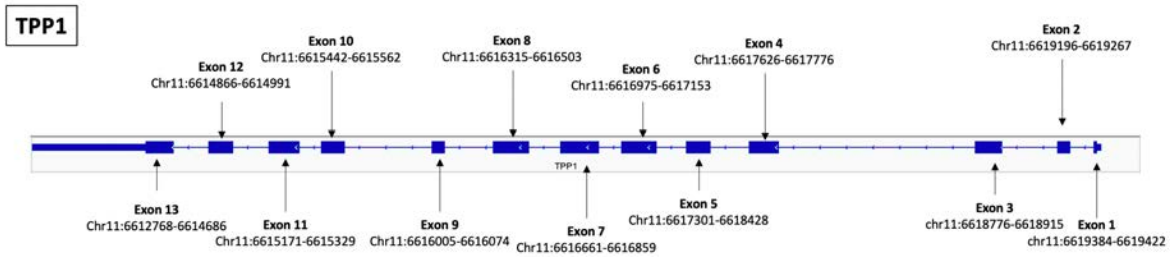
CLN2 is a specific subtype of Batten disease caused by mutations in the lysosomal enzyme tripeptidyl peptidase 1 (TPP1) gene. [17]. Deficiencies in the TPP1 enzyme and mutations in its gene serve for clinical diagnosis of CLN2 Batten disease [11]. Clinically, CLN2 is the classic late infantile-onset form of Batten disease. Children will typically appear healthy and develop normally until CLN2 symptoms present themselves. Symptoms typically will begin between the ages of 2 and 4 years old. CLN2 disease leads to a slowing of development and to psychomotor regression. Photosensitivity is also an early indicator of CLN2 disease, which can be identified using electroencephalography [18]. Symptoms progressively worsen with age, so effective CLN2 disease management requires timely diagnosis. However, irreversible neurodegeneration often occurs before a diagnosis is reached at around 5 years of age [19].

In 2017, cerliponase alfa (Brineura; BioMarin Pharmaceutical) was approved by the FDA as a treatment for CLN2 Batten disease. It is currently the only palliative treatment available for patients with CLN2 disease. As a form of enzyme replacement therapy, cerliponase alfa delivers a recombinant

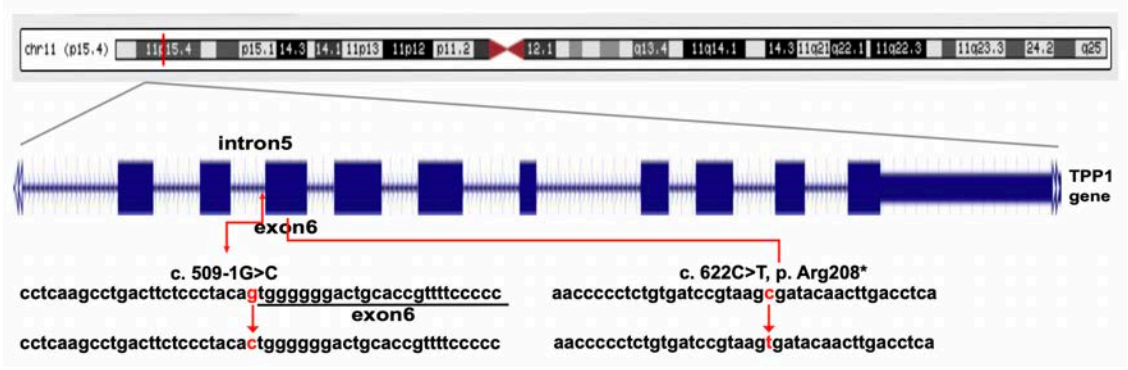
proenzyme containing a mannose-6-phosphate post-translational modification intraventricularly to the patient's brain. Cerliponase alpha has demonstrated efficacy in its ability to delay the progression of CLN2 Batten disease and slow the decline in motor and language function for patients affected [11].

## Tripeptidyl Peptidase I

The TPP1 gene, on which mutations occur in CLN2 Batten disease, is located on chromosome 11 and chromosome band 11p15. (TPP1: chr11:6616503-6617359). The gene is composed of 13 exons, 12 introns, and spans a total length of 6.65 kb [20]. Substantial allelic heterogeneity exists in the TPP1 gene in patients with CLN2 disease. To date, 131 unique disease-associated TPP1 variants have been identified and added to the UCL TPP1-specific database. Only 39 (30%) of these variants are recorded in the ClinVar database with an associated clinical classification [21]. In the spectrum of the 131 disease-associated variants, missense variants dominate (63, 48%), followed by frameshift variants (21, 16% each) and nonsense variants (17, 13%) all along the length of the TPP1 gene including in the propeptide domain [21]. Despite its heterogeneity, 60% of patients with CLN2 disease appear to have one of the two high incidence pathogenic variants. These high incidence variants are: (1) an intronic splice acceptor mutation (c.509-1 G>C), (2) a nonsense mutation (c.622C>T, p. Arg208*) [19, 21]. These mutations can be studied in human embryonic stem cells by inducing TPP1 mutations in human ESCs using CRISPR/Cas9 Homology Directed Repair technology [22].




**a.**




**b.**

**Figure 1a.** The TPP1 gene, labelled exons, and location on chromosome 11 as shown on the Integrated Genome Viewer (ivg.org) **Figure 1b.** The location of the TPP1 gene on chromosome 11 and the location of the two most common splice variants in CLN2 Batten disease [48].

## Alternative Splicing

Alternative splicing was first discovered in 1977 [23] and has now been found to occur in approximately 95% of the human genome [24]. This mechanism generates a large number of mRNA and protein isoforms from individual genes. Alternative splicing produces mature mRNA from pre-mRNA and splicing is an important regulatory mechanism for gene expression. It determines protein stability, enzymatic activity, posttranslational modifications, increases protein diversity and is closely linked with many diseases. The effects of alternative splicing range from subtle modifications to complete loss of function. As alternative splicing events are key to understanding the increased cellular and functional complexity in cells, a variety of computational methods have been developed to process short read RNA-sequencing data in order to study the varied expression of isoforms and splicing events [25]. By identifying splice events, a deeper understanding of the regulatory mechanism of alternative splicing can be obtained as well.

Splice mutations have a critical role in human genetic disease and the increased use of genome sequencing in research practice has led to the discovery of diverse, causal mutations in human genetic disease. Through exome sequencing, likely genetic diagnoses can be identified which underlie many

developmental disorders. Potential diagnostic *de novo* mutations in developmental disorder genes have been found in ~25% of probands [26]. In addition, it has been estimated that disruptions in the normal splicing patterns of mRNAs can cause human genetic disease in up to 15% of known single base pair substitutions. [27]

Alternative splicing programs aim to predict the alternative splice site positions on the reference transcriptome or the reference genome. Using the alignment, alternative splice sites are located based on a wide variety of methods. The resulting alternative splicing patterns are most often divided and identified by the following categories: Skipped Exon (SE), Alternative 5' splice site (A5SS), Alternative 3' splice site (A3SS), Mutually exclusive exons (MXE), and retained intron (RI) [28]. The cassette-type exon skipping pattern (SE) is most prevalent in vertebrates (~30%), in which an exon is spliced out of the transcript with its flanking introns, and the intron retention alternative splicing event is the rarest type [29].

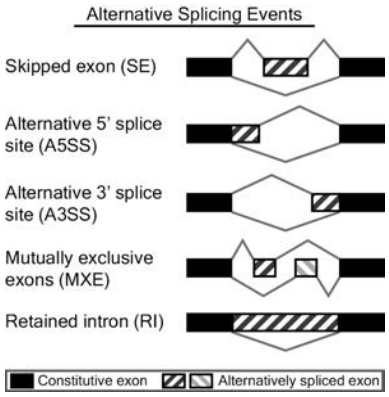



**Figure 2.** Alternative Splicing Event Patterns (from rMATS) [30]

Alternative splicing software tools differ in their approaches of uncovering alternative splicing events, as the short nature of sequencing reads results in multiple possible alignments of different transcripts to the same gene. Generally, tools differ in their mapping process of RNA-Sequencing reads to reference genomes or the transcriptome, their algorithms to predict splice junction locations, and their criteria to estimate positive and negative false rates [31].

Differential splicing tools can be divided into two major methodological categories: isoform-based and count-based methods [31]. Isoform-based methods aim to first reconstruct full-length transcripts into various isoforms and then estimate the relative abundances of the isoforms in each sample based on the sequencing reads. Statistical testing is applied following reconstruction to identify any significant differences in the relative transcript abundances between different experimental conditions. Common isoform-based differential splicing analysis methods include cuffdiff2 [32], DiffSplice [33], and MISO [34]. The second category, the count-based methods, configures genes into counting units and records counts based on the number of sequencing reads falling on each of these units. Analysis is then carried out on the defined differentially expressed counting units. These count-based methods can be further separated into two categories: either event-based methods or exon-based methods. For event-based counting methods, the individual splicing events, which are often quantified by percentage splice in (PSI) values, are used to count and measure the fraction of mRNA expressed from a gene for each specific alternative splicing event. Event-based counting programs include rMATS[30], SUPPA/SUPPA2[35], MAJIQ [36], Whippet [37] and dSpliceType. [38]. In exon-based counting methods, read counts are assigned to features that model differential exon usage, such as exons or splice junctions. These read counts are then used to inform alternative splicing patterns. Exon-based counting programs include DEXSeq [39], edgeR [40], JunctionSeq [41] and limma [42]. Finally, a more recently developed, count-based approach to alternative splicing is Leafcutter [43], which avoids the necessity of estimating isoforms or exon usage, and identifies differentially expressed regions based on intron usage.

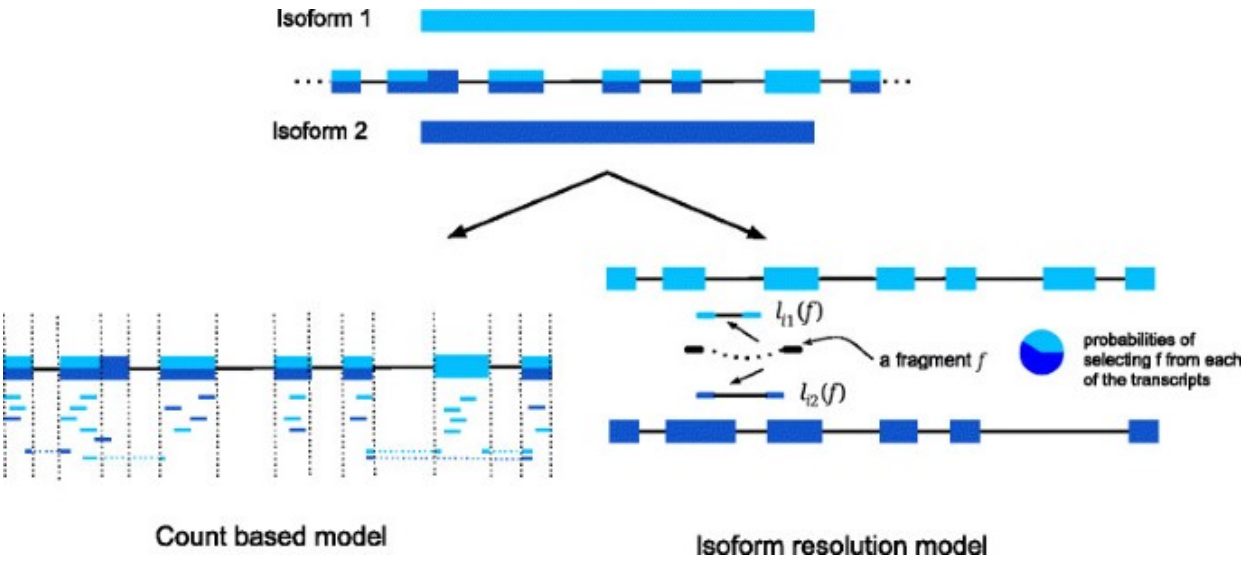



**Figure 3.** Visualization of the two major methods used in alternative splicing analysis adapted from [44]

## Alternative Splicing Literature Review

The following provides a brief overview of some commonly used alternative splicing software.

**ASGAL (Alternative Splicing Graph Aligner)** [45] – While other tools align spliced alignment of reads to reference genomes for differential analysis, ASGAL is the first tool for computing a splice-aware alignment of RNA-Sequencing data to a splicing graph. The splice-aware alignment to the graph is achieved through gap-factors that most often represent introns in relation to alternative splicing events. Pattern matching algorithms are used, which take into account these gap factors to align RNA-Sequencing reads to the splicing graph using the notion of Maximal Exact Match [46]. The constructed splicing graph is acyclic and directed with a vertex set defined as exons and an edge set as pairs $(v_i,v_j)$ such that $v_i$ and $v_j$ are consecutive in at least one transcript. The main goal of this event-based approach is to enrich a gene annotation with novel alternative splicing events supported by the RNA-sequencing samples. The study investigated an approach that directly aligns input reads against a splicing graph representing a gene annotation. The main motivation of this method is that, by using the splicing graph during the alignment phase, an alignment focused on enriching a gene annotation with alternative splicing events that produce

novel isoforms can be obtained. The output includes a file that contains the alignment of input reads to the constructed splicing graph and another file that contains all alternative splicing events detected in the RNA-Sequencing sample.

**DiffSplice [33]-** DiffSplice is an isoform-based approach which first reconstructs the transcriptome based on aligned reads, then quantifies the abundance of alternative paths through the graph and finally identifies alternative splicing modules (ASMs). ASMs encompass genomic regions where alternative transcripts diverge and have at least two possible paths. The method localizes the difference between transcriptomes and performs alternative splicing analysis on these more localized ASMs rather than the entire transcript. The abundance of alternative splicing isoforms in each ASM is used to estimate and compare across sample groups. DiffSplice does not depend on transcript or gene annotations, circumventing the need for full transcript interference and quantification. However, DiffSplice did not gain popularity in usage and literature.

**Leafcutter [43]:** As opposed to isoform- and exon- quantification approaches that rely on transcript models and statistically challenging methods of estimating isoform abundance from short-read data, Leafcutter does not require read assembly or inference on which isoforms are supported by ambiguous short RNA-Sequencing reads. Rather, Leafcutter studies variation in intron splicing to find events of high complexity. Leafcutter identifies alternatively excised introns and pools all mapped reads to find overlapping introns demarcated by split reads. Leafcutter then constructs a graph that connects all overlapping introns sharing a donor or an acceptor splice site. The nodes of this graph are introns and edges represent shared splice junctions between two introns. The connected components of this graph are clusters, representing alternative intron excision events. Leafcutter then filters rarely used introns, based on the proportion of reads supporting an intron compared to other introns in that cluster and re-clusters. Overall, leafcutter identifies alternatively excised intron clusters, and summarizes intron usage as a count.

Leafcutter works hand in hand with LeafViz, a Shiny app, which creates a visualization of the intron clusters and spliced gene. Leafcutter poses a large advantage in scalability and memory usage.

**RMATS [30] -** rMATS is an extension of MATS (multivariate analysis of transcript splicing) which was published two years prior [47]. rMATS accounts for extension of the previous analysis method to include replicate RNA-sequencing data. These programs use a hierarchical framework built upon the idea that variability within a sample set represents differences of levels of exon inclusion among replicates. The workflow of MATS and rMATS begins by using counts of RNA-Sequencing reads for each exon to map reads to exon-exon junctions of its inclusion or skipping isoform. These allow for an estimation of the exon inclusion levels in the two samples. The exon inclusion levels of all alternatively spliced exons are then used to construct a multivariate uniform prior that models the overall similarity in alternative splicing profiles of the two input samples. Using a binomial likelihood model to calculate the Bayesian posterior probability for the splicing difference between the two input samples, MATS then uses a Markov chain Monte Carlo (MCMC) to calculate a posterior probability for splicing difference. Finally, MATS calculates a P-value for each exon by comparing the observed posterior probability with a set of simulated posterior probabilities from the null hypothesis. Adjustment for multiple testing is used to obtain the FDR value.

Extending upon the framework of MATS, rMATS simultaneously models variability among replicates and estimates uncertainty of isoform proportion in individual replicates. To model the variability among replicates, rMATS incorporates a logit-normal distribution model integrating variations among replicates within the sample set. To estimate the uncertainty of isoform proportion in individual replicates, rMATS uses a binomial distribution model, considering the total read number and effective lengths of the exon inclusion isoforms it models. rMATS starts from either a RNA-Seq dataset or an alignment file produced by a spliced aligner, and produces an output a file for each considered AS event obtained from the annotation and the samples. rMATS also has an analysis model for unpaired replicates and paired replicates, allowing it to analyze over any user-defined magnitude of splicing change and

detect all major types of AS patterns. For paired replicates it uses a covariance structure to calculate the exon-specific correlation for each exon.

**SplAdder** [48] - SplAdder builds an augmented splicing graph starting from a given annotation and enriches it by exploiting the spliced alignments. To identify the alternative splicing events, it requires that all the isoforms involved in the event are supported by reads in the experiment. The main steps of the SplAdder workflow consist of (1) integrating annotation information and RNA-Seq data (2) generating an augmented splicing graph from the integrated data (3) extracting the splicing events from the generated augmented splicing graph (4) quantifying the extracted events and (5) differential analysis between samples and producing visualizations. The focus of the program is on usability, performance and accuracy.

**SUPPA2** [35] - SUPPA2 is an event-based counting method that uses transcript quantification to compute inclusion values of alternative splicing events across multiple samples. Transcript abundances are determined using the RSEM tool. [49]. The software implements density-based clustering of differentially spliced events which has an added advantage that the number of clusters do not need to be specified. These methods produce higher accuracy than other methods compared to it in the paper, especially at low sequencing depth and short read length. SUPPA2 also assesses two additional event types in conjunction with the five standard alternative splicing event types previously noted, alternative first exon (AF) and alternative last exon (AL).

**Whippet** [37] - Whippet provides an accurate, rapid method of quantifying both complex and simple alternative splicing events. It models transcriptome structure by building “contiguous splice graphs” (CSGs). Splice graphs allow single isoforms to be represented as paths through nodes in the graph. These are directed graphs, whose nodes are non-overlapping exonic sequences and whose edges (i.e., connections between nodes) represent splice junctions or adjacent exonic regions [36, 50, 51]. From these

CSGs, alternative splicing graphs are constructed that consist of paths through the alternative splicing event to determine the percent spliced in (PSI) values as proportional to the abundance of paths containing the node. PSI is quantified through convergence of the expectation-maximization (EM) algorithm. Furthermore, Whippet uses Shannon's entropy as a metric for the formalized analysis of AS complexity. This entropic measure reflects the total number of possible outcomes for an event and the degree of its proportional contribution to the transcriptome in a read-depth- and read-length- independent manner. This attribute of entropy contributes to Whippet's high degree of accuracy.

# MATERIALS AND METHODS

Preparation of the materials and assistance in analysis has been performed by members of the Morrow and Uzun lab groups. The subsequent analysis will reference contributions to these researchers accordingly. The TPP1 mutations in human stem cells were generated by Li Ma et al. [22] Processing of the RNA sequencing data and its alignment preparation has been done by Qing Wu.

## Generation of TPP1 Mutants in human ESCs

Using CRISPR/Cas9 homology directed repair knock-in technology, TPP1 mutations have been successfully induced in H9 human embryonic stem cells in Eric Morrow's laboratory in Providence, RI [22]. The intron mutation c.509-1G>C was targeted in producing homozygous mutant clones. These clones are identified as clone #23 (hPSCreg name: EMe-TPint5GC23) and clone #42 (hPSCreg name: EMe-TPint5GC42) (Figure 4). The edit was confirmed using Sanger sequencing, stem cell morphology imaging, and analysis of TPP1 protein and enzyme activity [22]. PCA analysis further supports that these clones are composed of separate subtypes (Figure 5a). The TPP1-mutant lines showed the expected characteristics of mutation and showed diminished enzyme activity. Western blot of lysates showed two new bands that indicate new splicing variants in c.509-1G>C clones (Figure 4). No proenzyme and mature enzyme bands can be observed in the mutant lines, as the intensity of TPP1 proenzyme and TPP1

mature enzyme bands in the c.509-1 G>C columns indicates that characteristic TPP1 enzyme activity was knocked out. New TPP1 variant 1 and variant 2 bands suggest that there are alternative splicing mechanisms at play in these TPP1 mutant lines which produce the apparent new protein products. Splicing abnormalities may arise due to the CLN2-related TPP1 mutation. The RNA-Sequencing data used in this alternative splicing analysis is directly from the clones #23 (SA23) and #42 (SA42) shown below, containing the mutated TPP1 splice acceptor site.

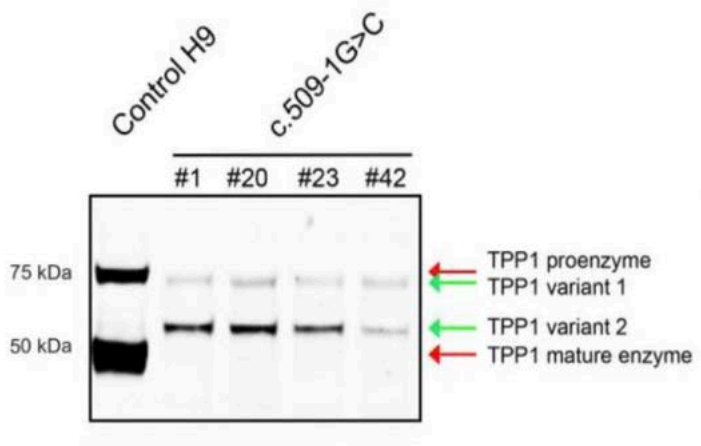



**Figure 4.** Western blot of H9 and TPP1-mutant lysates [22]

## Generation of Sequence Alignment Files

RNA-Sequencing of H9 human embryonic stem cell control lines and ES TPP1 mutant cell lines #23 and #42 differentiated neurons were obtained from GENEWIZ. Cut_adapter was used from the Trimglorae software in order to remove the adaptors and generate a quality check report that indicates all adaptors were removed. This produced .fq files which were further aligned to the reference genome using HISAT2 [52]. HISAT creates a global, whole-genome index in addition to thousands of small, local indexes during alignment. The HISAT system uses some Bowtie2 [53] code in generating the same Burrows-Wheeler Transform [54] indices and Ferragina–Manzini indexes [55] as Bowtie, but efficient data structures and algorithms allow HISAT to generate alignments faster than Bowtie while using only twice as much memory. The SAM files produced from HISAT2 were converted to BAM files using

samtools. Samtools was also used to sort the reads based on chromosome and location. The reference genome used in analysis was the human genome build hg38 (http://www.genome.ucsc.edu/). The gene annotation file used from the UCSC hg38 genome as well.

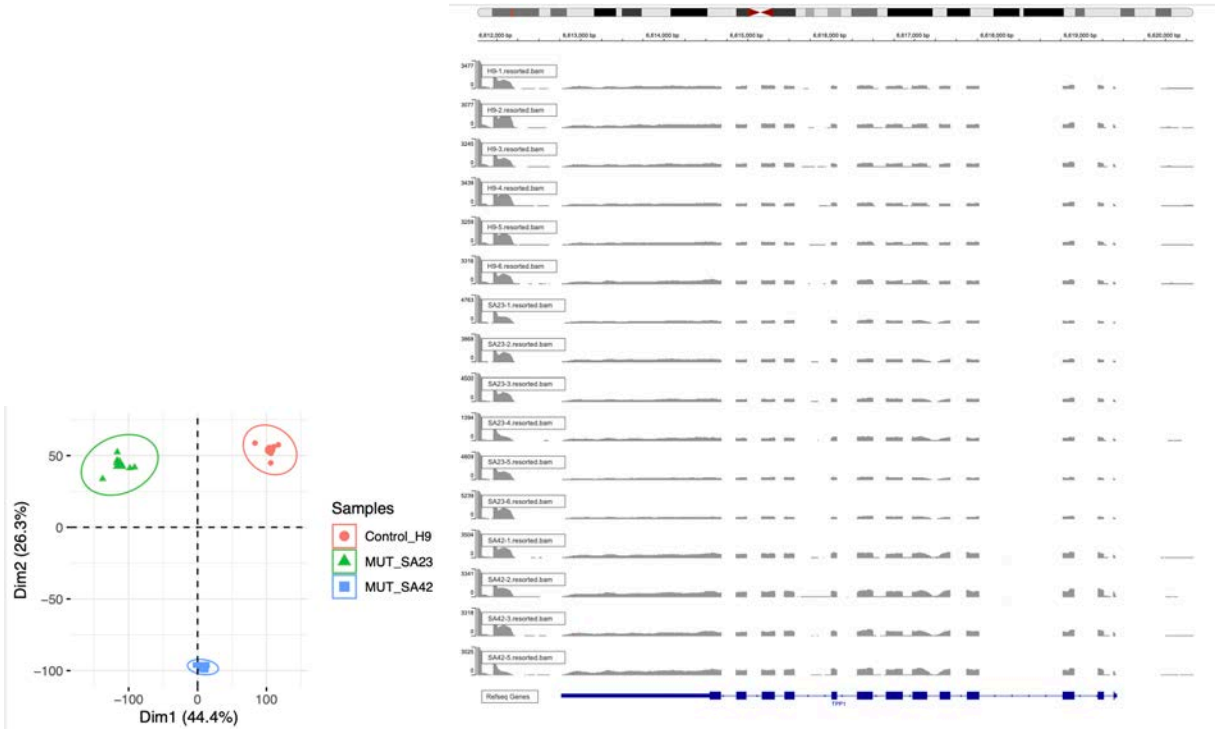



**Figure 5a.** PCA analysis of mutant SA23 and SA42 samples to indicate that they originated from different subclones (Qing Wu). **Figure 5b.** IGV coverage for the twelve SA samples.

## Leafcutter Analysis Pipeline

Leafcutter was installed with all required dependencies and modules, including R, gcc, v8 python, and gsl. The process below was performed twice: once to analyze all six H9 files with the six SA23 files, and once to analyze all six H9 files with SA42 files.

First, all bam files were converted to .junc files in a batch script using samtools and regtools. The list of .junc files were used to define intron clusters in the leafcutter_cluster_regtools.py script provided by leafcutter. The minimum intron length required was 50nt, and 50 split reads were required to support each intron cluster. Introns of up to 500kb were supported. After generating a list of intron clusters,

differential splicing analysis of the introns was performed using the leafcutter_ds.R script. A tab-separated file split the samples into their respective H9 and SA groups: the first column with sample names, and the second column with a grouping variable (H9 or SA). The leafcutter_ds.R script was run using 4 threads, with the groups file and intron clusters file as inputs. An optional exon_file derived from GENCODE v.31 was included to annotate clusters according to their corresponding genes. This exon_file is from hg38 and was converted to the correct format for Leafcutter using the provided gtf_to_exons.R script.

This process produced two relevant output files. A leafcutter_ds_cluster_significance.txt. file that includes the p-values for each cluster indicating whether or not there was significant differential intron excision for each intron cluster between the two groups tested, H9 and SA42. The leafcutter_ds_effect_sizes.txt file was produced, listing the difference in usage for introns in each group and the corresponding deltaPSI (difference in percent spliced in) for each of the listed introns. These two files were further used in the LeafViz application to produce the splicing visualizations.

## Leafcutter Visualization: Leafviz

LeafViz, an interactive, browser-based visualization application, was used to generate gene and cluster plots from the Leafcutter output files. Though Leafcutter and Leafviz produced genome-wide splicing analysis, the mutation in the TPP1 gene was of importance to this analysis, so all significant clusters with positions on the TPP1 gene on chromosome 11 were selected for visualization in Leafviz. The gene level plot for the TPP1 gene was selected, which shows the positions of the intron clusters identified that fall on the TPP1 gene. The prepare_results.R script was used to generate an Rdata object from the intron count file produced in Leafcutter and UCSC hg38 annotation files. The FDR p value threshold was specified to be 0.05. This Rdata object was loaded with ./run_leafviz.R, which ran the Leafcutter Visualization App in a browser popup. This process was repeated twice, for the H9 comparison with the SA23 group and with the SA42 group.

## RMATS Analysis Pipeline

rMATS turbo v4.1.0 was installed for further alternative splicing analysis of SA23 and SA42 files after Leafcutter. rMATS was built and installed in an anaconda environment that included the required Python and R dependencies. To pass the three groupings of H9, SA23, and SA42 inputs into rMATS, .txt files were created. These .txt files listed all 6 replicates in each group and were used in the input specifications for the rMATS batch script. The rmats.py script was executed with the following additional inputs: a GTF file with annotation of the genes and transcripts from the hg38 UCSC Homo sapiens database, a groups txt file of all 6 H9 replicates, a groups txt file of all 6 SA replicates, and the variable-read-length flag to accommodate differing read lengths of the input files. Paired-end data was used in the analysis. A minimum intron length of 50nt and maximum exon length of 500nt were used. The script was run twice: with specifications for comparing the H9 groups txt file and SA23 groups file, and another for comparison of the H9 and SA42 sample files. In each analysis, rMATS generated an output file corresponding to each alternative splicing event type assessed. These [AS_Event].MATS.JC.txt files count only reads that span each alternative splicing event junction as defined by rMATS (Figure 6).

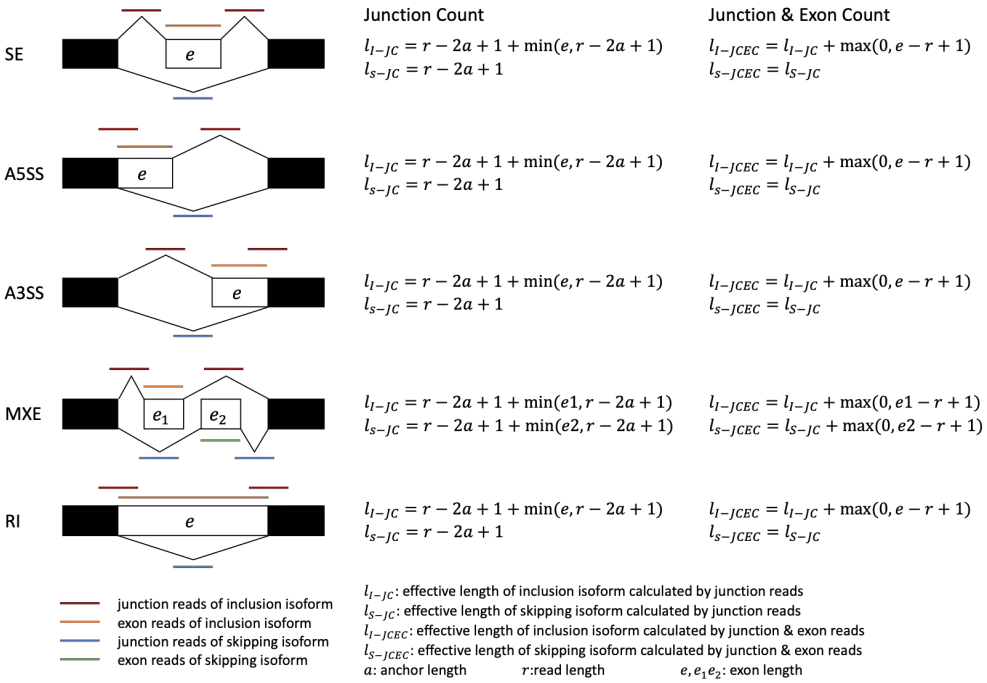



**Figure 6.** Junction count specifications for alternative splicing events as defined by rMATS. [47]

### RMATS visualization: rmats2sashimiplot

rMATS companion program, rmats2sashimiplot (https://github.com/Xinglab/rmats2sashimiplot/), was used to produce visualizations from the rMATS output files. Samtools sorted and indexed the bam files for SA23, SA42, and H9 before use of rmats2sashimiplot. Bam files with rMATS event outputs were used to run rmats2sashimiplot for the SA23 and H9 comparison and for the SA42 and H9 comparison. All six replicates were included in the rmats2sashimiplots script. The Skipped Exon (SE) rMATS event type and output data from SE.MATS.JC.txt were selected for visualization because skipped exon alternative splicing events were identified on the TPP1 gene. Mutually Exclusive Exon (MXE) rMATS output data was also visualized because this event was identified on the TPP1 gene. Exons were scaled down to 1 and introns were scaled to 5 to produce visualizations.

### ASGAL Analysis Pipeline

ASGAL takes the annotation of a single gene and its relative chromosome as input. The ASGAL pipeline was run on the TPP1 gene to build the splicing graph, format a SAM file with alignments of the splicing graph to the reference genome, and finally to analyze the alignments to detect possible alternative splicing events. Each RNA-Seq fastq file sample (each of the six SA23 and each of the six SA42 replicates) was aligned to the gene TPP1 by using a hg38 annotation file from ensembl for TPP1. The allevents flag was included in the ASGAL command to obtain all alternative splicing events in the output. The ASGAL output event files were converted to BED files and uploaded to IGV with the hg38 reference genome.

# RESULTS

Initial analysis of the SA23 and SA42 replicates indicates a mechanism that is likely disrupting the normal intron excision between exons 5 and 6. When the replicates are inputted into the UCSC genome browser as bigwig files for visualization, significant retention of DNA between exons 5 and 6 in

the mutant samples is visible that is not present in the wild type samples. This observation and site (red box below) is used to inform subsequent alternative splicing program analyses.

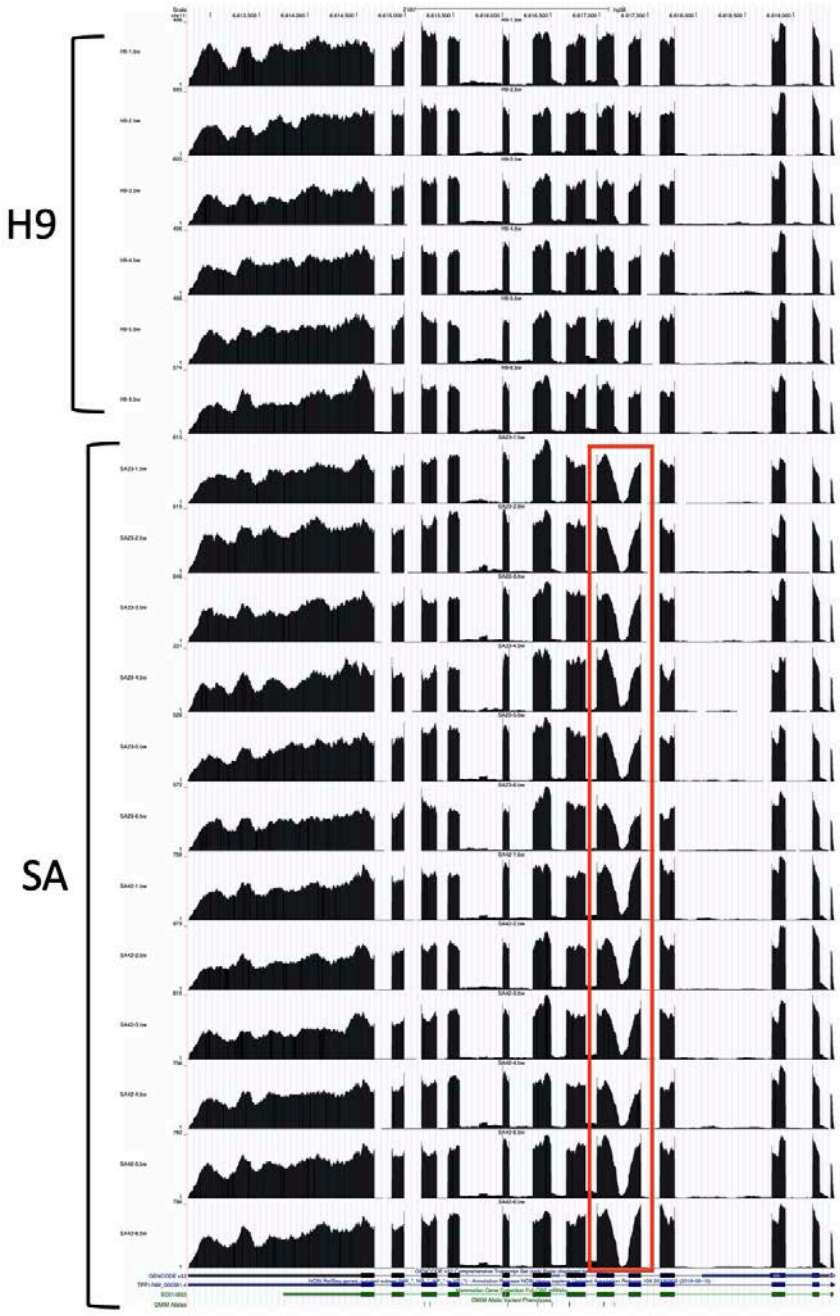



**Figure 7.** UCSC genome browser visualization of the TPP1 mutant samples using bigwig files. In the WT samples, exon 5 and exon 6 show two separate bars while the mutant samples (SA), with the red box indicated, show a valley shape between exon 5 and 6.

## Leafcutter: H9 vs. SA23

Leafcutter assigned 5 intron clusters to the TPP1 gene for the SA23 samples. The gene-level visualization (Figure 8) maps the clusters to their locations on the TPP1 gene. All TPP1 exons are plotted as black rectangles and taken from the provided annotation file. Each cluster is labelled with its change in percent spliced in (dPSI) value corresponding to its inclusion value. Significant splice junctions are colored according to their estimated dPSI values and according to their up or down regulation. Based on the gene-level visualization, clusters 48659 and 48658 were selected for further analysis due to their significant ($p < 0.05$) dPSI values: Cluster 48659 is significantly downregulated in mutant samples, while cluster 48658 is significantly upregulated. Clusters 48660 and 48657 were not up or down regulated, and therefore were not used in subsequent closer analysis.

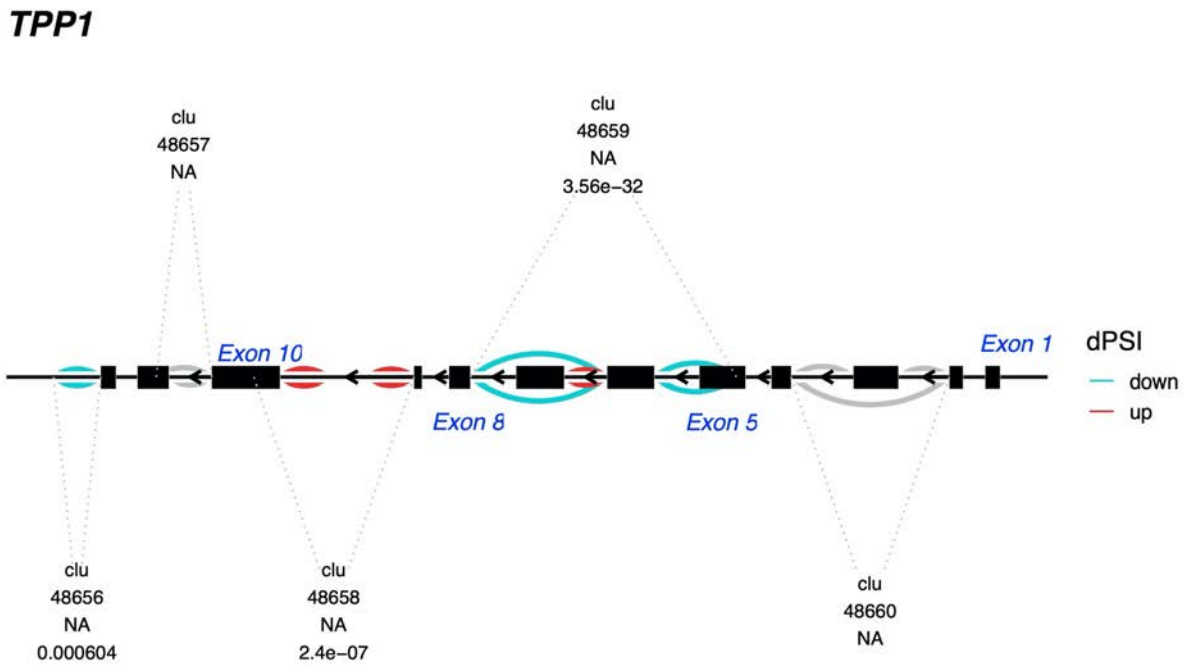
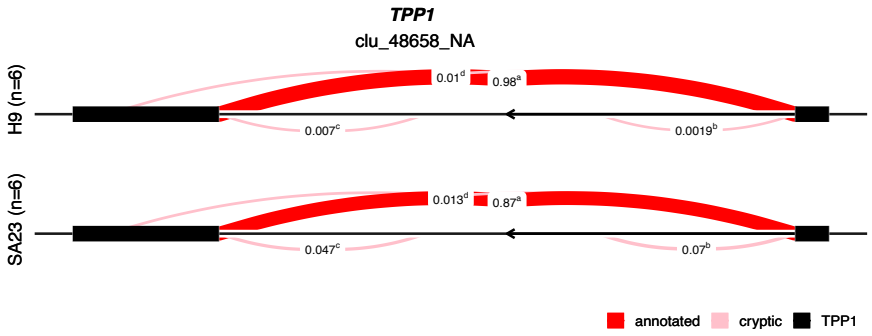



**Figure 8.** All clusters assigned by Leafcutter and Leafviz to the TPP1 gene with SA23 sample files. Exons are derived from the provided annotation and labelled for reference. Junctions are plotted as curved lines with uniform thickness. Significant junctions are colored based on estimated dPSI.

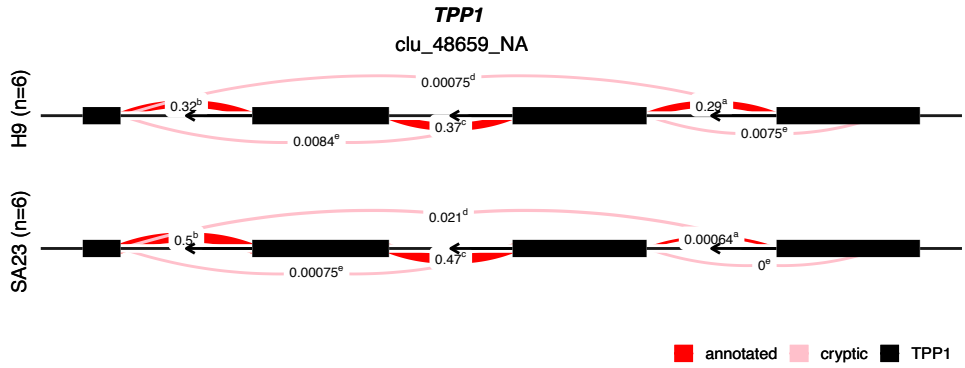
Cluster 48658 indicates an intron junction between exon 9 and exon 10 in TPP1 (Figure 9). Though annotated splice junction *a* between these two exons dominates, Leafcutter identified three significant alternative splicing junctions from genomic coordinates 6615542 to 6616005 that show increased percent spliced in for the SA mutant samples. These junctions, *b, c,* and *d* in the table below stem from two types of intron retention patterns, *cryptic_fiveprime*, in which the 3' splice site is found in the annotation but the 5' splice site is not, and *cryptic_threeprime*, in which the 5' splice site is found in the annotation but the 3' is not. Junction *b* begins at the annotated 5' end of exon 9 and ends at a cryptic 3' location in the middle of the annotated intron. Junction *c* begins at a cryptic 5' location within the annotated intron and ends at the annotated 3' start of exon 10. Junction *d* begins at the annotated 5' end of exon 9 and spans the entire length of the annotated intron, ending in the middle of exon 10 at a cryptic 3' location. These alternative junctions would produce alternative splicing variants and likely abnormal intron retention.




| | chr | start | end | verdict | dPSI |
|---|---|---|---|---|---|
| *a* | chr11 | 6615562 | 6616005 | annotated | −0.115 |
| *b* | chr11 | 6615848 | 6616005 | cryptic_threeprime | 0.067 |
| *c* | chr11 | 6615562 | 6615765 | cryptic_fiveprime | 0.041 |
| *d* | chr11 | 6615545 | 6616005 | cryptic_threeprime | 0.007 |

**Figure 9.** Analysis of intron cluster 48658 in Leafviz. The mean number of splice junctions supporting each intron is used to calculate a normalized fraction of the total counts. Introns are plotted as arcs connecting start and end coordinates with thicknesses proportional to the normalized count.

Cluster 48659 (Figure 10) spans the annotated region from exon 5 to exon 8 of the TPP1 gene. Leafcutter identified three annotated splice junctions at high absolute dPSI values in both samples. Notably, the annotated junction *a* has a negative dPSI value, indicating a significant decrease in this junction in the SA mutation lines. This is the junction between exons 5 and 6, and matches the retained intron region indicated in the bigwig files (Figure 7). In this cluster, two novel annotated pairs were also identified, in which both splice sites of these junctions are contained in the provided annotation file, but they are not annotated as being paired. Novel annotated pair junction *d* spans the end of exon 5 to the start of exon 8, removing exons 6 and 7 as a novel excised intron. This junction is significant because it increases percent splice in for the SA mutation lines. This region could be the origin of one of the observed splicing variants in SA mutants.




| | chr | start | end | verdict | dPSI |
|---|---|---|---|---|---|
| *a* | chr11 | 6617153 | 6617301 | annotated | −0.294 |
| *b* | chr11 | 6616503 | 6616661 | annotated | 0.185 |
| *c* | chr11 | 6616859 | 6616975 | annotated | 0.104 |
| *d* | chr11 | 6616503 | 6617301 | novel annotated pair | 0.020 |
| *e* | chr11 | 6616503 | 6616975 | novel annotated pair | −0.008 |
| *f* | chr11 | 6617153 | 6617359 | cryptic_fiveprime | −0.008 |

**Figure 10.** Cluster analysis of intron cluster 48659 in Leafviz.

## Leafcutter: SA42 vs. H9

An identical analysis pipeline for SA23 mutation lines was performed on the SA42 mutation lines. The gene-level visualization (Figure 11) shows significant clusters 49615, 49616, 49617, and 49618 on the TPP1 gene. Junctions in cluster 49618 are primarily downregulated while those in clusters 49616 and 49615 are primarily upregulated. Cluster 49617 contains both up and down regulated junctions.

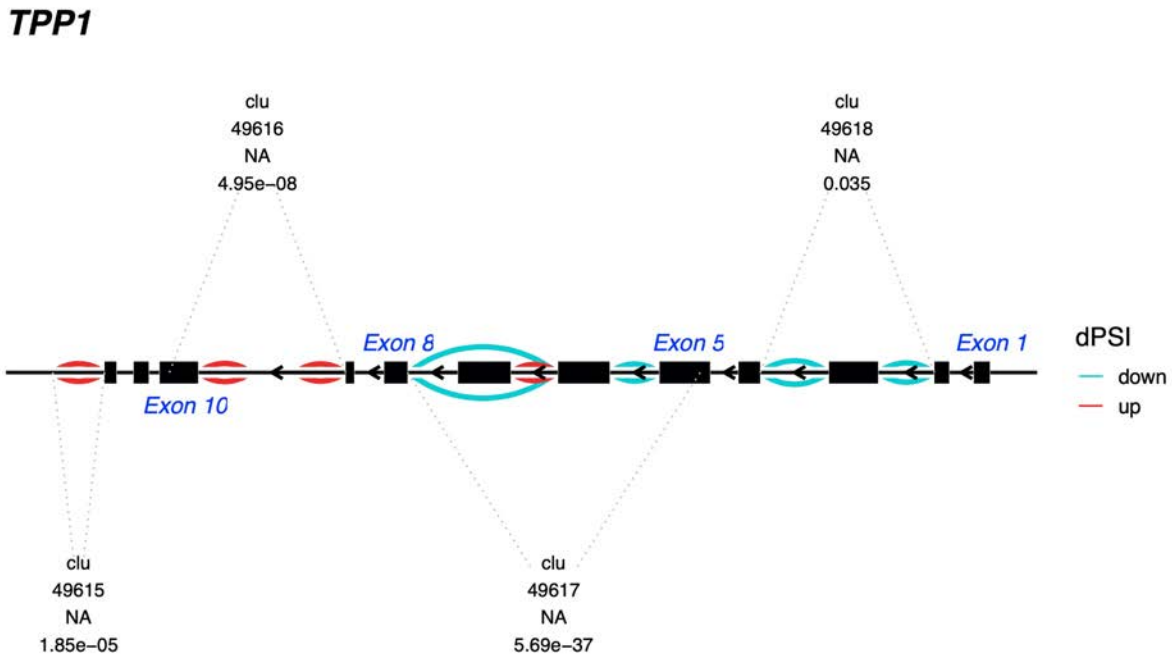
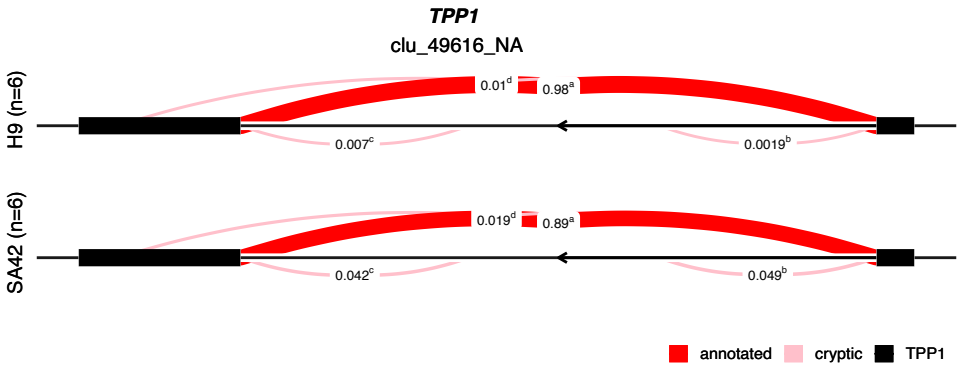



**Figure 11.** All clusters assigned by Leafcutter to the TPP1 gene with SA42. Exons are derived from the provided annotation. Junctions are plotted as curved lines with uniform thickness. Significant junctions are colored based on estimated dPSI.

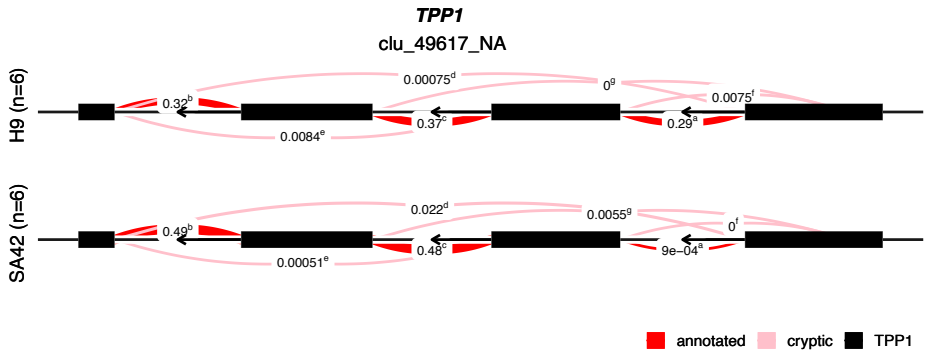
Cluster analysis of intron cluster 49616 in SA42 (Figure 12) shows that it spans the annotated exons 9 and 10 of the TPP1 gene. Leafcutter identified junctions in this region in the SA42 lines as it did in the SA23 lines (Figure 9). Specifically, presence of new junctions *b, c,* and *d* as cryptic 5' and 3' splice sites with positive dPSI values indicates a potential increase in the inclusion genomic regions that lie on the annotated intron between exons 9 and 10 in the mutant lines.




| | chr | start | end | verdict | dPSI |
|---|---|---|---|---|---|
| *a* | chr11 | 6615562 | 6616005 | annotated | −0.094 |
| *b* | chr11 | 6615848 | 6616005 | cryptic_threeprime | 0.047 |
| *c* | chr11 | 6615562 | 6615765 | cryptic_fiveprime | 0.035 |
| *d* | chr11 | 6615545 | 6616005 | cryptic_threeprime | 0.012 |

**Figure 12.** Cluster analysis of intron cluster 49616 from Leafviz.

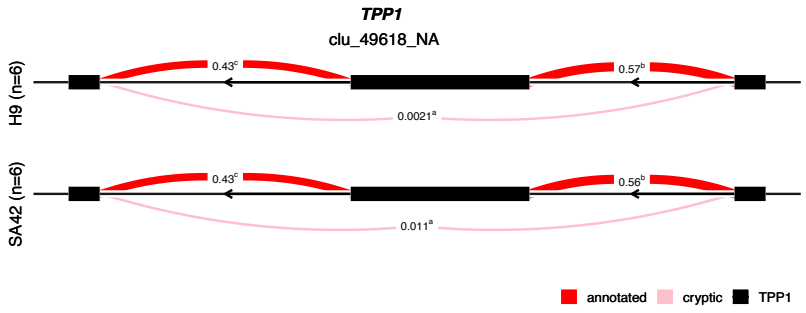
Cluster 49617 (Figure 13) spans exons 5 to 8 in the annotated TPP1 gene. The results indicate similar junctions as the SA23 cluster 48659 (Figure 10). This cluster also contains the region with the SA mutation at intron 5. Splice junction *f*, starting at chr11:66187153 begins next to the TPP1 SA mutation being studied and decreases from 0.0075 to 0 in the mutated replicates. Most significantly, the annotated junction *a* decreases in proportion on the SA mutant lines, with a dPSI of -0.294, which once again agree with the bigwig visualization showing higher intron retention in that region for SA mutants (Figure 7). The novel annotated pair junction *d* that spans from the end of exon 5 to the start of exon 8 increases in the SA mutant lines with dPSI of 0.022, supporting the skipping and excision of exons 6 and 7. Unlike with the SA23 cluster though, the SA42 cluster below identified a seventh possible splice junction, junction *g*, which spans from a cryptic 5' splice site just after exon 5 to an annotated 3' splice site on exon 7. The result of junction *g* is exon skipping of exon 6, in which the resulting transcript would splice out just exon 6 completely.




| | chr | start | end | verdict | dPSI |
|---|---|---|---|---|---|
| *a* | chr11 | 6617153 | 6617301 | annotated | −0.294 |
| *b* | chr11 | 6616503 | 6616661 | annotated | 0.170 |
| *c* | chr11 | 6616859 | 6616975 | annotated | 0.113 |
| *d* | chr11 | 6616503 | 6617301 | novel annotated pair | 0.022 |
| *e* | chr11 | 6616503 | 6616975 | novel annotated pair | −0.009 |
| *f* | chr11 | 6617153 | 6617359 | cryptic_fiveprime | −0.007 |
| *g* | chr11 | 6616859 | 6617359 | cryptic_fiveprime | 0.005 |

**Figure 13.** Cluster analysis of intron cluster 49617.

Cluster 49618 (Figure 14) spans exons 2 to 4 in the TPP1 gene. Unlike the results for the SA23 clones in which this region was not up or down regulated, the alternative splicing variant in this region was upregulated in SA42 mutants. The leafcutter results indicate a novel annotated pair, junction *a*, that increases in abundance in the SA42 mutant with a dPSI of 0.009. The annotated junctions, junctions *b* and *c*, show a negative dPSI, indicating that the typical splicing could be disrupted by the novel annotated pair splicing in the mutant cells. Junction *a* splicing eliminates exon 3 in the final transcript of SA42 mutants.




| | chr | start | end | verdict | dPSI |
|---|---|---|---|---|---|
| *a* | chr11 | 6617776 | 6619196 | novel annotated pair | 0.009 |
| *b* | chr11 | 6618915 | 6619196 | annotated | −0.005 |
| *c* | chr11 | 6617776 | 6618776 | annotated | −0.004 |

**Figure 14.** Cluster analysis of intron cluster 49617.

PCA plots (Figure 15) were produced by Leafcutter for the H9 control samples as compared to both the SA23 and the SA42 samples. Clustering of the H9 apart from these SA clusters is expected and confirms the expected dissimilarity of the two groups.

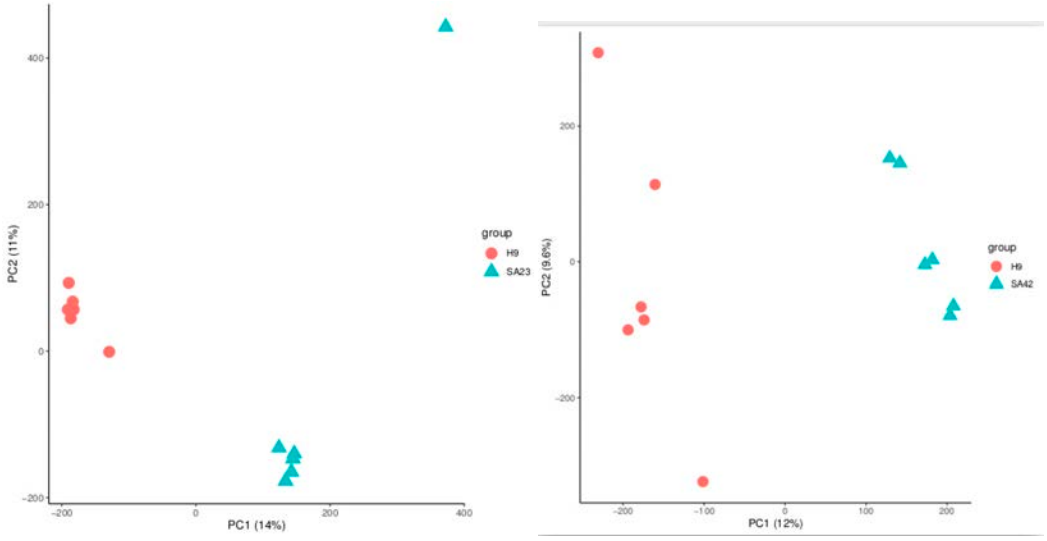



**Figure 15.** PCA plots for SA23 and SA42 based on splice junction counts.

## rMATS

For each analysis, rMATS generates an output file corresponding to each alternative splicing event type defined by the program. The [AS_Event].MATS.JC.txt files counts reads that span each alternative splicing event junction, and identified alternative splicing events of types Exon Skipping (ES) and Mutually Exclusive Exons (MXE) within the TPP1 gene of chromosome 11. Other alternative splicing event types were not identified within the TPP1 region and therefore not used for further visualization in rmats2sashimiplot.

## rMATS: H9 vs SA23

rMATS identified 14 distinct exon skipping events on the TPP1 gene based on the H9 and SA23 sample groups and rmats2sashimiplots produced visualizations for each of these events. The graphs below highlight three distinctive and relevant locations of the TPP1 gene for further analysis, based on the presence of significant differences in the junctions and the location of the junction. Because Leafcutter identified significant alternative splicing variation within exons 5 and 8 of TPP1 in both SA23 and SA42 groups, these three following visualizations focus on this region of TPP1.

The sashimi plot visualizations generated by rmats2sashimiplots display the splicing junctions on all six H9 control samples and all six SA mutant samples. The Sashimi plot is plotted against an axis of calculated, modified RPKM, as the read density. The exact calculation is performed using the equation:

$$rmats2sashimiplot = \frac{numReads}{\frac{queryLength}{1{,}000} \times \frac{totalNumRead}{1{,}000{,}000}}$$

In the following visualization (Figure 16), rMATS identified potential skipping of exon 5. However, the most interesting aspect of this segment is the minimal junction between exons 5 and 6 in the SA mutants. This is indicative of an intron retention, in which the intron between exons 5 and 6 is preserved in the final transcript. This pattern of intron inclusion supports the read depth shown inclusion

(Figure 7) produced by UCSC genome browser bigwig file visualization of the H9 and SA files before alternative splicing analysis.

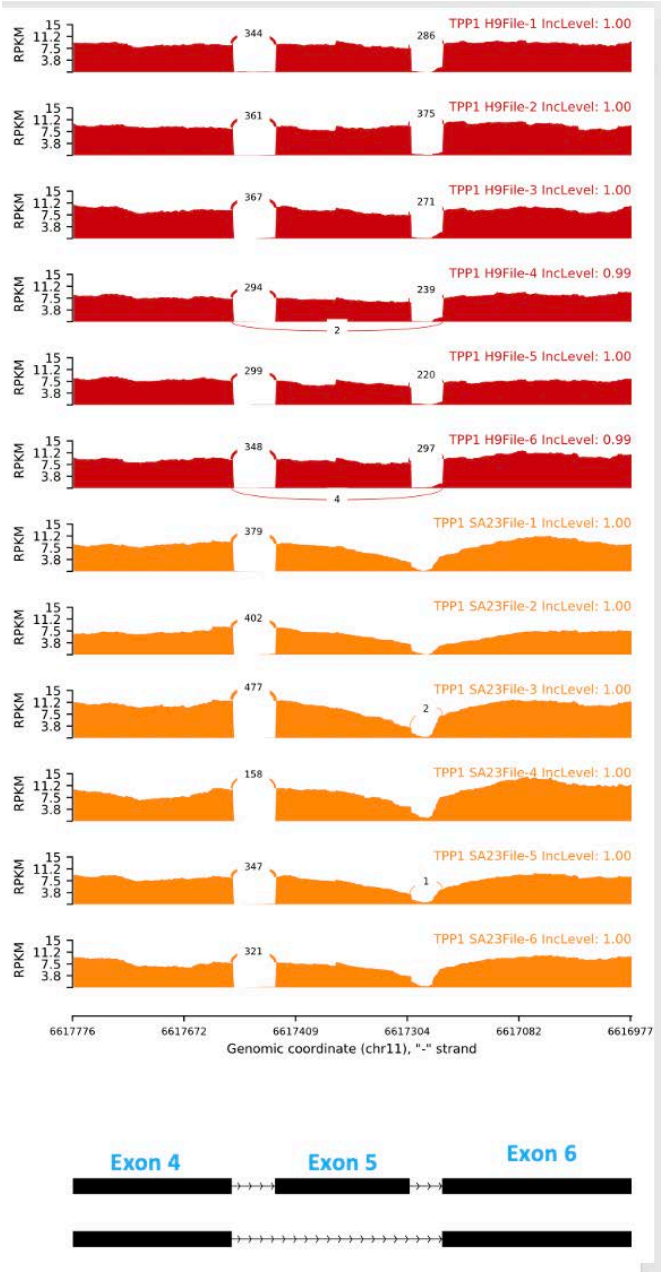



**Figure 16.** rMATS visualization of skipped exon event involving skipping of exon 5 and elimination of the splice junction from exon 5 to exon 6.

rMATS identified a junction involving exon skipping of exons 6 and 7. The splice junction spans from exon 5 to exon 8 in TPP1 and is similar to the junction identified in clusters 49617 and 48659 by Leafcutter (Figures 10 and 13). The exon skipping event (Figure 17) shows an increase in a splice junction that extends from the end of exon 5 to the start of exon 8, in which exons 6 and 7 would be spliced out of the final transcript.

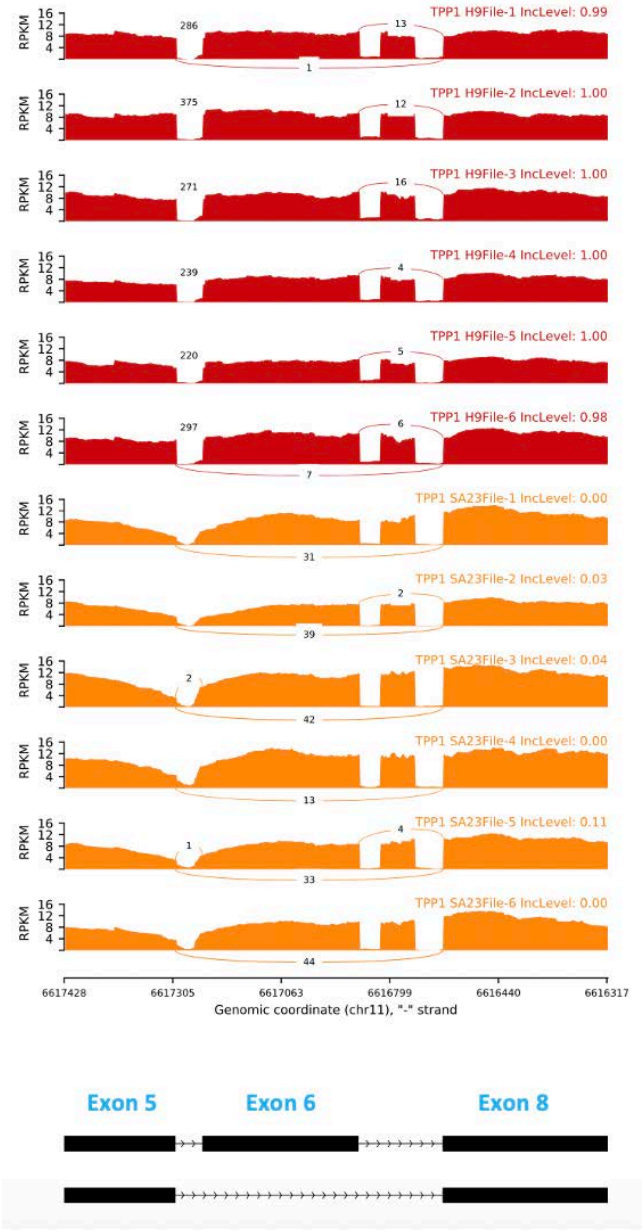



**Figure 17.** rMATS visualization of skipped exon event involving skipping of exon 6 and exon 7.

Another second skipped exon junction produced by rmats2sashimiplots continues to support the junction between exon 5 and 8 in the SA23 file samples (Figure 18). As the junction between the two exons is mapped again in the visualization. This visualization additionally indicates even greater potential exon skipping in the SA23 samples from the end of exon 5 to the start of exon 9, supported primarily by replicates 2 and 6. This alternative splicing event would cause three TPP1 exons, 6, 7, and 8 to all be spliced out in SA mutants.

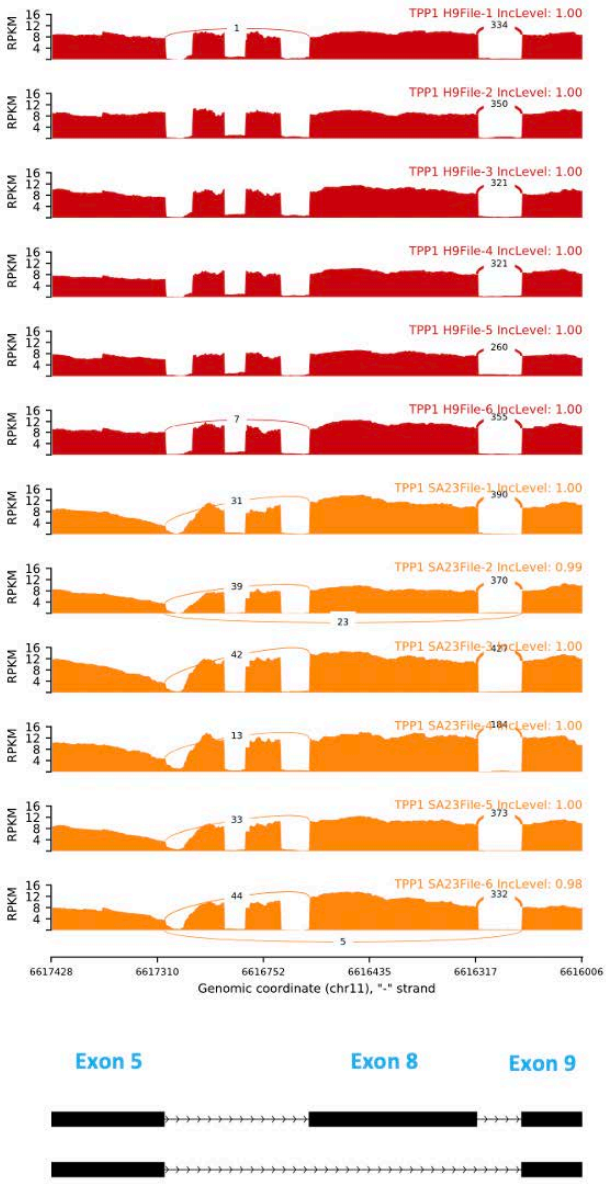



**Figure 18.** rMATS visualization of skipped exon event involving skipping of exon 6, exon 7, and exon 8.

rMATS identified 3 instances of mutually exclusive exon alternative splicing events. The most greatly supported of these events is shown in Figure 19 below. This event supports an alternative junction from exon 6 to 8 in the control H9 replicates at a higher degree than skipping of exon 7 in the SA replicates.

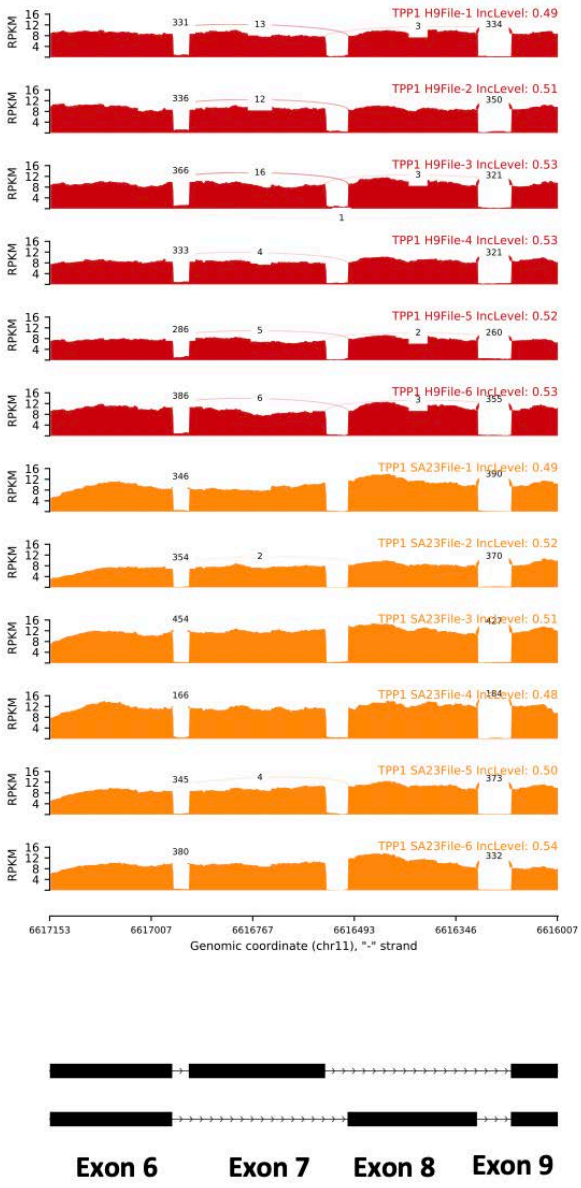



**Figure 19.** rMATS visualization of Mutually Exclusive Exon event in SA23 replicates indicating a greater skipping of exon 7 in the H9 replicates as opposed to the SA23 replicates.

## rMATS: H9 vs SA42

rMATS identified 17 distinct exon skipping event visualizations based on the H9 and SA42 sample groups. The data below highlights some of the most relevant locations of the TPP1 gene for further analysis, based on the significant differences in alternative splicing at junctions in the SA mutant lines and the location of the splicing event. Similar to SA23, because Leafcutter identified significant alternative splicing variation with exons 5 and 8 of TPP1, the following visualizations have been selected to focus on this region of TPP1.

Figure 20 supports exon skipping from the end of the TPP1 exon 5 to the start of TPP1 exon 8, in which exons 6 and 7 would be removed from the final transcript.

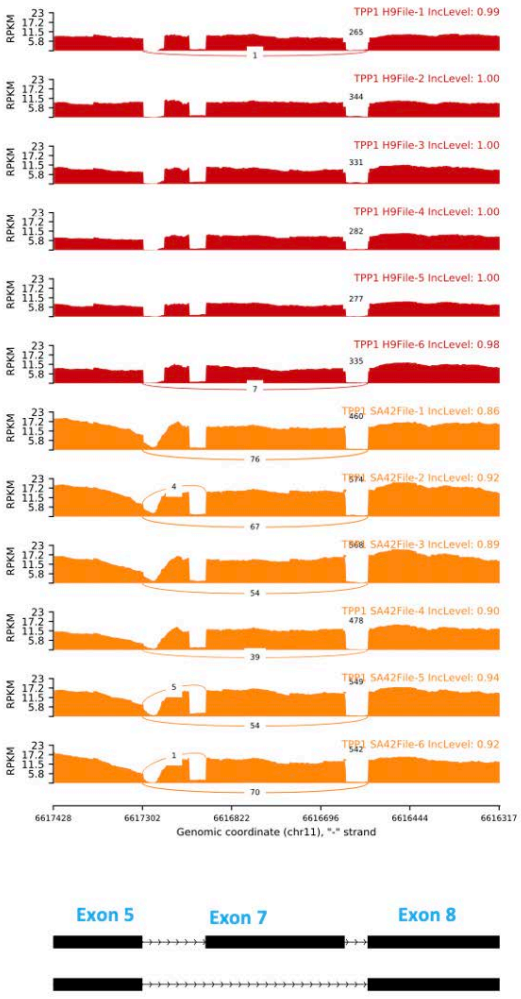



**Figure 20.** rMATS visualization of skipped exon event involving skipping of exon 6 and exon 7 in SA42.

Figure 21 supports exon skipping from exon 5 to exon 9 of TPP1, primarily by the SA42-1 file which has a read count of 30 in support of this junction. This skipped exon event appears in the SA23 sample group as well, shown previously in Figure 18.

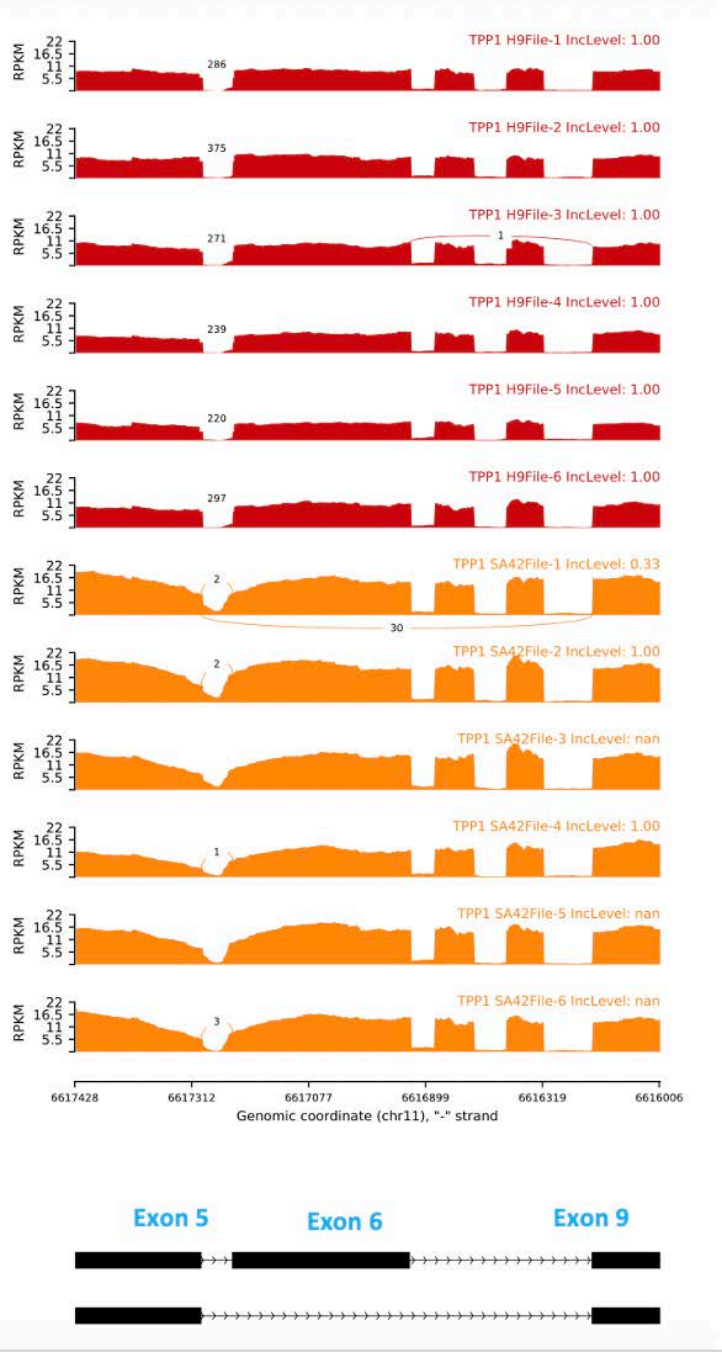



**Figure 21.** rMATS visualization of skipped exon event involving skipping of exon 6, exon 7, and exon 8 in SA42 samples.

rMATS also identified 6 Mutually Exclusive Exon alternative splicing events on TPP1 in the SA42 sample group. Of these, the most differentially spliced event is shown below in Figure 22. This visualization highlights the inclusion or exclusion of exons 6 and 7, indicating the previously observed splicing event from exon 5 to 8 and exon skipping of exons 6 and 7 is likely, as well as a higher number of instances where exon 7 is skipped in the H9 controls. This indicates, perhaps, that the most novel and characteristic aspect of the alternative splicing in the SA group is the skipping and splicing out of exon 6.

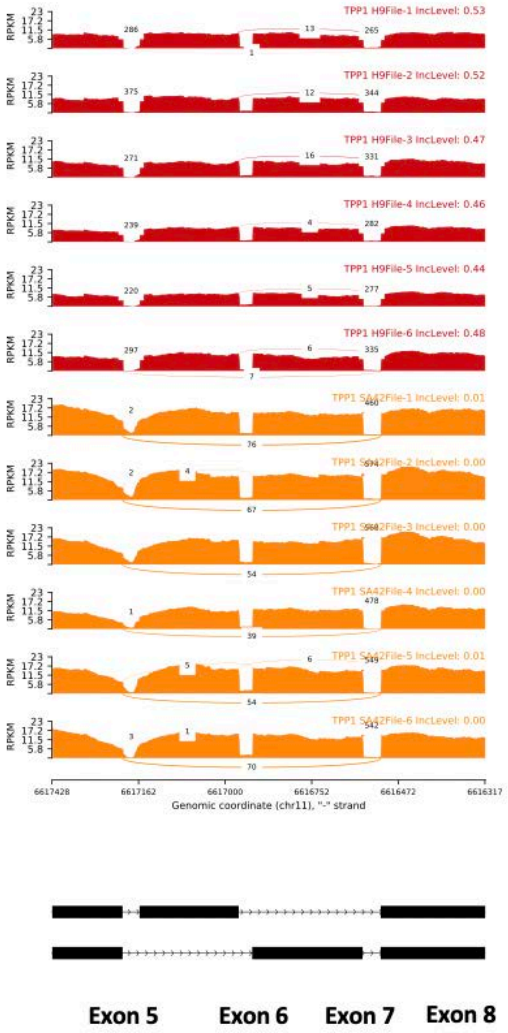



**Figure 22.** Mutually Exclusive Exon event identified by rMATS for the SA42 sample group which indicates skipping of exons 6 and/or 7. The H9 control replicates appear to retain all exons or skip only exon 7, whereas the SA42 replicates skip both exons 6 and 7 to a higher proportion.

### ASGAL

ASGAL bed files indicate exon skipping events in the exon 5 to exon 8 region, as identified through Leafcutter and rMATS. In addition, ASGAL files support the intron retention around exon 5 shown initially in the bigwig files.

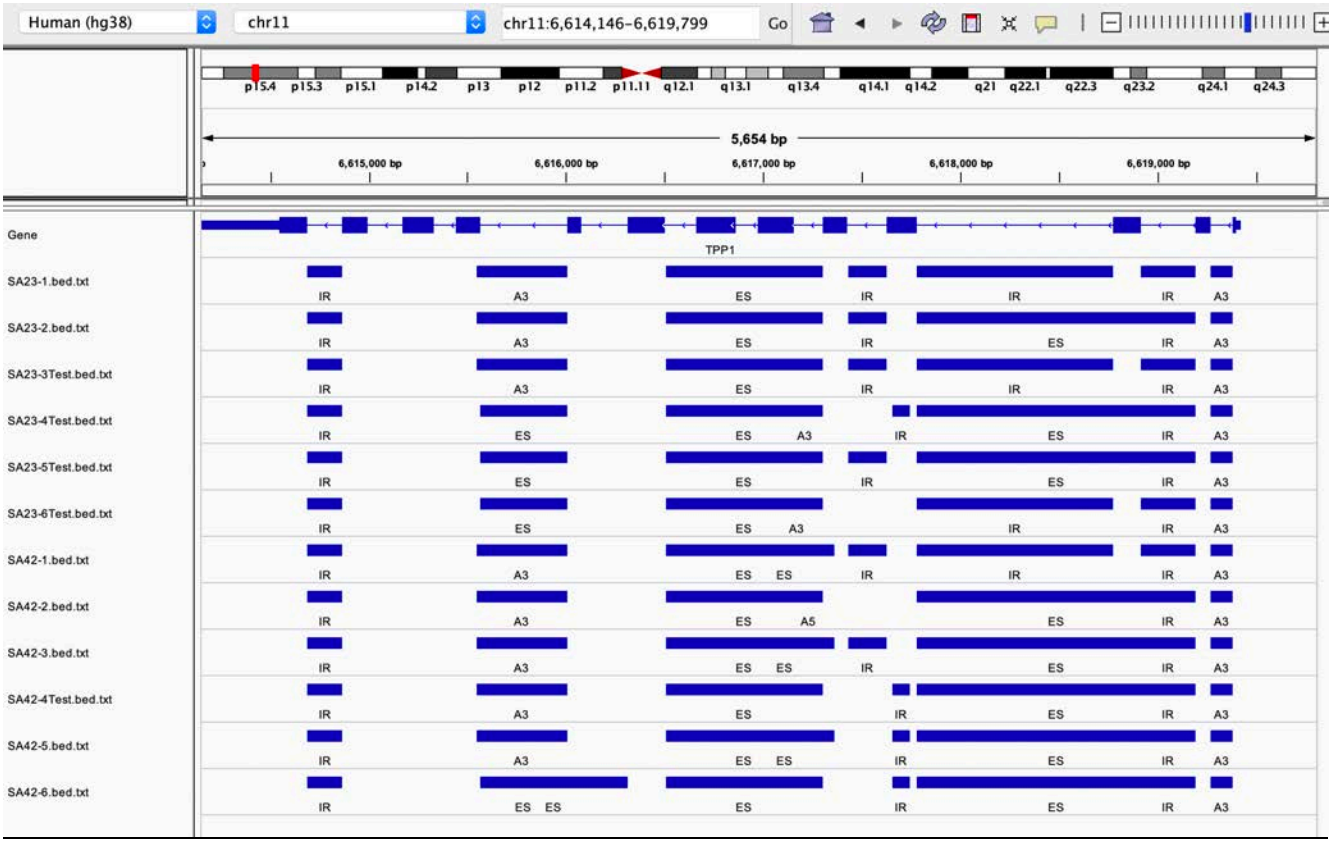



**Figure 23.** ASGAL identified alternative splicing exon skipping events (ES), intron retention events (IR), and alternatively spliced 5' and 3' sites (A5 and A3) along the TPP1 gene.

# DISCUSSION

The aim of this work was to perform alternative splicing analysis on the TPP1 gene using differential splicing software. Based on this work, several alternative splicing patterns were identified on

the TPP1 splice acceptor mutants and several of these patterns were further supported as they occurred more than once across different SA subclones and/or different alternative splicing software methods. The intronic splice-site, SA mutation studied in the analysis appears at high incidence in individuals with CLN2 Batten disease and is located on the 3' splice acceptor site of TPP1 intron 5 and at the genomic coordinate of chr11:6617154.

Leafcutter and rMATS were the primary software programs used in this analysis to study the impact of the SA mutation on splicing patterns of the TPP1 gene. All in all, Leafcutter and rMATS analysis both indicate that there is potential exon skipping in the presence of the splice acceptor mutation. The exon skipping event identified most commonly across both alternative splicing programs was the TPP1 exon 5 being spliced into TPP1 exon 8, which results in skipping of exon 6 and exon 7 in TPP1. Both SA23 and SA42 subclones produced this exon skipping pattern. ASGAL also identified exon skipping events in this region, further confirming the analysis. rMATS also identified a longer skipped exon event that may splice from exon 5 into exon 9, but this event was not highlighted in prior Leafcutter analysis. The exon skipping in this region generates in-frame mutant products which can result in a loss of function mutation, as seen in the TPP1 gene of CLN2 Batten disease.

Notably, variant 2 in the western blot of the TPP1 SA mutants in Figure 4 corresponds in size with this splicing isoform that result in skipped of exon 6 and 7. However, the size of the gene region that would need to be excised from the isoform to produce the other splicing variant, variant 1, is smaller than any complete full skipped exon and likely is the result of partial skipping of an exonic or intronic region of TPP1. Potentially due to the methods of the software programs, or the manner in which the RNA sequencing was obtained, Poly(A)-seq, potential isoforms corresponding with variant 1 were not clearly identified through this initial analysis.

Additional regions of the TPP1 gene may undergo alternative splicing and include the region from exons 2 and 4, the region from exons 9 to 10, and the region after at exon 12. The alternative splicing of these regions occurs at a lower proportion and has lower difference in percent spliced in (dPSI) when compared to controls than the recurring pattern observed from the exon 5 to 8 skipping

region. However, these all are splicing regions on the TPP1 gene selected to be intron clusters by Leafcutter, indicating up or down regulation of these patterns in both SA23 and SA42 mutants. rMATS supports the alternative splicing of these intron cluster locations identified by Leafcutter as well for the regions from exon 2 to 4 and exon 9 to 10. Though rMATS identified a vast majority of its alternative splicing events within the exon 5 to 11 region, there were 2 potential skipped exon patterns in each of the 14 and 17 total distinct exon skipping events identified from SA23 and SA42 mutants, respectively, that fell in the exon 2 to 4 region to support the splicing originally identified by Leafcutter. rMATS uses flanking exons in its methods and is therefore blind to events involving the first or last exon of the gene. This is a limitation to the rMATS program for analyses of alternative splicing that may involve these regions of the gene and is a potential cause for the absence of events following exon 12 in rMATS outputs.

Analysis of genome-wide short read RNA-Sequencing data is computationally intensive. Accordingly, all tools used in this analysis were run on a computer cluster with a Redhat 7.3 operating system and on Oscar (Ocean State Center for Advanced Resources), Brown University's supercomputer, maintained and supported by the Center for Computation and Visualization (CCV). The figures below show the time required to run rMATS and Leafcutter with the corresponding values listed in the table below.

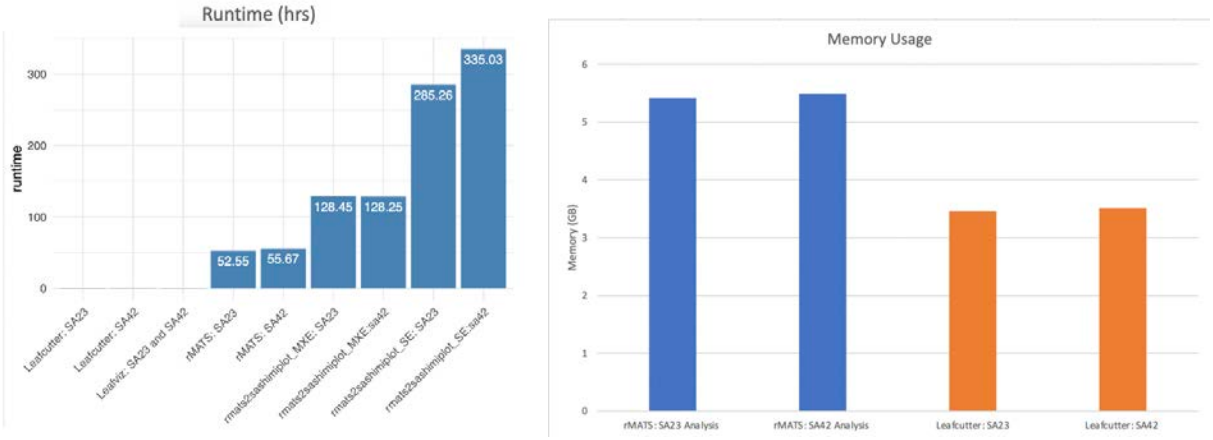



**Figure 24.** Runtime and memory usage of Leafcutter and rMATS

| Job | Job Clock<br><br>Wall time | Memory<br><br>Utilized | Notes |
|---|---|---|---|
| rmats2sashimiplot: SA23 MXE<br><br>event | 5-8:00:27 | 1.12 GB | Full visualization plots would have taken longer to produce. The job exited due to timeout as it did not complete visualizations for entire genome in allocated time, but did complete necessary visualizations on the TPP1 gene for this analysis |
| rmats2sashimiplot: SA42 MXE event | 5-8:00:27 | 1.09 GB | Full visualization plots would have taken longer to produce. The job exited due to timeout as it did not complete visualizations for entire genome in allocated time, but did complete necessary visualizations on the TPP1 gene for this analysis |
| Rmats2sashimiplot: SA23 SE event | 11-21:16:57 | 1.03 GB | Full visualization plots would have taken longer to produce. The job exited due to timeout as it did not complete visualizations for entire genome in allocated time, but did complete necessary visualizations on the TPP1 gene for this analysis |
| Rmats2sashimiplot: SA42 SE event | 13-23:02-25 | 1.02 GB | Full visualization plots would have taken longer to produce. The job exited due to timeout as it did not complete visualizations for entire genome in allocated time, but did complete necessary visualizations on the TPP1 gene for this analysis |
| rMATS: SA23 analysis | 2-04:13:45 | 5.41 GB | |
| rMATS: SA42 analysis | 2-07:40:44 | 5.49 GB | |
| Leafcutter differential splicing analysis step: SA23 | 00:22:10 | 3.46 GB | |
| Leafcutter differential splicing analysis step: SA23 | 00:30:55 | 3.51 GB | |
| Leafviz: SA23 and SA42 | < 60 s | N/A | Leafviz loads a Shiny app popup in the browser |

**Table 1.** Leafcutter and rMATS runtimes and memory usage for both the differential analysis component and the visualization component.

Based on the results show in the table and plots, Leafcutter shows a significant advantage in both runtime and memory usage for both the differential splicing step and the visualization software. Though Leafcutter required a few more steps and commands for the user to implement, the Leafcutter package is very clearly documented and user-friendly. In contrast, the rMATS software requires only one primary script but at the expense of a more extensive runtime and memory allocation. Though all steps are documented as well with rMATS, the installation of the program was more involved and required more time to implemented, based on the experience in implementing these alternative splicing tools for this analysis.

A diverse selection of alternative splicing programs and reconstruction methods exist, but there are limitations to many of the available tools. Selection of alternative splicing tool depends on the goal of analysis and computational capability. For example, ASGAL focuses heavily on identifying novel splice sites. MISO, while efficient and offering high complexity and extensive visualization modeling, cannot handle replicates or groups of samples. CuffDiff2 can return a list of differentially spliced genes, associated p values, and false discovery corrections, but does not indicate any of the exons or junctions involved. When computational performance is of concern, Leafcutter, the newer program, is highly recommended for efficient alternative splicing analysis. Whippet, also a newer program, can similarly provide fast and efficient alternative splicing analysis on laptop-compatible hardware. rMATS is more computationally expensive, but allows for robust analysis that simultaneously models variability among replicates and estimates uncertainty of isoform proportion in individual replicates. rMATS has high precision at the expense of some sensitivity. Another weakness in the current features of the Leafcutter and rMATS software programs is a lack of flexibility in localizing the alternative splicing event analysis to process only single chromosomes or genes, like the TPP1 gene in this case. Specifically, both tools and their corresponding visualization software perform analysis on all genes across all chromosomes of the annotated genome, which adds greatly to their expense both computationally and in memory allocation. The TPP1 gene is only located on a short region of a single chromosome, chromosome 11, and is the most relevant region initially in this the study of the splice acceptor TPP1 mutation in CLN2 Batten disease. However, analysis of all other regions in the genome had to be performed in the job scripts during the splicing analysis despite this localized region.

A major challenge to alternative splicing analysis is accurate reconstruction of full-length isoforms. In order to be certain that all possible isoforms are considered, either full knowledge of all possible isoforms or full-length RNA-Sequenced reads must be available. All possible alternatively spliced isoforms are often unknown, and the identification of these isoforms is the end goal of many alternative splicing software. In addition, to this date, given current computational methods and sequencing costs which do not allow for extensive full-length RNA-Sequencing, statistically robust

detection of differential splicing is biased toward the more abundant transcripts [56]. This problem arises in part due to the fact that, as the number of possible isoforms increases, the number of degrees of freedom and confidence intervals increase as well. Many genes are subsequently filtered out due to low coverage or due to wide confidence intervals of isoform prediction, contributing to the bias for more abundant transcripts and absence of the detection of potentially significant isoforms.

Sensitivity is another major challenge in alternative splicing software and other RNA-Sequencing analysis methods. This problem can be partially overcome by increasing the RNA-Sequencing read depth or length [57]. Other alternatives to improve upon the problem of read depth and sensitivity include using a secondary analysis based on pooled data in samples or conditions that show strong correlation to inform and improve read depth. Another alternative is to use paired end read data that allows for analyses to infer greater information about exon usage [56].

One aspect of the alternative splicing analysis pipeline to consider is the alignment program used prior to the differential splicing analysis tool. The software in this analysis were all run based on alignments that HISAT2 produced. The most common tools used for genome alignment are Tophat, STAR and HISAT. Earlier pipelines, including that in the protocol paper for the alternative splicing program Cufflinks, call for alignment using Tophat or Tophat2. However, the computational inefficiency of this program has made it transition into becoming an obsolete tool and brought it recently to a low maintenance and low support stage. It is now superseded by HISAT2, an alignment program that follows similar core functionality to Tophat, but produces alignments in a much more efficient and accurate manner. Though HISAT2 uses fewer resources than an aligner like STAR, a broader range of alternative splicing software may have been applicable if another aligner like STAR had been used, since some current alternative splicing software methods are only compatible with alignment output from one or another aligner.

Further research on the TPP1 gene in these SA mutants should be performed to characterize and confirm potential CLN2 disease alternative splicing patterns. Other differential splicing programs with alternate approaches should be run on the SA mutants to identify new splicing patterns or confirm those

identified by rMATS and Leafcutter. Another useful direction to pursue is to run another alignment program on the RNA-Sequencing files. If a program, like the STAR aligner, were used, a greater range of alternative splicing software will be compatible and can be implemented for future analysis. Finally, it may be worthwhile to investigate the alternative splicing patterns of similar genes involved in CLN2 Batten disease or other genes linked with TPP1 activity to achieve a broader analysis of the alternative splicing patterns involved in CLN2 Batten disease.

# CONCLUSION

Overall, alternative splicing patterns are an important regulatory mechanism in understanding the cellular and functional complexity of cells. In this analysis, alternative splicing patterns were characterized in human ESC clones induced with a high incidence splice acceptor mutation derived from CLN2 Batten disease. Alternative splicing tools, including Leafcutter, rMATS and ASGAL suggest that significantly alternatively excised exons within the exon 5 to exon 8 region of the TPP1 gene may be at play. In addition, numerous other exon skipping events and mutually exclusive exon events on the TPP1 region were identified. Significant intron retention was also seen across files between exons 5 and 6, just prior to the location of the splice acceptor mutation on TPP1. Alternative splicing is key to understanding the role of TPP1 and this splice acceptor mutation in the overarching CLN2 disease mechanism and progression. These results may help guide further research into CLN2 Batten disease and provide a gateway to therapeutic pathways for the disease.

# ACKNOWLEDGEMENTS

I would like to thank my mentors Eric Morrow and Ece Uzun for their invaluable guidance and support throughout the past year. I would like to thank Qing Wu for providing data and guidance in the software analysis to make this project possible. I also acknowledge Li Ma for her review and feedback, and work

in providing the TPP1-mutant stem cell lines. I would like to thank the Uzun lab group for our weekly discussions. I would like to thank my academic advisor and second reader, Sorin Istrail, for his continuous support and mentorship. I would also like to thank Brown's Center for Computation and Visualization and Brown's Center for Computational Molecular Biology. I would like to acknowledge the support of the Sandra Nusinoff Lehrman and Stephen Lehrman Family Fund provided to the Morrow Laboratory at Brown University that led to the development of the TPP1-mutant stem cell lines.

**KEYWORDS**

NCL = neuronal ceroid lipofuscinoses

CLN2 = ceroid lipofuscinosis, neuronal 2

ESC = embryonic stem cells

TPP1 = tripeptidyl peptidase 1

SA = splice acceptor

PSI = percent spliced in

FDR = false discovery rate

**REFERENCES**

1. Johnson, T.B., et al., *Therapeutic landscape for Batten disease: current treatments and future prospects.* Nat Rev Neurol, 2019. **15**(3): p. 161-178.
2. Mole, S.E., *Neuronal ceroid lipofuscinoses (NCL).* Eur J Paediatr Neurol, 2006. **10**(5-6): p. 255-7.
3. Goebel, H.H. and K.E. Wisniewski, *Current state of clinical and morphological features in human NCL.* Brain Pathol, 2004. **14**(1): p. 61-9.
4. Tyynela, J., et al., *Storage of saposins A and D in infantile neuronal ceroid-lipofuscinosis.* FEBS Lett, 1993. **330**(1): p. 8-12.

5. Elleder, M., et al., *Neuronal ceroid lipofuscinosis in the Czech Republic: analysis of 57 cases. Report of the 'Prague NCL group'.* Eur J Paediatr Neurol, 1997. **1**(4): p. 109-14.

6. Jalanko, A. and T. Braulke, *Neuronal ceroid lipofuscinoses.* Biochim Biophys Acta, 2009. **1793**(4): p. 697-709.

7. Persaud-Sawin, D.A., et al., *Neuronal ceroid lipofuscinosis: a common pathway?* Pediatr Res, 2007. **61**(2): p. 146-52.

8. Mole, S.E., *Batten disease: four genes and still counting.* Neurobiol Dis, 1998. **5**(5): p. 287-303.

9. Mole SE, G.E., Schulz A, Xin WW, *Molecular basis of CLN2 disease: A review and classification of TPP1 gene variants reported worldwide.* . Molecular Genetics and Metabolism, 2018. **123(2)**.

10. Haltia, M., *The neuronal ceroid-lipofuscinoses.* J Neuropathol Exp Neurol, 2003. **62**(1): p. 1-13.

11. Schulz, A., et al., *NCL diseases - clinical perspectives.* Biochim Biophys Acta, 2013. **1832**(11): p. 1801-6.

12. Williams, R.E. and S.E. Mole, *New nomenclature and classification scheme for the neuronal ceroid lipofuscinoses.* Neurology, 2012. **79**(2): p. 183-91.

13. Mole, S., *NCL mutation and patient database. Retrieved from https://www.ucl.ac.uk/ncl-disease/mutation and patient database*. 2017.

14. Mole, S.E. and S.L. Cotman, *Genetics of the neuronal ceroid lipofuscinoses (Batten disease).* Biochim Biophys Acta, 2015. **1852**(10 Pt B): p. 2237-41.

15. Neverman, N.J., et al., *Experimental therapies in the neuronal ceroid lipofuscinoses.* Biochim Biophys Acta, 2015. **1852**(10 Pt B): p. 2292-300.

16. Sands, M.S., *Considerations for the treatment of infantile neuronal ceroid lipofuscinosis (infantile Batten disease).* J Child Neurol, 2013. **28**(9): p. 1151-8.

17. Junaid, M.A., et al., *A novel assay for lysosomal pepstatin-insensitive proteinase and its application for the diagnosis of late-infantile neuronal ceroid lipofuscinosis.* Clin Chim Acta, 1999. **281**(1-2): p. 169-76.

18. Specchio, N., et al., *Photosensitivity is an early marker of neuronal ceroid lipofuscinosis type 2 disease.* Epilepsia, 2017. **58**(8): p. 1380-1388.

19. Nickel, M., et al., *Disease characteristics and progression in patients with late-infantile neuronal ceroid lipofuscinosis type 2 (CLN2) disease: an observational cohort study.* Lancet Child Adolesc Health, 2018. **2**(8): p. 582-590.

20. Sharp, J.D., et al., *Loci for classical and a variant late infantile neuronal ceroid lipofuscinosis map to chromosomes 11p15 and 15q21-23.* Hum Mol Genet, 1997. **6**(4): p. 591-5.

21. Gardner, E., et al., *Mutation update: Review of TPP1 gene variants associated with neuronal ceroid lipofuscinosis CLN2 disease.* Hum Mutat, 2019. **40**(11): p. 1924-1938.

22. Ma, L.P., A.; Schmidt, M.; Morrow, E.; , *Generation of Pathogenic TPP1 Mutations in Human Stem Cells as a Model for CLN2 Disease.* bioRxiv, 2021.

23. Chow, L.T., et al., *An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA.* Cell, 1977. **12**(1): p. 1-8.

24. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.

25. Alamancos, G.P., E. Agirre, and E. Eyras, *Methods to study splicing from high-throughput RNA sequencing data.* Methods Mol Biol, 2014. **1126**: p. 357-97.

26. Deciphering Developmental Disorders, S., *Prevalence and architecture of de novo mutations in developmental disorders.* Nature, 2017. **542**(7642): p. 433-438.

27. Cooper, D.N., *Human gene mutations affecting RNA processing and translation.* Ann Med, 1993. **25**(1): p. 11-7.

28. Wang, Y., et al., *Mechanism of alternative splicing and its regulation.* Biomed Rep, 2015. **3**(2): p. 152-158.

29. Kim, E., A. Magen, and G. Ast, *Different levels of alternative splicing among eukaryotes.* Nucleic Acids Res, 2007. **35**(1): p. 125-31.

30. Shen, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proc Natl Acad Sci U S A, 2014. **111**(51): p. E5593-601.

31. Mehmood, A., et al., *Systematic evaluation of differential splicing tools for RNA-seq studies.* Brief Bioinform, 2020. **21**(6): p. 2052-2065.

32. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nat Biotechnol, 2013. **31**(1): p. 46-53.

33. Hu, Y., et al., *DiffSplice: the genome-wide detection of differential splicing events with RNA-seq.* Nucleic Acids Res, 2013. **41**(2): p. e39.

34. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.

35. Trincado, J.L., et al., *SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions.* Genome Biol, 2018. **19**(1): p. 40.

36. Vaquero-Garcia, J., et al., *A new view of transcriptome complexity and regulation through the lens of local splicing variations.* Elife, 2016. **5**: p. e11752.

37. Sterne-Weiler, T., et al., *Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop.* Mol Cell, 2018. **72**(1): p. 187-200 e6.

38. Zhu, D., N. Deng, and C. Bai, *A generalized dSpliceType framework to detect differential splicing and differential expression events using RNA-Seq.* IEEE Trans Nanobioscience, 2015. **14**(2): p. 192-202.

39. Anders, S., A. Reyes, and W. Huber, *Detecting differential usage of exons from RNA-seq data.* Genome Res, 2012. **22**(10): p. 2008-17.

40. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

41. Hartley, S.W. and J.C. Mullikin, *Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq.* Nucleic Acids Res, 2016. **44**(15): p. e127.

42. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

43. Li, Y.I., et al., *Annotation-free quantification of RNA splicing using LeafCutter.* Nat Genet, 2018. **50**(1): p. 151-158.

44. Liu, R., A.E. Loraine, and J.A. Dickerson, *Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems.* BMC Bioinformatics, 2014. **15**: p. 364.

45. Denti, L., et al., *ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events.* BMC Bioinformatics, 2018. **19**(1): p. 444.

46. Beretta S, B.P., Denti L, Previtali M, Rizzi R, *Mapping rna-seq data to a transcript graph via approximate pattern matching to a hypertext.* In: International Conference on Algorithms for Computational Biology. Berlin: Springer: p., 2017: p. 49–61.

47. Shen, S., et al., *MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data.* Nucleic Acids Res, 2012. **40**(8): p. e61.

48. Kahles, A., et al., *SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data.* Bioinformatics, 2016. **32**(12): p. 1840-7.

49. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.

50. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

51. Heber, S., et al., *Splicing graphs and EST assembly problem.* Bioinformatics, 2002. **18 Suppl 1**: p. S181-8.

52. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.

53. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

54. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

55. Ferragina, P.M., G. , *Opportunistic data structures with applications.* Proceedings 41st Annual Symposium on Foundations of Computer Science., 2000.

56. Hooper, J.E., *A survey of software for genome-wide discovery of differential splicing in RNA-Seq data.* Hum Genomics, 2014. **8**: p. 3.

57. Liu, Y., et al., *Evaluating the impact of sequencing depth on transcriptome profiling in human adipose.* PLoS One, 2013. **8**(6): p. e66883.