SNF5/SMARCB1 Perturbation Results in Alternative Splicing for Specific Genes Rather Than Global Genomic Splicing Changes

By

Sarah Garcia

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Biology

Middle Tennessee State University

September 2022

Thesis Committee:

Dr. Rebecca Seipelt-Thiemann, Thesis Advisor

Dr. Mary Farone

Dr. April Weissmiller

I dedicate this body of work to my father,

for whom I wished could have seen me finish it.

ACKNOWLEDGEMENTS

This project would not have reached completion without all the love and support I received from friends, family, and faculty at Middle Tennessee State University.

I would like to thank my thesis advisor, Dr. Rebecca Seipelt-Thiemann, for her endless patience with me and all my pushed-back deadlines. Her attention to detail and wealth of knowledge made her an invaluable resource during the construction and completion of this project. Without her enthusiasm and wonderful organization skills, this project would not have been possible.

Next, I would like to thank my friends for their willingness to stick with me through the late-night freakouts and lending an ear to me whenever I needed to vent. Throughout this project, they've helped me to grow more as a person and become a better friend.

And lastly, I would like to thank my family for all their emotional and financial support they gave willingly to me when I needed it most. Whenever I needed her, my mother was there to believe in me even when I didn't believe in myself. She was my voice of reason during the most difficult days of my life, and she never gave up on me. Since day one, she's been teaching me how to think like her, a scientist, and now I can say I have proof that she's a good teacher, even if she may say otherwise.

ABSTRACT

Alternative splicing is a major source of protein diversity in cells of higher eukaryotes by having several different mRNAs potentially produced from the same pre-mRNA. Chromatin structure has been reported to be a regulator of alternative splicing, but much of the evidence is correlative in nature. To investigate this connection more directly, we analyzed alternative splicing events in RNA sequencing data from human cells expressing a regulated, degradable SWI/SNF chromatin remodeling complex subunit (SNF5). A total of 65 genes with significant isoform switching were identified, with 53 showing alternative splicing events. Exon skipping was the most prevalent alternative splicing event. In addition, examination of splice site strength indicated a majority of alternative 5' and 3' splice sites switched from a stronger site to a weaker site. This small proportion of alternative splicing effects support chromatin as a gene-specific rather than genome-wide regulator of alternative splicing. With this in mind, genes with alternative isoforms in response to SNF5 degradation were combined with a traditional differential expression gene set to explore function and protein interaction networks for a comprehensive gene expression set. This analysis revealed a majority of genes had functions related to zinc finger transcriptional pathways, glycoproteins, cancer pathways, cadherins and cell adhesion, axon guidance, differentiation, metallopeptidase, transcription/RNA pol II binding, and chromatin and nucleosome structure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Prior to completing the sequencing of the human genome in 2000, scientists estimated that humans would have >200,000 genes encoded in their DNA. Once the annotation of the genome began, each estimate was reduced to the current estimate of approximately 20,000 genes (Pertea et al. 2018). Despite having only 20,000 genes, which are common to each and every human cell in the body, each cell produces about 100,000 different proteins that vary from tissue to tissue during development and in response to external stimuli (Smith et al. 2021). This extreme protein diversity, functional diversity, and cellular diversity is generated by multiple interconnected levels of gene expression regulation including chromatin remodeling, transcription regulation, alternative splicing, alternative mRNA decay, translation regulation, post-translational modification, protein degradation, protein-protein interactions, and protein localization. While all of these processes are integral to the cell and organism's ability to function and adapt, the focus of this study is the effect of chromatin remodeling on alternative splicing, which has been proposed in the literature (Luco et al. 2011).

Eukaryotic DNA mainly exists in nuclear chromosomes which are comprised of tightly coiled chromatin fibers with many levels of structure that are dynamic. A more open conformation and higher levels of localized transcription describes euchromatin, whereas a more compact conformation that is harder to access and/or transcribe describes heterochromatin (Bannister and Kouzarides 2011). DNA is primarily associated with a

protein complex into a structure called a nucleosome. Nucleosomes are composed of two turns of DNA wrapped around a histone octamer. This octamer, named for the eight proteins of which it is comprised, contains two each of its core proteins H3, H4, H2A, and H2B. The turns of DNA are held in place by an additional histone, the H1 histone linker protein (Zhao 2019; Zhou 2019). In addition to the heterochromatin/euchromatin effects on gene expression, nucleosome density and position along the DNA strand is dynamic and can change during development, differentiation, and in response to environmental stimuli. Nucleosome positioning and density affect the DNA's accessibility to the transcription machinery, and thus influence transcription levels of genes in the affected region (Pagliaroli and Trizzino 2021). Specialized molecules known as chromatin remodelers facilitate these changes in DNA accessibility.

Chromatin remodelers can affect nucleosome density and chromatin state by unwrapping DNA from the histone complex, sliding the complex along the DNA strand, and/or breaking down the complex entirely (Kobayashi and Kurumizaka 2019). One such chromatin remodeler, the SWI/SNF complex was initially discovered in yeast and has been the focus of many recent studies since the discovery that mutations in genes encoding subunits of this complex have been identified to be commonplace in nearly 25% of all cancers (Mittal and Roberts 2020). In mammals, SWI/SNF is not a single complex, but three different complexes with overlapping and unique components: BRG1/BRM-associated factor canonical complex (cBAF), polybromo containing complex (pBAF), and non-canonical complex (ncBAF) (Pagliaroli and Trizzino 2021). Each complex also has distinct genomic localizations. The cBAF complex is found at promotor-distal enhancers, while the pBAF and ncBAF complexes are enriched at

promoter-proximal regions (Reddy and Workman 2018). Euskirchen et al (2011) identified >49,000 SWI/SNF high confidence binding sites in HeLa cells. In this study, we focused on core subunit SNF5/SMARCB1 that is present in both cBAF and pBAF complexes. Loss of SNF5/SMARCB1 in rhabdoid tumors disturbed SWI/SNF complex formation at enhancers, but not super-enhancers (Wang et al. 2017). Localization studies showed >14,000 SNF5 binding sites in the human genome (Euskirchen et al. 2011).

Another aspect of this study is splicing which occurs or co-occurs with transcription. Briefly, after DNA region is accessible due to the action of chromatin remodelers such as SWI/SNF, it can be transcribed via a complex process involving RNA pol II, general and specific transcription factors, as well as a variety of RNA binding sites specific in the gene's promoter region (Hahn 2004). Pre-messenger RNA is transcribed using the coding DNA strand in a 3' to 5' direction. RNA maturation includes several steps: capping of the 5' end, cleavage, polyadenylation at the 3' end, and removal of intron sequences (splicing). Spliceosomes, which are comprised of RNA and protein, initiate splicing by recognition of complementary sequences between the pre-mRNA (splice sites) and non-coding RNAs within the spliceosome complex. These splice site sequences are located at the intron-exon borders with the donor splice site (GURAGU) located at the 5' end of the intron and the acceptor splice site (AG) located at the 3' end of the intron (Fig. 1). In higher eukaryotes, which have large introns and small exons, exon definition occurs across the exon which thus requires the small exon size found in humans; the average internal size in 150bp. Which pre-mRNA sequences are included in the mature mRNA is also influenced by splicing silencers and enhancers located in the pre-mRNA and by RNA-binding proteins that bind them and are produced in a tissue-

specific manner (De Conti et al. 2013). In higher eukaryotes, the more closely a splice site sequence conforms to consensus, the more efficiently it is spliced. Additionally, moderately sized exons are included more often than unusually sized exons (>300bp;50bp) (De Conti et al. 2013). This phenomenon of using different portions of the pre-mRNA to produce more than one RNA and thereby more than one protein is alternative splicing. Alternative splicing (AS), which occurs in approximately 95% of all human genes (Zhu et al. 2018) can produce many different RNAs (also known as isoforms) from the same pre-mRNA and is a major source of protein diversity in cells. AS types include exon skipping, alternative 5' and 3' splice sites, mutually exclusive exons, and intron retention (Fig. 2).

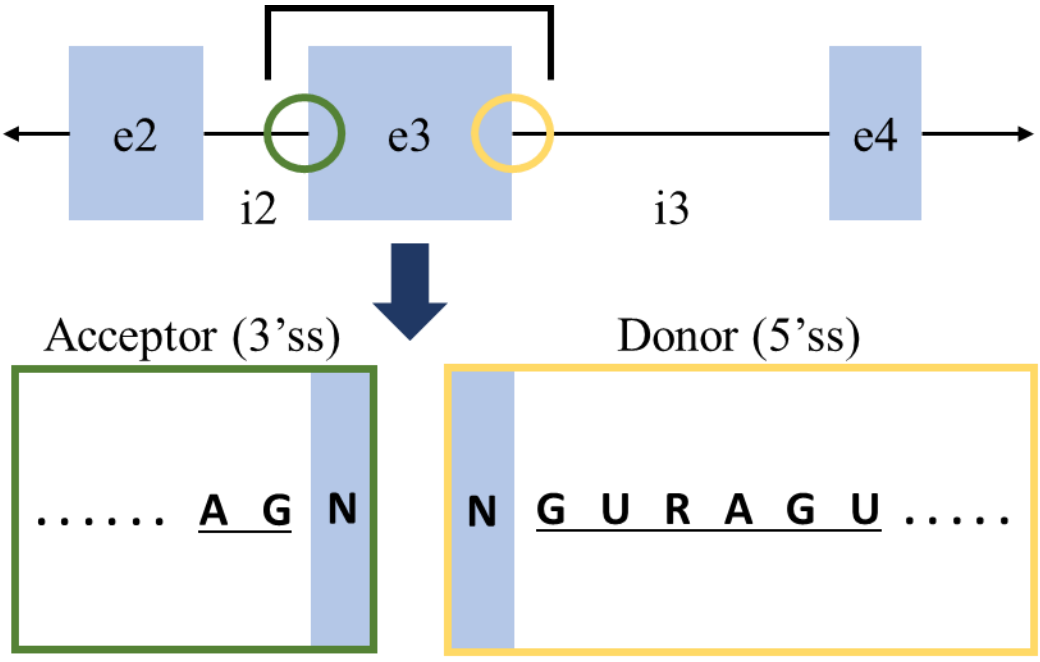


Fig. 1 – pre-mRNA Splice Consensus Sequence Diagram. Boxes represent exons, lines represent introns, and the region of interest is indicated by the black bracket. During splicing, spliceosomes identify consensus sequences at 5' and 3' ends of introns.

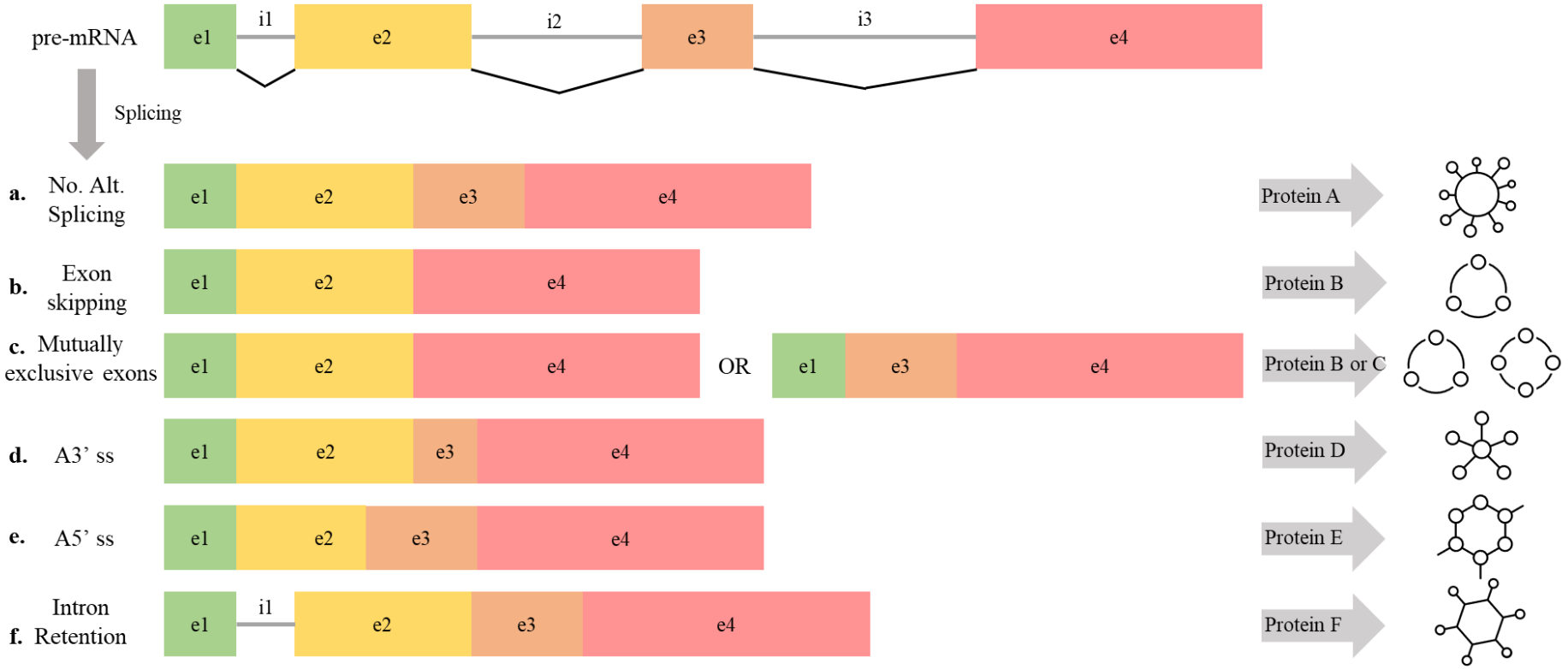


Fig. 2 – pre-mRNA Splicing in Eukaryotes. Boxes represent exons with connecting lines representing introns. Diagonal lines indicate splicing pattern. Introns in the pre-mRNA sequence are excised via spliceosomes during splicing, and exons are joined together. Five types of alternative splicing can occur, resulting in many possible mRNAs. Exon skipping leaves an exon out of the final mRNA. Mutually exclusive exons produce different mRNAs where one exon cannot exist in the mRNA if the other is present. An alternative 3' splicing site causes a portion of exon 3 to be skipped, whereas A5'ss skips a portion of exon 2. Intron retention leaves an intron in the mRNA sequence.

Exon skipping produces an mRNA in which a potential exon sequence was not identified by the spliceosome, and thus became part of a larger intron by default (2b). This produces a shorter RNA that is missing some coding region. Mutually exclusive exons produce different mRNAs where only one of several possible exon sequences are defined as an exon (2c). This produces an RNA that has some different coding region. For example, the human *FGFR2* gene produces two different tissue-specific splicing products. In epithelial cells, exon III-b is retained, while in mesenchymal cells, exon III-c is retained (Luco et al. 2011). Alternative 5' splice sites involve recognition and use of a 5' splice site near another 5' splice site (2d), while alternative 3' splice sites involve recognition and use of a 3'splice site near another 3' splice site (2e). Both of these effectively change the mRNA coding potential by increasing or decreasing exon size. Intron retention occurs when two potential exons with an intervening intron are recognized as a single exon (2f). This produces an RNA with more coding potential. However, any change can introduce a reading frame change, such that a new translation stop codon is now in frame. In the event these alternative splicing events produce mRNA transcripts with severely premature stop codons, nonsense-mediated mRNA decay will ensure transcript degradation before protein synthesis of the truncated protein can occur (Kurosaki and Maquat 2016).

For this study, we focused on the connection between chromatin remodeling and alternative splicing that has been suggested in the last twenty years (Luco et al. 2011). Many corelative pieces of evidence suggest a connection between chromatin and splicing. First, the amount of DNA involved in a nucleosome structure is similar to the average

mammalian internal exon (150bp) (Schwartz et al. 2009) which is defined by exon definition rather than intron definition in lower eukaryotes. Exons in long introns, a feature of many higher eukaryotic exons, are enriched for nucleosomes compared to exons in short introns (Spies et al. 2009) which have fewer nucleosomes or pseudo-exons, which have very few nucleosomes (Tilgner et al. 2009). Furthermore, included exons are enriched for nucleosomes compared to excluded exons (Schwartz et al. 2009) and nucleosome density is high at included exons with weak splice site strength (Spies et al. 2009). Two mechanisms have been proposed to explain this putative relationship: kinetics and recruitment. First, dense chromatin structures slow rates of transcription which also correlates with exon inclusion (Hodges et al. 2009). Slow transcription rates or RNA pol II pausing may allow splicing regulatory protein to be recruited (Luco et al. 2010). With these ideas in mind, we theorized that perturbation of a core subunit of the SWI/SNF chromatin remodeling complex, SNF5/SMARCB1, would allow us to identify genes, whose alternative splicing is influenced by chromatin remodeling, and to examine which genomic/splicing features might be important. Finally, given the prevalence of SNF5 mutations in cancer cells, we hypothesize that these genes would have functions in transcription regulation, cell cycle, cell signaling, and/or cancer development.

MATERIALS AND METHODS

Data Acquisition

RNA sequencing data from Human Embryonic Kidney (HEK) cells that express a version of the SNF5 core protein in the SWI/SNF chromatin remodeling complex that can be acutely degraded was obtained. HEK293-DTSNF5 cell line was engineered by first removing *SNF5* by CRISPR to create a SNF5-knock out. Then cells were transduced with a lentiviral version of SNF5 containing a N-terminal HA-FKBP12F36V dTAG module. Six total samples, three replicates per treatment (Table 1), were treated with dTAG47 for initiating SNF5 degradation or dimethyl sulfoxide for 24 hours before RNA isolation (Weissmiller pers. comm.) and checked via western blot. RNA sequencing was performed on an Illumina NovaSeq 6000 sequencing lane with 150bp paired end reads.

Table 1. Description of Experimental Samples RNA sequencing was performed following dTAG47/DMSO addition for 24 hours. Sequencing data was obtained on an Illumina NovaSeq 6000 with 150 bp paired end reads.

Sample ID	Sample Description	SNF5 State
5293-AW-7	HEK293-DTSNF5-DMSO-1	Present
5293-AW-8	HEK293-DTSNF5-DTAG-1	Degraded
5293-AW-9	HEK293-DTSNF5-DMSO-2	Present
5293-AW-10	HEK293-DTSNF5-DTAG-2	Degraded
5293-AW-11	HEK293-DTSNF5-DMSO-3	Present
5293-AW-12	HEK293-DTSNF5-DTAG-3	Degraded

Analysis Pipelines

Analysis of RNA sequencing data was conducted using tools at Cyverse Discovery Environment and Galaxy as well as stand-alone bioinformatic analysis sites: IuPred, SignalP, CPC2, and Pfam. Additionally, R and RStudio were used with several R packages. Each workflow is visualized (Fig. 3-5) and described in detail below. Some tools and output files are used in multiple pipelines, as noted in each pipeline. The quality pipeline (Fig. 3) was utilized first to determine library quality and identify outlier replicates. The DEG pipeline (Fig. 4) was then utilized to identify differentially expressed genes. Lastly, the isoform identification pipelines (Fig. 5) were utilized to identify differentially expressed isoforms (DEI) with functional consequences and DEI regardless of functional consequences.

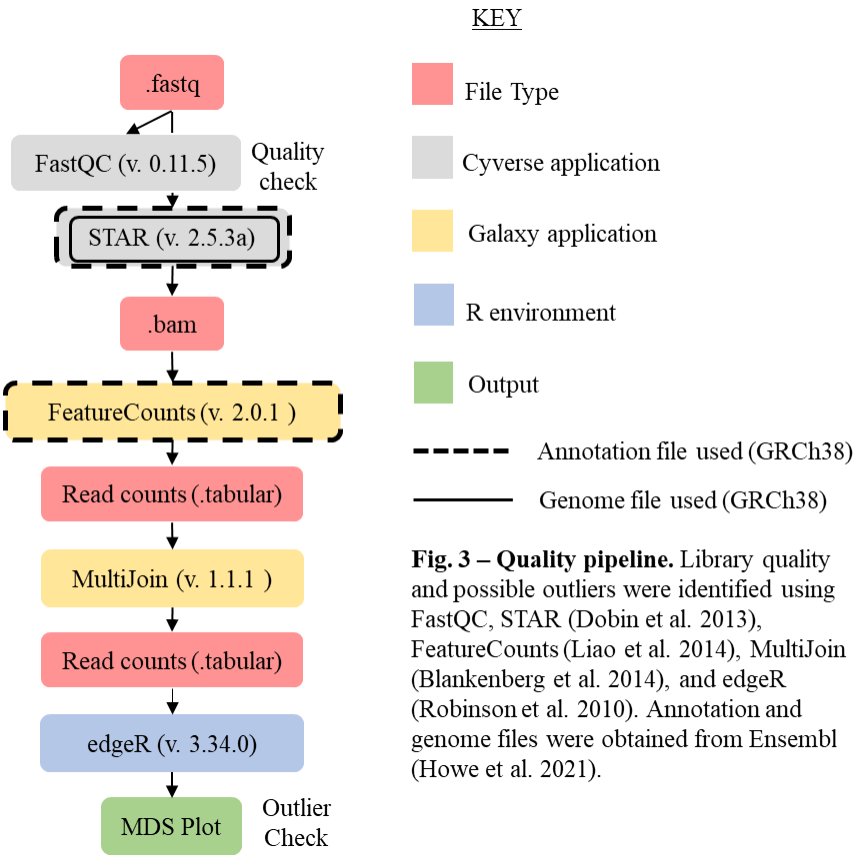
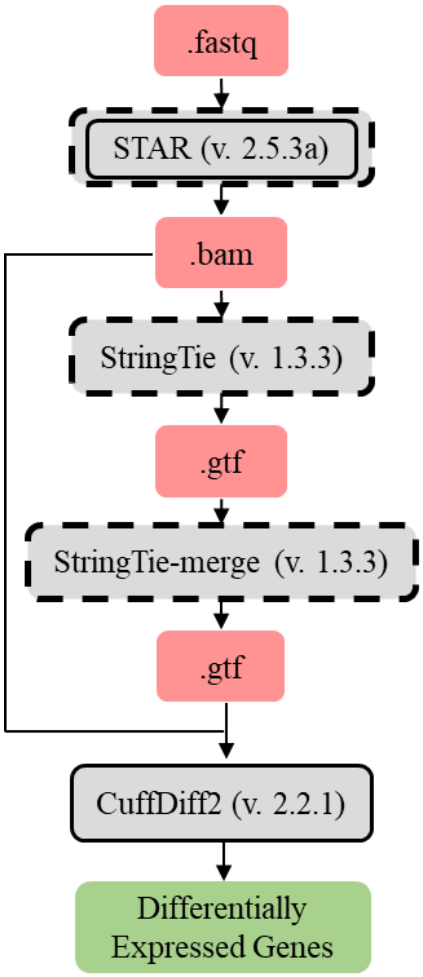


Fig. 3 – Quality pipeline. Library quality and possible outliers were identified using FastQC, STAR (Dobin et al. 2013), FeatureCounts (Liao et al. 2014), MultiJoin (Blankenberg et al. 2014), and edgeR (Robinson et al. 2010). Annotation and genome files were obtained from Ensembl (Howe et al. 2021).



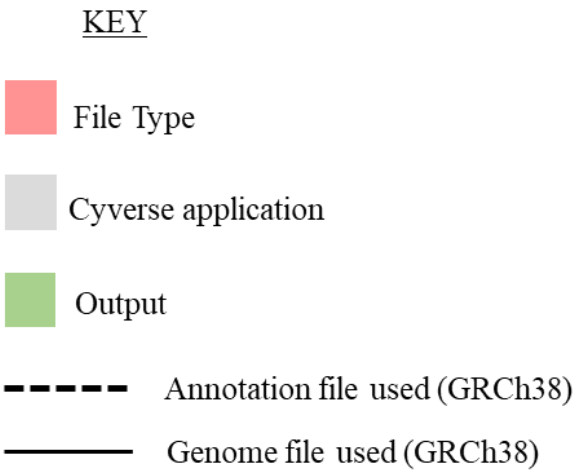


Fig. 4 – Differentially Expressed Gene Identification pipeline. Differentially expressed genes were identified using STAR (Dobin et al. 2013), StringTie and StringTie-merge (Kovaka et al. 2019), and CuffDiff (Trapnell et al. 2013). Annotation and genome files were obtained from Ensembl (Howe et al. 2021).

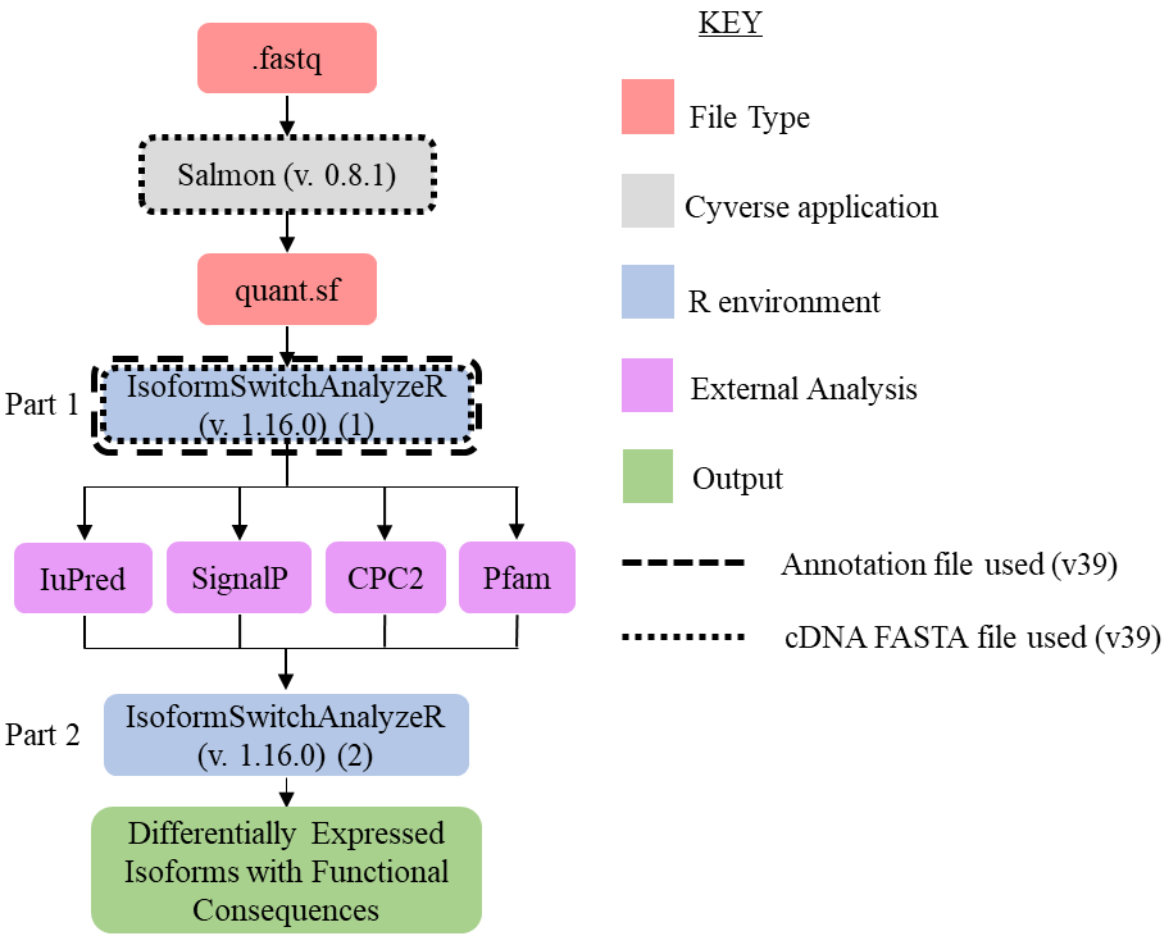


Fig. 5 – Isoform Identification pipeline. Differentially expressed isoforms with functional consequences were identified using Salmon (Patro et al. 2017) and IsoformSwitchAnalyzeR (Vitting-Seerup and Sandelin 2019). External analysis was done using IuPred (Erdős and Dosztányi 2020), SignalP (Almagro et al. 2019), CPC2 (Kang 2017), and Pfam (Potter 2018). Annotation and cDNA files were obtained from GENCODE (Frankish et al. 2021). For identification of differentially expressed isoforms regardless of functional consequence, external analysis is not needed.

Quality Pipeline

FastQC ver. 0.11.5 (Babraham Bioinformatics) was used to verify library quality before beginning major pipelines. STAR ver. 2.5.3 (Dobin et al. 2013) was used to align the paired end reads to a reference genome using annotation and genome files (GRCh38) acquired from Ensembl (Howe et al. 2021). Coordinates of reads are returned in a single file per sample (.bam). Within the Galaxy environment, FeatureCounts ver. 2.0.1 (Liao et al. 2014), with parameters set to reads per gene, was used to generate a feature count table per sample. Multijoin ver. 1.1.1(Blankenberg et al. 2014) was then used to combine all individual counts files. The feature count table was then imported into R. The code package edgeR ver. 3.34.0 (Robinson et al. 2010) and RStudio functions (Loraine et al. 2015) were used to output a multidimensional scaling (MDS) plot (Fig. 6) to determine the presence of outliers. Sample 8 was deemed an outlier due to unusual clustering and not included in any downstream analysis for any pipeline.

Differentially Expressed Genes Pipeline

Differentially expressed genes (DEG) were identified using the STAR-StringTie CuffDiff pipeline. Transcript annotations were performed using STAR output and StringTie ver. 1.3.3 and String-Tie-merge ver. 1.3.3 (Kovaka et al. 2019), which assembles RNA-seq alignments into potential transcripts resulting in a single annotation file (.gtf). CuffDiff ver. 2.2.1 (Trapnell et al. 2013) was used to perform differential transcript abundance analysis to produce a list of significant DEGs with a q-value less than 0.05.

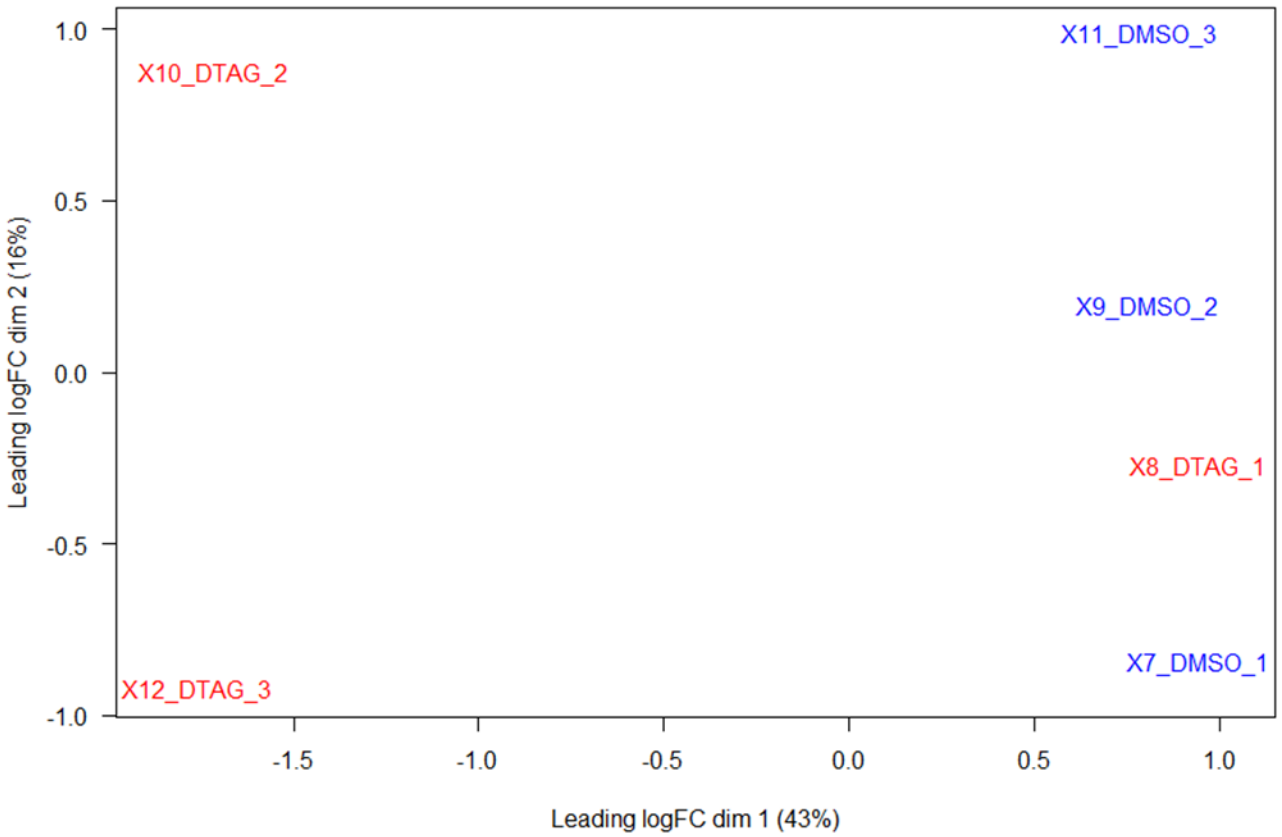


Fig. 6 – Gene Expression Clustering with Multidimensional scaling (MDS). Plot created in RStudio using edgeR ver. 3.34.0 (Robinson et al. 2010) and RStudio functions (Loraine et al. 2015). Control samples are shown in blue, and treatment samples are shown in red. Sample 8 (X8_DTAG_1) was determined to be an outlier due to abnormal clustering.

Isoform Identification Pipeline

The isoform identification pipelines were undertaken to determine statistically significant switches in transcripts generated by alternative splicing with functional consequences and regardless of functional consequences (Fig. 5). Transcript quantification was done a second time using .fastq files. Salmon ver. 0.8.1 (Patro et al. 2017) uses a pseudo-mapping technique to perform quantification that is more compatible with the IsoformSwitchAnalyzeR package. Quantification files were imported into R for use in code package IsoformSwitchAnalyzeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). A prefiltering step was added to remove single isoform genes, implement a gene expression cutoff (>5 TPM), and remove unused isoforms with an isoform expression cutoff (>0 FPKM). To control false discovery rates (FDR), a DEXSeq test (Anders et al. 2017) was done to test for differential exon usage. Prefiltering and DEXSeq was performed for both pipelines. The simplified pipeline (Fig. 5b) speeds up processing time by forgoing external analysis to output significant changes in isoform switching and allows for structural analysis of the switching isoforms. IsoformSwitchAnalyzeR determines significance using a q-value cutoff (<0.05) and a minimum change in isoform usage (>0.1).

For determining functional consequences, external analysis was needed. SignalP (Almagro et al. 2019) was used to predict signal peptides, CPC2 (Kang 2017) to predict coding potential, Pfam (Potter 2018) to predict protein domains, and IuPred (Erdős and Dosztányi 2020) to predict disordered regions. External analysis was then imported into RStudio, and isoform switch analysis was performed. The following consequences were checked for changes between switching isoforms: alternative transcription start site

(ATSS), alternative transcription termination site (ATTS), A3'ss, A5'ss, exon skipping (ES), intron retention (IR), coding potential, nonsense mediated decay (NMD) status, identified domains, intrinsically disordered regions, and identified signal peptides. A nucleotide cutoff (>1) was implemented to detect changes in length. Detailed switch plots are generated from both pipelines visualizing gene expression, isoform expression, isoform usage, and transcript changes with predicted protein domains being added to show functional consequences (Fig 7, Appendix A). Exon and sequence information was obtained for each switching isoform from Ensembl (Howe et al. 2021) and were checked for unusually large (>300) and small (<50) exons. Sequences in immediate proximity to alternative splicing events and unusually sized exons were scored using SpliceRover (Zuallaert et al. 2018) which uses a convolutional neural network (CNN) to predict splice sites and assigns a probability between zero and one to each possible site and is a measure of splice site strength. For A3'ss events, both acceptor sites were scored, A5'ss, both donor sites were scored. For exon skipping and unusually sized exons, both acceptor and donor sites of the exon were scored.

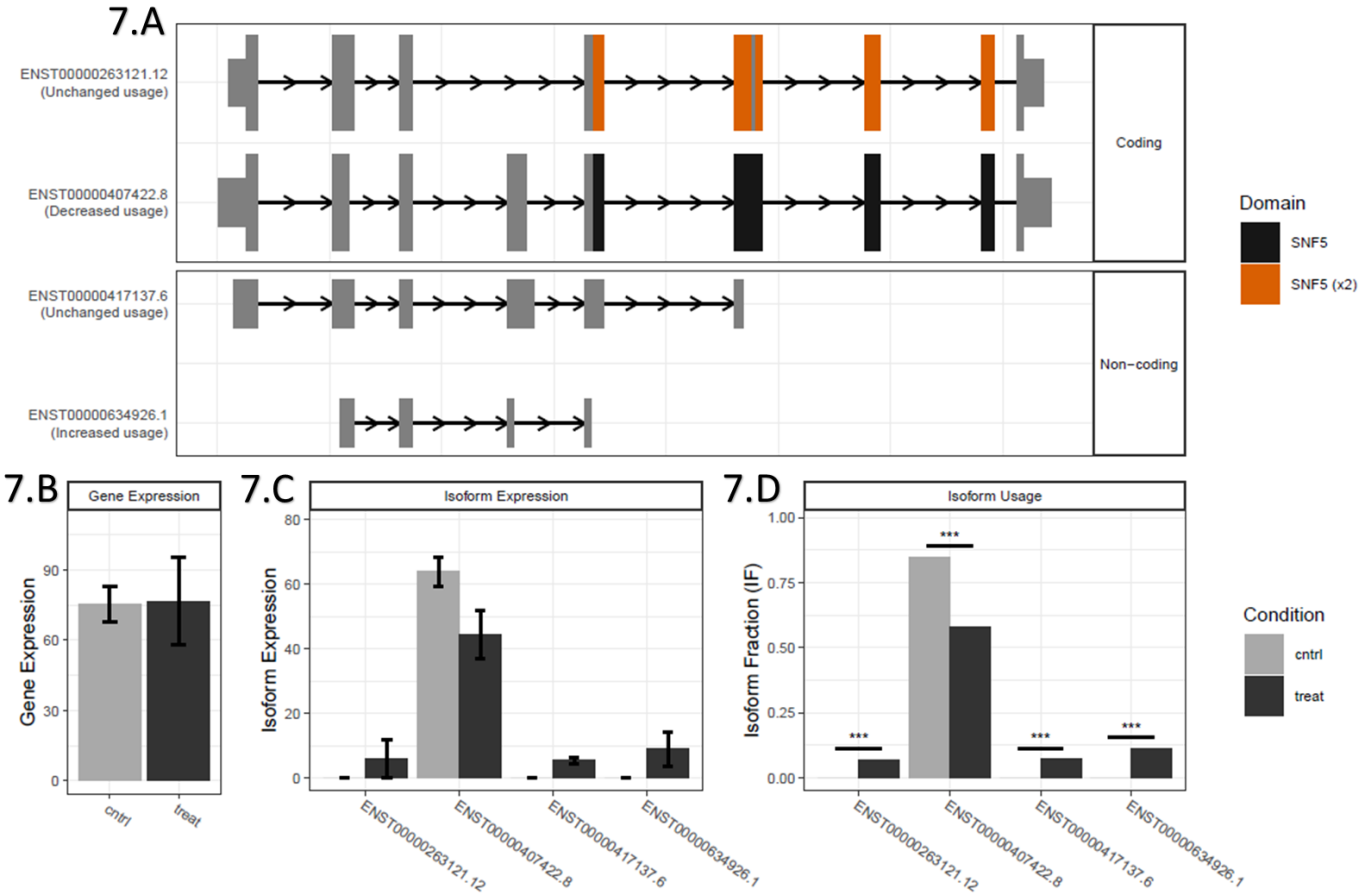


Fig. 7 – Gene Transcript and Expression Summary. Grouped plots visualize transcript structure, gene and isoform expression, and isoform usage for gene *SMARCB1* which codes for the SNF5 subunit of the SWI/SNF chromatin remodeling complex. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot A visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$ cutoff. Plots B and C visualize gene and isoform expression, respectively. Error bars indicate a 95% confidence interval. Bar color indicates condition (control, treatment). The y-axis for plot C represents transcripts per million (TxPM); the y-axis for plot D represents the sum of all TxPM values for each condition. Plot D notes isoform usage by comparing isoform fraction values for each transcript in all conditions. Asterisks represent significance level with (*) being significant with $q < 0.05$ and (***) being highly significant with $q < 0.001$. No significance is indicated by "ns" with $q > 0.05$.

RESULTS

Guiding Questions

The goal of this project was to explore the mechanistic connection between chromatin structures and alternative splicing, which has been postulated in the literature (Luco 2011). While at least three mammalian chromatin remodeling complexes are known (cBAF, ncBAF, pBAF), one subunit (SNF5/SMARCB1) that is common to two of the complexes (cBAF and pBAF) is the focus of this study. We aimed to identify genes whose alternative splicing is altered when SNF5 is specifically targeted for degradation, and then explore gene features and RNA levels for affected genes that may provide clues for the alternative splicing-chromatin relationship. In addition, examining all the genes that are differentially expressed, not simply differential RNA levels, as is commonly done, may provide a more comprehensive view of genes affected in *SNF5*-deficient pediatric tumors.

Initial Findings

Before analysis, we conducted initial steps to check library quality and identify outliers. All samples were of high quality (data not shown). Clustering of overall gene expression was visualized using multidimensional scaling (MDS) plot and showed sample 8 to be an outlier due to unusual clustering (Fig. 6) That sample was not included in downstream analysis for any pipeline.

Alternative Splicing and Transcription Events Within Switching Isoforms

To determine the effect of *SNF5* perturbation on alternative splicing, we identified changes in isoform structure using the IsoformSwitchAnalyzeR pipeline detailed in the methods. We removed 111,312 transcripts that were from single isoform genes and low expressed genes and isoforms. Sixty-five of approximately 19,000 (0.342%) human genes (Frankish, Diekhans et al, 2021) had significant switching isoforms in response to *SNF5* perturbation. The output of this pipeline includes one detailed visualization with isoform structure, transcript level, isoform expression, and isoform fraction per gene (Appendix A). Statistically significant isoform usage is noted on the isoform fraction histogram. Only isoform switching, not unregulated isoform use, is the focus, so isoforms with unchanged levels are not considered in these analyses. Switch plots for each gene were visually inspected and then isoform switching categorized as alternative splicing events (exon skipping, intron retention, alternative 5’ splice site, alternative 3’ splice site) and alternative transcription events (alternative transcription start site, alternative transcription termination site). Of the splicing events, exon skipping occurred most frequently. A total of 53 of the 65 genes identified showed alternative splicing events. Of these 53, eighteen showed exon skipping, three showed intron retention, twelve showed A5’ss located in a terminal exon, five showed internal 3’ alternative splicing, and a further twelve showed 3’ alternative splicing in a terminal exon (Table 2). Eight additional genes showed complex splicing/terminal exon events.

Gene	ATSS	ATTS	ES	IR	A5'SS	A3'SS
AMN1	1	1	0	0	0	C
APEX2	1	1	1	0	0	0
ARHGEF1	1	0	0	0	1 (T)	0
C1orf50	1	1	0	0	0	1
CCDC82	1	0	0	0	0	C
CD2BP2	1	1	0	0	0	C
COMMD2	1	1	0	0	0	C
ELK1	1	0	1	0	0	0
ERV3-1	1	1	0	0	0	1 (T)
FAM189B	1	0	3	0	0	0
FKBP11	1	1	0	5	0	0
GABARAPL1	1	1	0	1	0	0
GHRL	0	0	1	0	0	0
HS3ST3A1	0	1	0	0	1 (T)	0
IFITM3	1	1	0	0	1 (T)	1 (T)
IGFBP2	1	1	0	0	1 (T)	0
JADE3	1	0	0	0	1 (T)	0
LAMB3	1	0	0	0	0	C
LINC00467	1	1	0	0	1 (T)	C
LPGAT1	1	1	1	0	0	0
MYL6	1	1	1	0	0	0
NDUFAF8	1	0	0	0	1 (T)	0
OAF	1	1	0	0	1 (T)	1 (T)
POFUT2	1	1	0	0	0	1 (T)
POU6F1	1	1	1	0	0	1 (T)
PPP1R18	1	0	0	0	0	C
PPT2	1	0	1	0	0	0
RDM1	1	1	0	1	0	0
REEP5	1	1	1	0	0	0
RNF114	0	1	1	0	0	0
RPP21	1	0	0	0	0	1
SELENOO	1	0	0	0	0	1
SEPTIN9	1	0	0	0	1 (T)	0
SMARCB1	1	1	0	0	1	0
SMARCC1	0	1	1	0	0	0
SNX18	0	1	0	0	1 (T)	0
SRGAP2B	1	1	1	0	0	0
STX18-AS1	1	1	0	0	1 (T)	0
SUDS3	1	1	1	0	0	0
SUOX	1	1	0	0	0	1 (T)
SYNJ2BP	1	1	1	0	0	0
TGFB1	1	1	0	0	0	1 (T)
TKT	0	0	1	0	0	1
TMEM167A	1	1	0	0	1 (T)	0
TMEM184C	1	1	2	0	0	1 & 1 (T)
TNFRSF12A	1	0	1	0	0	0
TXLNA	1	0	0	0	0	1 (T)
U2AF1L5	1	1	2	0	0	0
UST	1	1	0	0	0	1 (T)
WDR4	0	1	0	0	1	C
ZNF251	1	1	0	0	0	C
ZNF362	1	1	0	0	0	1 (T)
ZNRF2	1	1	0	0	0	1 (T)
Total	43	37	22	7	13	17

Table 2 – Events Per Gene. Summaries of alternative splicing and alternative transcription events noted in columns. Splicing events include alternative 3' splice sites (A3'ss), alternative 5' splice sites (A5'ss), exon skipping (ES), and intron retention (IR). Transcription events include alternative transcription start sites (ATSS) and alternative transcription termination sites (ATTS). Events notated with (T) involve events relating to terminal exons. Events labeled as "C" are complex events involving a single splice site and a terminal exon. Thirty-two genes noted in blue are near putative SNF5 binding sites (Euskirchen et. al 2011).

Functional Consequences of Identified Isoform Switching

To gain a better understanding of how these identified splicing changes affected the encoded proteins, we incorporated data from several tools to predict functional consequences. SignalP (Almagro et al. 2019) was used to predict signal peptides, CPC2 (Kang 2017) to predict coding potential, Pfam (Potter 2018) to predict protein domains, and IuPred (Erdős and Dosztányi 2020) to predict disordered protein regions. Visual inspection of switch plots of the 65 genes showed 41 genes with altered predicted protein domains as a result of the switch, 16 genes with unaffected domains, and 8 genes that did not contain any predicted protein domains (Appendix A).

Functional Summary

After identifying the 32 genes near putative *SNF5* target sites from the genes with alternative splicing events (Table 2), we took a closer look at individual gene function to gain more insight on the functional consequences predicted (Table 3). Gene function information was acquired from DAVID (Huang da W et al. 2009). Five major functional groups were identified: Transcription, Chromatin and/or Histone, Cell Signaling /Receptor Activity, Protein Binding and/or Transport, and Health/Disease Progression, with a sixth group assigned to miscellaneous gene functions not pertaining to the other groups. Along with gene function, data obtained from IsoformSwitchAnalyzeR (Vitting-Seerup and Sandelin 2019) was used to provide context into the consequence of the resulting isoform switch (Table 3). Six genes had functions relating to transcription factors and/or transcription regulation: *ELK1, POU6F1, OAF, TGTB1, ZNF251,* and *ZNF362.* Three of the six gained protein domains in the isoform with increased usage, one lost all domains, and two had no domain change. Four genes had functions relating to

chromatin and/or histones: *MPHOSPH8, SMARCC1, SMARCB1,* and *SUDS3*. Three genes resulting in complete or partial protein domain loss, and one gene gained protein domains. Three genes had functions relating to cell signaling and/or cell receptor activity: *CD2BP2*, *TNFRSF12A,* and *SYNJ2BP*. One gene gained domains, one lost all domains, and one had no domain changes. Four genes had functions relating to protein binding and/or transport: *VBP1, REEP5, TXLNA,* and *C1orf50*. Two had complete domain loss, and two had no domain changes. Three genes had functions relating to health and/or disease progression: *IGFBP2, IFITM3,* and *LAMB3*. Two genes had complete or partial domain gains, and one gene had no domain changes. Ten genes were assigned to the miscellaneous function group, from functions ranging from DNA damage repair to appetite-regulating hormones (Table 3).

SWI/SNF Binding Targets

To determine whether these alternative splicing events are primary or secondary effects, we determined which of our identified genes were near putative *SNF5* binding sites. Using a list (Euskirchen et. al 2011) of known binding regions for components of the SWI/SNF complex, we identified 32 genes with alternative splicing events to be near *SNF5* target sites (Table 2).

Gene	Function - Transcription	Consequence
ELK1	transcription factor	no change
COMMD2	transcription regulation	complete domain loss
POU6F1	transcription regulation	unaffected
OAF	transcription regulation	complete domain gain
TGFB1	recruitment of transcription factors	complete domain gain
ZNF251	transcription factor activity, transcription regulation	complete domain loss
ZNF362	transcription factor activity, transcription regulation	complete domain gain
Gene	**Function - Chromatin and/or Histone**	**Consequence**
SMARCC1	chromatin remodeling SWI/SNF	partial domain loss
SMARCB1	chromatin remodeling SWI/SNF	complete domain loss
SUDS3	HDAC-dependent corepressor complex subunit	complete domain loss
Gene	**Function - RNA and/or DNA-damage repair**	**Consequence**
CD2BP2	mRNA processing, mRNA splicing	no change
RPP21	ribonuclease subunit – tRNA precursor, RNA-binding	partial domain gain
WDR4	tRNA processing, DNA damage	no change
APEX2	DNA damage repair	complete domain loss
Gene	**Function - Protein Binding and/or Transport**	**Consequence**
REEP5	ER organization, intracellular transport regulation, receptor	complete domain loss
TXLNA	syntaxin binding activity, exocytosis	no change
C1orf50	identical protein binding regulation	unaffected
Gene	**Function - Health and/or Disease Progression**	**Consequence**
IGFBP2	inhibit tumor supression, growth regulation	partial domain gain
IFITM3	antiviral protein, immunity	complete domain gain
LAMB3	cell adhesion, epidermolysis bullosa, protein subunit	no change
Gene	**Function - Cell Signaling and/or Receptor Activitiy**	**Consequence**
TNFRSF12A	apoptotic cell signaling, receptor	complete domain gain
SYNJ2BP	receptor binding activity, mitochondrial outer membrane component	complete domain loss
Gene	**Function - Miscellaneous Enzyme Activity**	**Consequence**
FKBP11	isomerase, rotamase	complete domain loss
LPGAT1	acyltransferase, metabolism	complete domain loss
SUOX	oxidoreductase	unaffected
ZNRF2	ubiquitin protein ligase activity, protein ubiquitination, transferase	complete domain gain
TKT	glycolysis, transferase	partial domain loss
UST	transferase	complete domain loss
Gene	**Function - Other**	**Consequence**
AMN1	protein catabolic processes	unaffected
GHRL	energy homeostasis, appetite-regulating hormone	unaffected
MYL6	motor protein, muscle protein, myosin	unaffected
ARHGEF1	GTPase activation, guanine-nucleotide releasing factor	no change

Table 3 – Functional and Consequence Summary of Genes with alternative splicing and/or transcription events near putative SNF5 binding sites in response to *SNF5* perturbation. Gene functions are summarized and grouped from information acquired from the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang da W et al. 2009). All genes noted are near putative SNF5 binding sites (Euskirchen et. al 2011). The switch consequence listed notes if the isoform switch identified by IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019) affected protein domains within the isoforms involved. No change refers to a switch between two isoforms with unaffected domains. Unaffected refers to a switch between two isoforms with no domains present. Complete domain gain/loss refers to a switch between an isoform with no domains and an isoform with one or more domains. Partial domain gain/loss refers to a switch resulting in increased/decreased number and/or size of domains present.

Gene Structure with Splicing Events

Now that we had identified genes with altered splicing and transcription events in response to SNF5 degradation, we next examined structural, sequence, and level characteristics known to affect alternative splicing. We focused on identifying unusually sized exons (>300bp;<50bp), estimating splice site strength, and RNA levels for the 53 genes and alternative events identified in the isoform switch pipeline. Splice site strength, as predicted by SpliceRover showed that most of the exons involved in exon skipping had strong splice sites, including exons that were skipped in response to SNF5 perturbation (Table 4). Predicted splice site strength for alternative 5' and 3' splice site events indicated that for the most part SNF5 degradation correlated with a switch from a stronger splice site to a weaker splice site (Table 5). Alternative splicing can also be influenced by the internal exon size – small exons (<50bp) and large exons (>300bp) are spliced less efficiently due to steric hinderance and interactions of splicing marching across the exon. Very few small exons were identified among the regulated isoforms (Table 6). Five isoforms had small internal exons while only three had large exons.

Gene	isoform_id	event	5'ss	3'ss	exon size (bp)	conseq
APEX2	ENST00000374987.4	ES	0.62288	0.94314	181	loss
ELK1	ENST00000376983.8	ES	0.922248	0.956303	106	loss
FAM189B	ENST00000361361.7	ES	0.676241	0.960272	57	gain
FAM189B	ENST00000361361.7	ES	0.631281	0.980103	111	gain
FAM189B	ENST00000361361.7	ES	0.978603	0.986856	120	gain
GHRL	ENST00000439975.6	ES	0.985005	0.988438	109	loss
LPGAT1	ENST00000488600.1	ES	0.991058	0.997969	148	gain
MYL6	ENST00000550697.6	ES	0.654154	0.619059	45	gain
POU6F1	ENST00000546685.5	ES	0.999669	0.998145	122	gain
PPT2	ENST00000395523.5	ES	0.828464	0.812259	406	loss
REEP5	ENST00000511865.6	ES	0.668483	0.892076	368	gain
RNF114	ENST00000244061.6	ES	0.984892	0.99995	108	loss
SMARCC1	ENST00000254480.10	ES	0.982338	0.999935	82	loss
SRGAP2B	ENST00000467933.2	ES	0	0	85	
SUDS3	ENST00000543473.2	ES	0.952488	0.978043	22	loss
SYNJ2BP	ENST00000256366.6	ES	0.966016	0.997261	137	loss
TKT	ENST00000423516.5	ES	0.845665	0.010153	24	loss
TMEM184C	ENST00000296582.8	ES	0.996494	0.932329	100	loss
TMEM184C	ENST00000296582.8	ES	0.922334	0.857169	172	loss
TNFRSF12A	ENST00000326577.9	ES	0.984911	0.994698	135	gain
U2AF1L5	ENST00000610664.5	ES	0.821122	0.966056	67	gain
U2AF1L5	ENST00000610664.5	ES	0.651492	0.986942	67	loss

Table 4 – Exon Skipping. This table lists exon skipping events observed in genes found to exhibit alternative splicing events after SNF5 perturbation. Donor (5') and acceptor (3') splice sites surrounding the exon of interest were scored via SpliceRover (Zuallaert et al 2018) which uses a convolutional neural network (CNN) to predict the likelihood of splice site usage. Scores were taken from the upregulated transcript as noted in the associated switch plot for each gene referenced. Scores of zero represent no intron present on one or both sides of splice event. Eleven genes noted in blue are near putative SNF5 binding sites (Euskirchen et. al 2011).

Gene	isoform_id	event	upregulated score	downregulated score	diff	exon size (bp)
C1orf50	ENST00000691126.1	A3'ss	0.976016	0.12244	0.853576	87
RPP21	ENST00000442966.7	A3'ss	0.067513	0.6613	-0.593787	83
SELENOO	ENST00000380903.7	A3'ss	0.695463	0.047426	0.648037	181
SMARCB1	ENST00000634926.1	A5'ss	0.450366	0.51896	-0.068594	27
TKT	ENST00000423516.5	A3'ss	0.814421	0.999273	-0.184852	82
TMEM184C	ENST00000296582.8	A3'ss	0.047849	0.139993	-0.092144	48

Table 5 – Alternative 3' and 5' Splice Site Events. This table lists alternative splice site events observed in genes found to exhibit alternative splicing events after SNF5 perturbation. Splice sites from the upregulated and downregulated transcripts were scored via SpliceRover (Zuallaert et al 2018) which uses a convolutional neural network (CNN) to predict the likelihood of splice site usage. Four genes noted in blue are near putative SNF5 binding sites (Euskirchen et. al 2011).

Gene	isoform_id	event	5'ss	3'ss	exon size (bp)
MYL6	ENST00000550697.6	XS	0.654154	0.619059	45
SMARCB1	ENST00000634926.1	XS	0.438659	0.999388	27
SUDS3	ENST00000543473.2	XS	0.952488	0.978043	22
TKT	ENST00000423516.5	XS	0.845665	0.010153	24
TMEM184C	ENST00000296582.8	XS	0.994524	0.027422	48
PPP1R18	ENST00000399199.7	XL	0.996305	0.996037	1639
PPT2	ENST00000395523.5	XL	0.828464	0.812259	406
REEP5	ENST00000511865.6	XL	0.668483	0.892076	368

Table 6 – Unusual Exon Sizes. This table lists unusual exon sizes observed in genes found to exhibit alternative splicing events after SNF5 perturbation. Donor (5') and acceptor (3') splice sites surrounding the exon of interest were scored via SpliceRover (Zuallaert et al 2018) which uses a convolutional neural network (CNN) to predict the likelihood of splice site usage. Scores were taken from the upregulated transcript as noted in the associated switch plot for each gene referenced. Five genes noted in blue are near putative SNF5 binding sites (Euskirchen et. al 2011).

Effects on RNA Levels

Since transcription rates were implicated as a mechanism by which chromatin might impact splicing, we next determined how RNA levels changed in response to *SNF5* perturbation. Initial findings from CuffDiff (Trapnell et al. 2013) showed 332 differentially expressed transcripts from 325 annotated genes with 151 transcripts downregulated (45.5%) and 181 transcripts upregulated (54.5%) after initiation of SNF5 degradation. There were 36 unannotated transcripts that were removed from the DEG list. Three-hundred twenty six differentially expressed genes were identified. Summaries of the top twenty-five upregulated and downregulated identified genes are shown (Table 7). *CARD11* had the highest upregulated fold change of 82.85. *NR0B1* had the highest downregulated fold change of 28.96. Having these data gathered, we then compared splicing events, splice site strength features, and RNA levels for switching isoform genes.

gene_id	gene_name	fold_change	direction	total_fpkm	q-value
ENST00000396946	CARD11	82.85	UP	5.02	0.014472
ENST00000265162	ENPEP	41.45	UP	3.13	0.014472
ENST00000409204	PCDH20	35.29	UP	2.24	0.014472
ENST00000304623	CTNND2	32.75	UP	3.93	0.014472
ENST00000378970	NR0B1	28.96	DOWN	6.83	0.014472
ENST00000244573	H1-1	27.1	UP	28.16	0.014472
ENST00000334197	ZNF347	20.27	DOWN	0.62	0.014472
ENST00000261233	IRAK3	17.65	UP	0.98	0.0351024
ENST00000398743	PRAME	16.75	UP	3.63	0.014472
ENST00000329047	SEPTIN9	16.43	UP	16.62	0.0351024
ENST00000427401	ZNF737	13.49	DOWN	1.15	0.014472
ENST00000319420	MSTRG.41150,SHISA2	11.98	UP	23.87	0.014472
ENST00000648973	ZNF600	11.94	DOWN	1.61	0.014472
ENST00000219070	MMP2	11.74	UP	3.53	0.014472
ENST00000373034	PCDH19	10.45	DOWN	5.69	0.014472
ENST00000318627	FIBIN	9.94	UP	1.52	0.014472
ENST00000402377	ZNF681	9.59	DOWN	2.27	0.014472
ENST00000254321	AC008770.2,ZNF700	9.1	DOWN	3.04	0.014472
ENST00000609111	ADIRF-AS1	8.71	UP	2.04	0.014472
ENST00000379221	DNAJC15	8.49	DOWN	1.32	0.0260325
ENST00000368222	CRABP2	8.48	UP	10.58	0.014472
ENST00000421239	AC010332.3,ZNF578	8.45	DOWN	0.82	0.014472
ENST00000541777	ZNF83	8.45	DOWN	4.84	0.014472
ENST00000355426	EFEMP1	8.41	UP	3	0.014472
ENST00000670294	ZNF667-AS1	8.27	DOWN	5.71	0.014472

Table 7 – Top 25 Differentially Expressed Genes. The list of differentially expressed genes was obtained using CuffDiff ver. 2.2.1 (Trapnell et al. 2013). A summary of the top twenty-five genes sorted by fold change is shown, along with their direction (up or down regulated), total fragments per kilobase of exon per million mapped fragments (fpkm) and associated q-value.

First, after having identified thirty-two genes located near a SNF5 binding site from our initial group of 65 genes with significant isoform switching, only one gene, *LAMB3* (Fig 11), was found to also be differentially expressed at the gene level when SNF5 is perturbed, with an upregulated fold change of 4.92. *LAMB3* has been implicated in the progression of epidermolysis bullosa and may have functions relating to cell adhesion (Table 3). No domain changes were noted between the two switching isoforms (ENST00000356082.9, ENST00000391911.5).

Secondly, twelve of the thirty-two genes located near a SNF5 binding site were found with exon skipping events (Table 2). *APEX2* (Fig 8), whose function is related to DNA damage repair (Table 3), exhibited increased skipping of a 181 bp exon with moderately strong splice sites (Table 4) which occurred alongside total domain loss. *ELK1* (Fig 10) has functions relating to transcription factors (Table 3) and showed increased skipping of a 106 bp exon with strong splice sites (Table 4) with negligible effect on domains. *GHRL* (Fig 10) which codes for an appetite-regulating hormone showed increased skipping of a 109 bp exon with strong splice sites (Table 4) with no effect on domains. *LPGAT1* (Fig 12) which plays a role in metabolism (Table 3) showed more inclusion of a 148 bp exon with strong splice sites (Table 4) with no effect on domains. *MYL6* (Fig 12) with function relating to myosin and muscle proteins (Table 3) showed higher inclusion of a small 45 bp exon with moderately strong splice sites (Table 4) with no effect on domains. *POU6F1* (Fig 13) plays a role in transcription regulation (Table 3) and showed increased inclusion of a 122 bp exon with strong splice sites (Table 4) with no affected domains. *REEP5* (Fig 13), implicated in protein binding and/or transport (Table 3) showed more inclusion of a large 368 bp exon with moderately strong

splice sites (Table 4) alongside complete domain loss. *SMARCC1* (Fig 14) which codes for a subunit of the SWI/SNF chromatin remodeling complex (Table 3) showed increased skipping of an 82 bp exon with strong splice sites (Table 4) with major domain loss. *SUDS3* (Fig 14) codes for a subunit of a histone deacetylase-dependent corepressor complex (Table 3) and showed increased skipping of a small 22 bp exon with strong splice sites (Table 4) alongside with total domain loss. *SYNJ2BP* (Fig 15) is implicated in receptor binding activity (Table 3) and exhibited increased skipping of a 137 bp exon with strong splice sites (Table 4) alongside complete domain loss. *TKT* (Fig 16), with functions relating to glycolysis (Table 3), showed increased skipping of a small 24 bp exon with a strong 5'ss and very weak 3'ss alongside partial domain loss. *TNFRSF12A* (Fig 16) is implicated in apoptotic cell signaling (Table 3) and showed increased inclusion of a 135 bp exon with strong splice sites (Table 4) alongside complete domain gain. Of these exon skipping events, a majority involved moderately strong to strong splice sites. Five exons showed increased inclusion, and seven showed increased skipping of the exon of interest.

Thirdly, there were nine genes of the thirty-two genes near SNF5 binding sites with alternative 3' and/or 5' splice sites. Two genes, *ARHGEF1* (Fig 8) and *IGFBP2* (Fig 11), showed alternative 5' splice sites in terminal exons which are located at the terminal ends of the affected transcript. *ARHGEF1* is implicated in GTPase activation (Table 3) and showed no change in domains. *IGFBP2* is implicated in tumor suppression and growth regulation and showed partial domain gain. Seven genes, *POU6F1, SUOX, TGFB1, TXLNA, UST, ZNF362,* and *ZNRF2,* showed alternative 3' splice sites in terminal exons. *POU6F1* which also exhibited exon skipping had unaffected domains.

SUOX (Fig 15) is implicated to be involved with oxidoreductase (Table 3) and had unaffected domains. *TGFB1* (Fig 15) plays a role in recruitment of transcription factors (Table 3) and exhibited complete domain gain. *TXLNA* (Fig 16) is implicated in syntaxin binding activity (Table 3) and exocytosis and showed no change in domains. *UST* (Fig 17) is implicated in transferase activity (Table 3) and showed complete domain loss. *ZNF362* (Fig 18) is implicated in transcription factor activity and transcription regulation (Table 3) and showed complete domain gain. *ZNRF2* (Fig 18) is implicated in ubiquitin protein ligase activity (Table 3) and showed complete domain gain. There were two genes, *IFITM3* and *OAF,* that exhibited both A5'ss and A3'ss in terminal exons. *IFITM3* (Fig 11) interacts with antiviral proteins and immunity (Table 3) and showed complete domain gain. *OAF* (Fig 12) plays a role in transcription regulation (Table 3) and exhibited complete domain gain.

Next, there were five genes, *AMN1, CD2BP2, COMMD2, WDR4,* and *ZNF251,* with complex events involving A3'ss that included multiple events and/or didn't fit well into other categories. *AMN1* (Fig 8) is involved in protein catabolic processes (Table 3) and had unaffected domains. *CD2BP2* (Fig 9) is implicated in mRNA processing and splicing (Table 3) and had no change to domains. *COMMD2* (Fig 9) plays a role in transcription regulation (Table 3) and exhibited complete domain loss. *WRD4* (Fig 17) plays a role in tRNA processing (Table 3) and had unaffected domains. *ZNF251* (Fig 17) is implicated in transcription factor activity and transcription regulation (Table 3) and showed complete domain loss.

Also, there were three genes, *C1orf50, RPP21,* and *TKT,* that showed alternative 3' splice sites in internal exons. *Clorf50* (Fig 9), implicated in identical protein binding

regulation (Table 3), showed increased usage of a stronger 3’ splice site (Table 5), resulting in a smaller exon but no affected domains. *RPP21* (Fig 13), which codes for a ribonuclease subunit (Table 3), showed increase usage of a weaker 3’ splice site (Table 5), resulting in a larger exon and partial domain gain. *TKT* (Fig 16), implicated in glycolysis (Table 3), showed increased usage of moderately strong 3’ splice site (Table 5), resulting in a smaller exon and partial domain loss. In addition, *SMARCB1* (Fig 14), which codes for the SNF5 subunit of the SWI/SNF5 chromatin remodeling complex (Table 3), showed increased usage of a moderately weak 5’ss (Table 5), resulting in a smaller exon and complete domain loss.

Finally, one gene, *FKBP11,* of the thirty-two genes located near SNF5 binding sites showed intron retention. *FKBP11* (Fig 10) may have functions relating to isomerase and rotamase (Table 3) and resulted in complete domain loss. The transcript with increased usage (ENST00000553027.1) retained all five introns present in the transcript with decreased usage, resulting in one large 4,957 bp exon. Taken together, these data indicate that chromatin remodeling via SNF5-containing complexes is not a global mechanism affecting alternative splicing or specific alternative events such as exon skipping, but rather supports an alternative hypothesis that it specifically influences a small number of genes and specific splicing events.

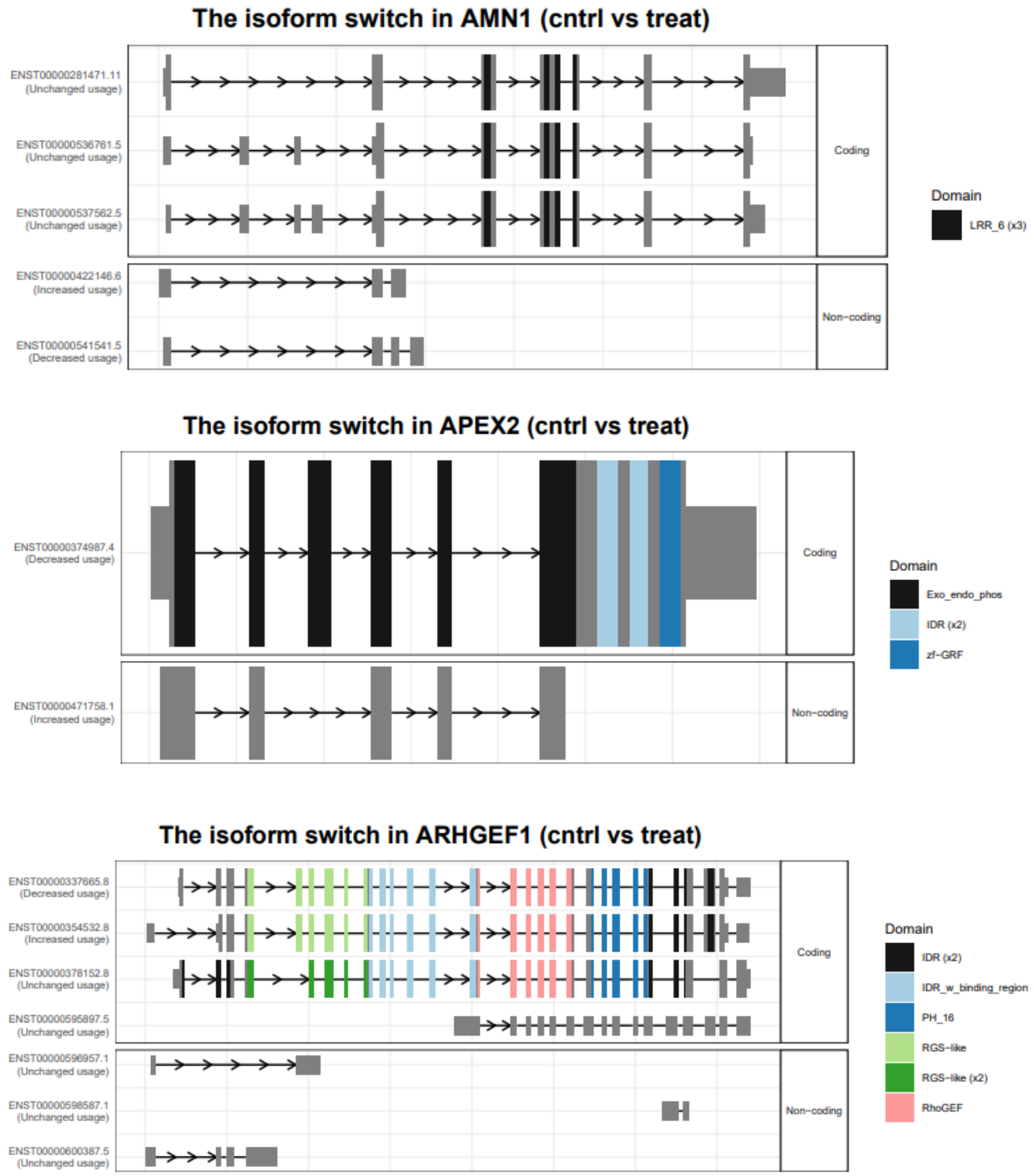


Fig. 8 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *AMN1, APEX2,* and *ARHGEF1* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot A visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$ cutoff.

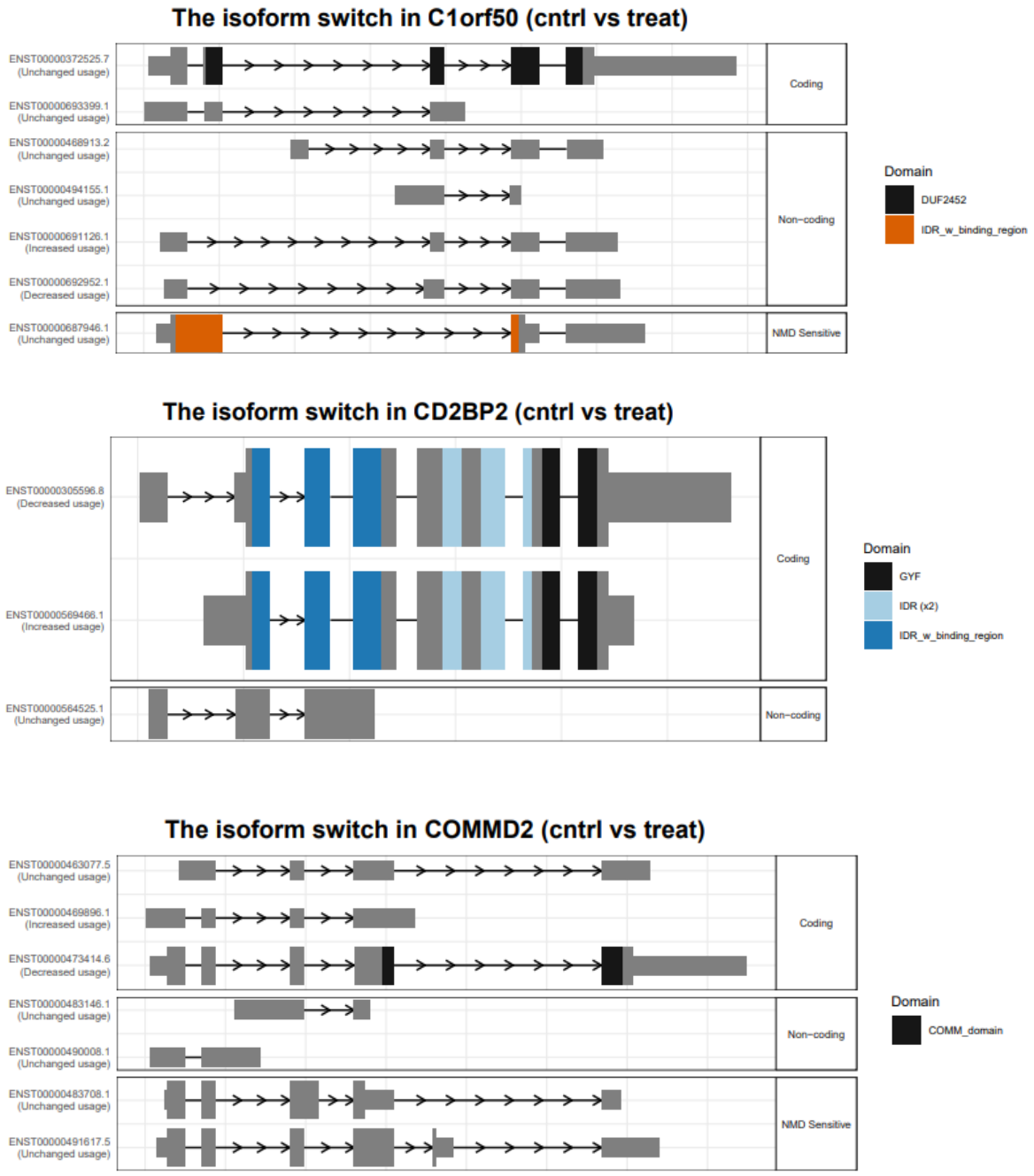


Fig. 9 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *C1orf50, CD2BP2,* and *COMMD2* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot A visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$ cutoff.

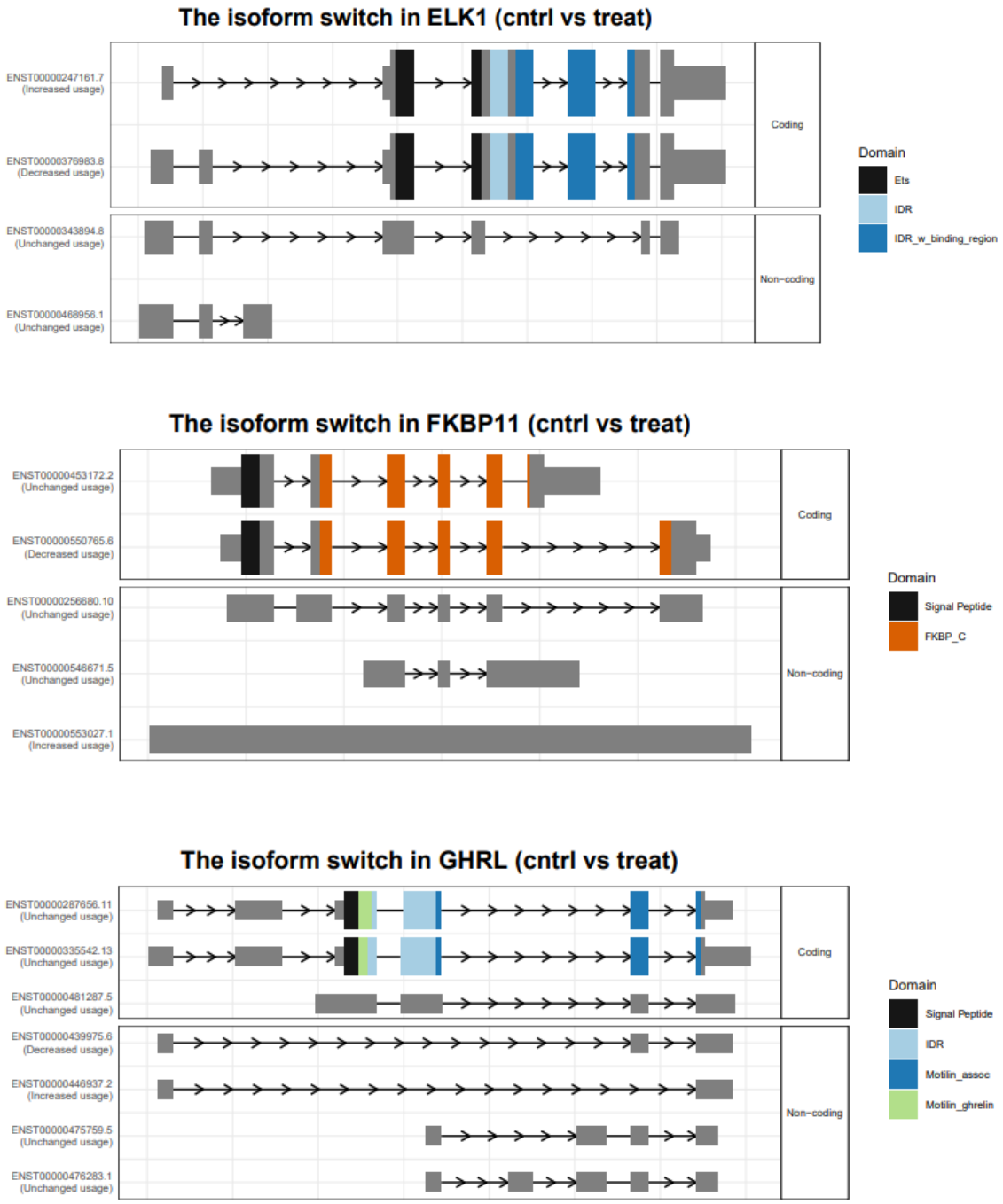


Fig. 10 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *ELK1, FKBP11,* and *GHRL* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

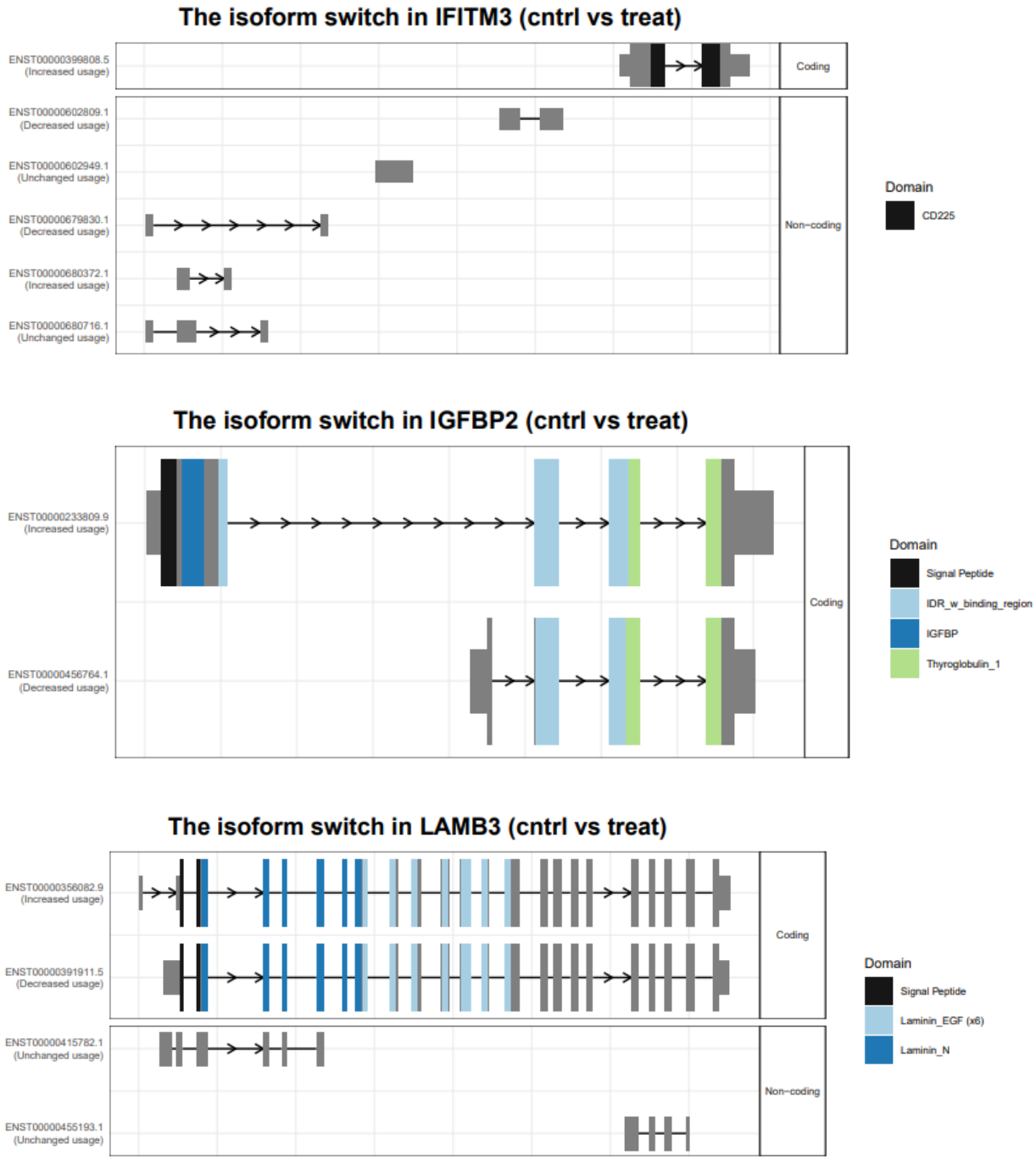


Fig. 11 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *IFITM3, IGFBP2,* and *LAMB3* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

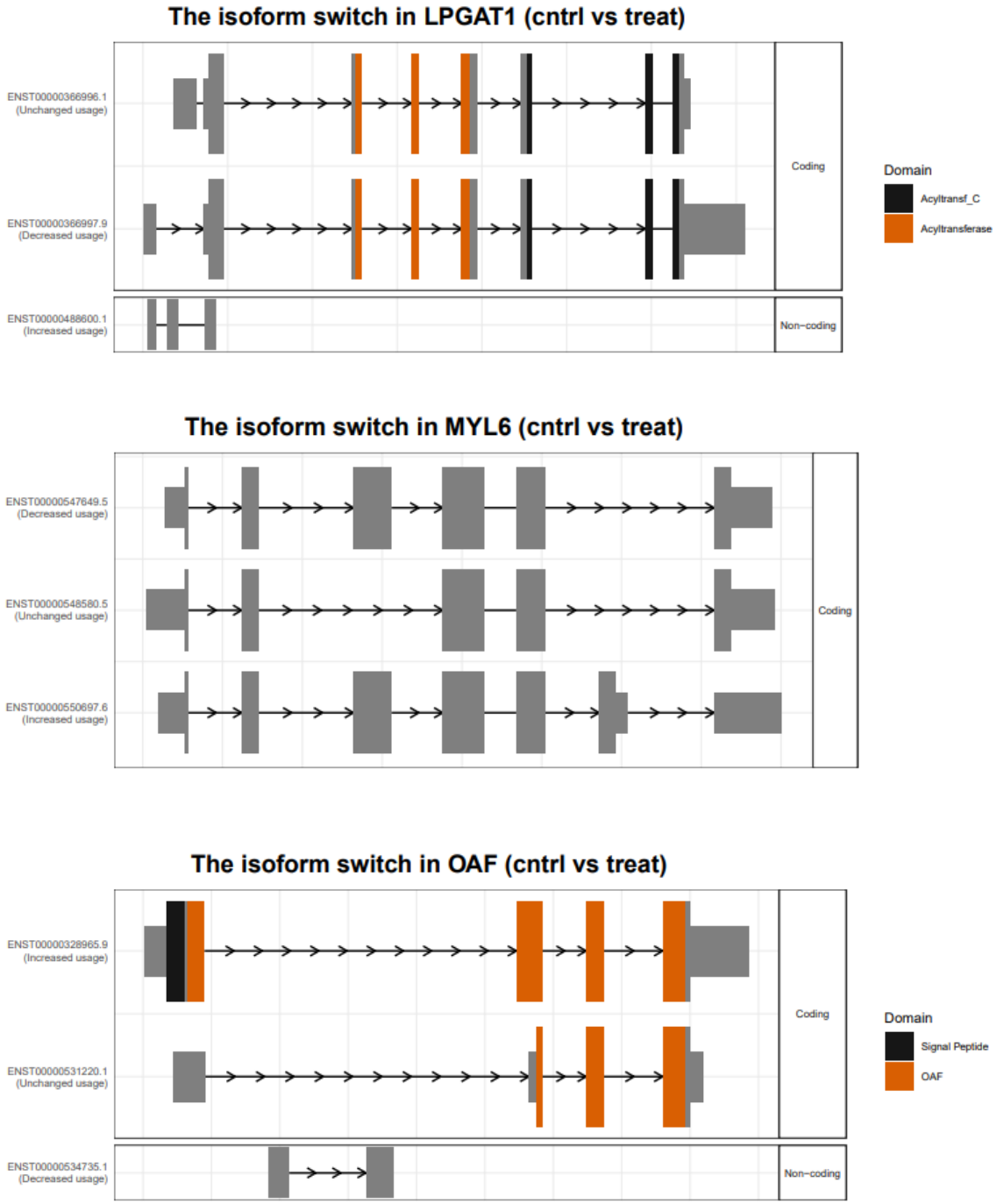


Fig. 12 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *LPGAT1, MYL6,* and *OAF* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

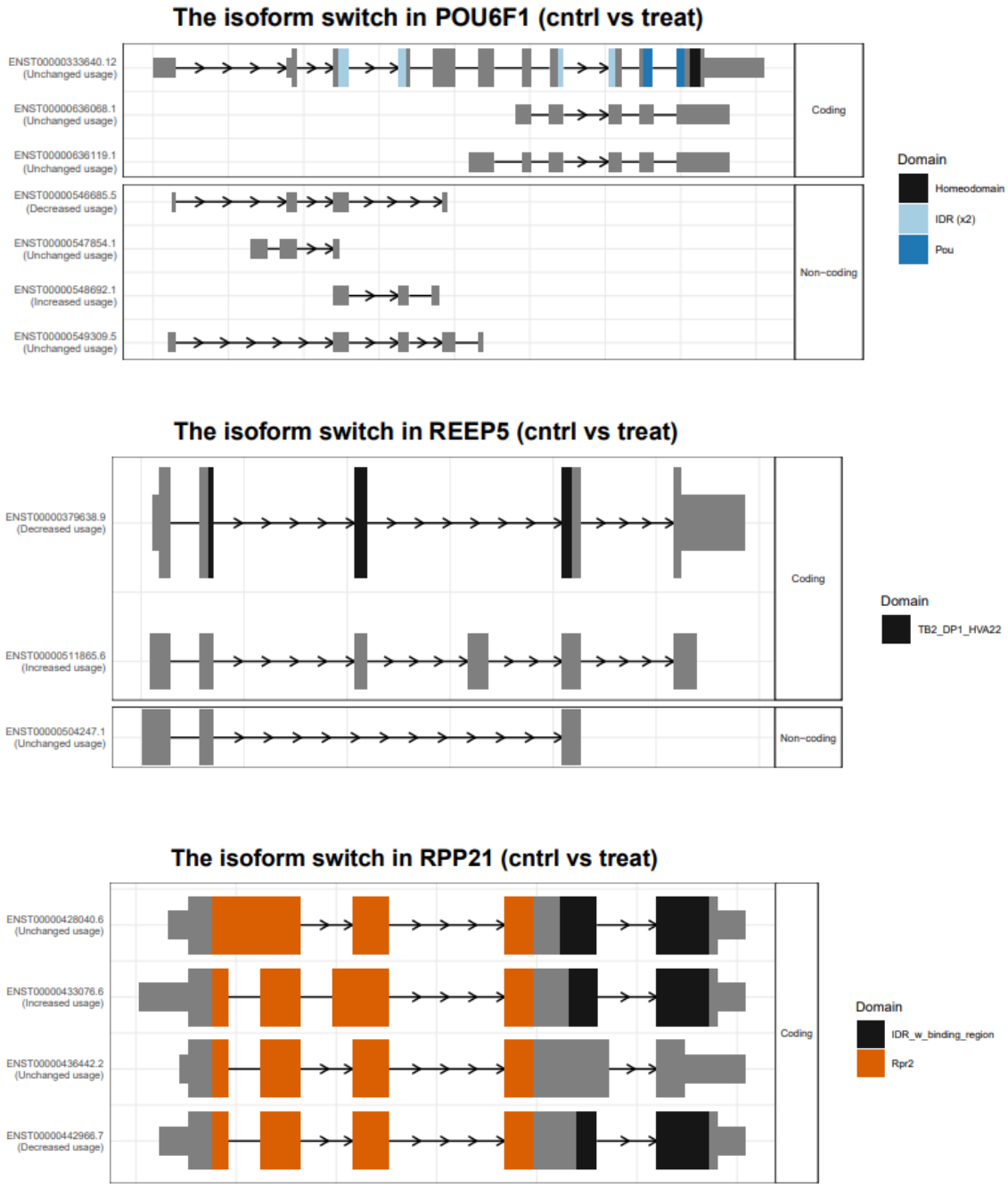


Fig. 13 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *POU6F1, REEP5,* and *RPP21* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

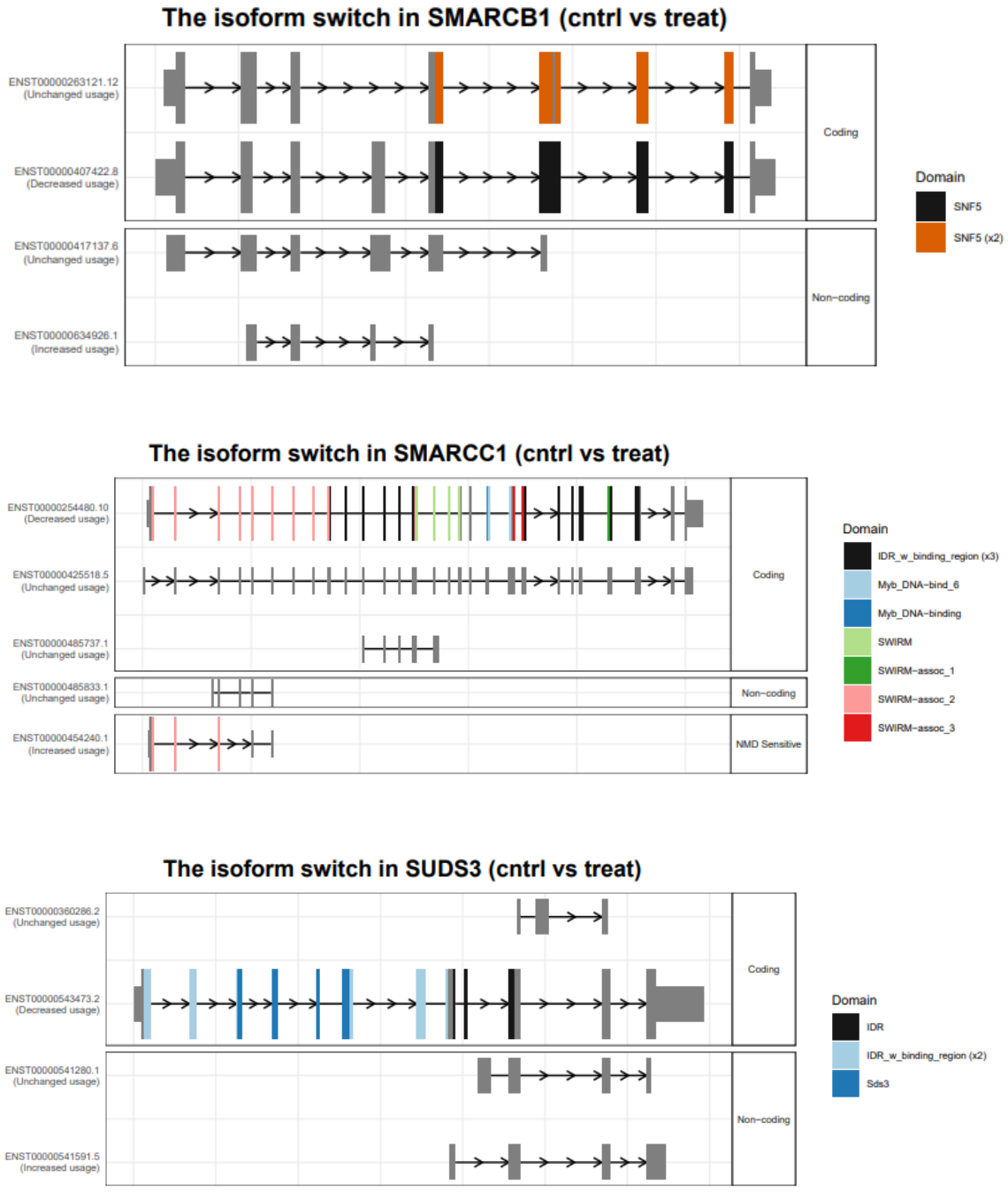


Fig. 14 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *SMARCB1, SMARCC1,* and *SUDS3* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$ cutoff.

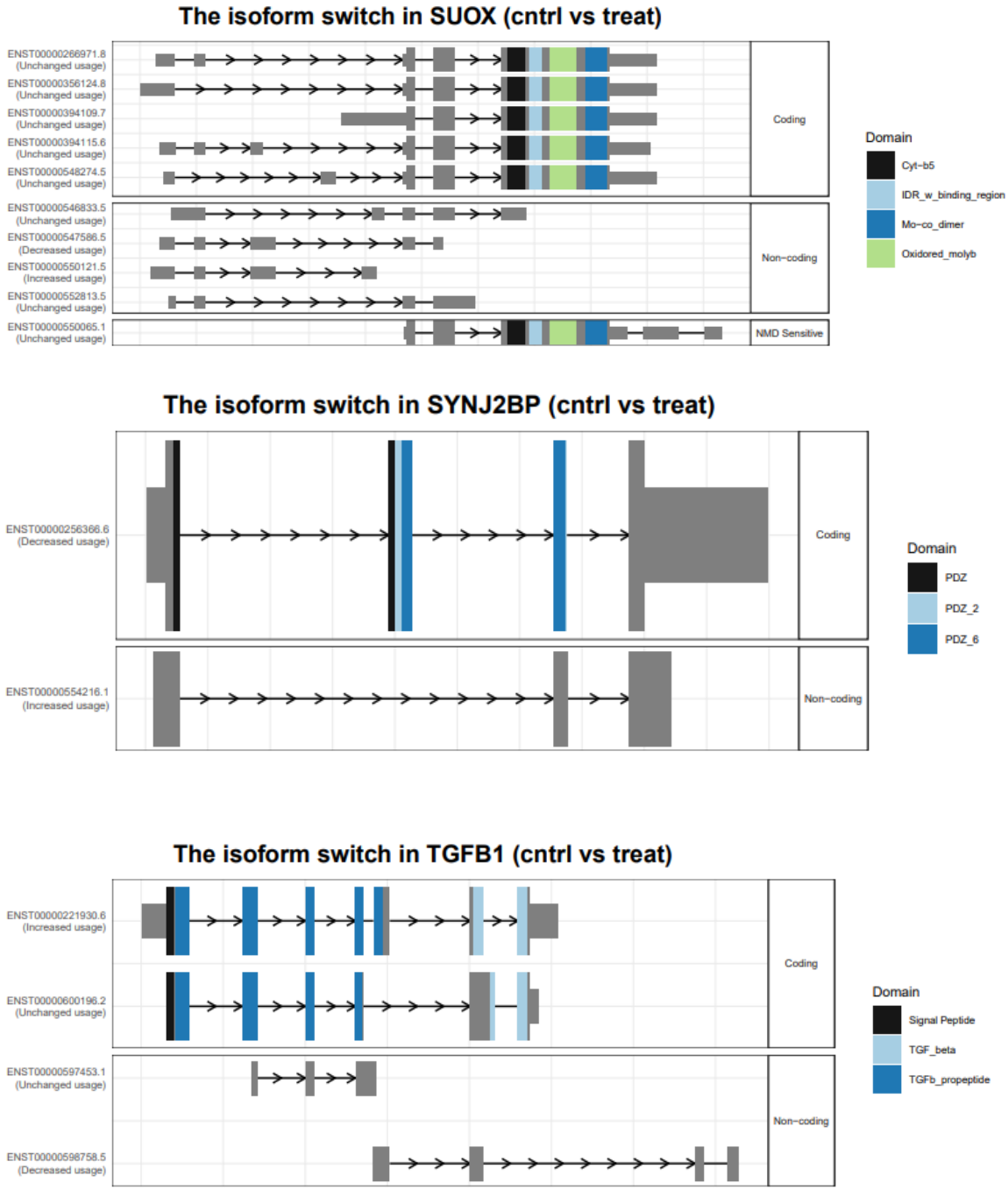


Fig. 15 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *SUOX, SYNJ2BP,* and *TGFB1* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

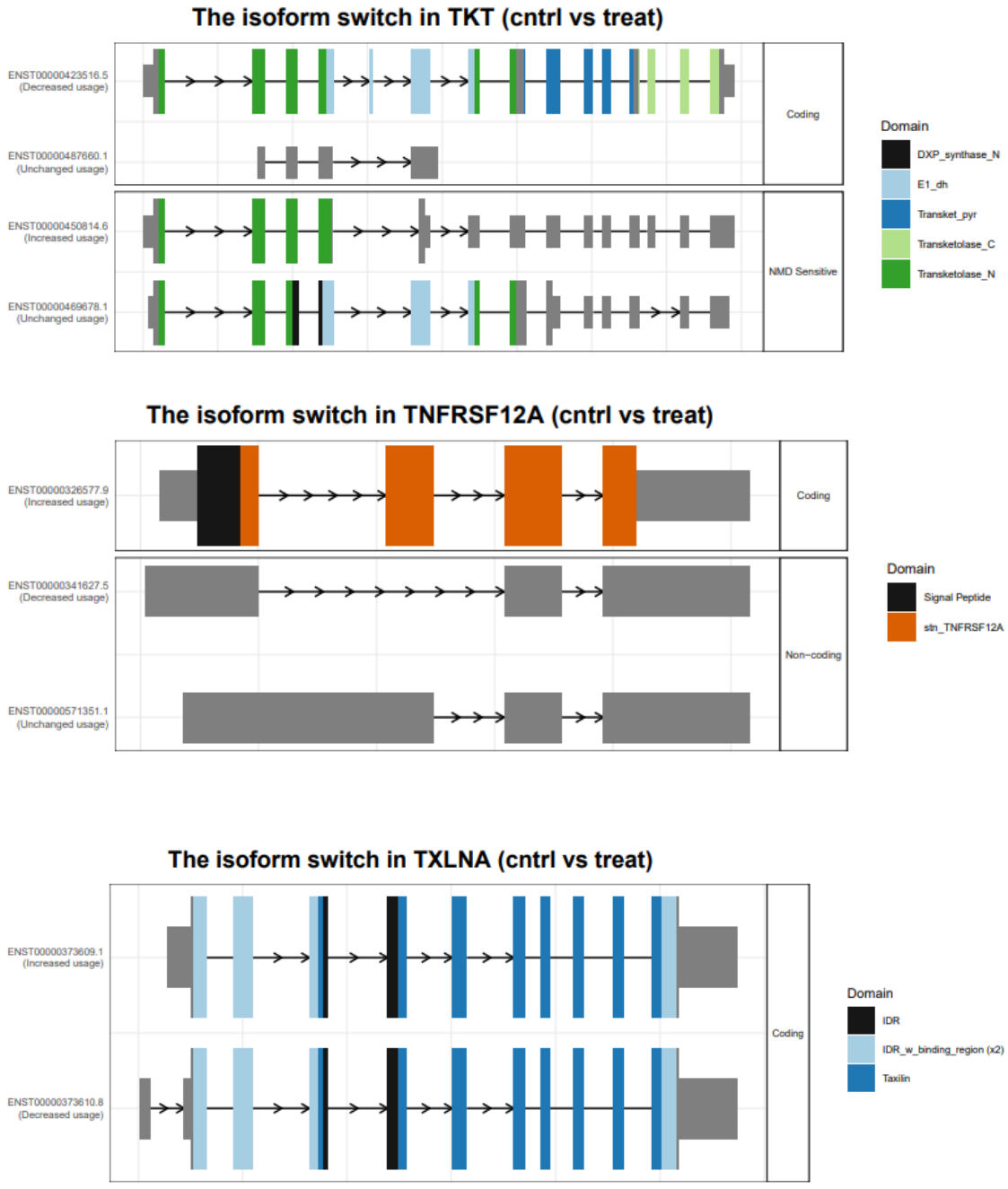


Fig. 16 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *TKT, TNFRSF12A,* and *TXLNA* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

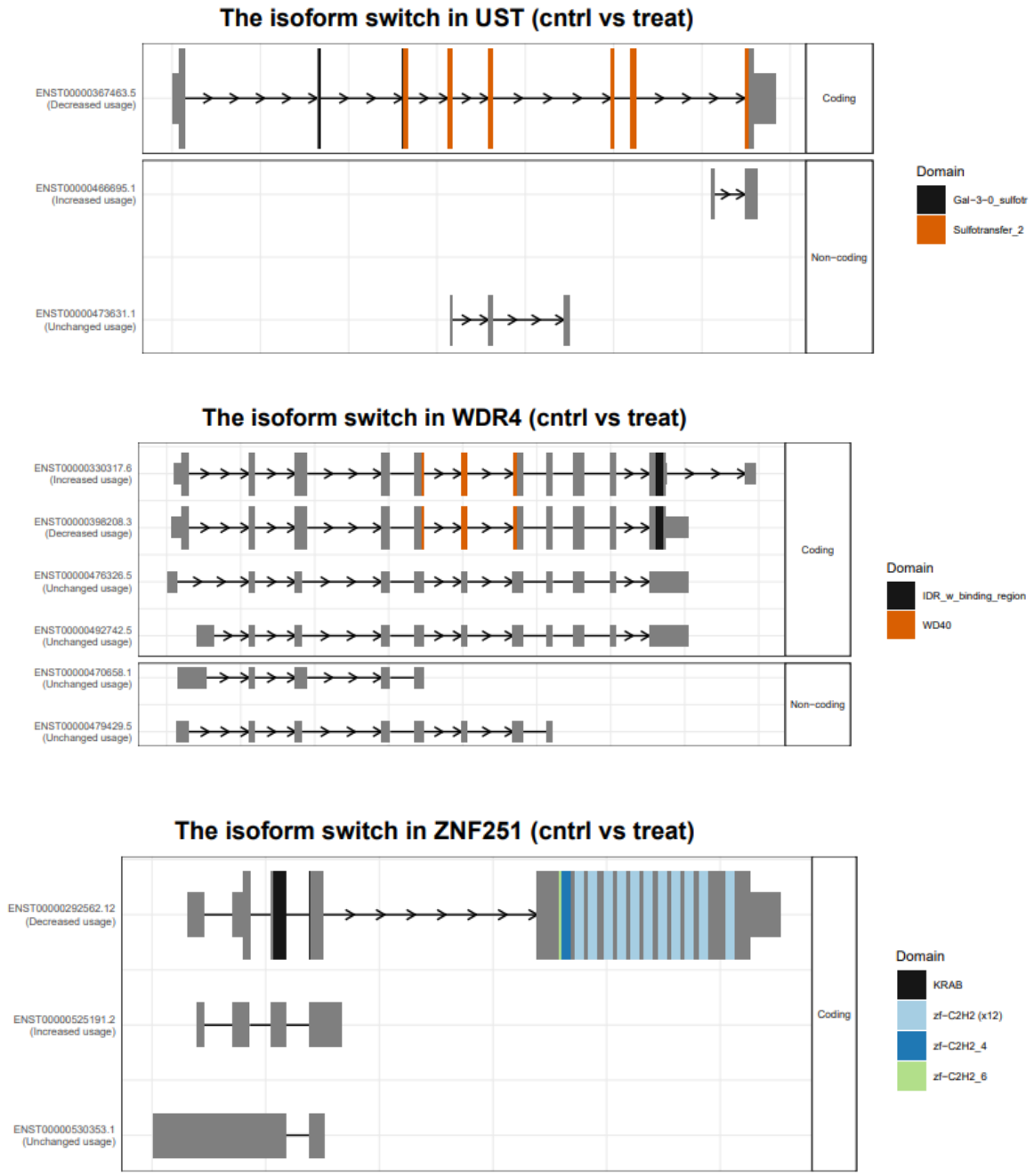


Fig. 17 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *UST, WDR4,* and *ZNF251* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$cutoff.

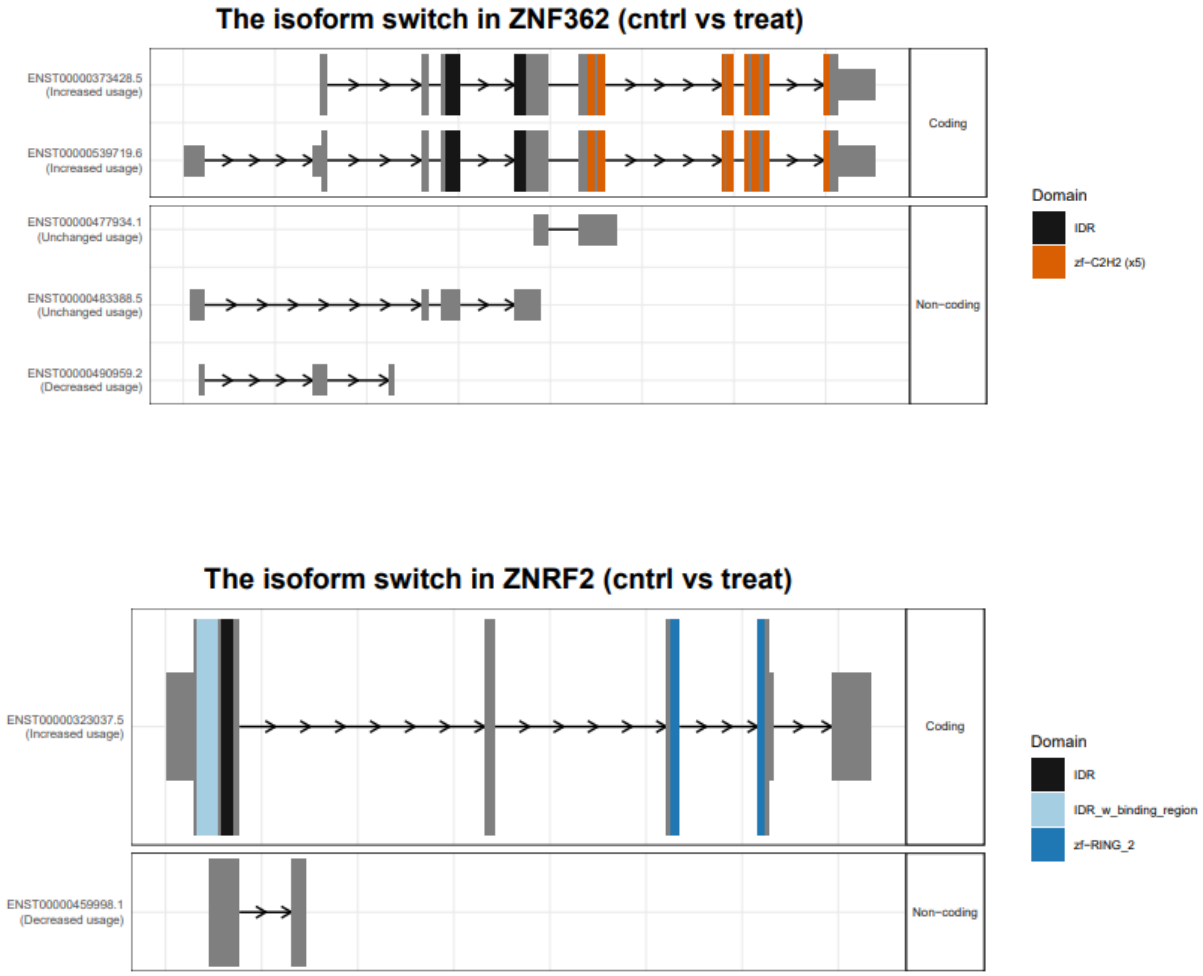


Fig. 18 – Gene Transcript Event Summary. A switch plot to visualize transcript structure, gene and isoform expression, and isoform usage for genes *ZNF362* and *ZNRF2* noted at the top of the figure. Plots created in Rstudio using IsoformSwitchAnalyeR ver. 1.16.0 (Vitting-Seerup and Sandelin 2019). To be plotted, transcripts need a minimum contribution to gene expression, defined as isoform fraction ($IF = \frac{iso_{exp}}{gene_{exp}}$) > 0.05. Plot visualizes transcript structure with boxes representing exons and lines representing introns. Transcripts are labeled with unique Ensembl (Howe et al. 2021) IDs and rescaled to the square root of their original size. Transcript status (coding, nonsense-mediated decay sensitive, non-coding) is noted on the right, as predicted using CPC2 (Kang 2017). Colors indicate predicted domain regions using Pfam (Potter 2018), signal peptides using SignalP (Almagro et al. 2019), and disordered regions using IuPred (Erdős and Dosztányi 2020). Change in isoform usage is noted underneath Ensembl ID and defined as $|(dIF = IF_2 - IF_1)| > dIF$ cutoff.

Consolidated Biological Consequences of SNF5-Degradation

Since our results did not indicate SNF5-depletion had global genomic effects on alternative splicing or specific alternative splicing events, we next wanted to consider these data in context of a more health-related question, SNF5 involvement in pathways related to development of pediatric cancer (Biegel JA 2014). Mutations in *SMARCB1* are common in 95% of pediatric rhabdoid tumors and loss of *SMARCB1* expression is a diagnostic hallmark of these tumors (Orlando et al. 2019). In addition, recent studies continue to provide more support to the theory that *SMARCB1* functions as a tumor suppressor gene and inactivation can drive development of other human cancers (Nakayama et al. 2017; Wang et al. 2017). To gain a more comprehensive list of genes whose expression is affected by SNF5 depletion, we combined our list of 214 differentially expressed genes with a fold change greater than or equal to two, for biological relevance, with our list of 65 genes found with significant switching among isoforms to form a group of 279 genes to be analyzed further, as there was very little overlap between the two groups. To determine biological pathways affected by *SNF5* perturbation, the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) were both used to investigate which biological pathways are enriched in the consolidated gene set (DAVID) and which proteins encoded by these genes form protein-protein interaction networks (STRING). The top nine most enriched pathways identified by DAVID show the zinc finger/transcription associated pathway was the most enriched with an enrichment score of 4.89 (Fig 19). The glycoprotein pathway and cancer pathway being second and third most enriched with scores of 2.68 and 2.40 respectively. In

descending order, the next enriched pathways were cadherins and cell adhesion, axon guidance, differentiation, metallopeptidase, transcription/RNA pol II binding, and chromatin and nucleosome structures (Fig 19).

Next, analysis of the protein-protein interaction network identified five major areas of protein crosstalk (Fig 20). The central node, shown in yellow (Fig 20), contains the majority of the protein-protein interactions and is comprised of 49 genes that, when inputted into DAVID (Huang da W et al. 2009), result in an enriched pathway (score of 6.92) relating to proteoglycans in cancer and hepatocellular carcinoma which may indicate a possible correlation with development of pediatric rhabdoid tumors . The bottom-left node, shown in cyan (Fig 20), contains 65 genes with an enriched pathway (score of 2.37) relating to zinc-finger proteins and transcription. The bottom-right node, shown in red (Fig 20), contains 52 genes with an enriched pathway (score of 2.26) relating to metabolism. The rightmost node, shown in green (Fig 20), contains 55 genes with an enriched pathway (score of 1.53) relating to transcription regulation and chromatin. The leftmost node, shown in blue (Fig 20), contains 33 genes with an enriched pathway (score of 0.80) relating to cell adhesion and cell membrane. Taken together, these consolidated results show SNF5 perturbation affects expression (RNA level and isoform switching) for genes involved in multiple pathways related to cancer, transcription, metabolism, chromatin, and cell adhesion.

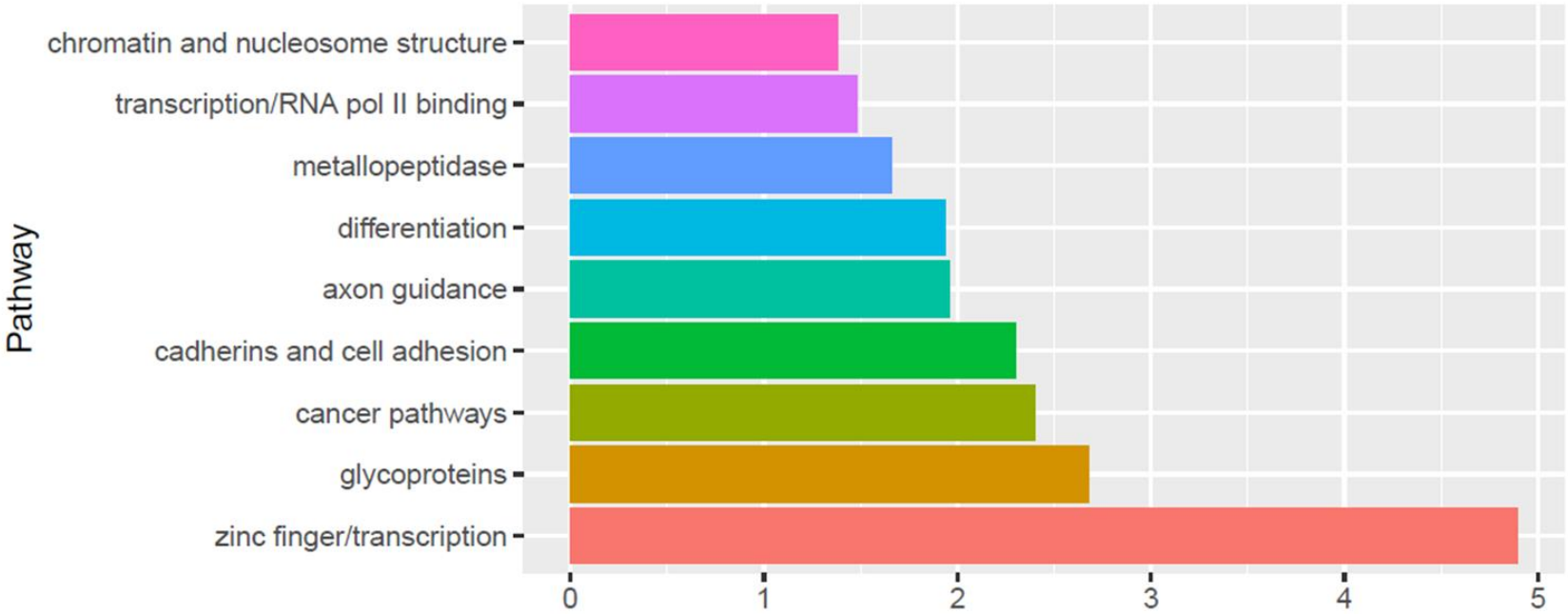


Fig. 19 – Summary of Enriched Pathways as Identified by DAVID. Enriched pathways and associated enrichment score obtained from the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang da W et al. 2009). Bars are sorted with the most enriched being the zinc finger/transcription pathway with an enrichment score of 4.89. A list of 279 genes, all either found to be differentially expressed with a fold change greater than or equal to two or found with significant switching among isoforms in response to *SNF5* perturbation were assigned functional annotation to produce the predicted enriched pathways. Numbers represent each pathways enrichment score.

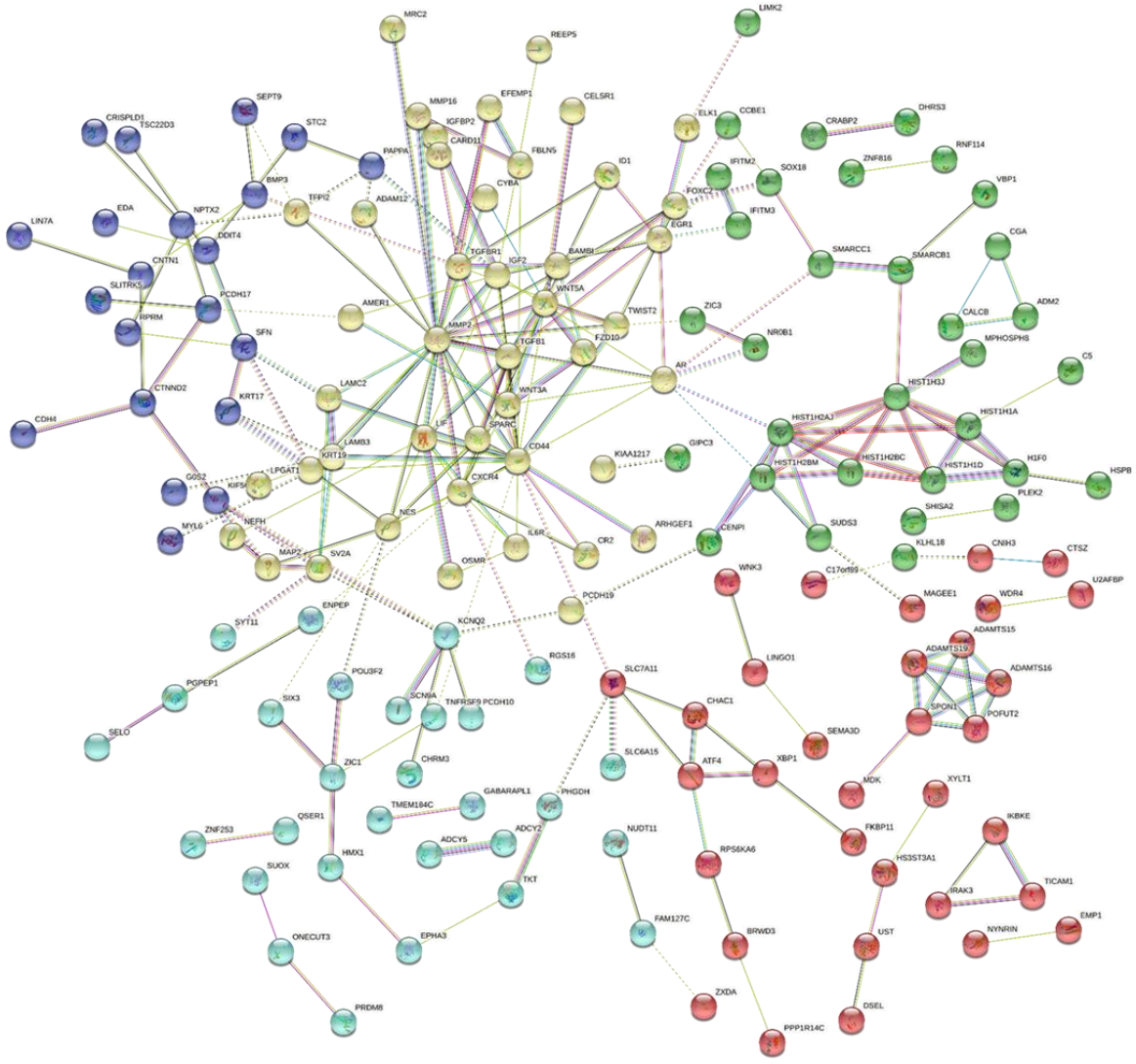


Fig. 20 – A visual network of protein-protein interactions between differentially expressed genes and genes with alternative splicing and/or transcription events as a result of *SNF5* perturbation, as identified by STRING . Image obtained via STRING (Szklarczyk D et al. 2019) showing protein-protein interactions between differentially expressed genes and genes with alternative splicing and/or transcription events. Five nodes were identified via k-means clustering and are color-coordinated.

CONCLUSIONS

SNF5, also known as SMARCB1/BAF47, was first identified in budding yeast and appeared to function in transcription (Laurent et al 1990). It has since been identified as a protein component of a chromatin remodeling complex that relieves repressive chromatin structures (Luco 2010). Mutation in the gene has been implicated in Coffin-Siris syndrome3 (Wieczorek et al 2013) and several types of cancer, including a majority of cases with pediatric rhabdoid tumors (Mittal and Roberts 2020). This gene and its encoding protein, therefore, have intriguing functions and broad ranging impacts. Our interest in this gene focused first on its role in chromatin remodeling. Prior correlative evidence suggested that chromatin could regulate alternative splicing (Luco 2010). In the first part of this study, we identified alternative splicing changes due to perturbation of the SNF5/SMARCB1 subunit of two known SWI/SNF chromatin remodeling complexes, cBAF and pBAF. We examined gene features known to affect splicing efficiency, unusual exon sizes (>300bp;<50bp), splice consensus sequence strength, and RNA levels. The number of genes we identified is small (65 of approx. 20,000) with eighteen showing exon skipping, three showing intron retention, twelve showing alternative 5’ splice sites located in a terminal exon, five showing internal 3’ alternative splicing, and a further twelve showing 3’ alternative splicing in a terminal exon, with eight additional genes showing complex splicing/terminal exon events. Thirty-two of these genes showing alternative splicing events were identified as near a SNF5 binding site, suggesting these may be primary effects. Recall that approximately 95% of human genes are postulated to

have at least one alternative isoform due to splicing (Zhu et al. 2018); if SNF5-based chromatin remodeling regulated alternative splicing more generally, we would have expected a much greater effect on splicing across the genome. Since we did not observe a high percentage of genes whose splicing is affected by SNF5 perturbation, we cannot conclude SNF5 remodeling plays a major role in global AS changes. Instead, we hypothesize gene-specific effects on splicing. These 32 genes were grouped into eight functional groups: seven genes had functions relating to transcription and/or transcription regulation, four genes had functions relating to chromatin and/or histones, four relating to RNA and/or DNA-damage repair, four relating to protein binding and/or transport, three relating to health and/or disease progression, two relating to cell signaling and/or receptor activity, six with miscellaneous enzyme activity, and four with functions unrelated to any other category.

Since alternative splicing due to SNF5 degradation likely affects specific genes, we next sought to combine this gene group with the genes identified in a traditional differential expression transcriptome gene set to get a comprehensive view of all measurable gene expression changes (splicing, transcription, RNA decay). Of the combined gene set including alternative event genes and differentially expressed genes, a total of 279, a total of nine enriched pathways were found, with the most enriched being the zinc finger/transcription pathway (with an enrichment score of 4.89). The eight other pathways are listed in descending order: glycoproteins, cancer, cadherins and cell adhesion, exon guidance, differentiation, metallopeptidase, transcription/RNA pol II binding, and chromatin and/or nucleosome structure. Many of the affected genes had

functions relating to transcription regulation (Table 3, Fig 19) which further strengthens the correlation between SNF5/SMARCB1 and transcription regulation.

Looking ahead, while SNF5-based chromatin remodeling may not play a role in global alternative splicing, other remodelers might. Investigating alternative splicing in response to perturbation of other known core subunits of the mammalian cBAF and pBAF complexes, such as SMARCC1 and SMARCC2 which also exist in the ncBAF complex, could provide a clearer understanding of the relationship between chromatin and alternative splicing. Furthermore, we suggest that the addition of alternatively spliced genes to a traditional differential expression gene set as we have done will provide a more comprehensive view of gene expression changes, and we propose this become a common part of transcriptome analysis pipelines.

LITERATURE CITED

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*. 2019;37(4):420–423. doi:10.1038/s41587-019-0036-z

Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 2012;22(10):2008-2017. doi:10.1101/gr.133744.111

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 2019. Babrahamacuk. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res*. 2011;21(3):381-395. doi:10.1038/cr.2011.22

Biegel JA, Busse TM, Weissman BE. SWI/SNF chromatin remodeling complexes and cancer. *Am J Med Genet C Semin Med Genet*. 2014;166C(3):350-366. doi:10.1002/ajmg.c.31410

Blankenberg D, Von Kuster G, Bouvier E, et al. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol*. 2014;15(2):403. Published 2014 Feb 20. doi:10.1186/gb4161

De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA*. 2013;4(1):49-60. doi:10.1002/wrna.1140

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635

Erdős G, Dosztányi Z. Analyzing Protein Disorder with IUPred2A. Current Protocols in *Bioinformatics*. 2020;70(1). https://onlinelibrary.wiley.com/doi/10.1002/cpbi.99. doi:10.1002/cpbi.99

Euskirchen GM, Auerbach RK, Davidov E, et al. Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet*. 2011;7(3):e1002008. doi:10.1371/journal.pgen.1002008

Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021. *Nucleic Acids Res*. 2021;49(D1):D916-D923. doi:10.1093/nar/gkaa1087

Hahn S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol*. 2004;11(5):394-403. doi:10.1038/nsmb763

Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science*. 2009;325(5940):626-628. doi:10.1126/science.1172926

Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res*. 2021;49(D1):D884-D891. doi:10.1093/nar/gkaa942

Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, Gao G. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*. 2017;45(W1):W12–W16. doi:10.1093/nar/gkx428

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*. 2019;20(1):278. doi:10.1186/s13059-019-1910-1

Kurosaki T, Maquat LE. Nonsense-mediated mRNA decay in humans at a glance. *J Cell Sci*. 2016;129(3):461-467. doi:10.1242/jcs.181008

Laurent BC, Treitel MA, Carlson M. The SNF5 protein of Saccharomyces cerevisiae is a glutamine- and proline-rich transcriptional activator that affects expression of a broad spectrum of genes. Mol Cell Biol. 1990;10(11):5616-5625. doi:10.1128/mcb.10.11.5616-5625.1990

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930. doi:10.1093/bioinformatics/btt656

Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and visualization of RNA-Seq expression data using RStudio, Bioconductor, and Integrated Genome Browser. *Methods Mol Biol*. 2015;1284:481-501. doi:10.1007/978-1-4939-2444-8_24

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. *Cell*. 2011;144(1):16-26. doi:10.1016/j.cell.2010.11.056

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. *Science*. 2010;327(5968):996-1000. doi:10.1126/science.1184208

Mittal P, Roberts CWM. The SWI/SNF complex in cancer - biology, biomarkers and therapy. *Nat Rev Clin Oncol*. 2020;17(7):435-448. doi:10.1038/s41571-020-0357-3

Nakayama RT, Pulice JL, Valencia AM, et al. SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. *Nat Genet*. 2017;49(11):1613-1623. doi:10.1038/ng.3958

Orlando KA, Nguyen V, Raab JR, Walhart T, Weissman BE. Remodeling the cancer epigenome: mutations in the SWI/SNF complex offer new therapeutic opportunities. *Expert Rev Anticancer Ther*. 2019;19(5):375-391. doi:10.1080/14737140.2019.1605905

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419. doi:10.1038/nmeth.4197

Pertea M, Shumate A, Pertea G, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*. 2018;19(1):208. Published 2018 Nov 28. doi:10.1186/s13059-018-1590-2

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Research*. 2018;46(W1):W200–W204. doi:10.1093/nar/gky448

Reddy D, Workman JL. Targeting BAF-perturbed cancers. *Nat Cell Biol*. 2018;20(12):1332-1333. doi:10.1038/s41556-018-0246-5

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–140. doi:10.1093/bioinformatics/btp616

Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*. 2009;16(9):990-995. doi:10.1038/nsmb.1659

Smith LM, Agar JN, Chamot-Rooke J, et al. The Human Proteoform Project: Defining the human proteome. *Sci Adv*. 2021;7(46):eabk0734. doi:10.1126/sciadv.abk0734

Spies N, Nielsen CB, Padgett RA, Burge CB. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*. 2009;36(2):245-254. doi:10.1016/j.molcel.2009.10.008

Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

Tilgner H, Nikolaou C, Althammer S, et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*. 2009;16(9):996-1001. doi:10.1038/nsmb.1658

Vitting-Seerup K, Sandelin A. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Berger B, editor. *Bioinformatics*. 2019;35(21):4469–4471. doi:10.1093/bioinformatics/btz247

Wang X, Lee RS, Alver BH, et al. SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. *Nat Genet*. 2017;49(2):289-295. doi:10.1038/ng.3746

Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. 2016. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wieczorek D, Bögershausen N, Beleggia F, et al. A comprehensive molecular study on Coffin-Siris and Nicolaides-Baraitser syndromes identifies a broad molecular and clinical spectrum converging on altered chromatin remodeling. Hum Mol Genet. 2013;22(25):5121-5135. doi:10.1093/hmg/ddt366

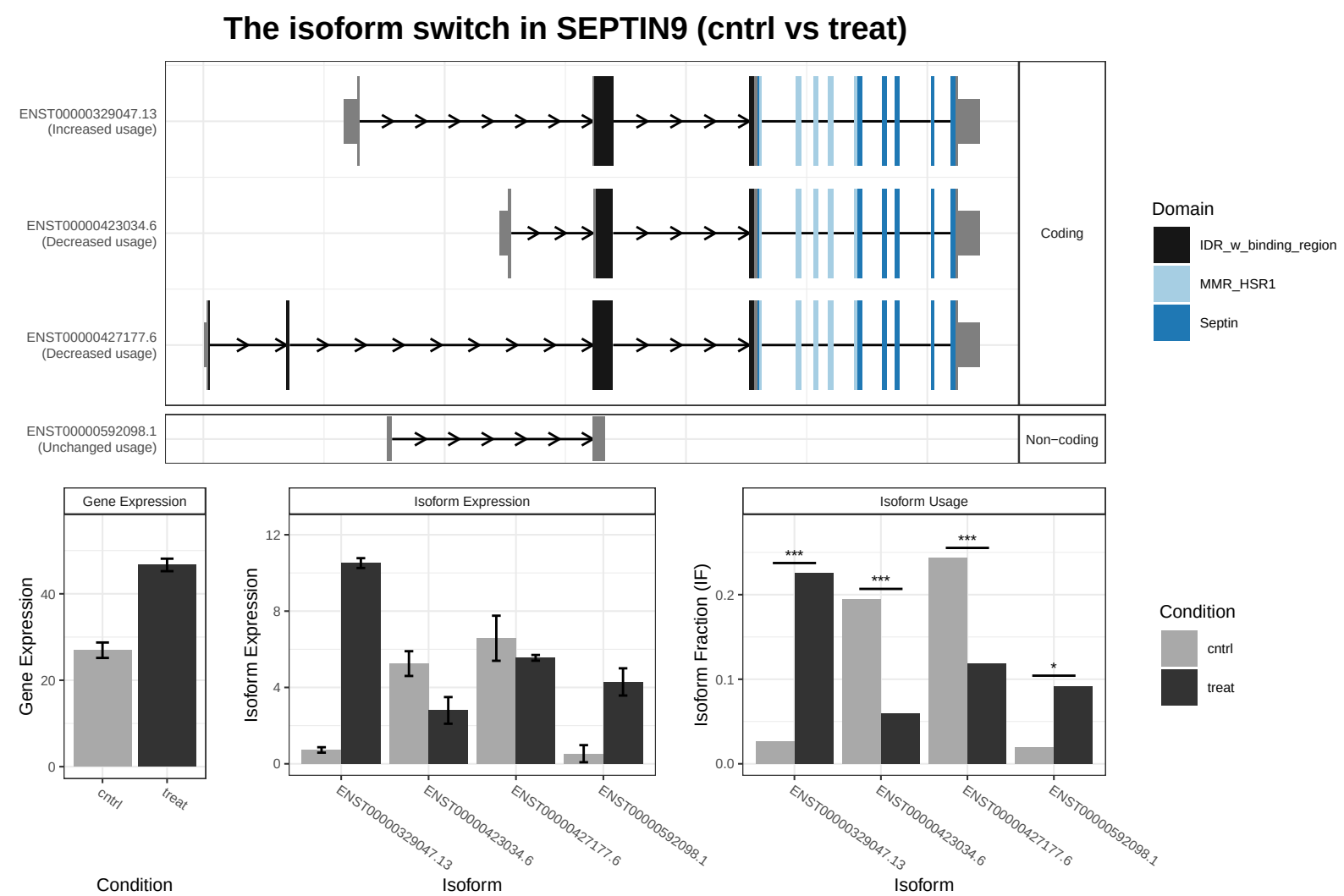
Zhao Z, Shilatifard A. Epigenetic modifications of histones in cancer. *Genome Biol*. 2019;20(1):245. Published 2019 Nov 20. doi:10.1186/s13059-019-1870-5

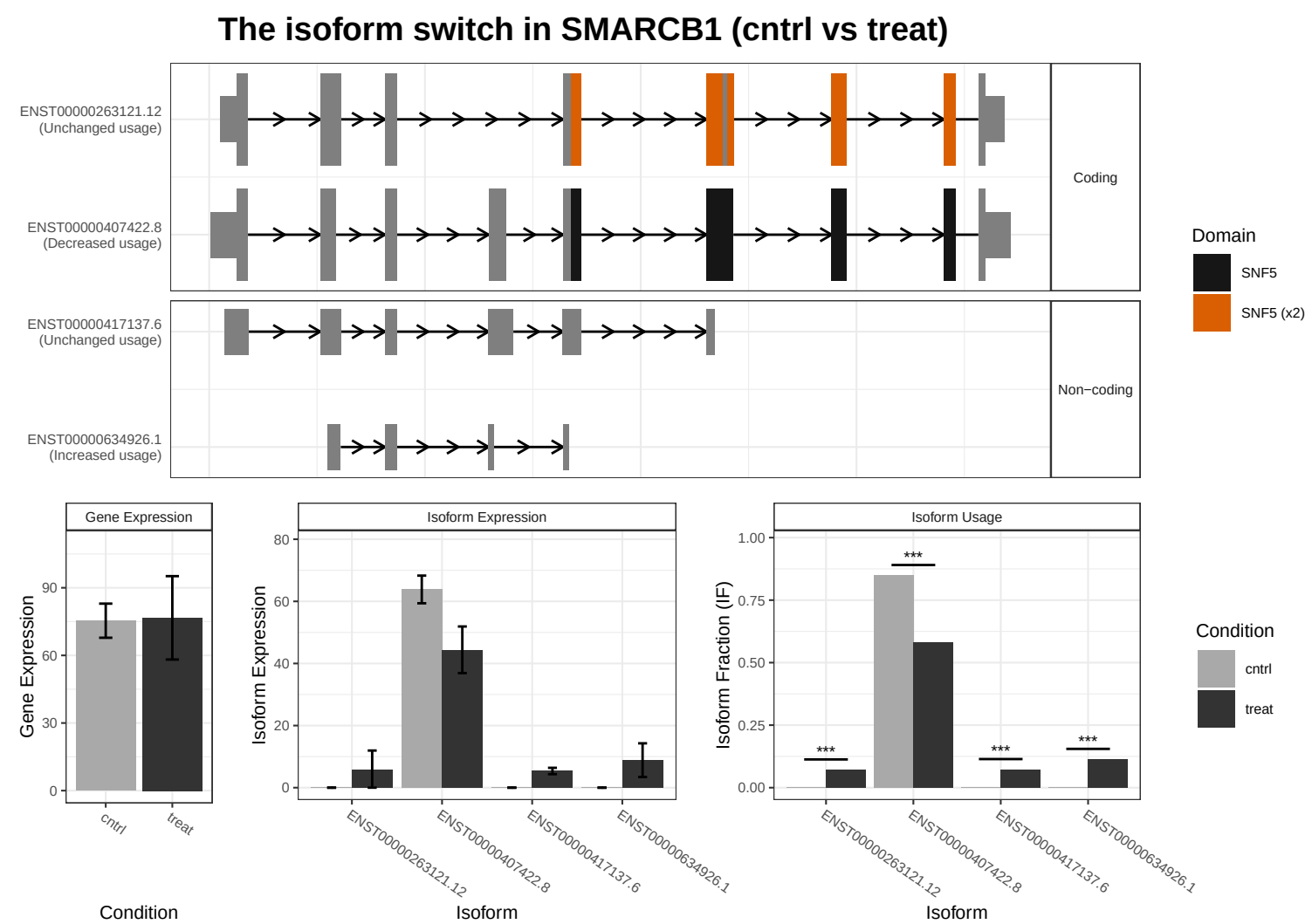
Zhou K, Gaullier G, Luger K. Nucleosome structure and dynamics are coming of age. *Nat Struct Mol Biol*. 2019;26(1):3-13. doi:10.1038/s41594-018-0166-x

Zhu LY, Zhu YR, Dai DJ, Wang X, Jin HC. Epigenetic regulation of alternative splicing. *Am J Cancer Res*. 2018;8(12):2346-2358. Published 2018 Dec 1.

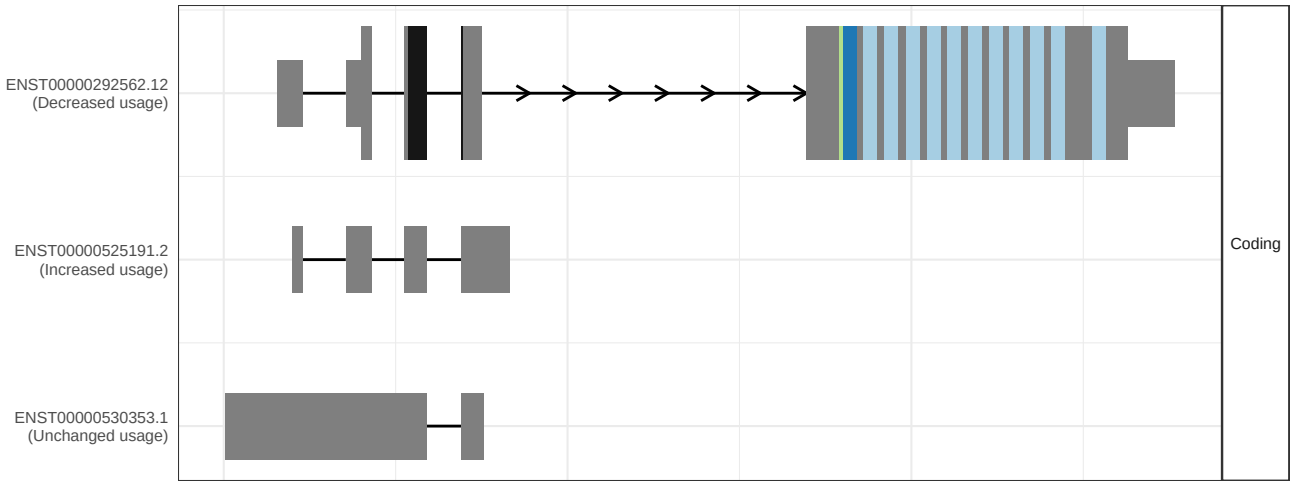
Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics*. 2018;34(24):4180-4188. doi:10.1093/bioinformatics/bty49

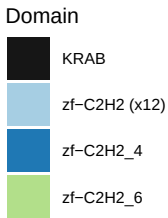
APPENDIX A

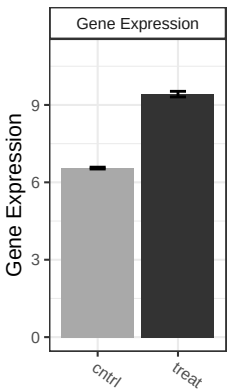

The isoform switch in SEPTIN9 (cntrl vs treat)
ENST00000329047.13
(Increased usage)
ENST00000423034.6
(Decreased usage)
ENST00000427177.6
(Decreased usage)
ENST00000592098.1
(Unchanged usage)
Coding
Non−coding
Domain
IDR_w_binding_region
MMR_HSR1
Septin
Gene Expression
Gene Expression
Condition
cntrl
treat
Isoform Expression
Isoform Expression
Isoform
ENST00000329047.13
ENST00000423034.6
ENST00000427177.6
ENST00000592098.1
Isoform Usage
Isoform Fraction (IF)
***
*
Condition
cntrl
treat

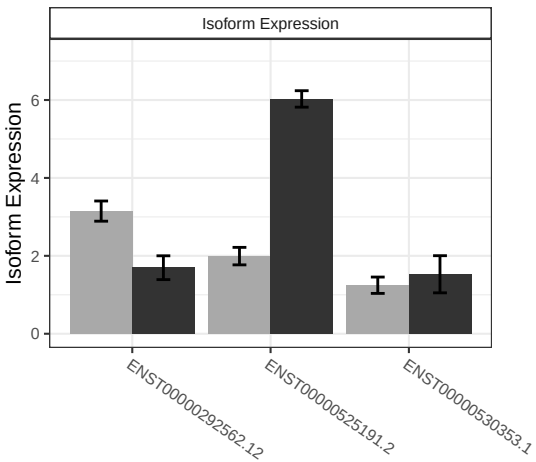

The isoform switch in SMARCB1 (cntrl vs treat)
ENST00000263121.12
(Unchanged usage)
ENST00000407422.8
(Decreased usage)
ENST00000417137.6
(Unchanged usage)
ENST00000634926.1
(Increased usage)
Coding
Non−coding
Domain
SNF5
SNF5 (x2)
Gene Expression
Gene Expression
90
60
30
0
cntrl
treat
Condition
Isoform Expression
Isoform Expression
80
60
40
20
0
ENST00000263121.12
ENST00000407422.8
ENST00000417137.6
ENST00000634926.1
Isoform
Isoform Usage
Isoform Fraction (IF)
1.00
0.75
0.50
0.25
0.00
***
ENST00000263121.12
ENST00000407422.8
ENST00000417137.6
ENST00000634926.1
Isoform
Condition
cntrl
treat

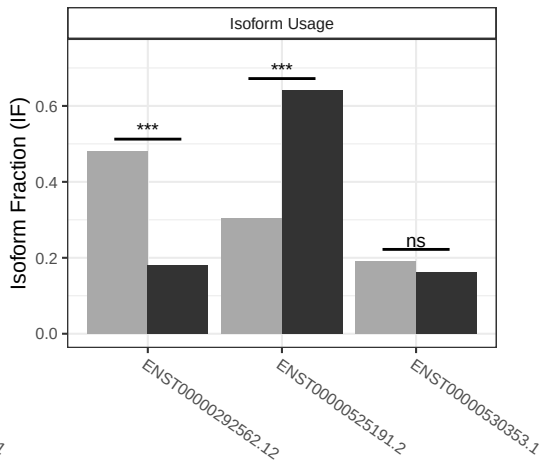
The isoform switch in ZNF251 (cntrl vs treat)

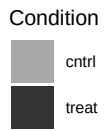




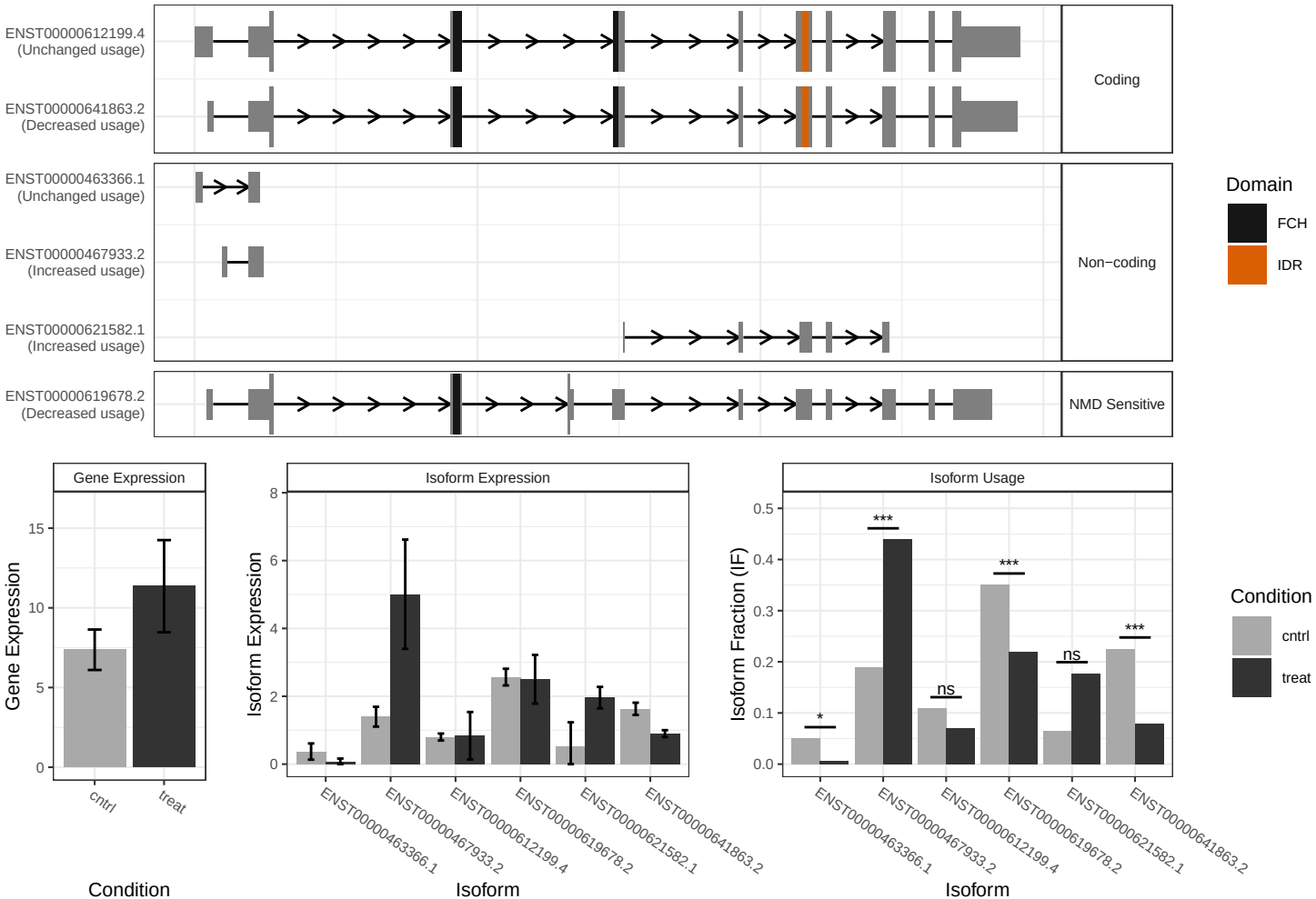




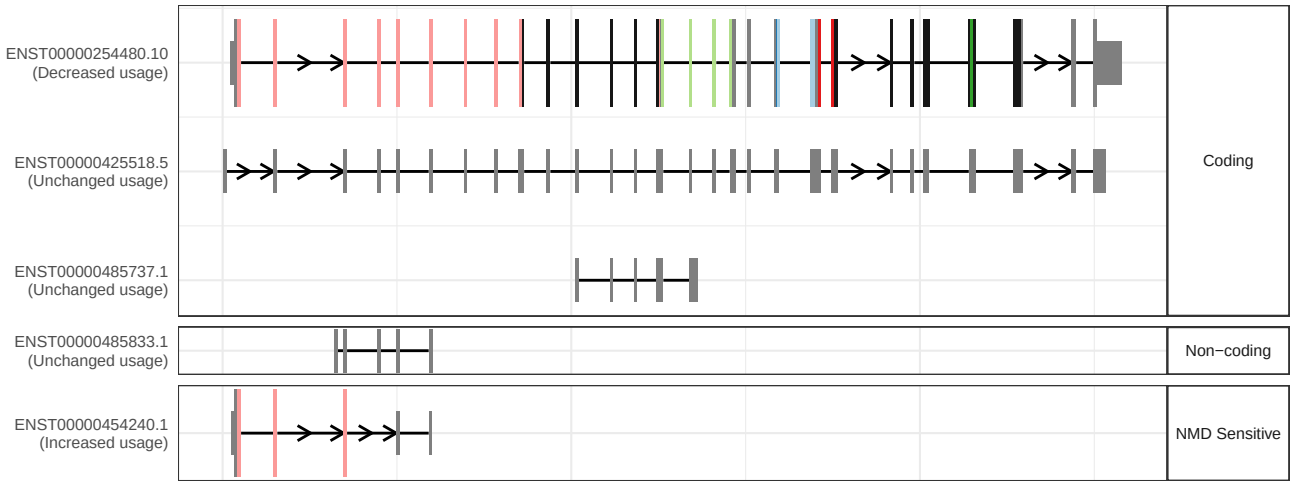


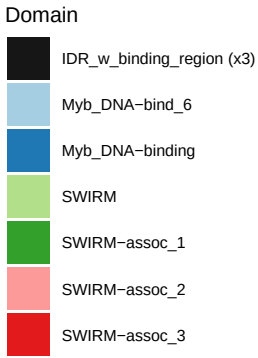


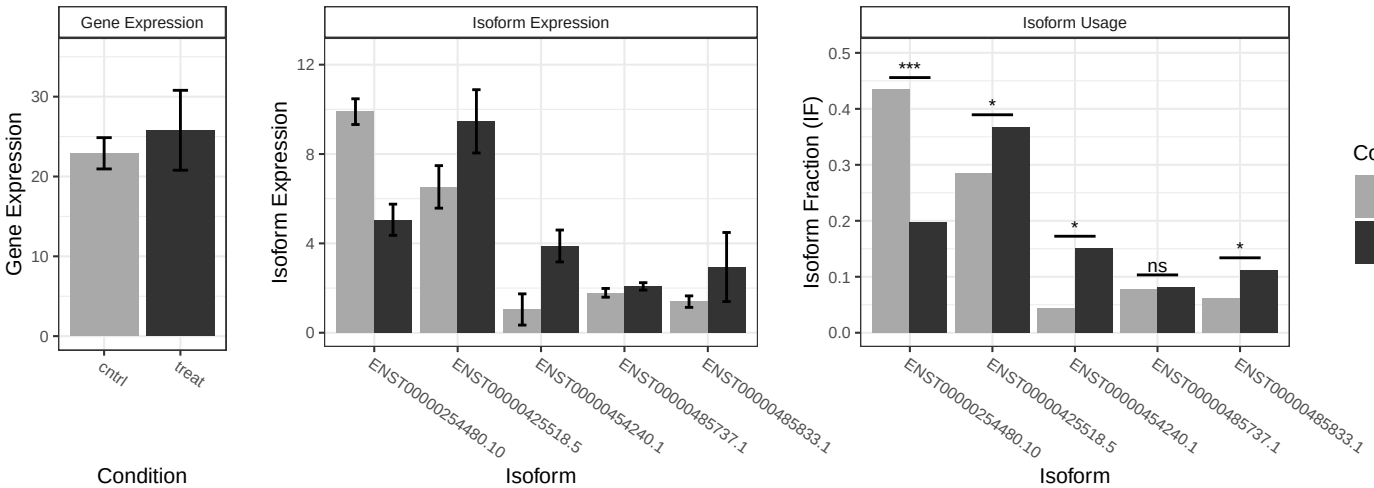
The isoform switch in SRGAP2B (cntrl vs treat)

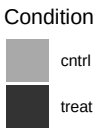


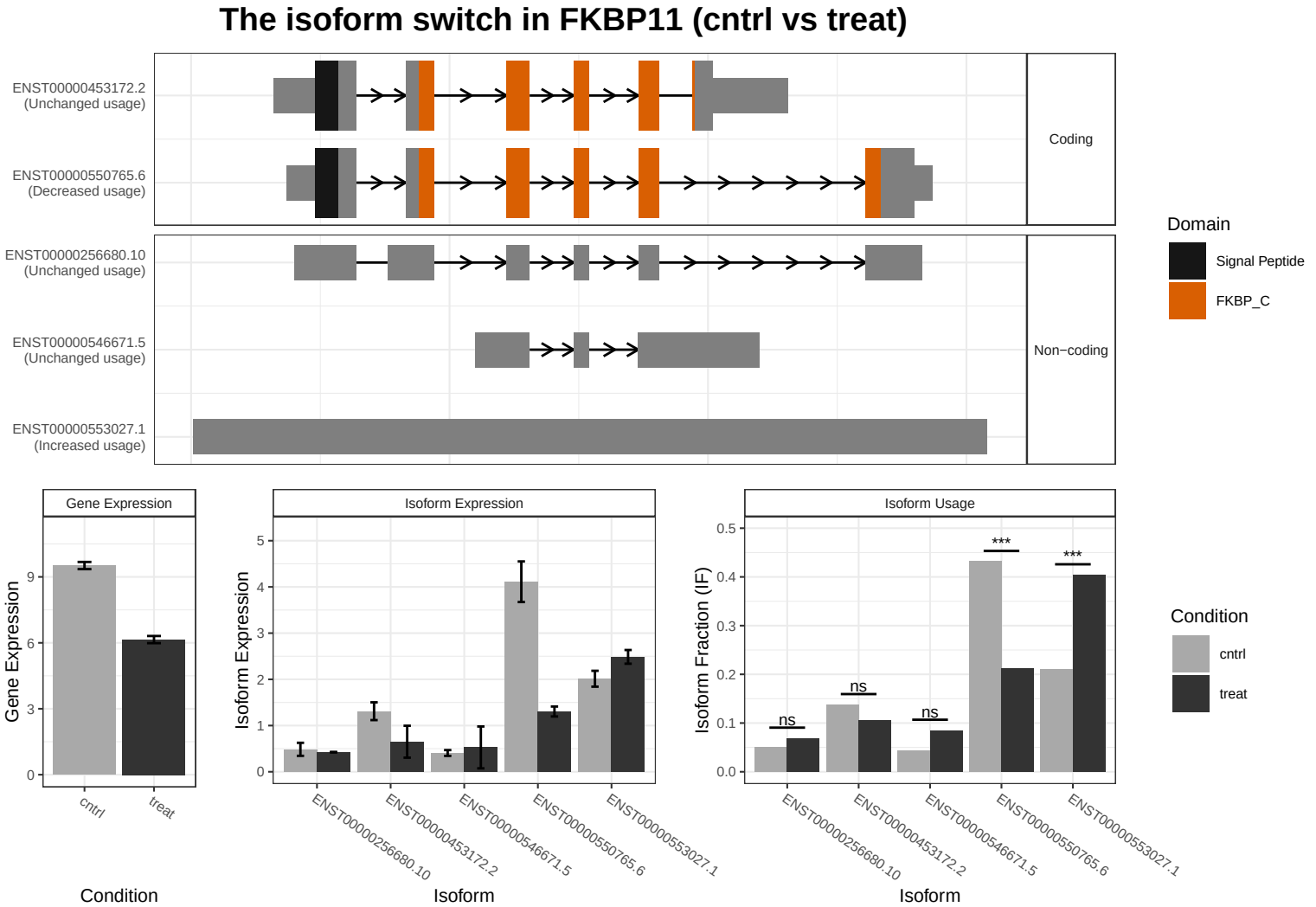
The isoform switch in SMARCC1 (cntrl vs treat)



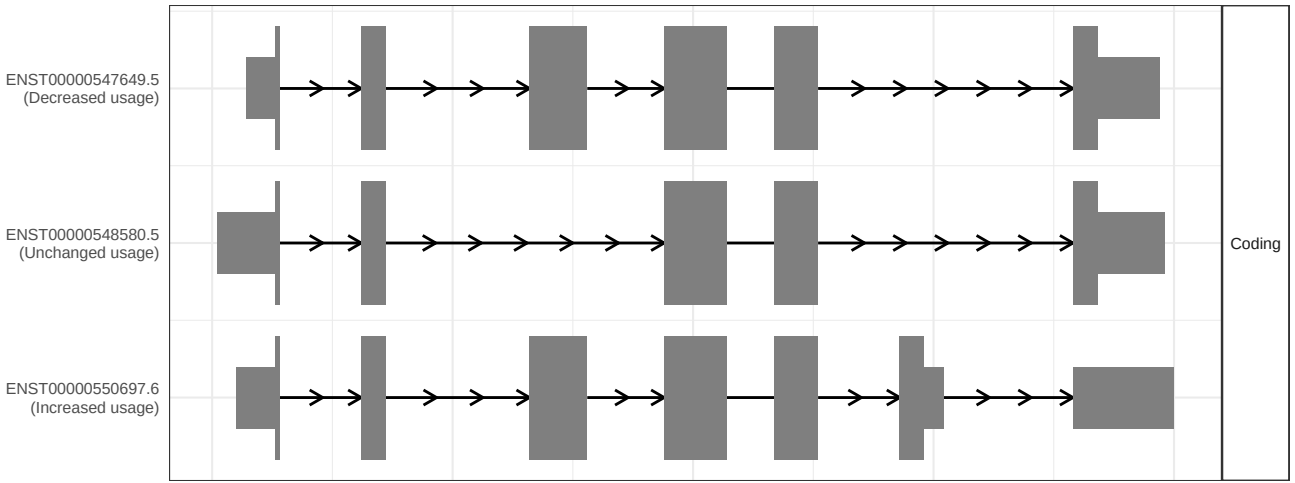


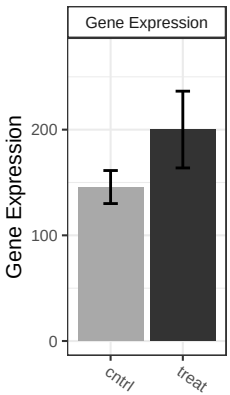


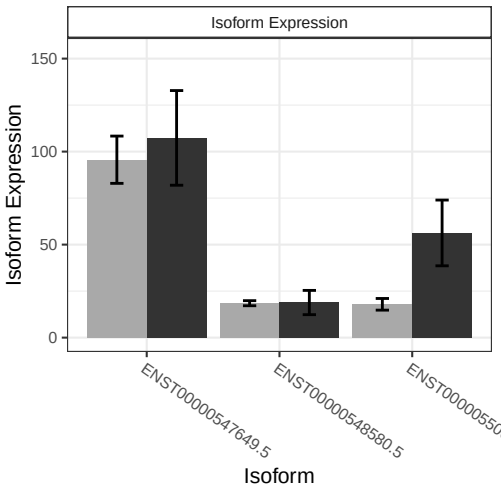


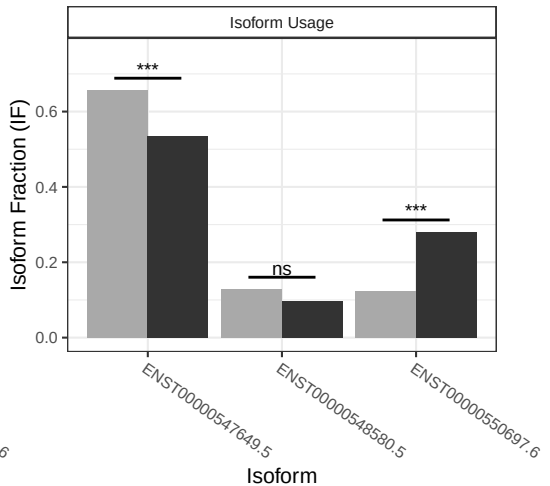

The isoform switch in FKBP11 (cntrl vs treat)
ENST00000453172.2
(Unchanged usage)
ENST00000550765.6
(Decreased usage)
ENST00000256680.10
(Unchanged usage)
ENST00000546671.5
(Unchanged usage)
ENST00000553027.1
(Increased usage)
Coding
Non−coding
Domain
Signal Peptide
FKBP_C
Gene Expression
Gene Expression
cntrl
treat
Condition
Isoform Expression
Isoform Expression
ENST00000256680.10
ENST00000453172.2
ENST00000546671.5
ENST00000550765.6
ENST00000553027.1
Isoform
Isoform Usage
Isoform Fraction (IF)
ns
***
Condition
cntrl
treat

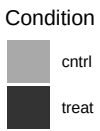
The isoform switch in MYL6 (cntrl vs treat)



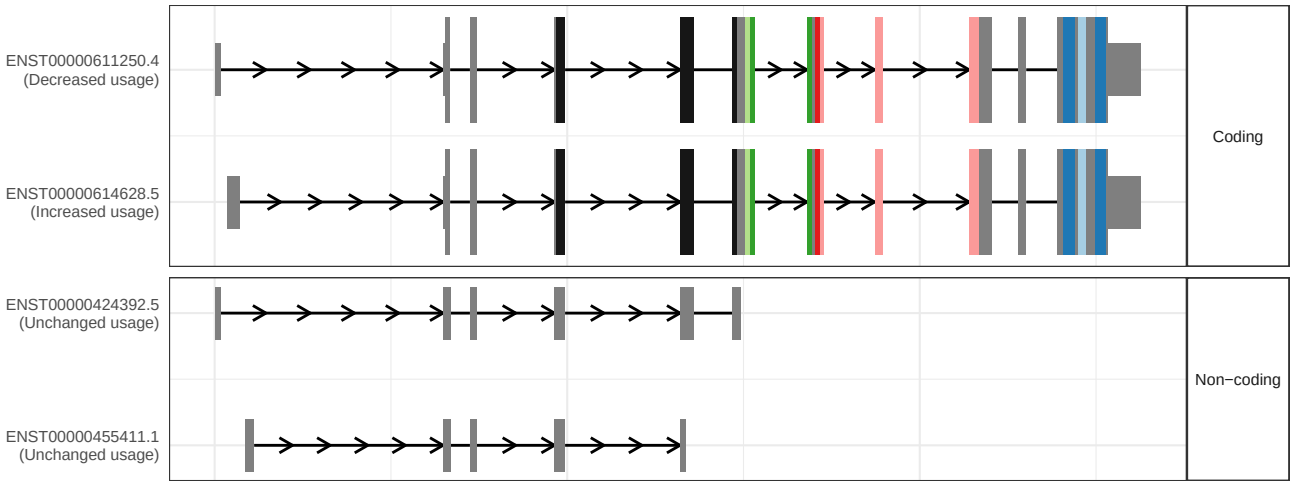


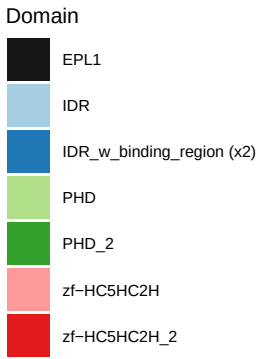


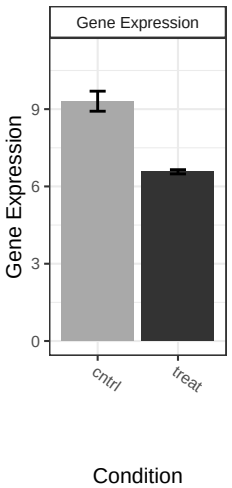


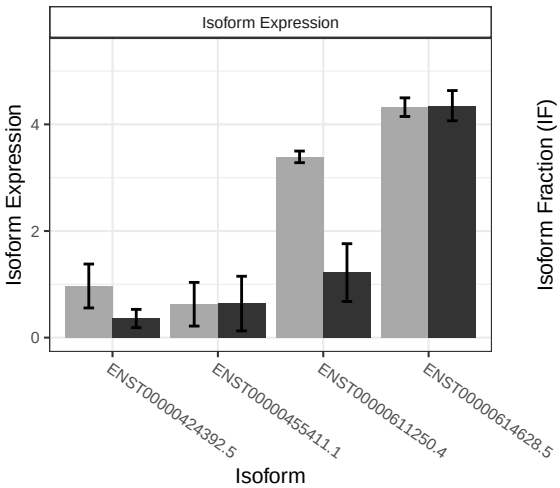


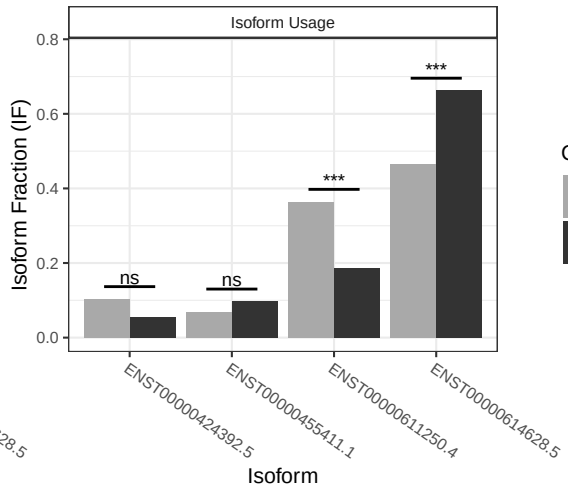
The isoform switch in JADE3 (cntrl vs treat)

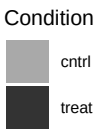


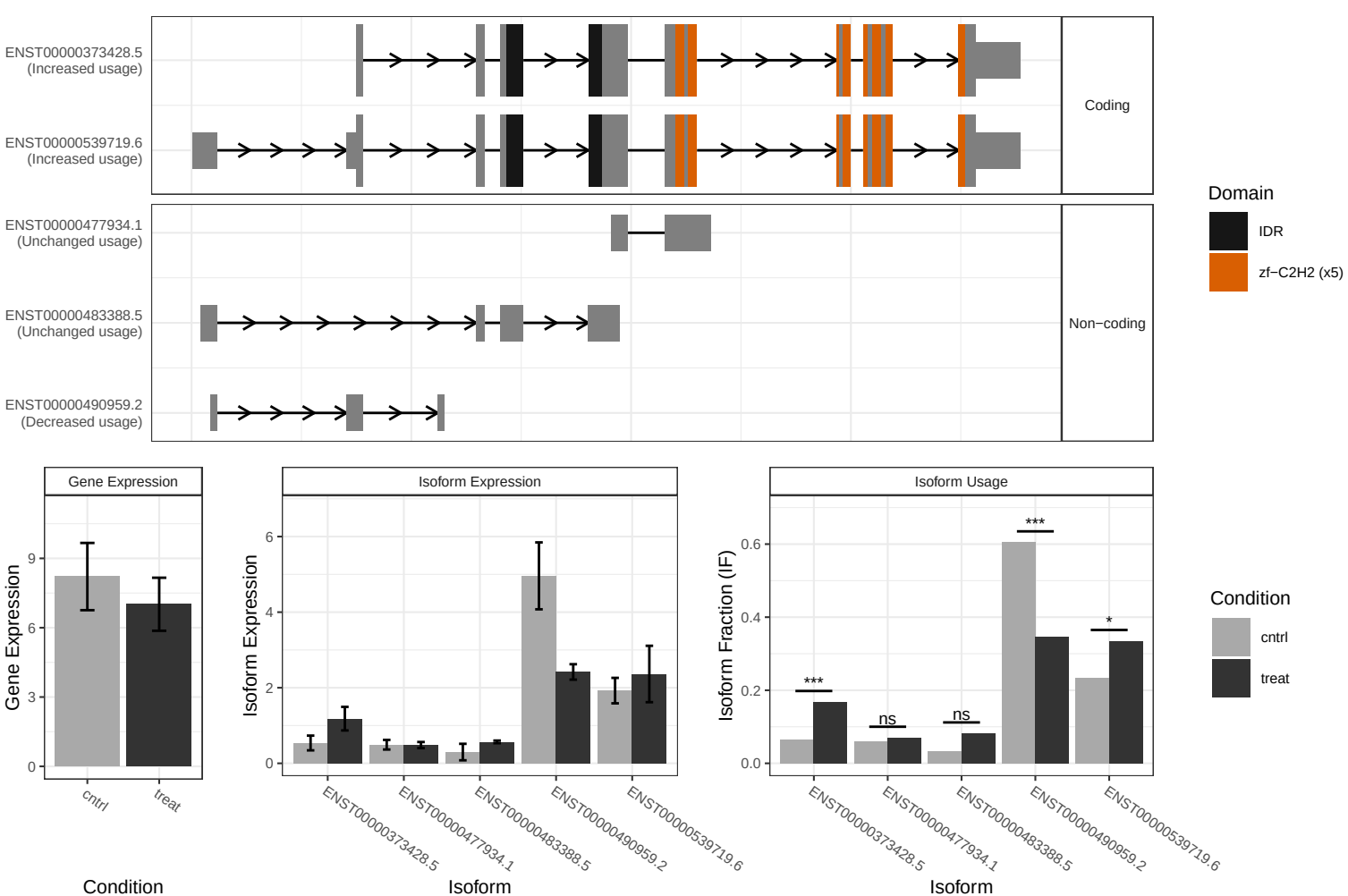




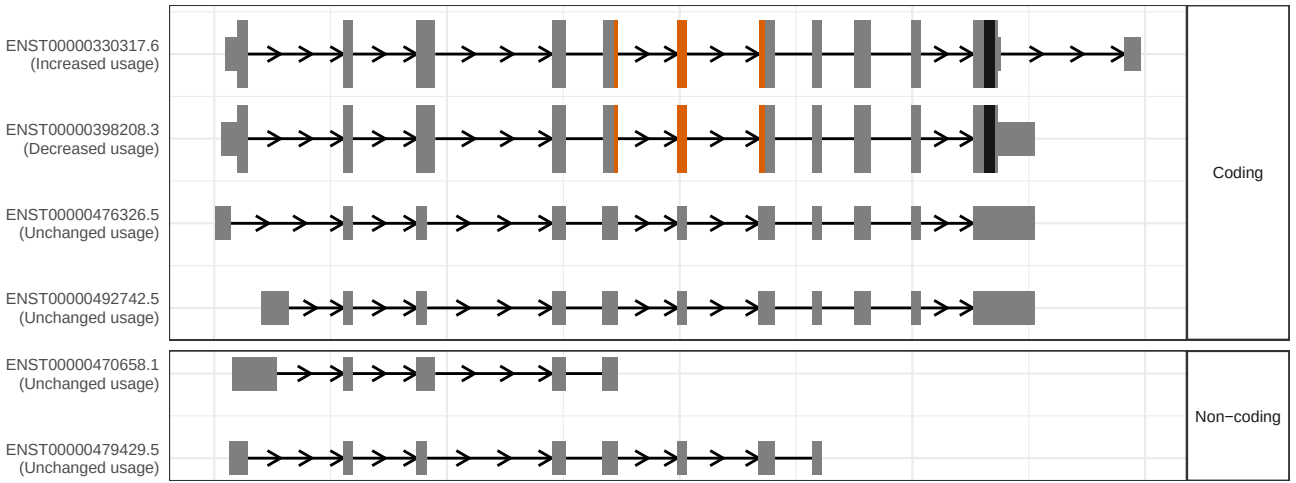


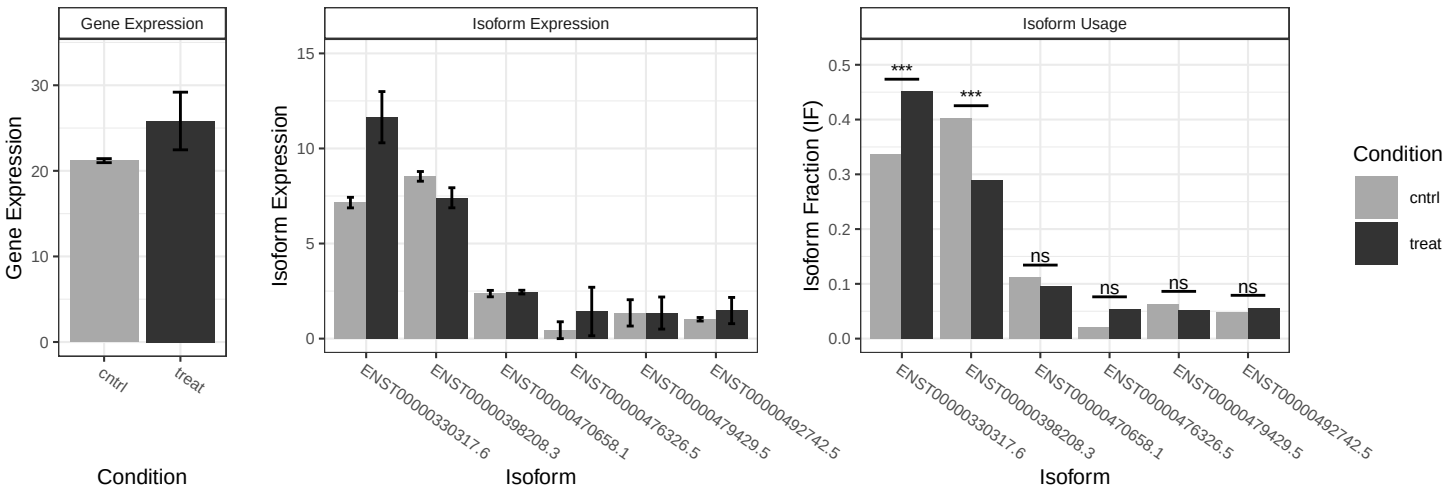




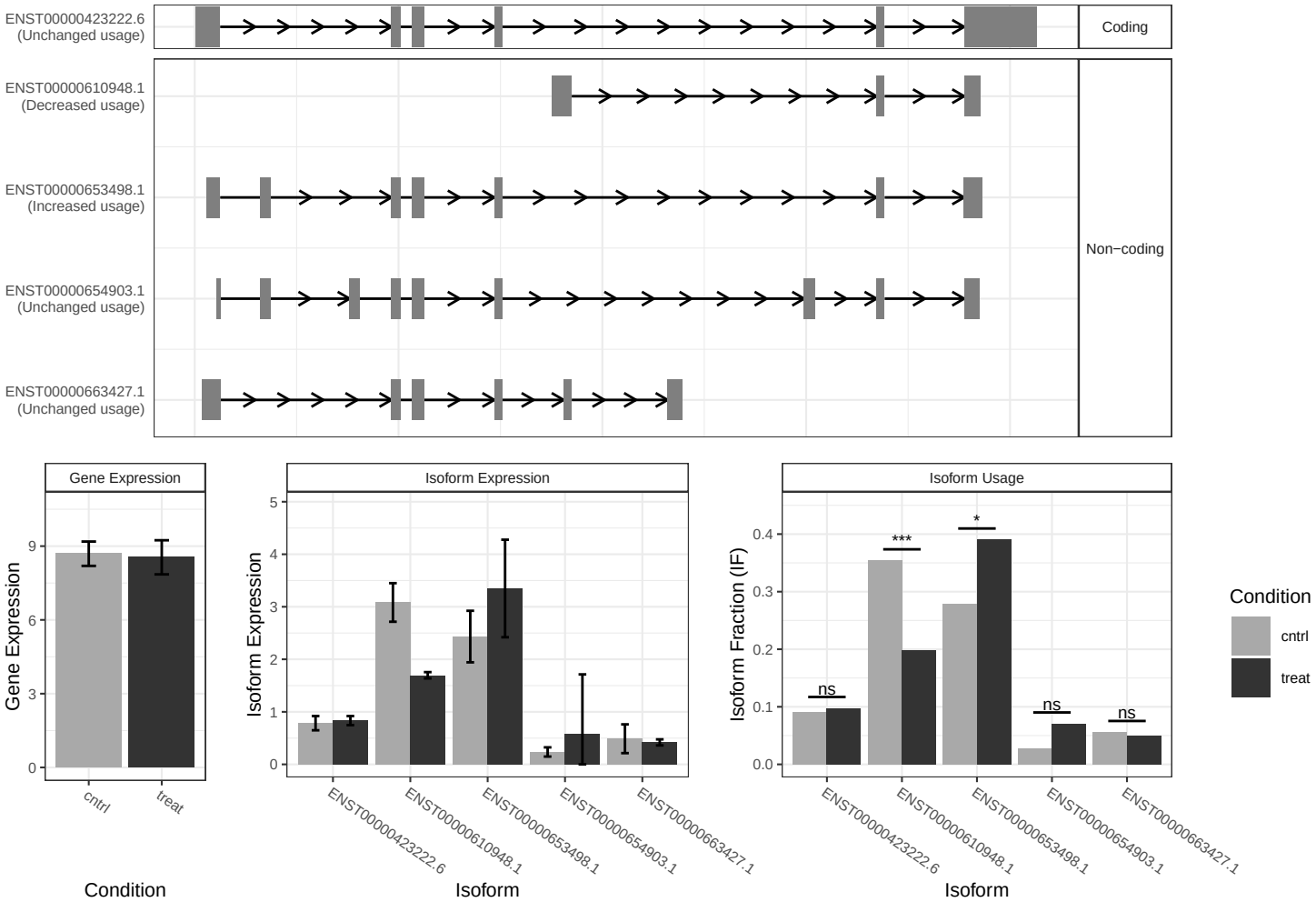

The isoform switch in ZNF362 (cntrl vs treat)
ENST00000373428.5
(Increased usage)
ENST00000539719.6
(Increased usage)
ENST00000477934.1
(Unchanged usage)
ENST00000483388.5
(Unchanged usage)
ENST00000490959.2
(Decreased usage)
Coding
Non−coding
Domain
IDR
zf−C2H2 (x5)
Gene Expression
Gene Expression
cntrl
treat
Condition
Isoform Expression
Isoform Expression
ENST00000373428.5
ENST00000477934.1
ENST00000483388.5
ENST00000490959.2
ENST00000539719.6
Isoform
Isoform Usage
Isoform Fraction (IF)
***
ns
ns
***
*
Condition
cntrl
treat

The isoform switch in WDR4 (cntrl vs treat)

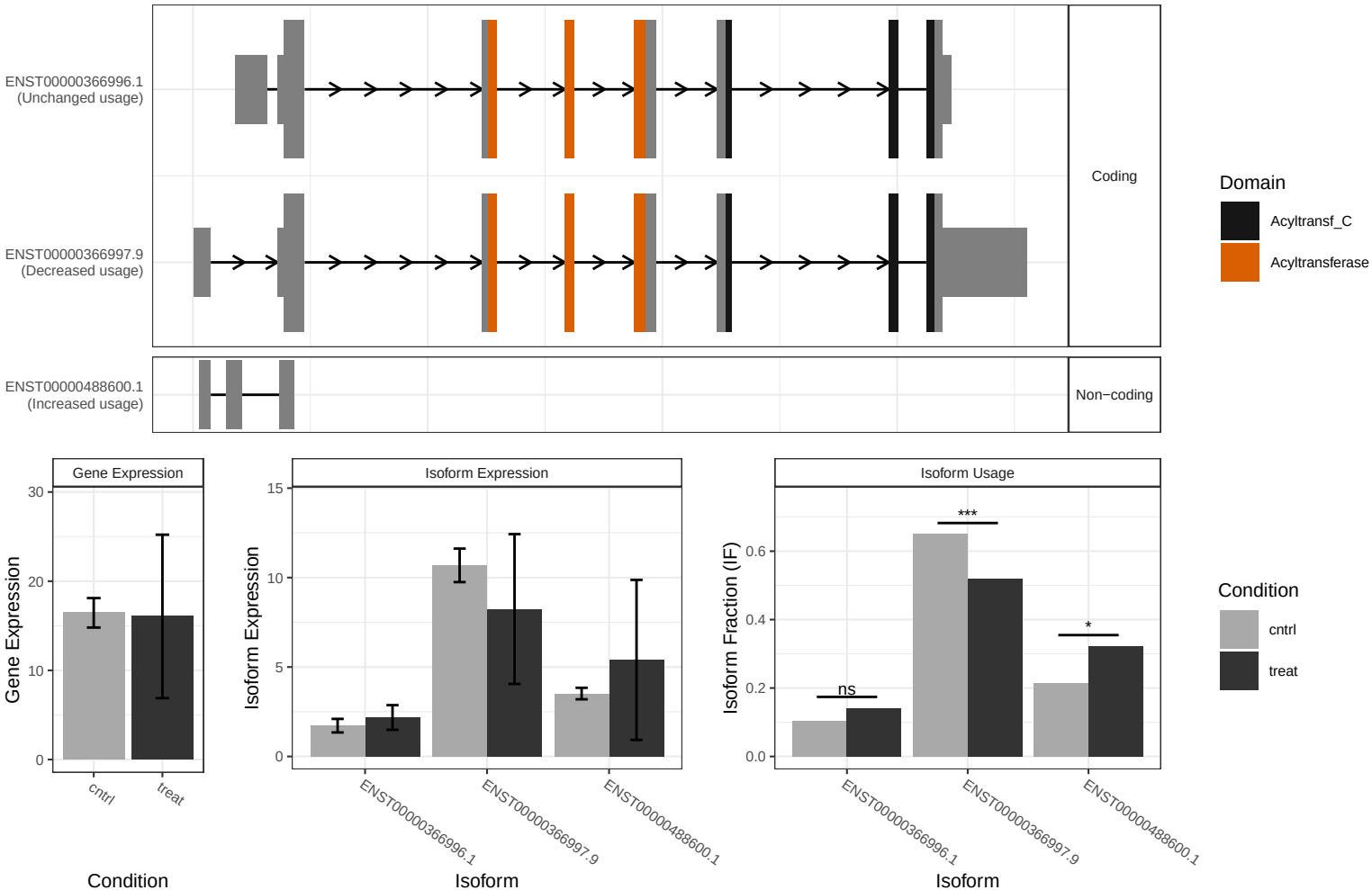




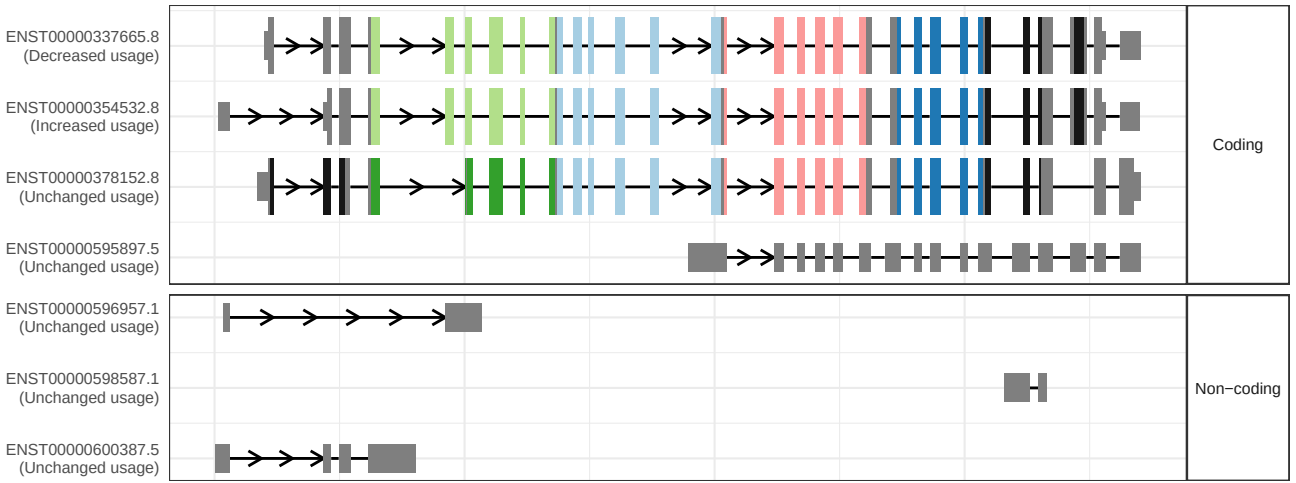
The isoform switch in LINC00467 (cntrl vs treat)

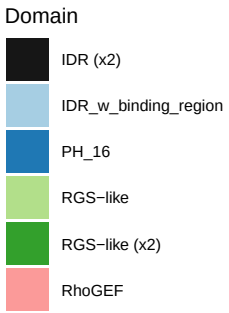


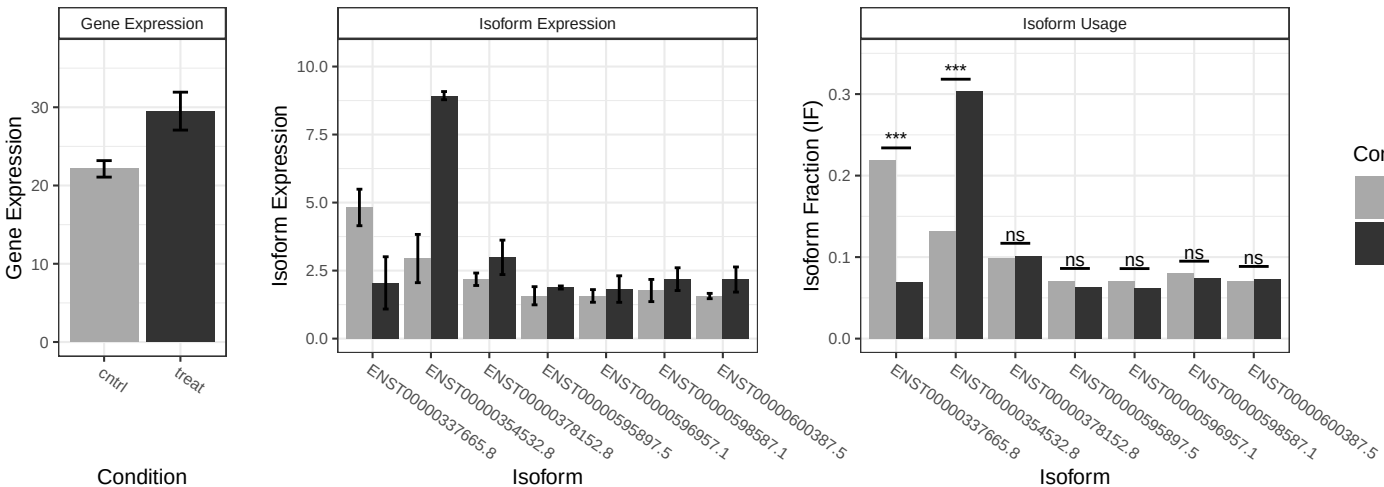
The isoform switch in LPGAT1 (cntrl vs treat)

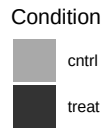


The isoform switch in ARHGEF1 (cntrl vs treat)

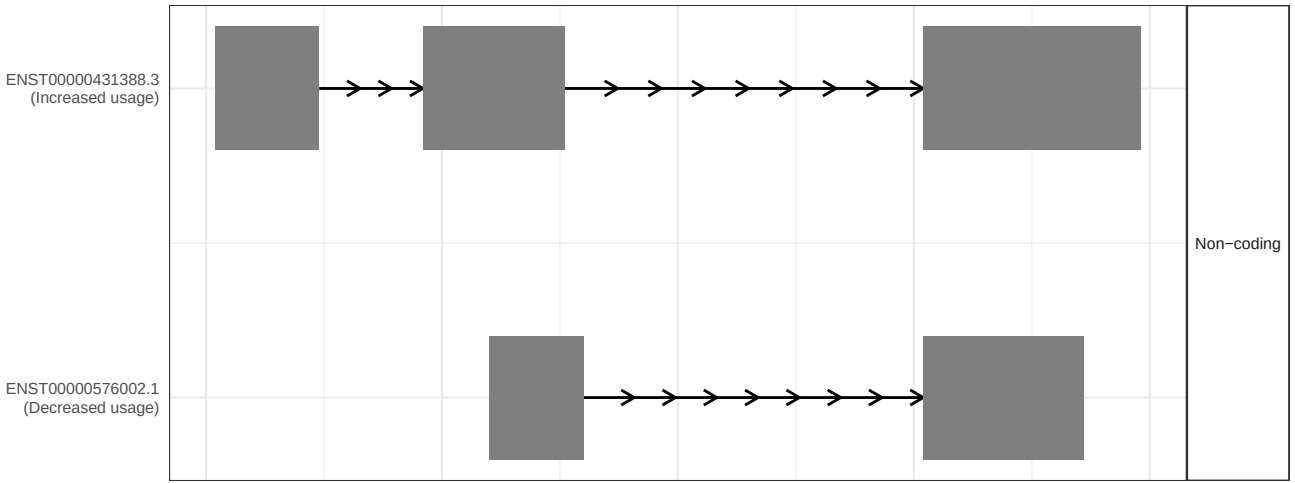


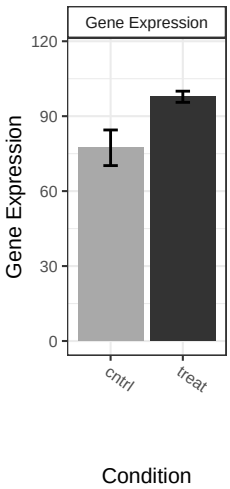


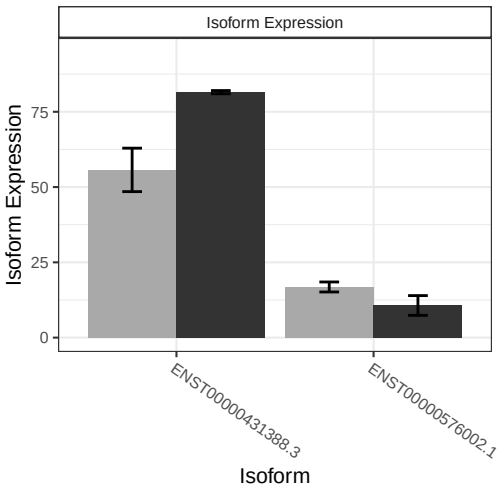


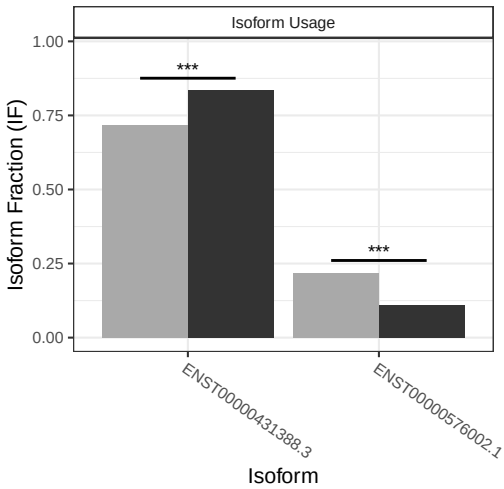


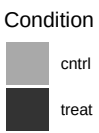
The isoform switch in NDUFAF8 (cntrl vs treat)



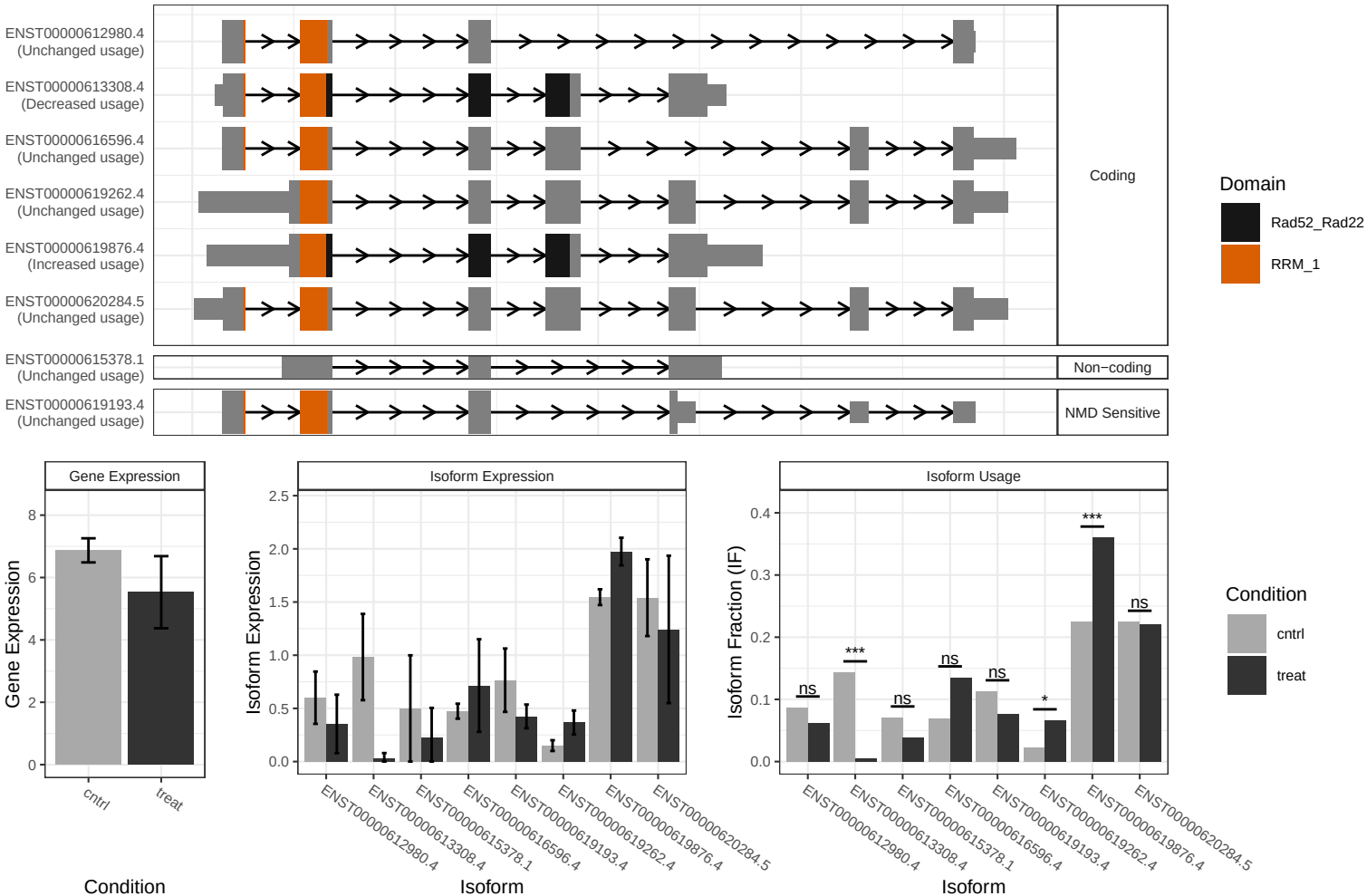




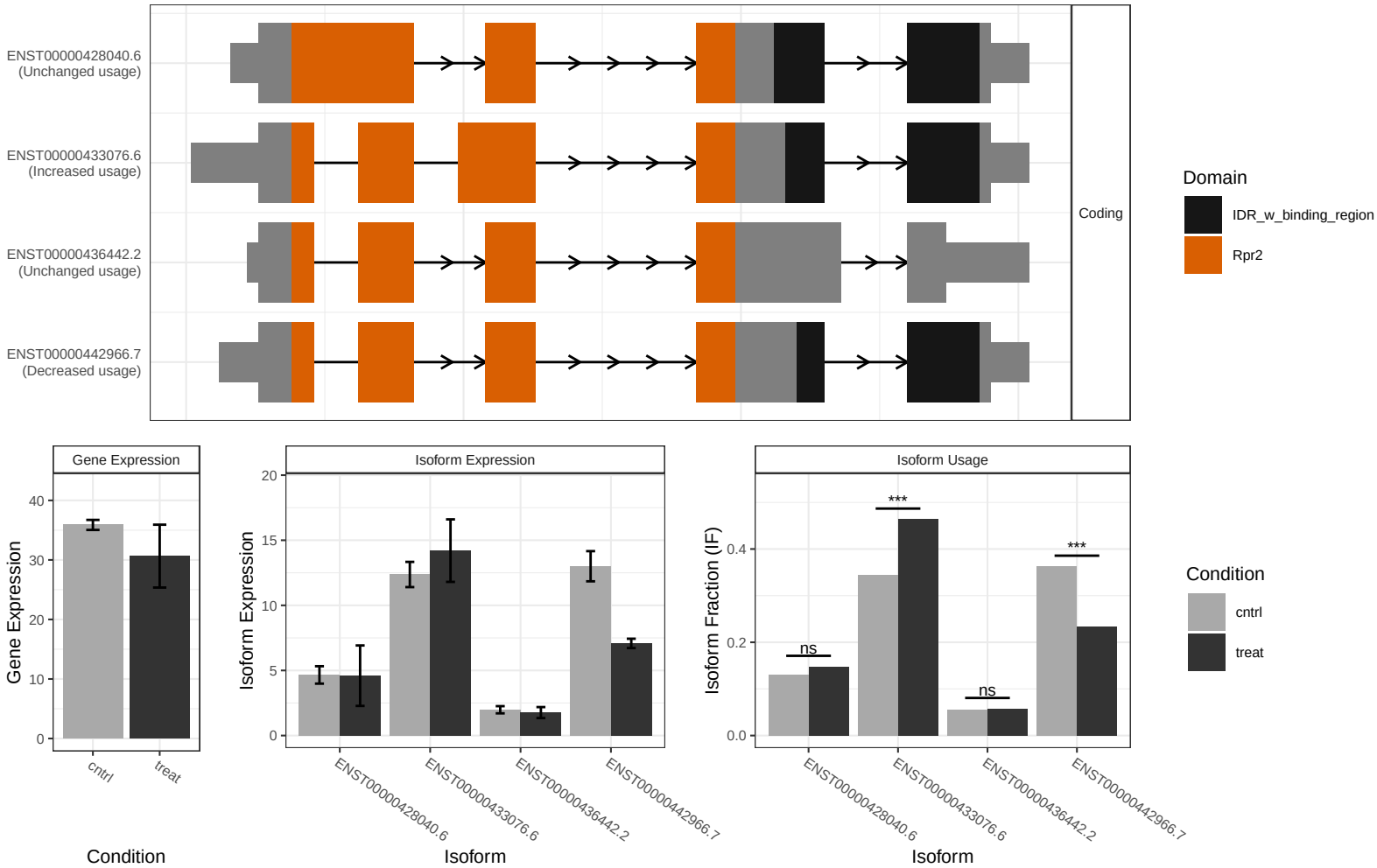




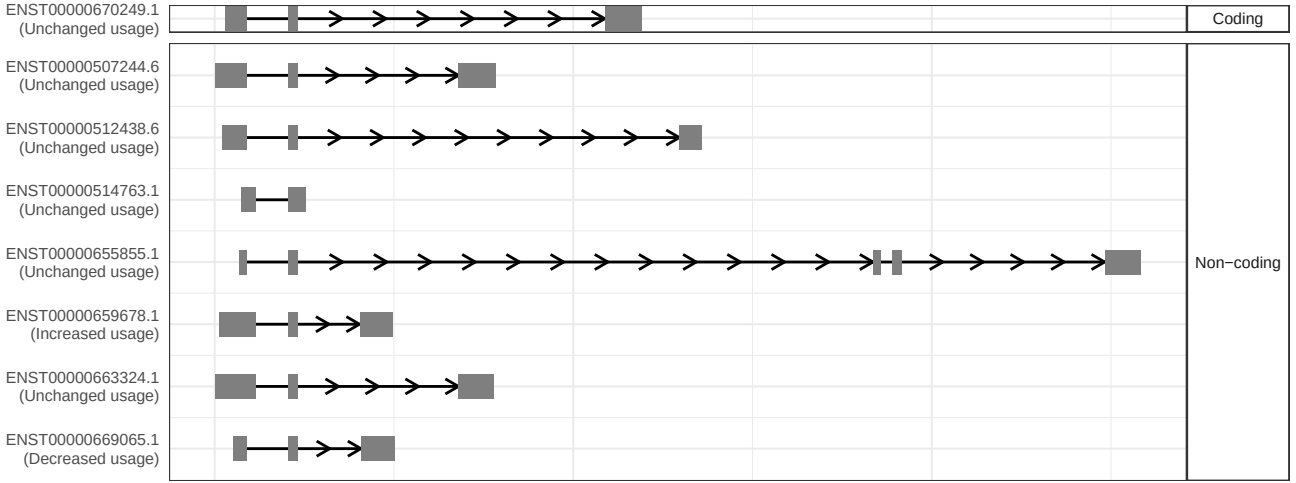
The isoform switch in RDM1 (cntrl vs treat)

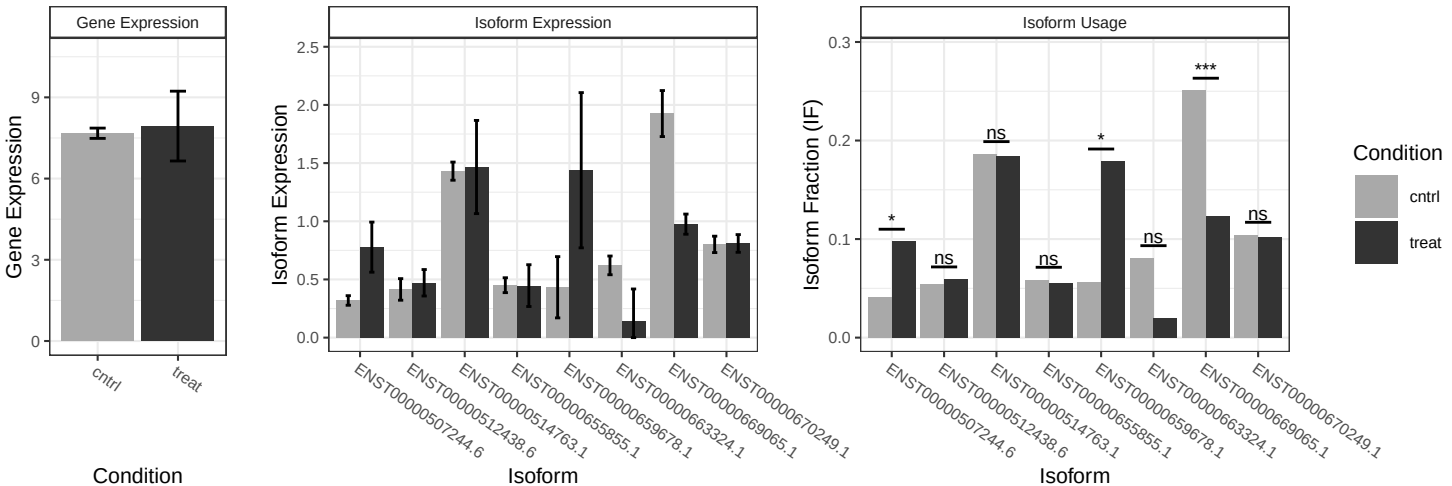


The isoform switch in RPP21 (cntrl vs treat)

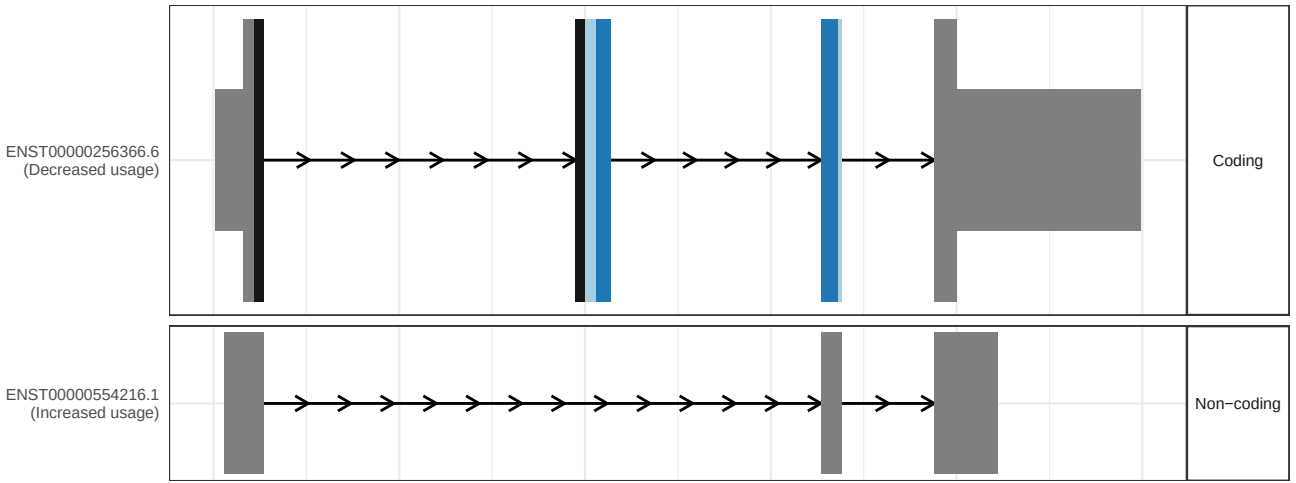


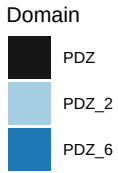
The isoform switch in STX18−AS1 (cntrl vs treat)

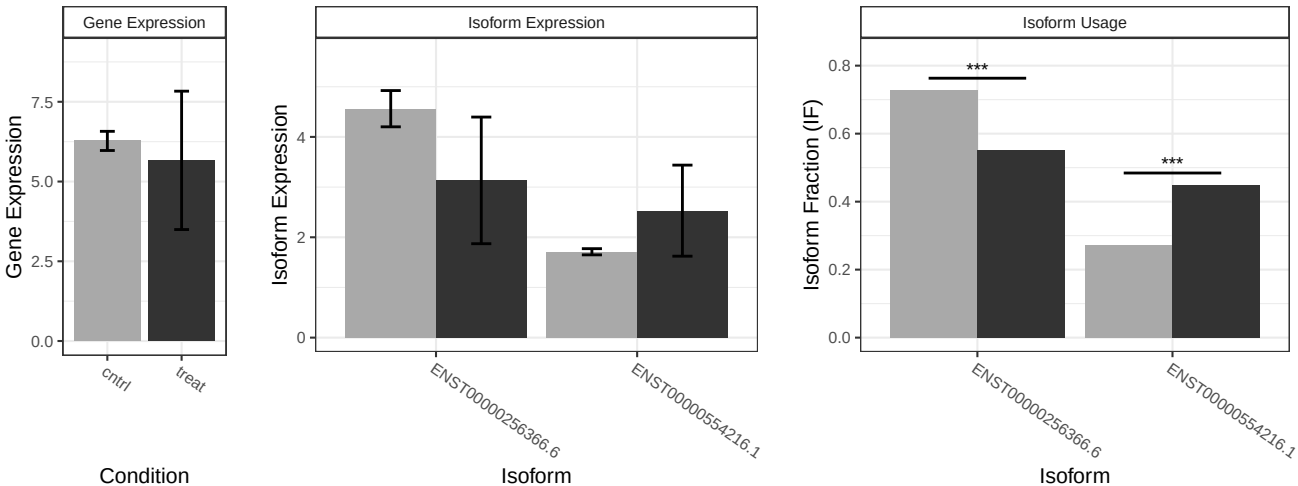


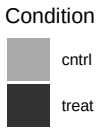


The isoform switch in SYNJ2BP (cntrl vs treat)

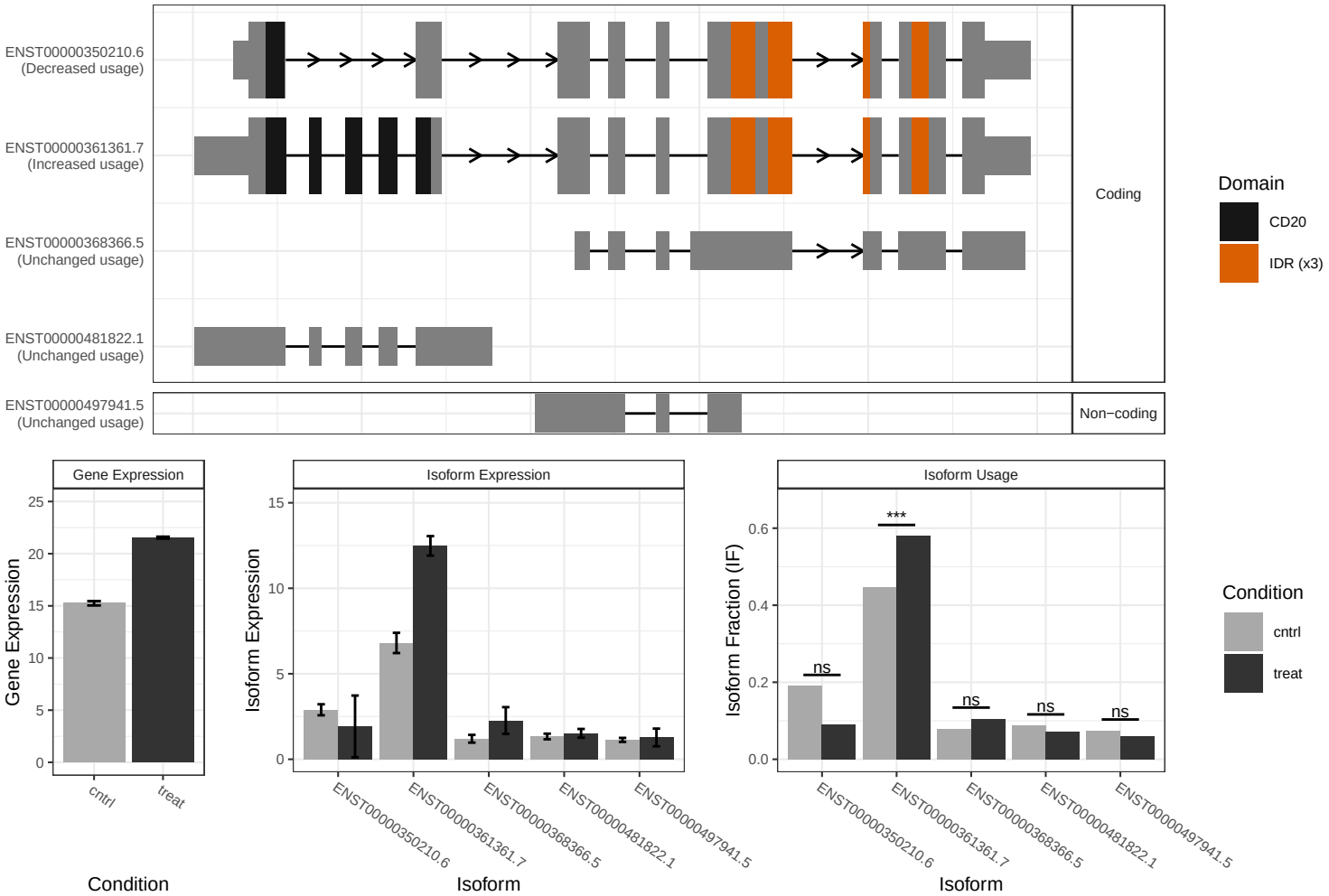




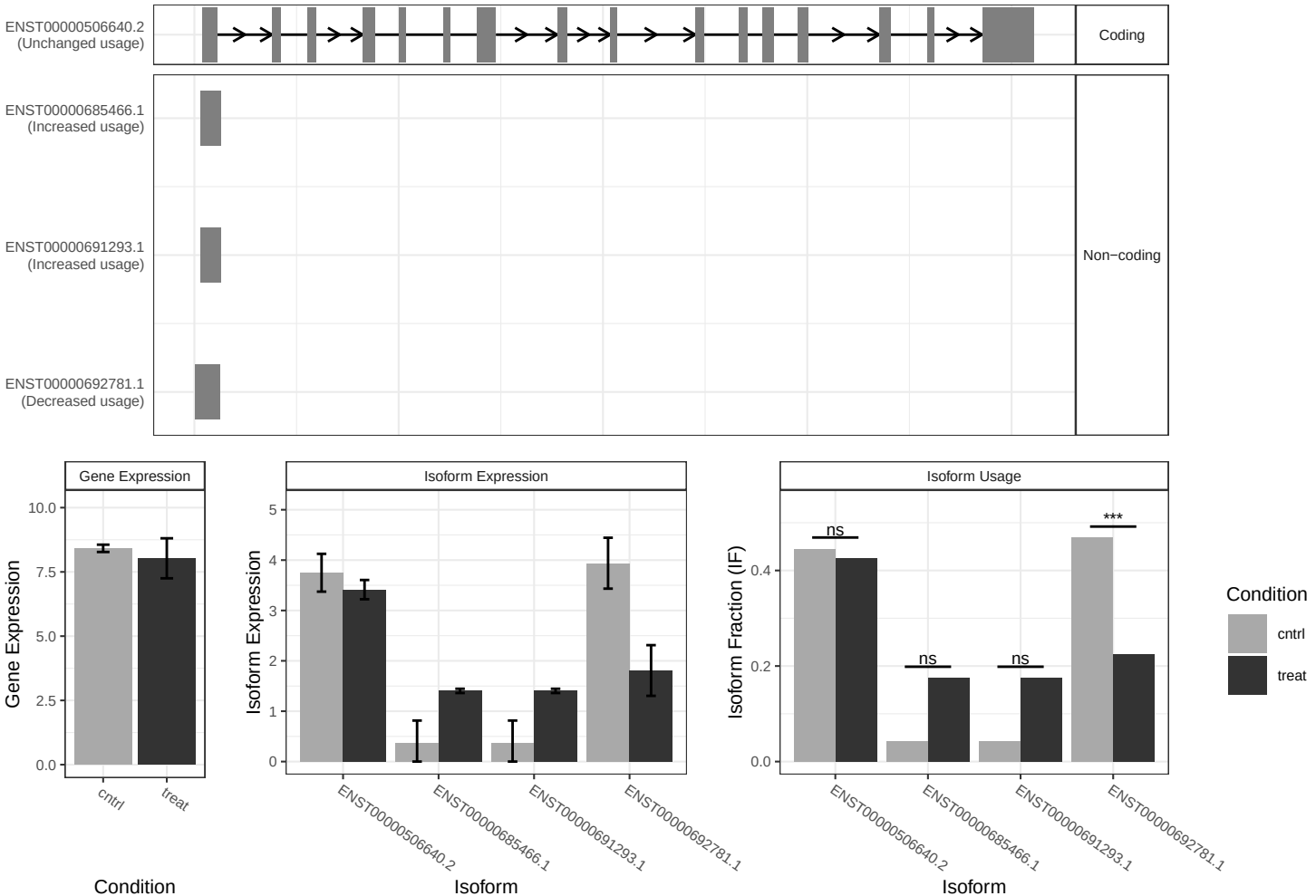




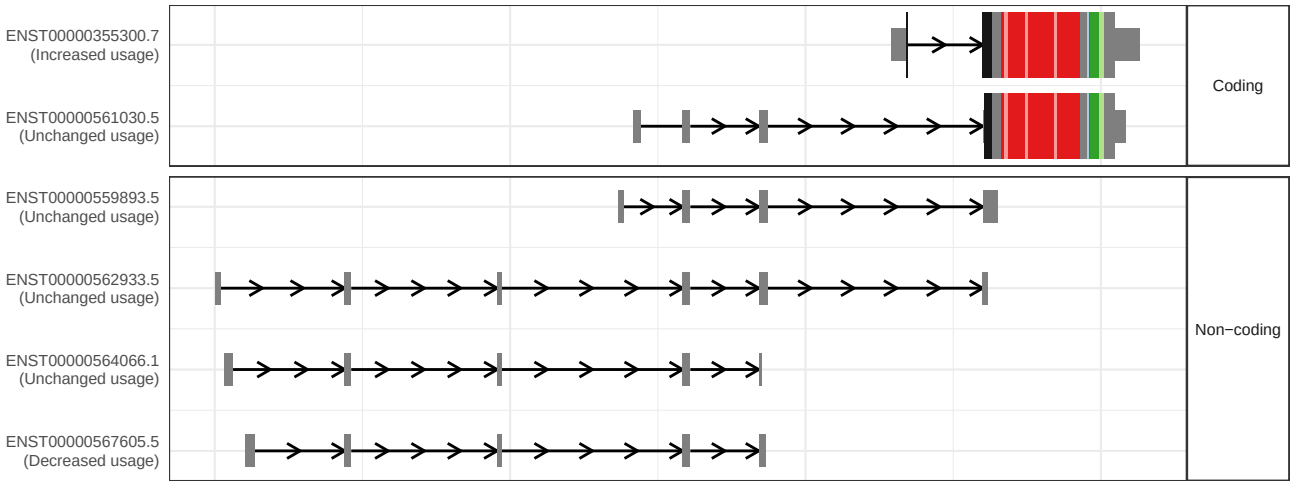
The isoform switch in FAM189B (cntrl vs treat)

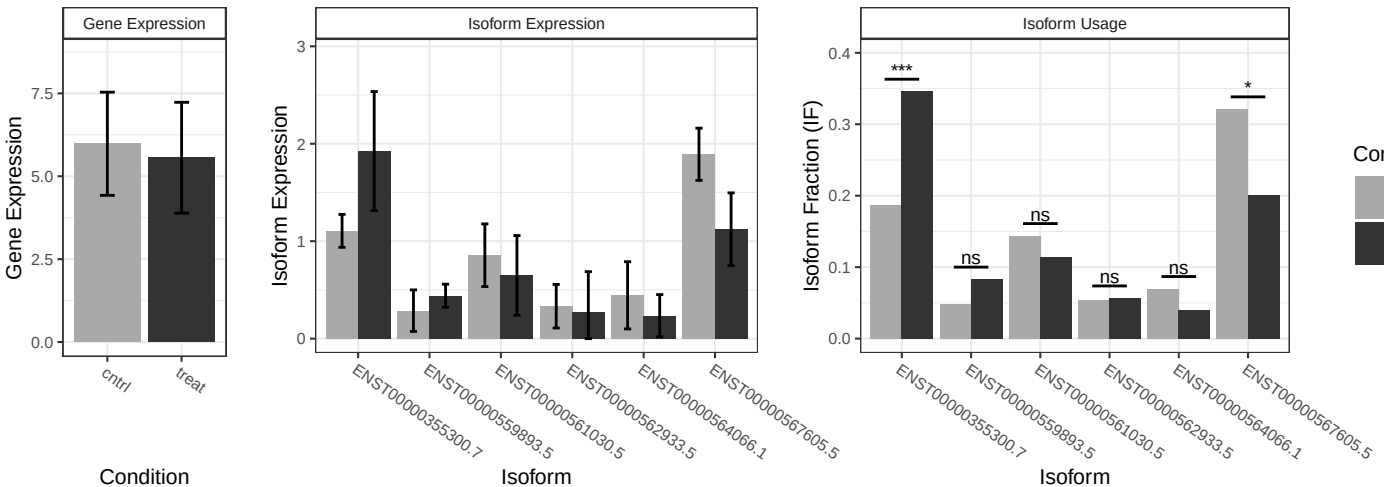


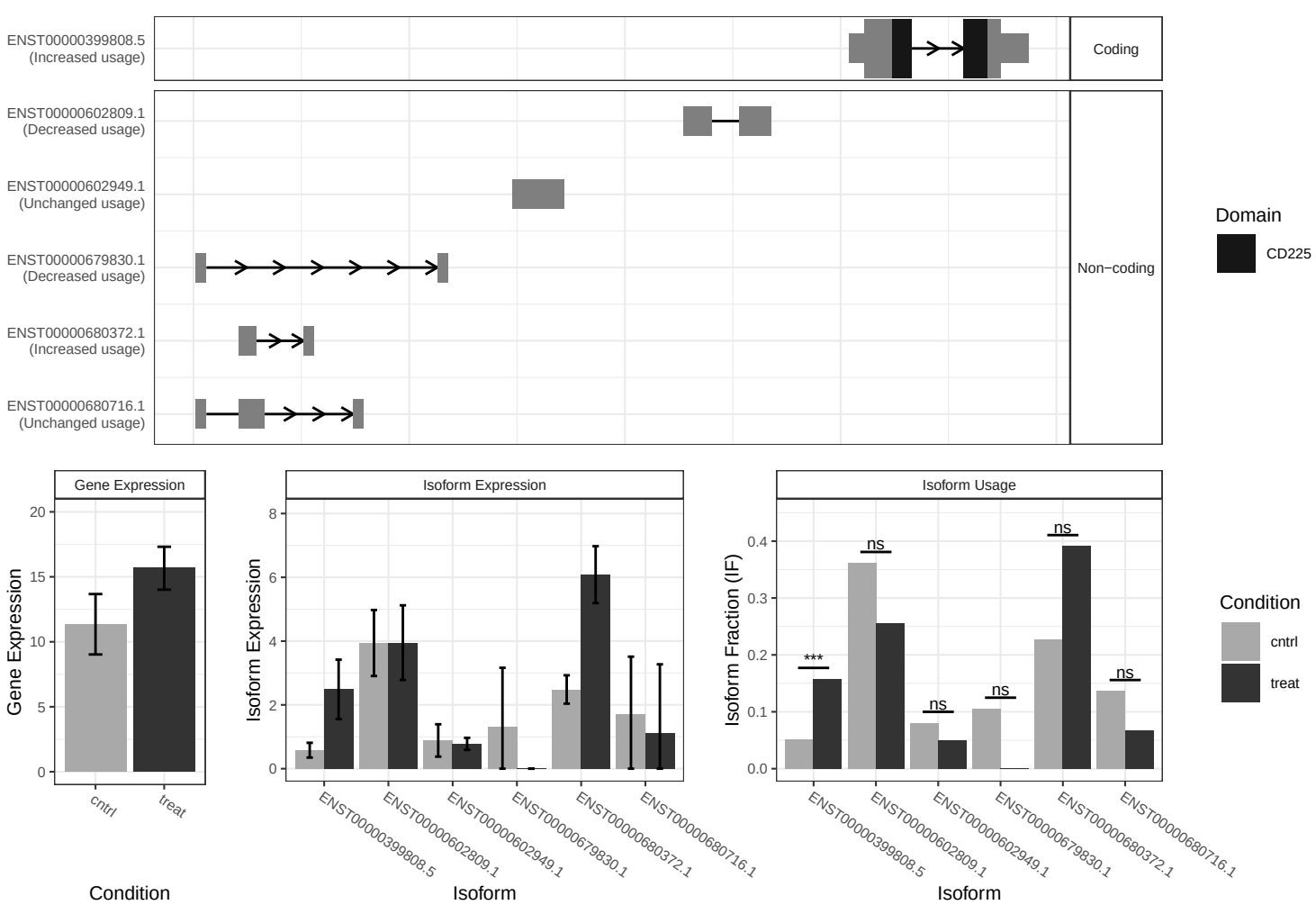
The isoform switch in ENSG00000228327 (cntrl vs treat)



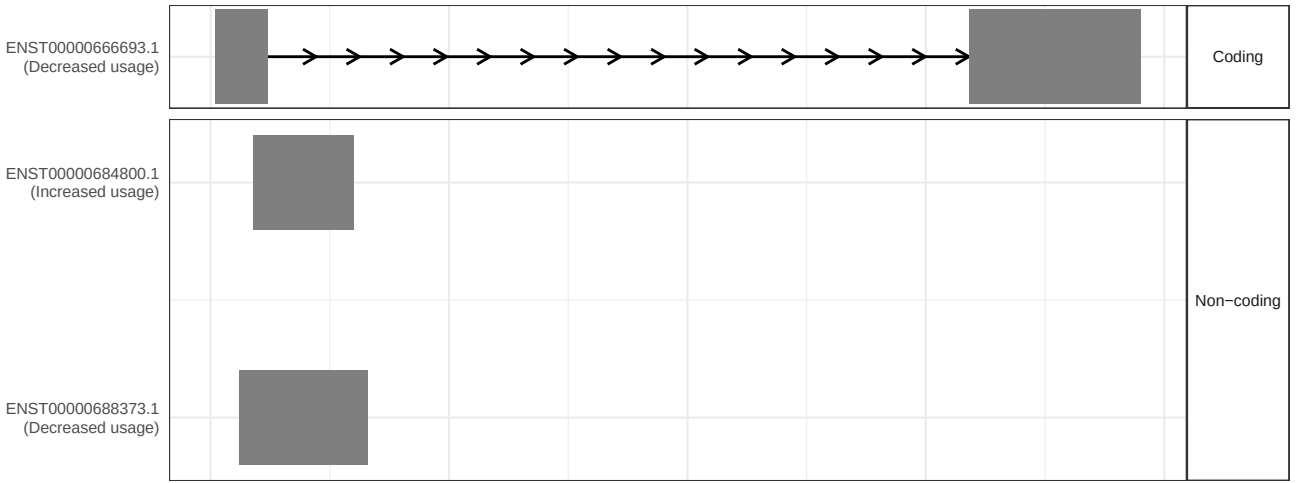
The isoform switch in LINGO1 (cntrl vs treat)

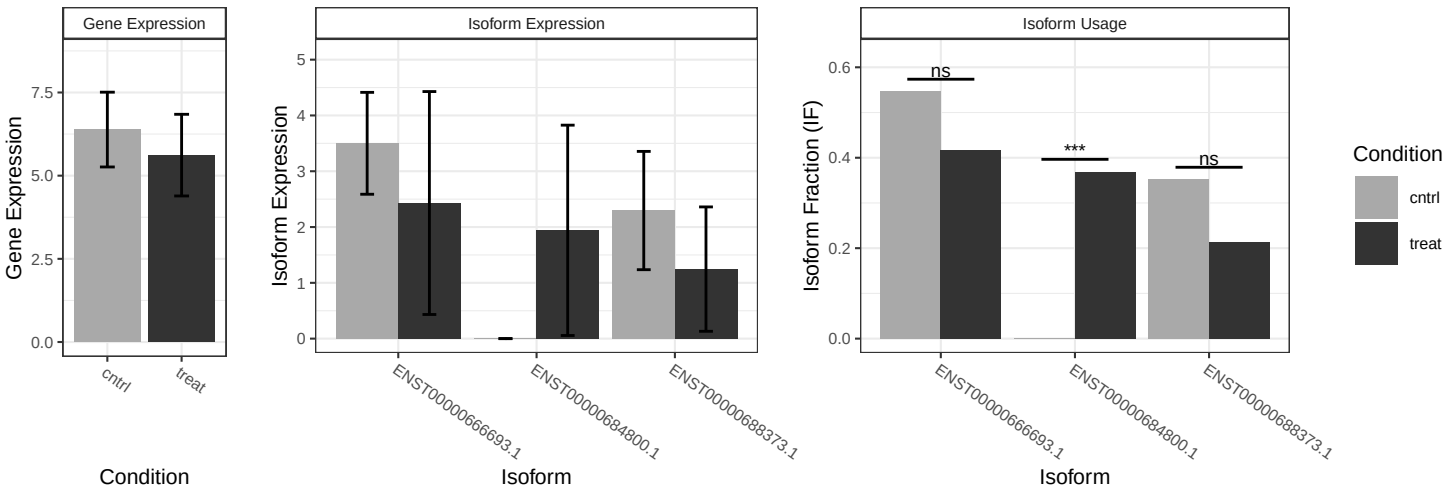




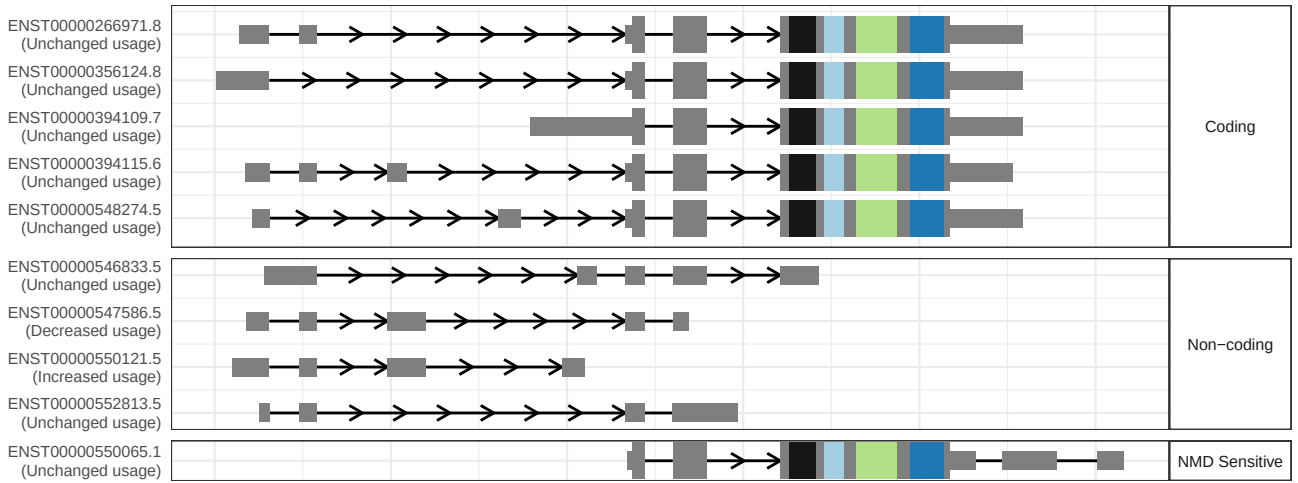

The isoform switch in IFITM3 (cntrl vs treat)
ENST00000399808.5
(Increased usage)
ENST00000602809.1
(Decreased usage)
ENST00000602949.1
(Unchanged usage)
ENST00000679830.1
(Decreased usage)
ENST00000680372.1
(Increased usage)
ENST00000680716.1
(Unchanged usage)
Coding
Non−coding
Domain
CD225
Gene Expression
Gene Expression
Condition
cntrl
treat
Isoform Expression
Isoform Expression
Isoform
ENST00000399808.5
ENST00000602809.1
ENST00000602949.1
ENST00000679830.1
ENST00000680372.1
ENST00000680716.1
Isoform Usage
Isoform Fraction (IF)
***
ns
Isoform
Condition
cntrl
treat

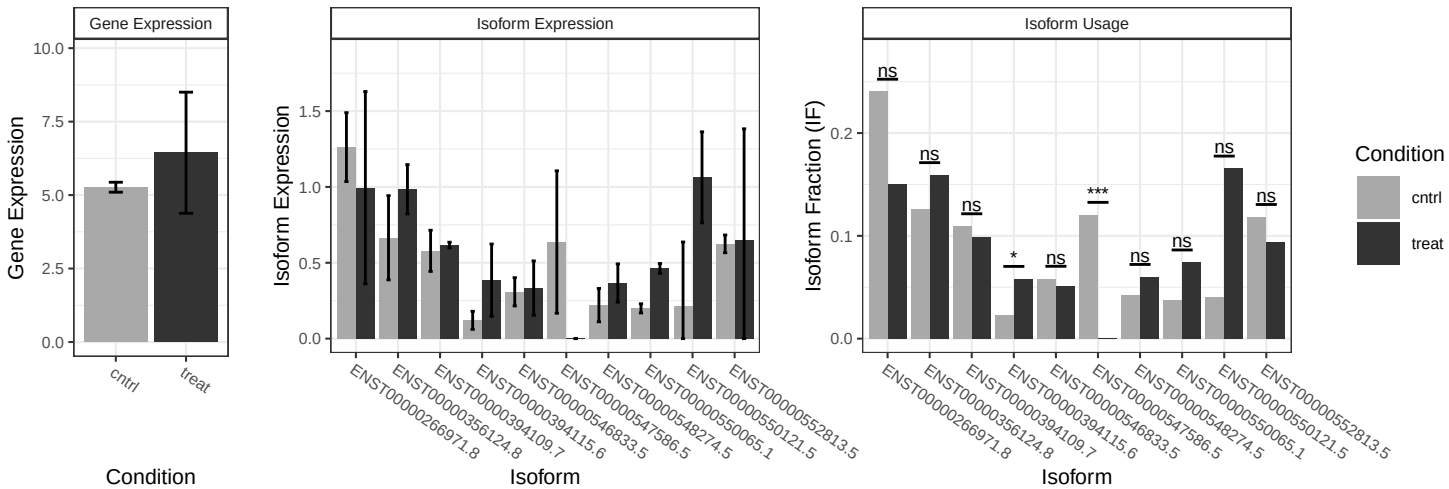
The isoform switch in ENSG00000286833 (cntrl vs treat)



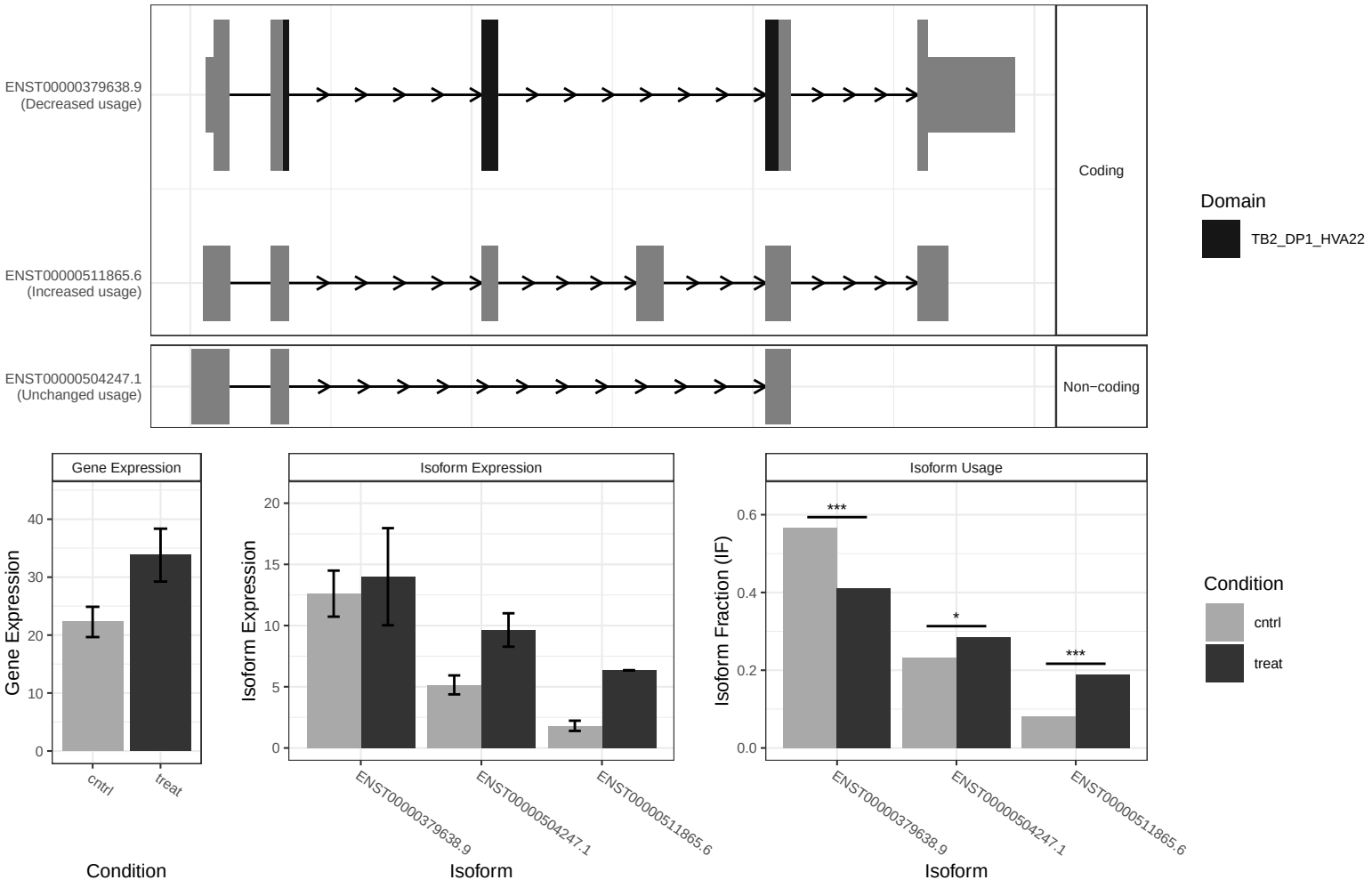


The isoform switch in SUOX (cntrl vs treat)

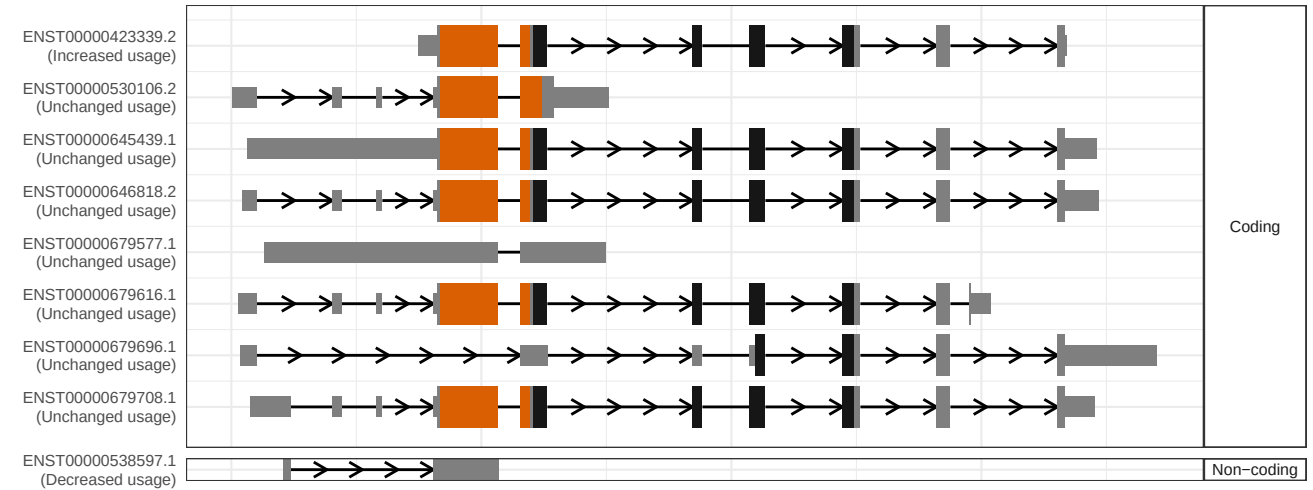


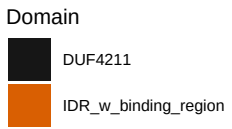


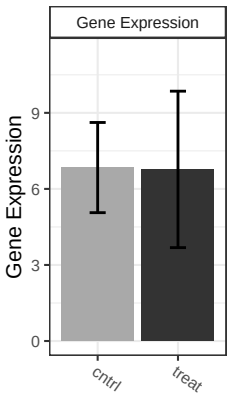
The isoform switch in REEP5 (cntrl vs treat)

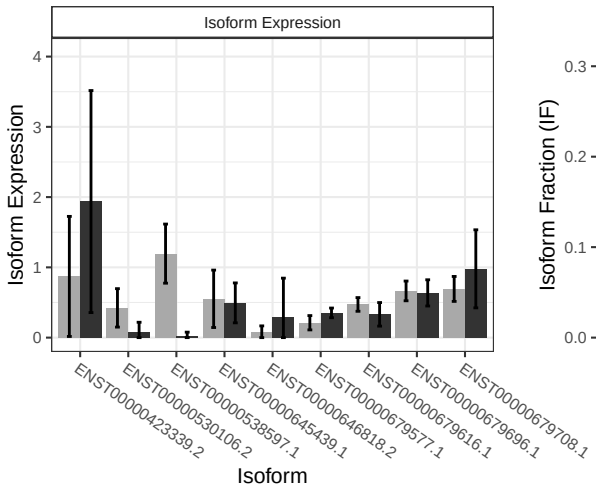


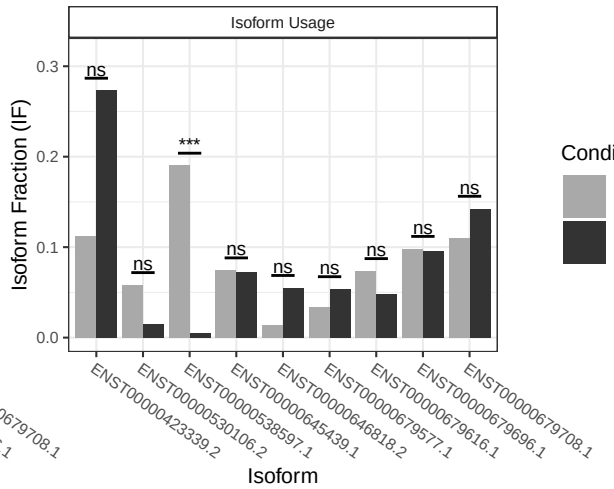
The isoform switch in CCDC82 (cntrl vs treat)

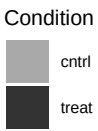




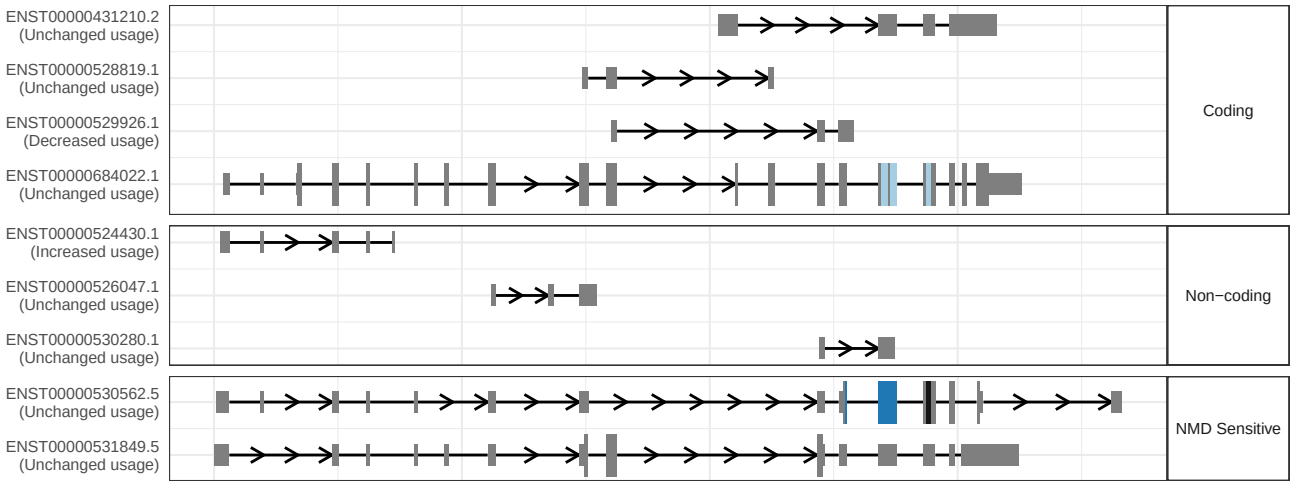


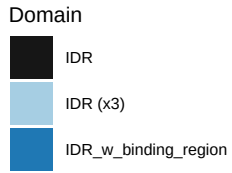


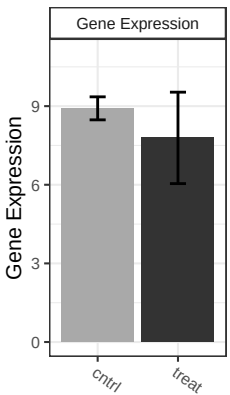


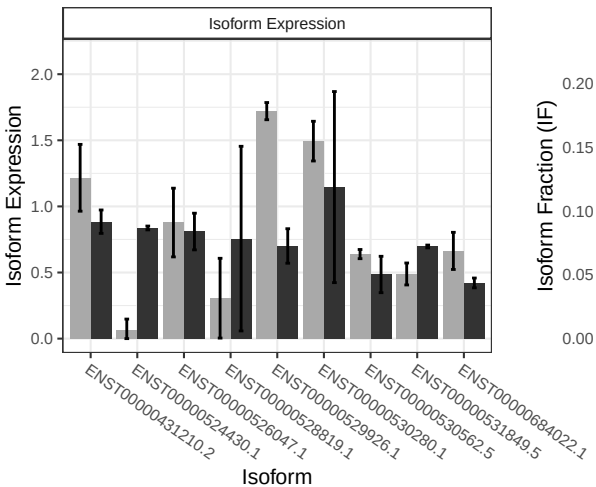


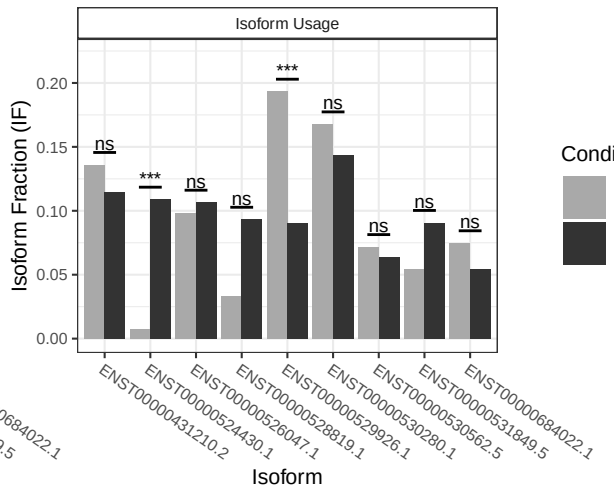
The isoform switch in XRRA1 (cntrl vs treat)

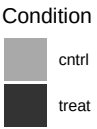




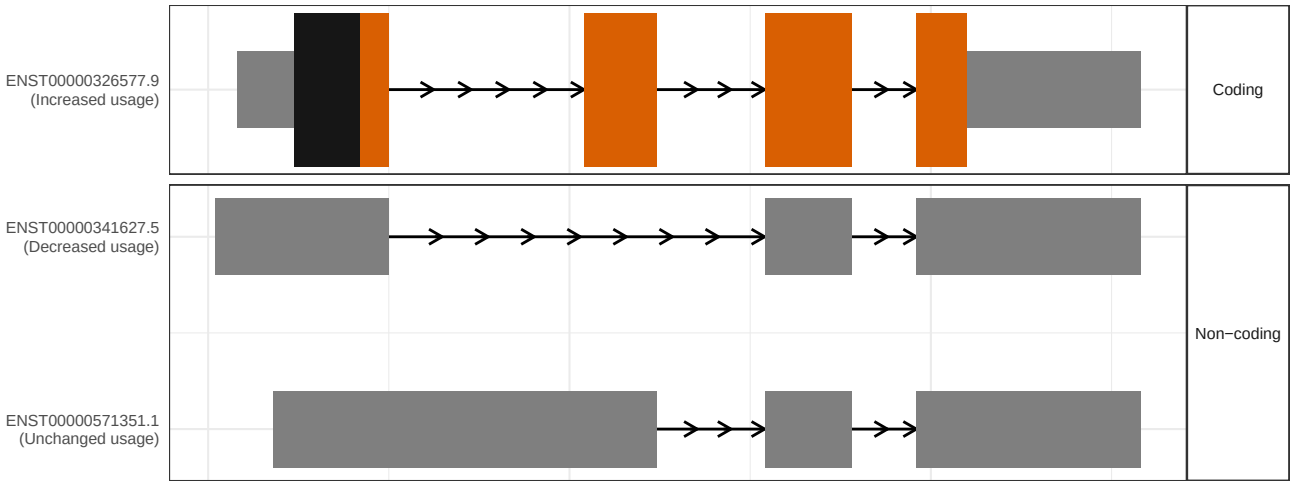


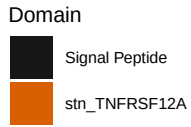


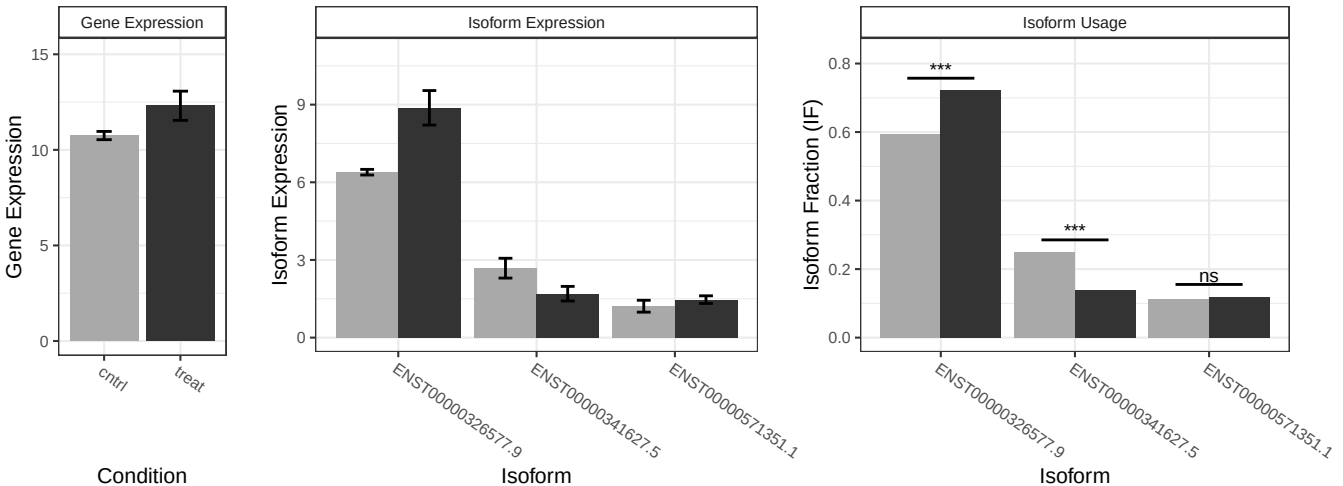


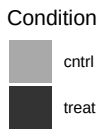


The isoform switch in TNFRSF12A (cntrl vs treat)

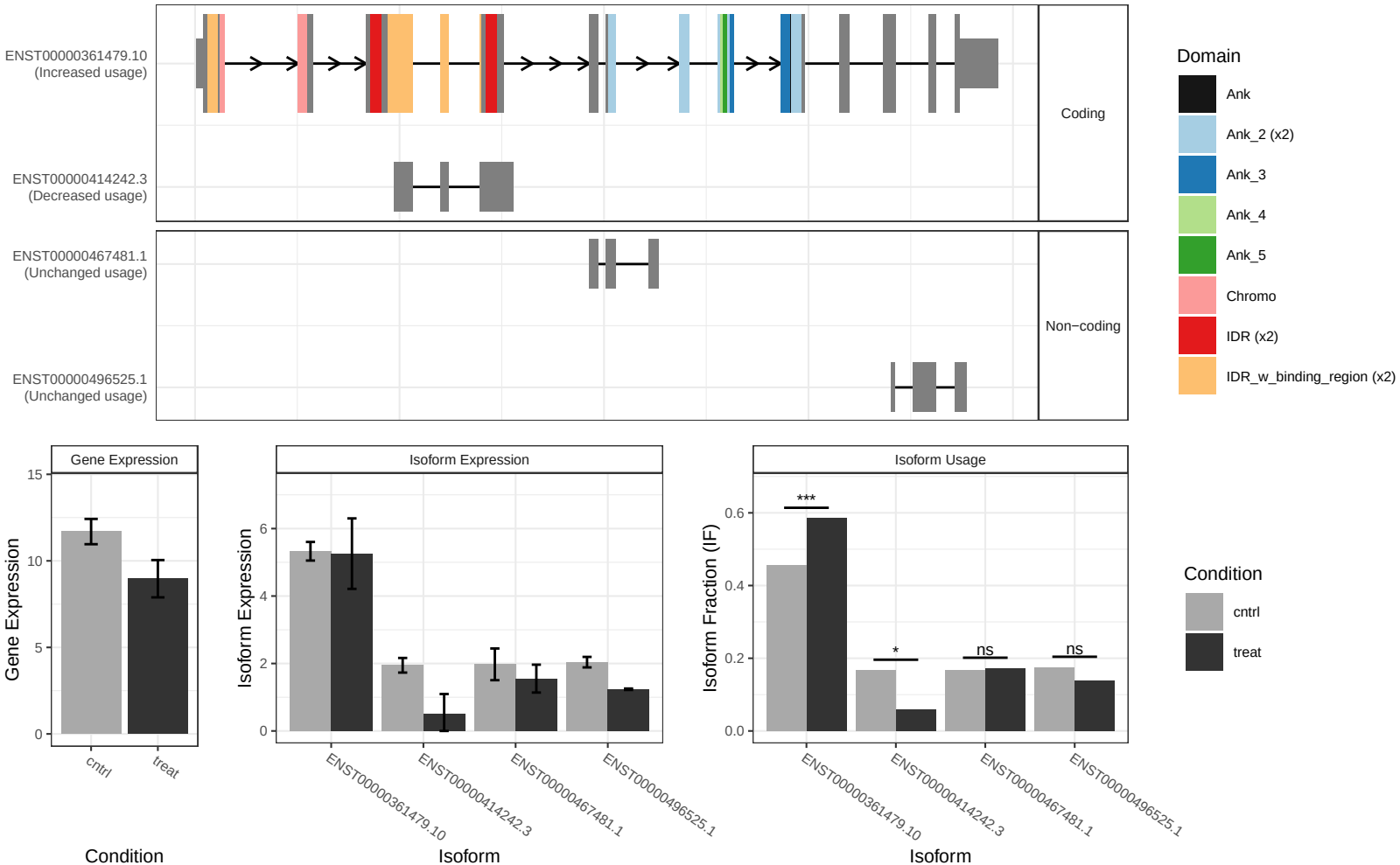




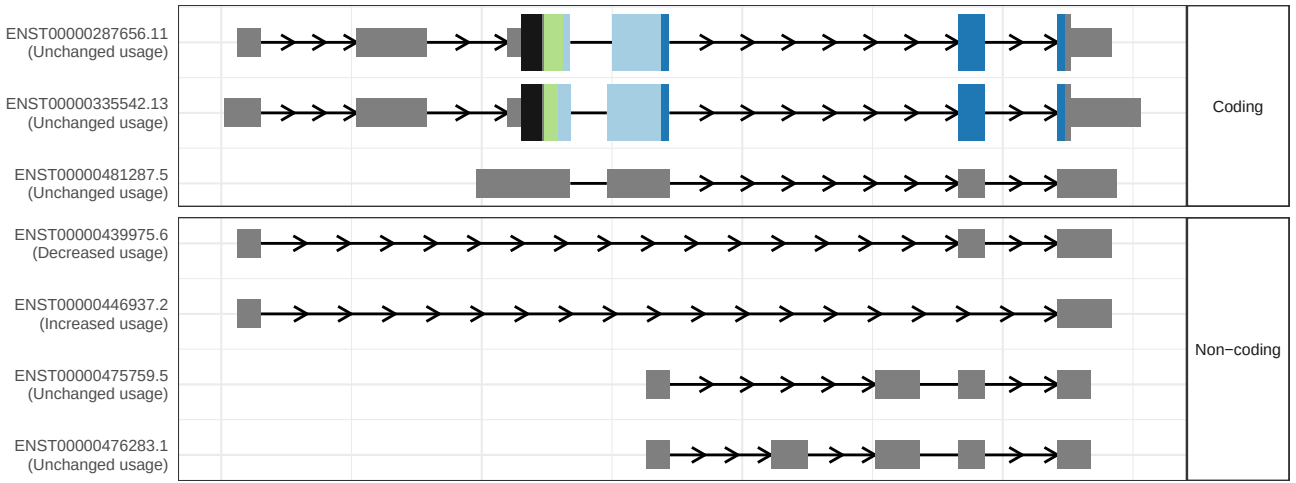


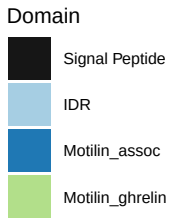


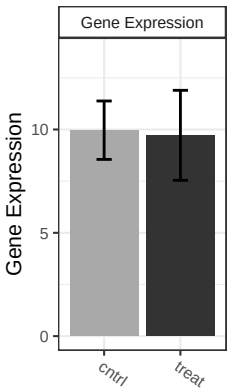
The isoform switch in MPHOSPH8 (cntrl vs treat)

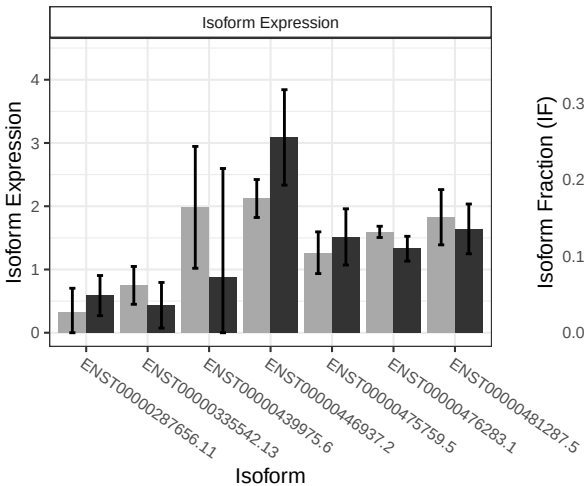


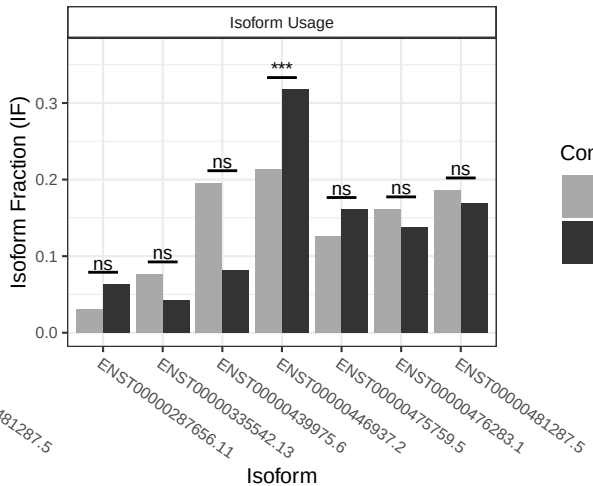
The isoform switch in GHRL (cntrl vs treat)

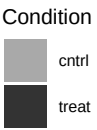




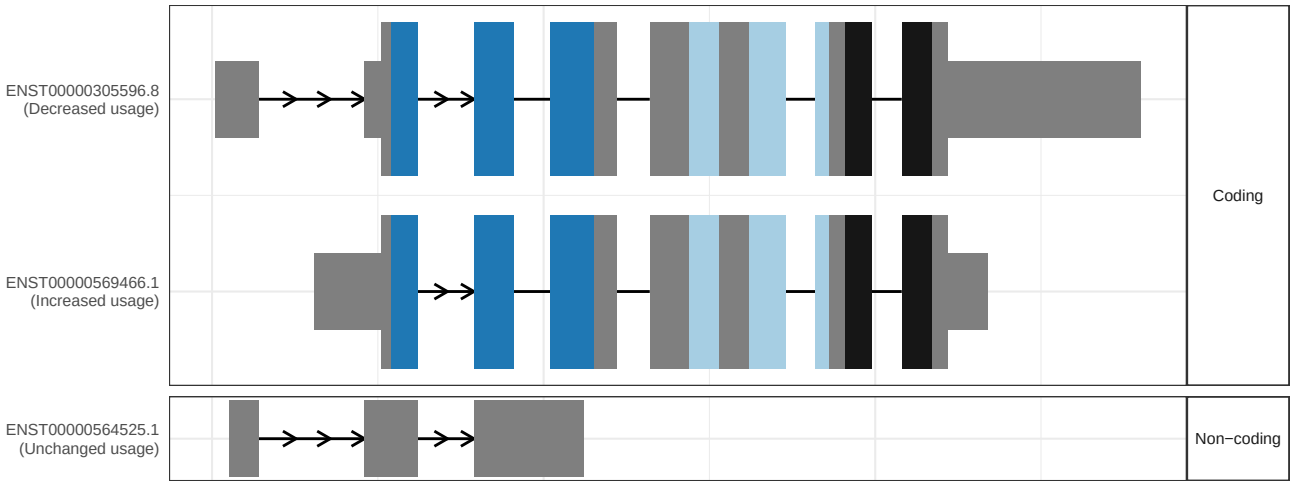


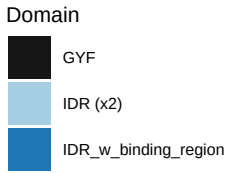


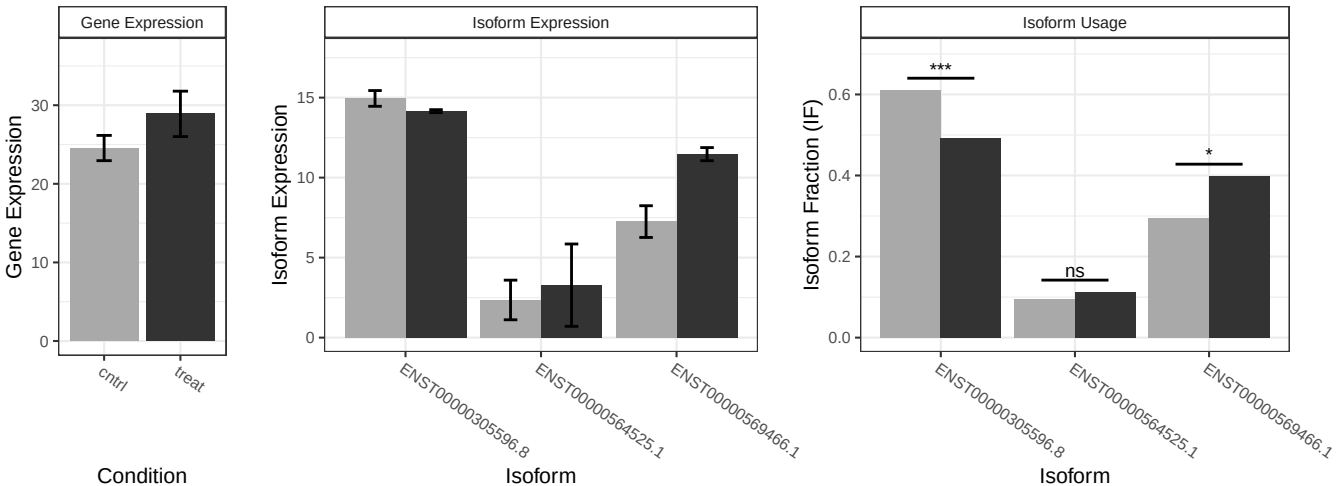


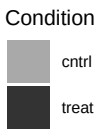


The isoform switch in CD2BP2 (cntrl vs treat)

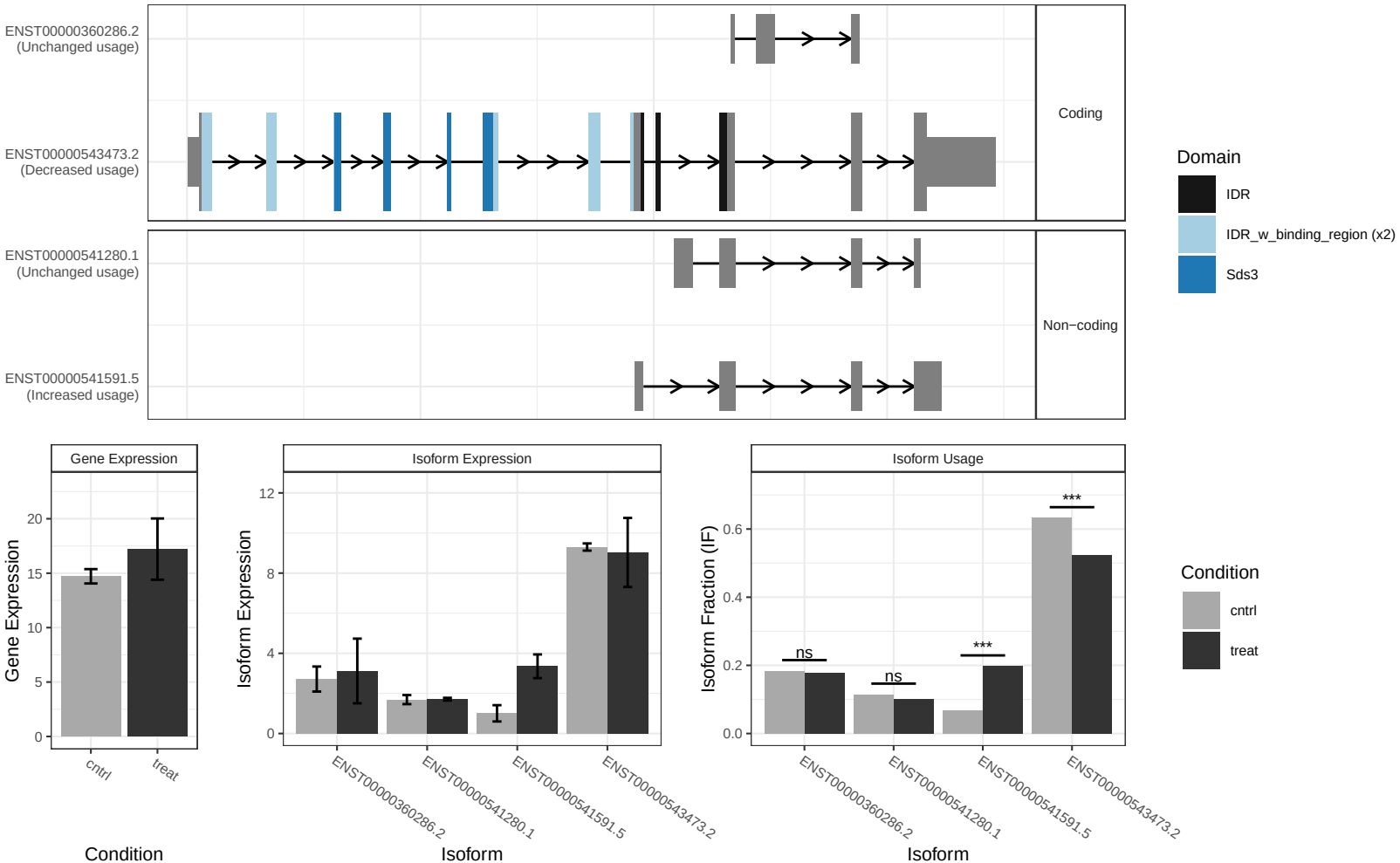




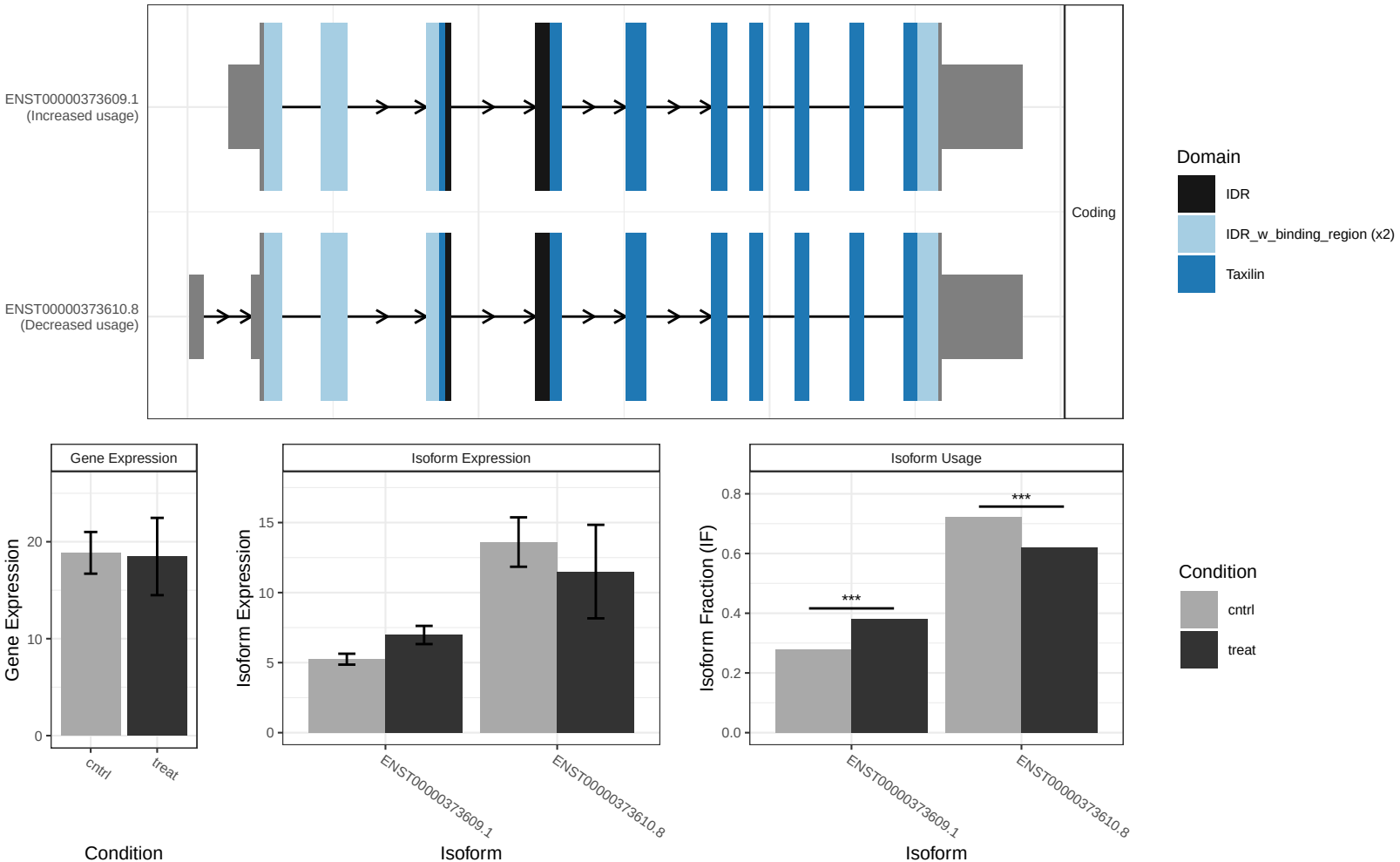




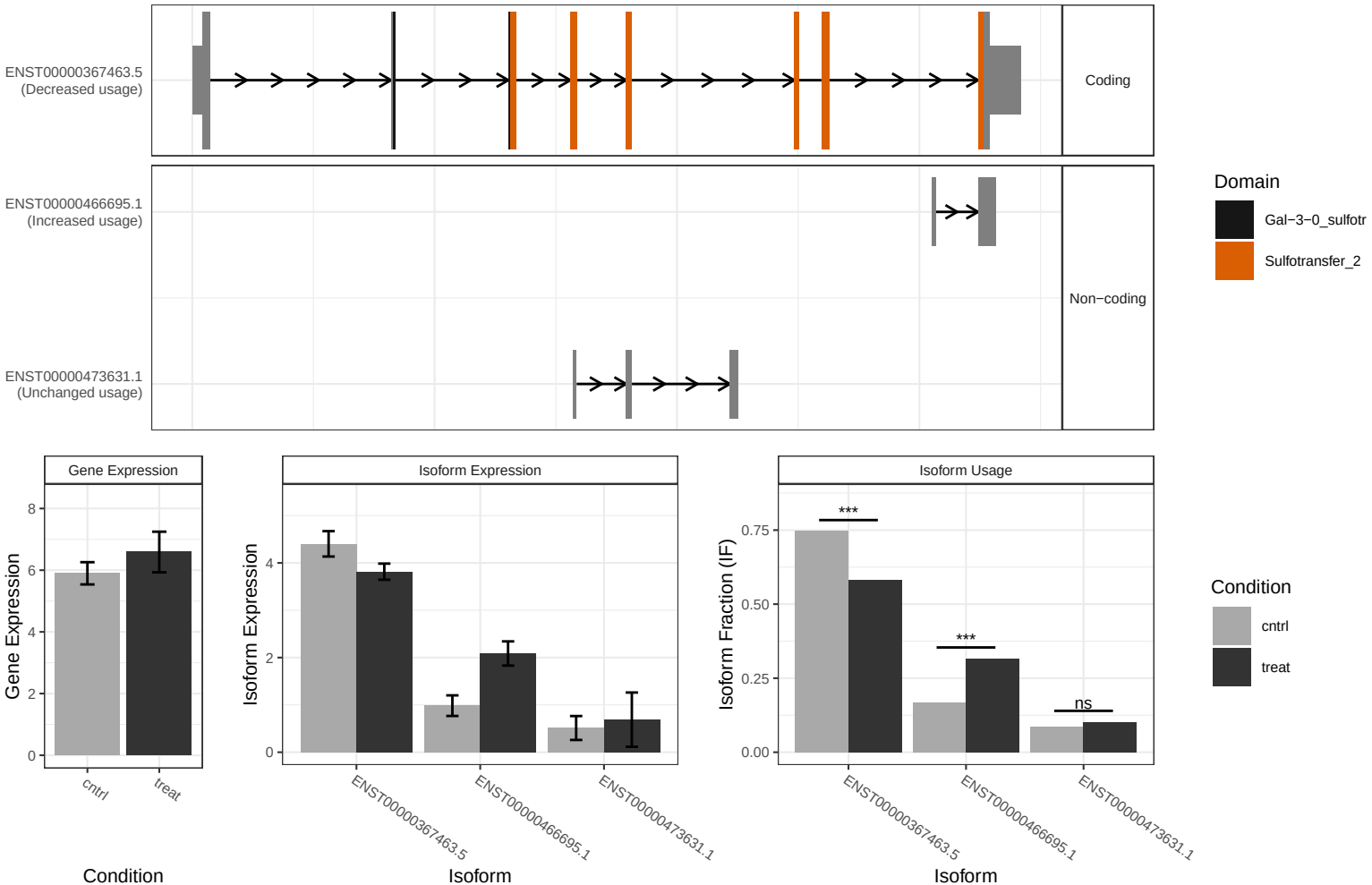
The isoform switch in SUDS3 (cntrl vs treat)

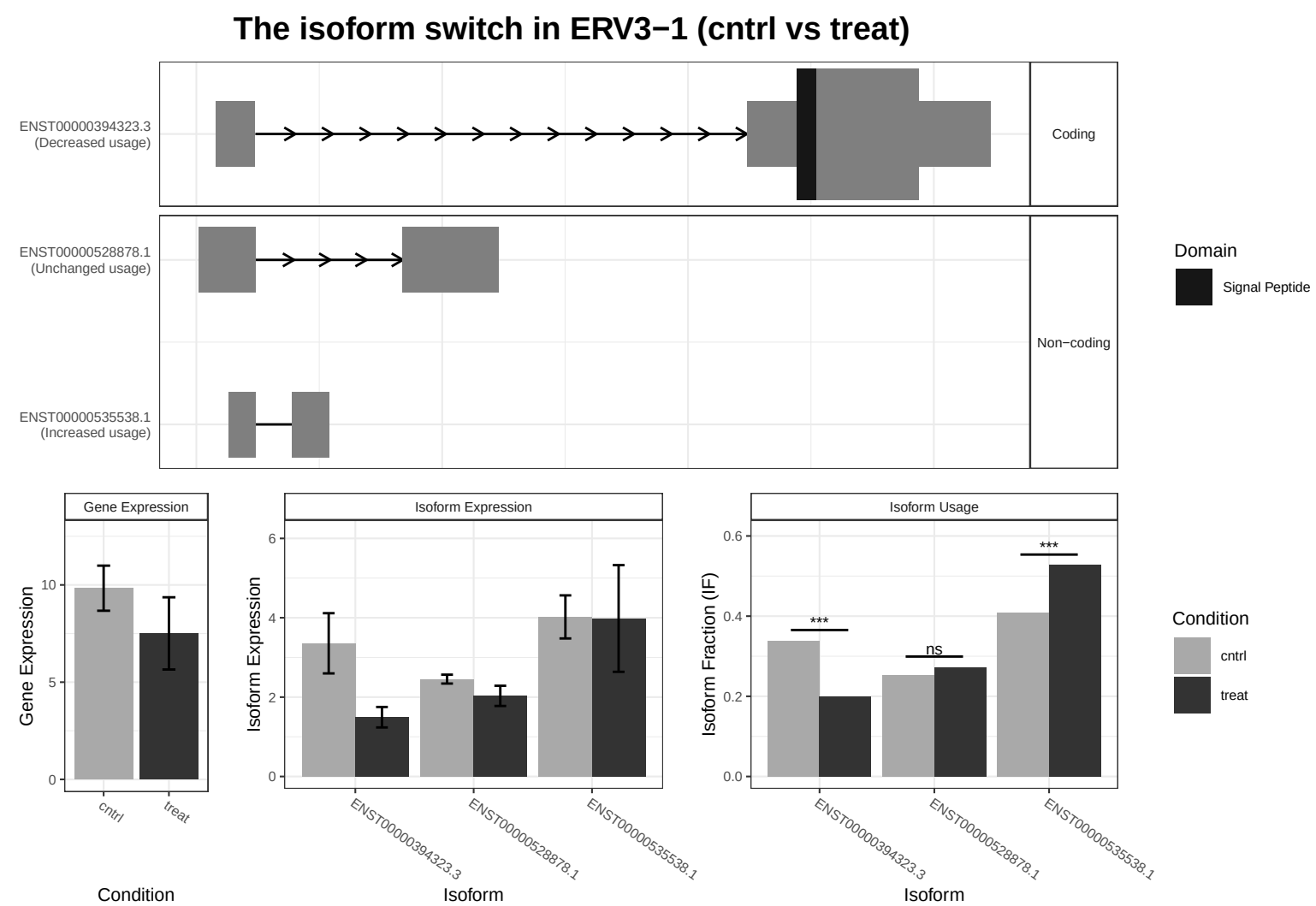


The isoform switch in TXLNA (cntrl vs treat)

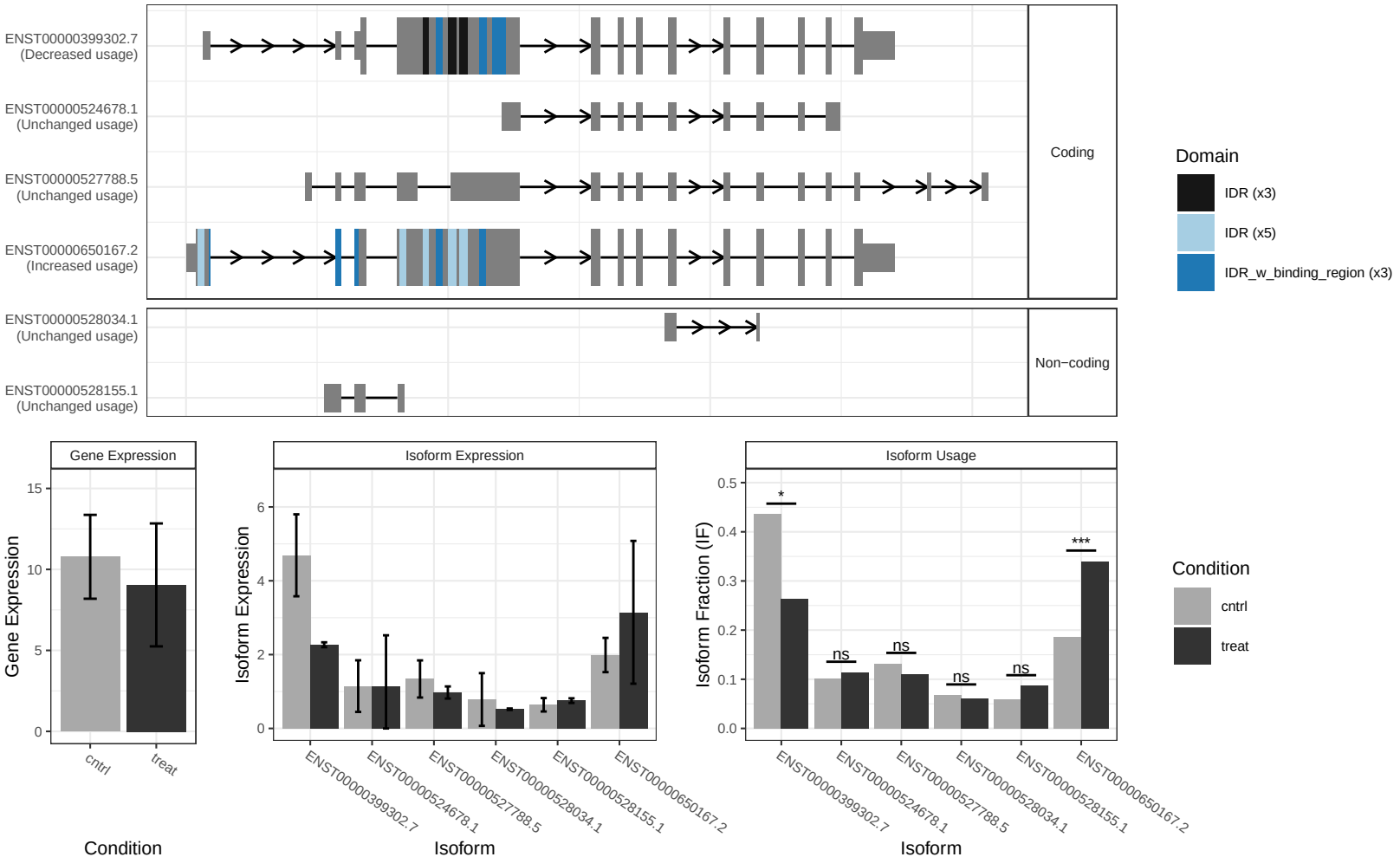


The isoform switch in UST (cntrl vs treat)

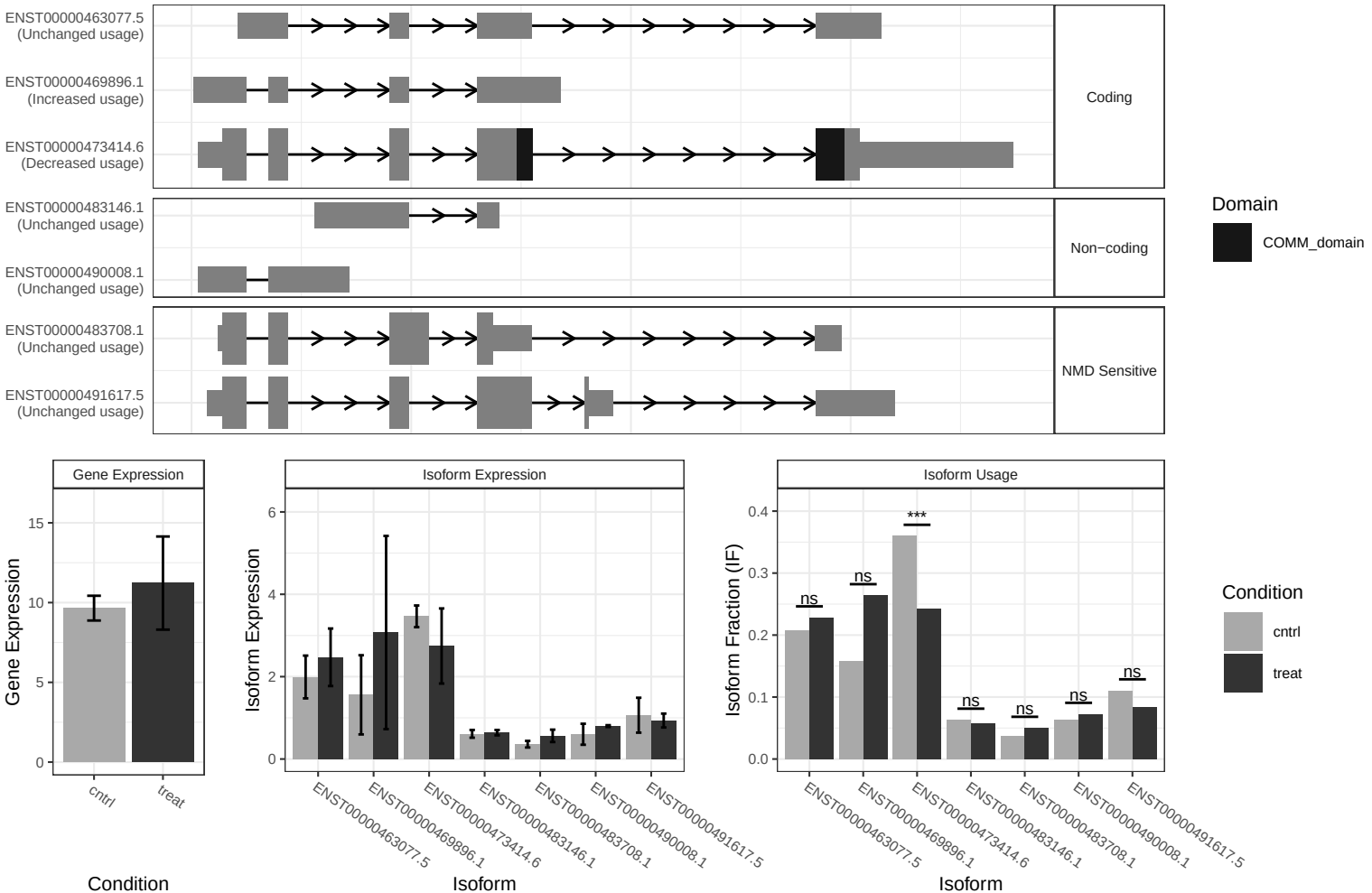


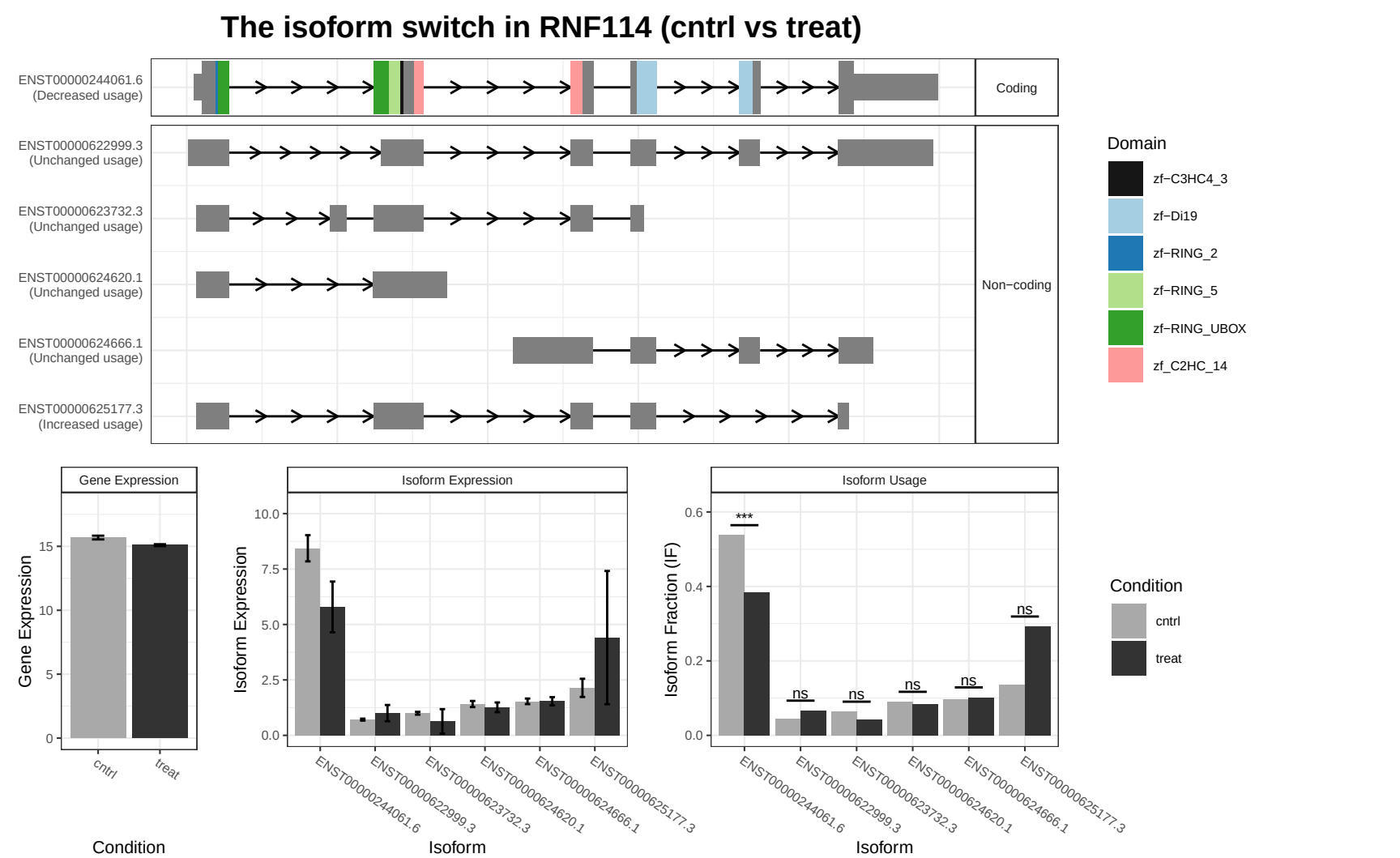

The isoform switch in ERV3−1 (cntrl vs treat)
ENST00000394323.3
(Decreased usage)
ENST00000528878.1
(Unchanged usage)
ENST00000535538.1
(Increased usage)
Coding
Non−coding
Domain
Signal Peptide
Gene Expression
Gene Expression
10
5
0
cntrl
treat
Condition
Isoform Expression
Isoform Expression
6
4
2
0
ENST00000394323.3
ENST00000528878.1
ENST00000535538.1
Isoform
Isoform Usage
Isoform Fraction (IF)
0.6
0.4
0.2
0.0
***
ns
***
ENST00000394323.3
ENST00000528878.1
ENST00000535538.1
Isoform
Condition
cntrl
treat

The isoform switch in QSER1 (cntrl vs treat)

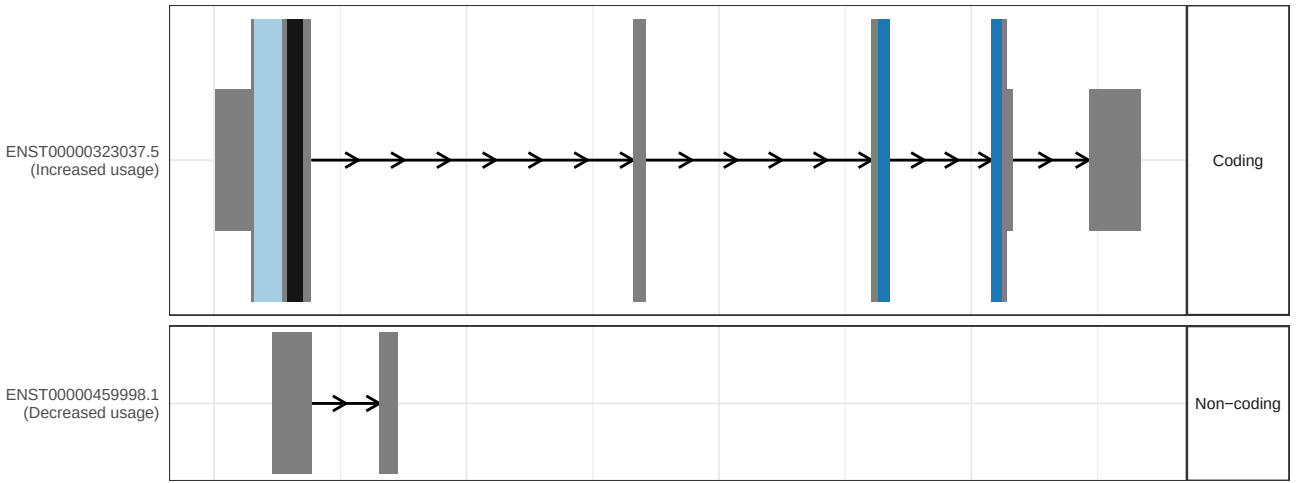


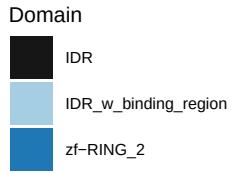
The isoform switch in COMMD2 (cntrl vs treat)

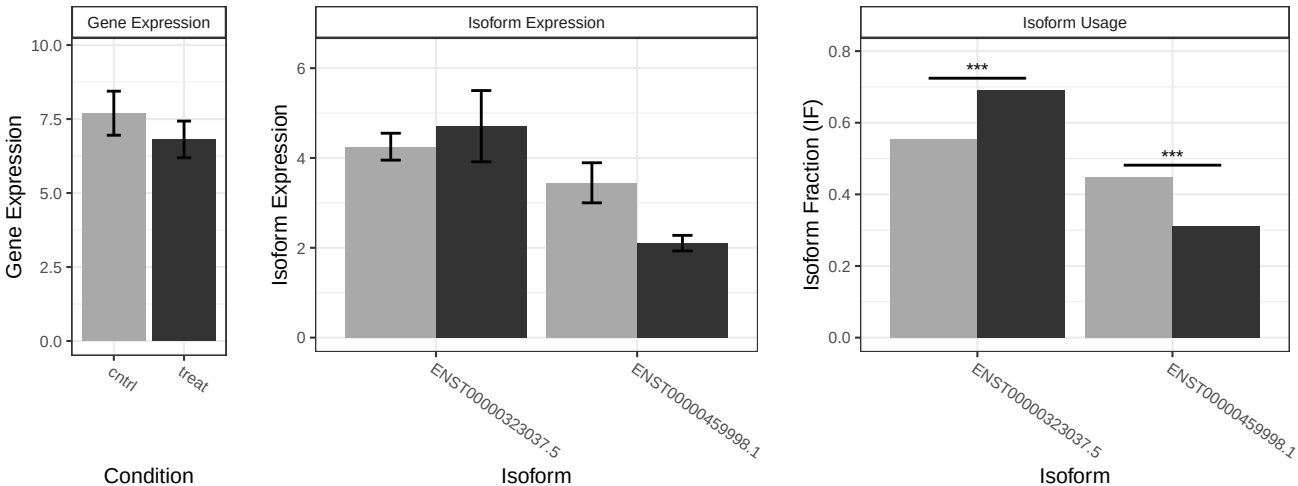


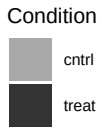

The isoform switch in RNF114 (cntrl vs treat)
ENST00000244061.6
(Decreased usage)
ENST00000622999.3
(Unchanged usage)
ENST00000623732.3
(Unchanged usage)
ENST00000624620.1
(Unchanged usage)
ENST00000624666.1
(Unchanged usage)
ENST00000625177.3
(Increased usage)
Coding
Non−coding
Domain
zf−C3HC4_3
zf−Di19
zf−RING_2
zf−RING_5
zf−RING_UBOX
zf_C2HC_14
Gene Expression
Gene Expression
Condition
cntrl
treat
Isoform Expression
Isoform Expression
Isoform
ENST00000244061.6
ENST00000622999.3
ENST00000623732.3
ENST00000624620.1
ENST00000624666.1
ENST00000625177.3
Isoform Usage
Isoform Fraction (IF)
***
ns
Isoform
Condition
cntrl
treat

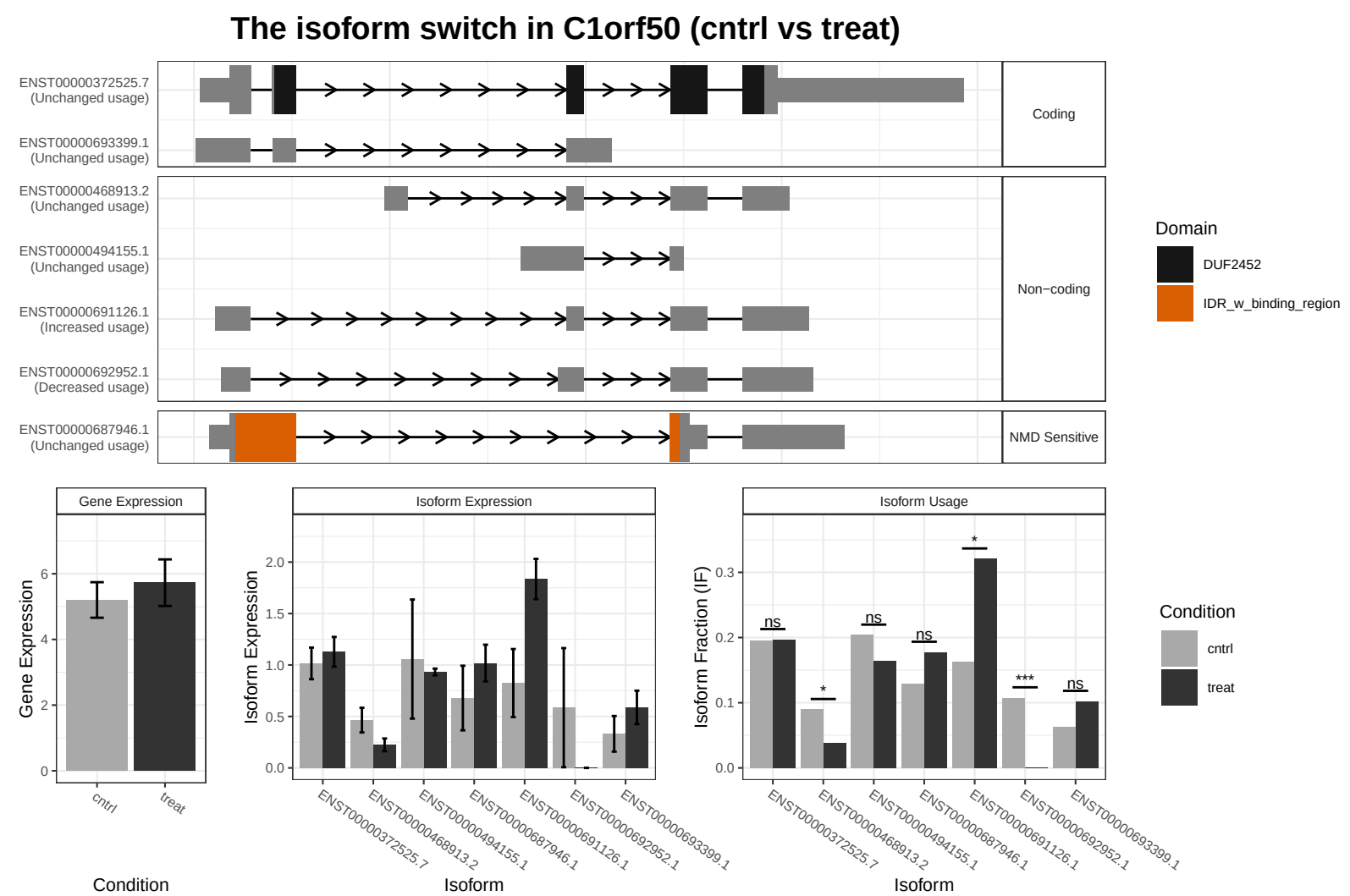
The isoform switch in ZNRF2 (cntrl vs treat)

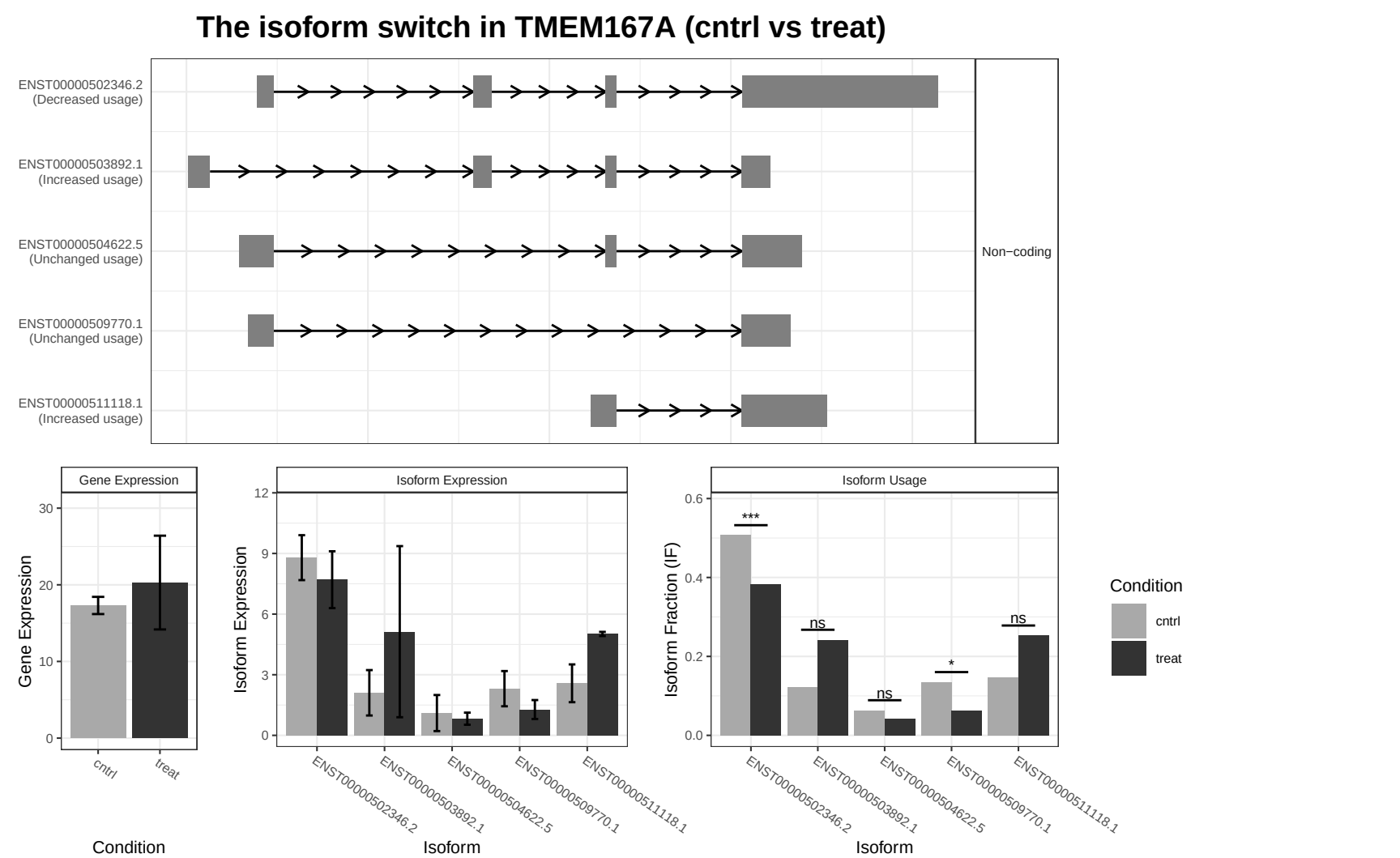




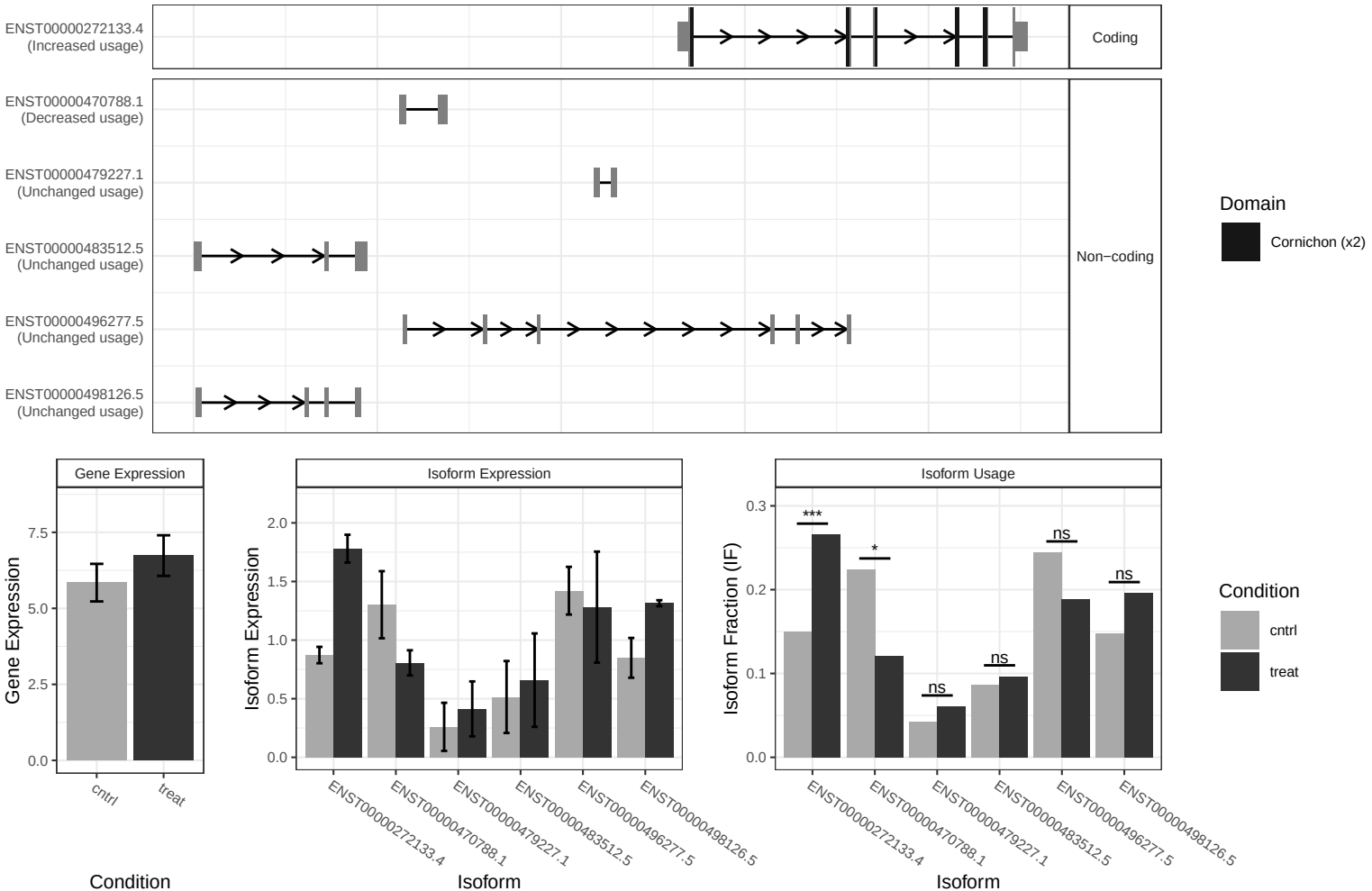




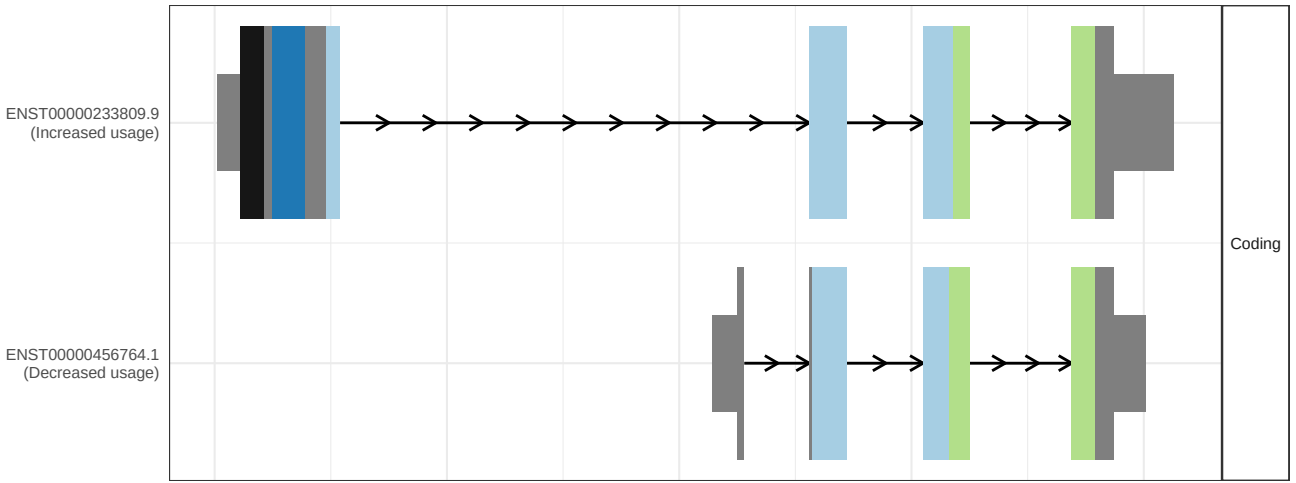

The isoform switch in C1orf50 (cntrl vs treat)
ENST00000372525.7
(Unchanged usage)
ENST00000693399.1
(Unchanged usage)
Coding
ENST00000468913.2
(Unchanged usage)
ENST00000494155.1
(Unchanged usage)
ENST00000691126.1
(Increased usage)
ENST00000692952.1
(Decreased usage)
Non−coding
ENST00000687946.1
(Unchanged usage)
NMD Sensitive
Domain
DUF2452
IDR_w_binding_region
Gene Expression
Gene Expression
cntrl
treat
Condition
Isoform Expression
Isoform Expression
ENST00000372525.7
ENST00000468913.2
ENST00000494155.1
ENST00000687946.1
ENST00000691126.1
ENST00000692952.1
ENST00000693399.1
Isoform
Isoform Usage
Isoform Fraction (IF)
ns
*
ns
ns
*
***
ns
Isoform
Condition
cntrl
treat

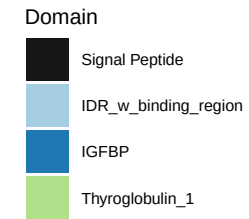

The isoform switch in TMEM167A (cntrl vs treat)
ENST00000502346.2
(Decreased usage)
ENST00000503892.1
(Increased usage)
ENST00000504622.5
(Unchanged usage)
ENST00000509770.1
(Unchanged usage)
ENST00000511118.1
(Increased usage)
Non−coding
Gene Expression
Gene Expression
30
20
10
0
cntrl
treat
Condition
Isoform Expression
Isoform Expression
12
9
6
3
0
ENST00000502346.2
ENST00000503892.1
ENST00000504622.5
ENST00000509770.1
ENST00000511118.1
Isoform
Isoform Usage
Isoform Fraction (IF)
0.6
0.4
0.2
0.0
***
ns
ns
*
ns
ENST00000502346.2
ENST00000503892.1
ENST00000504622.5
ENST00000509770.1
ENST00000511118.1
Isoform
Condition
cntrl
treat

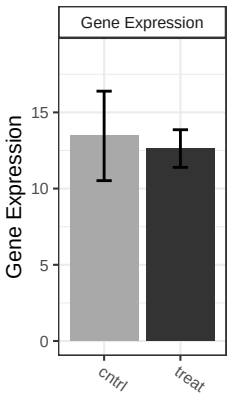
The isoform switch in CNIH3 (cntrl vs treat)

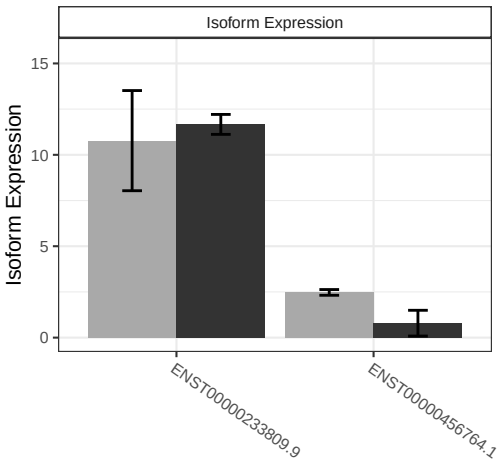


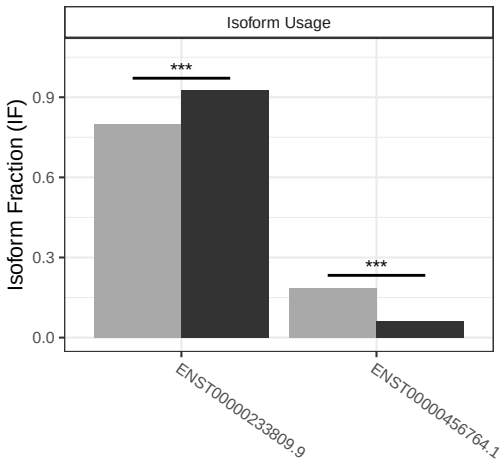
The isoform switch in IGFBP2 (cntrl vs treat)

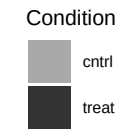




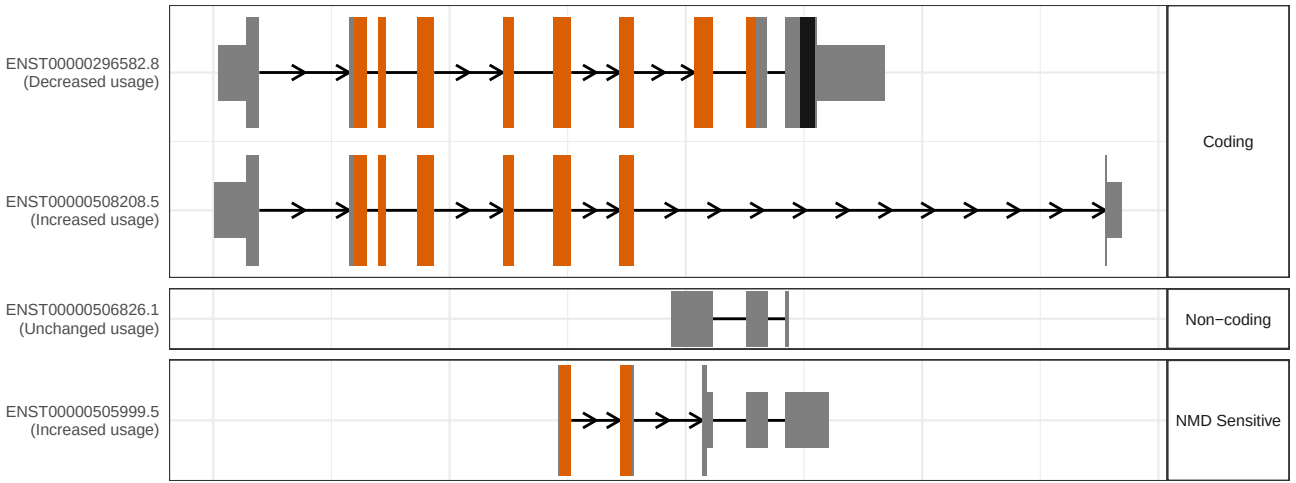




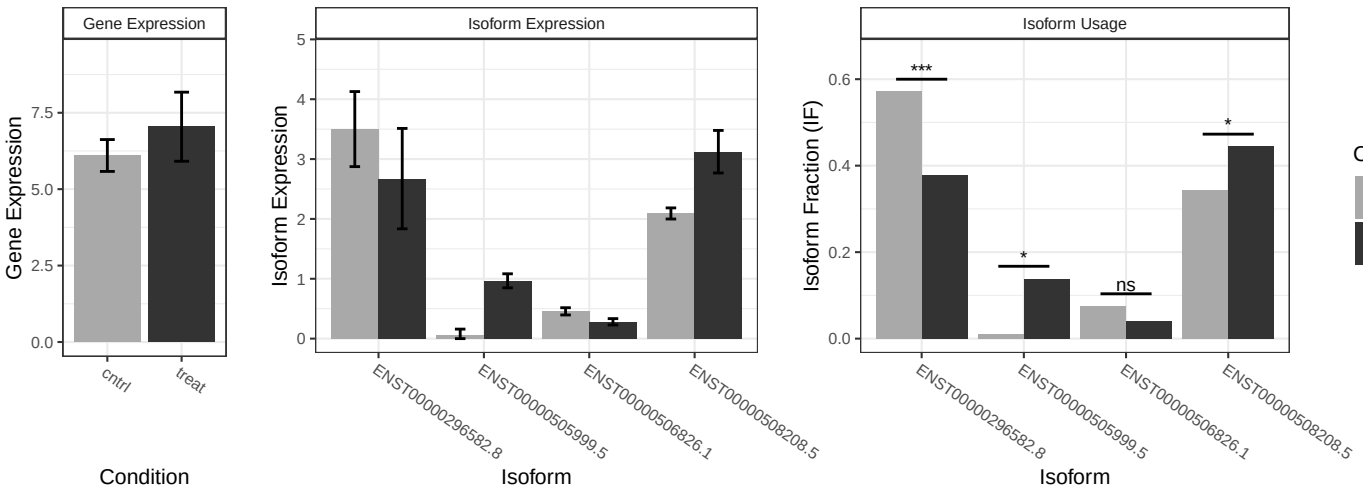




The isoform switch in TMEM184C (cntrl vs treat)

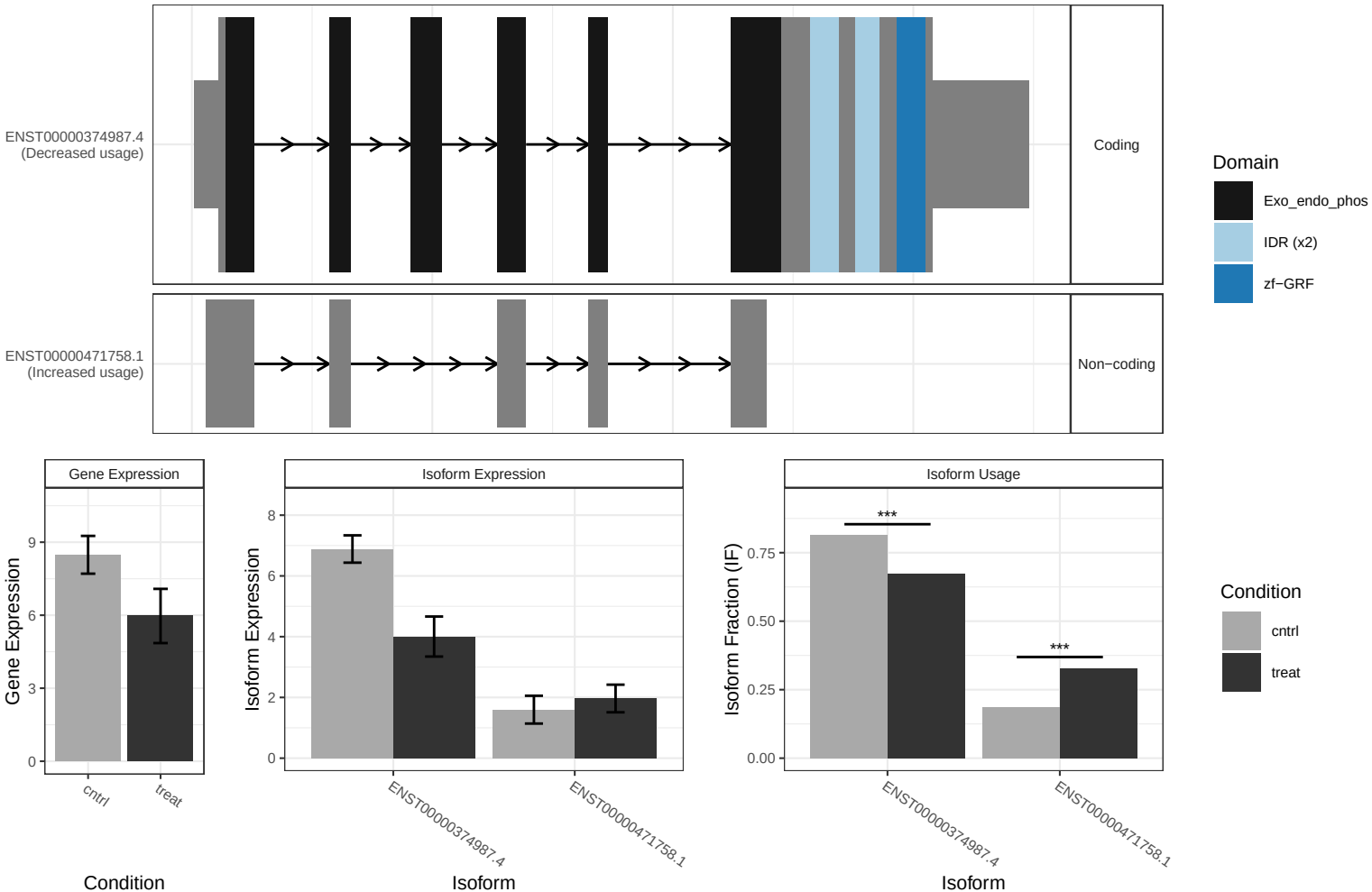


Domain
IDR_w_binding_region
Solute_trans_a

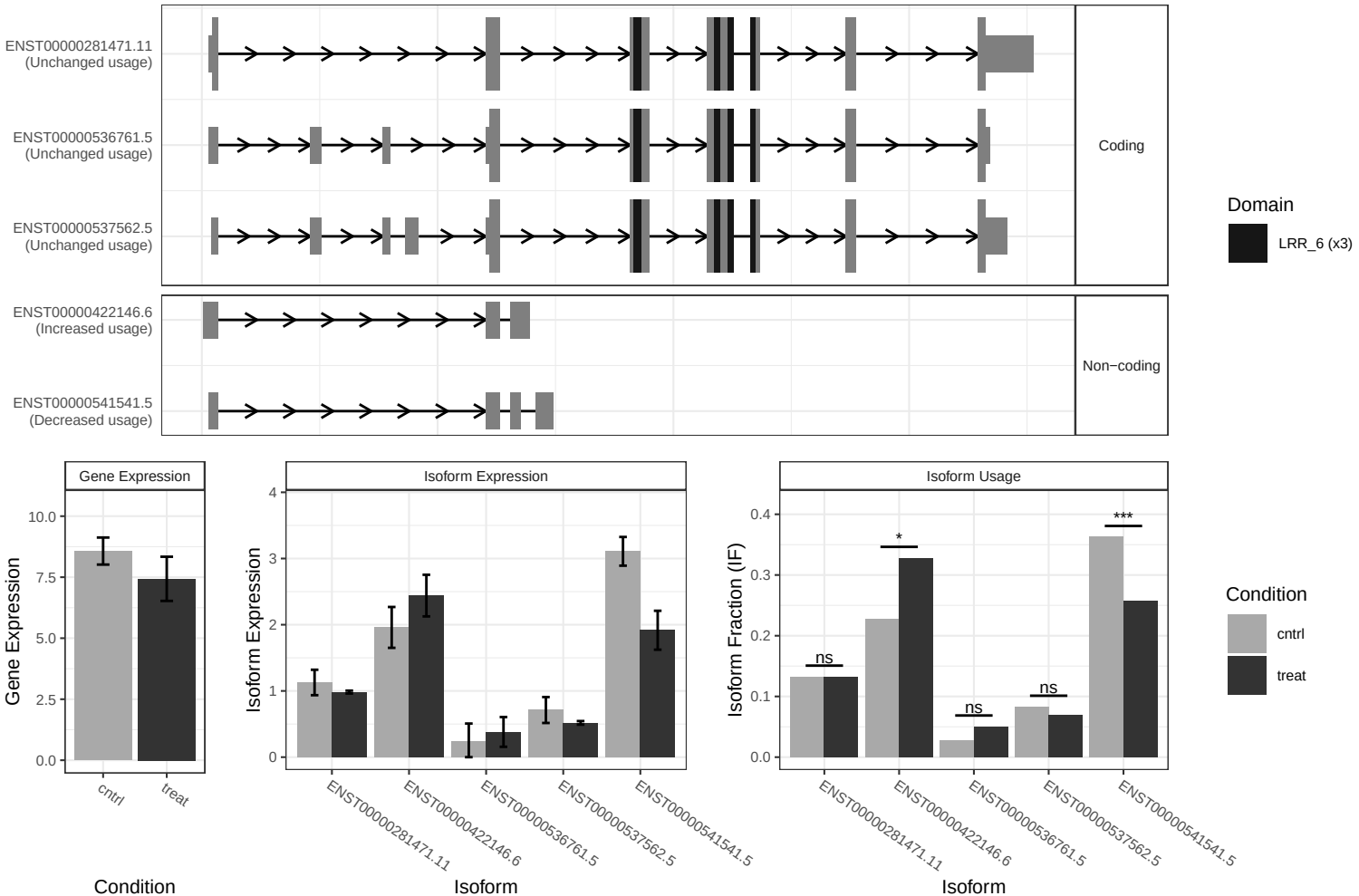


Condition
cntrl
treat

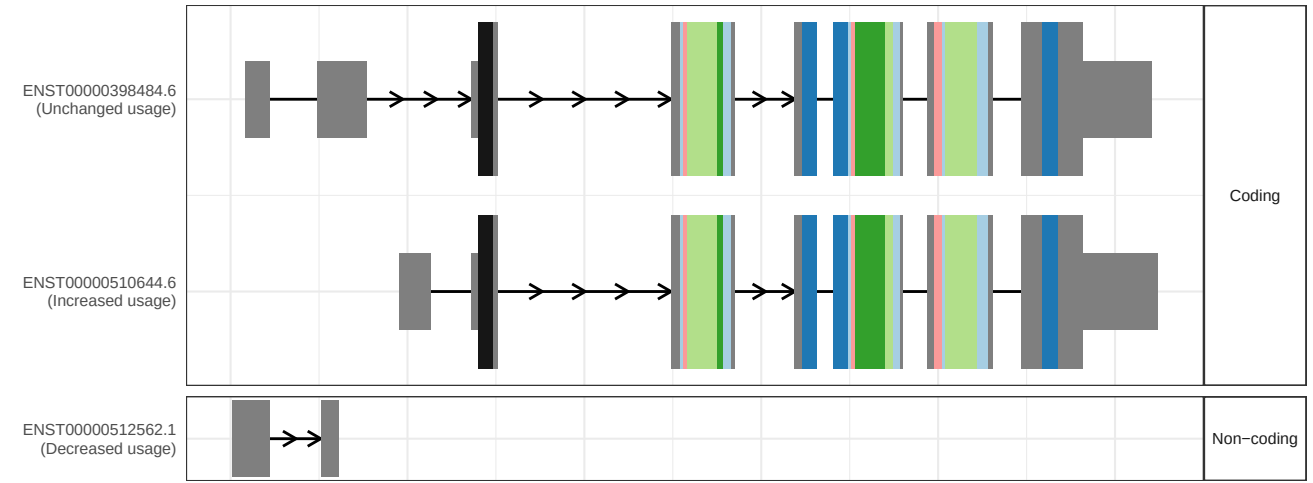
The isoform switch in APEX2 (cntrl vs treat)

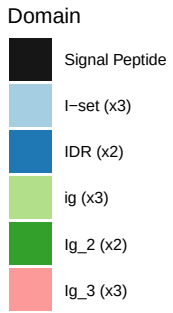


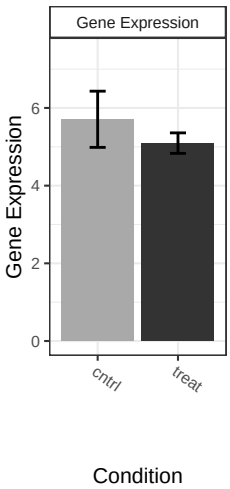
The isoform switch in AMN1 (cntrl vs treat)

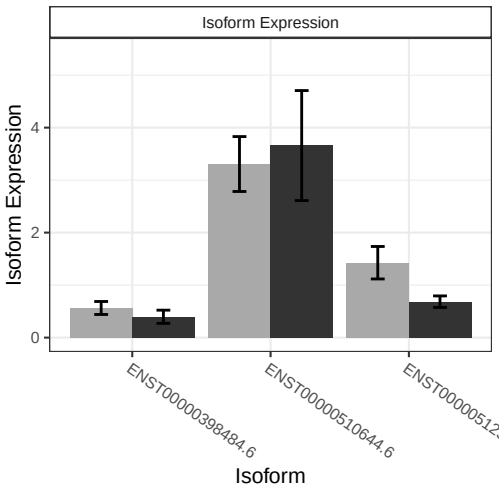


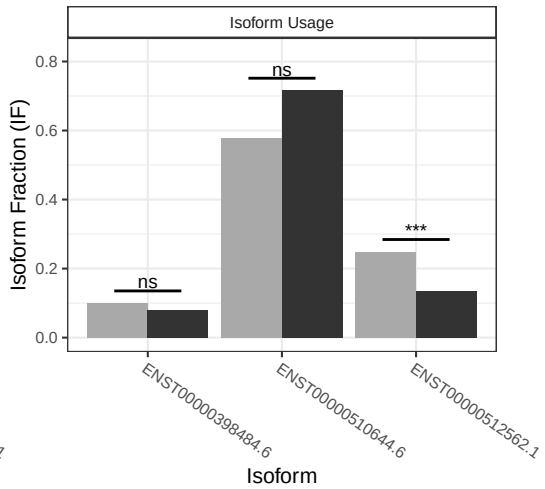
The isoform switch in FGFRL1 (cntrl vs treat)

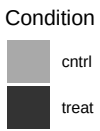




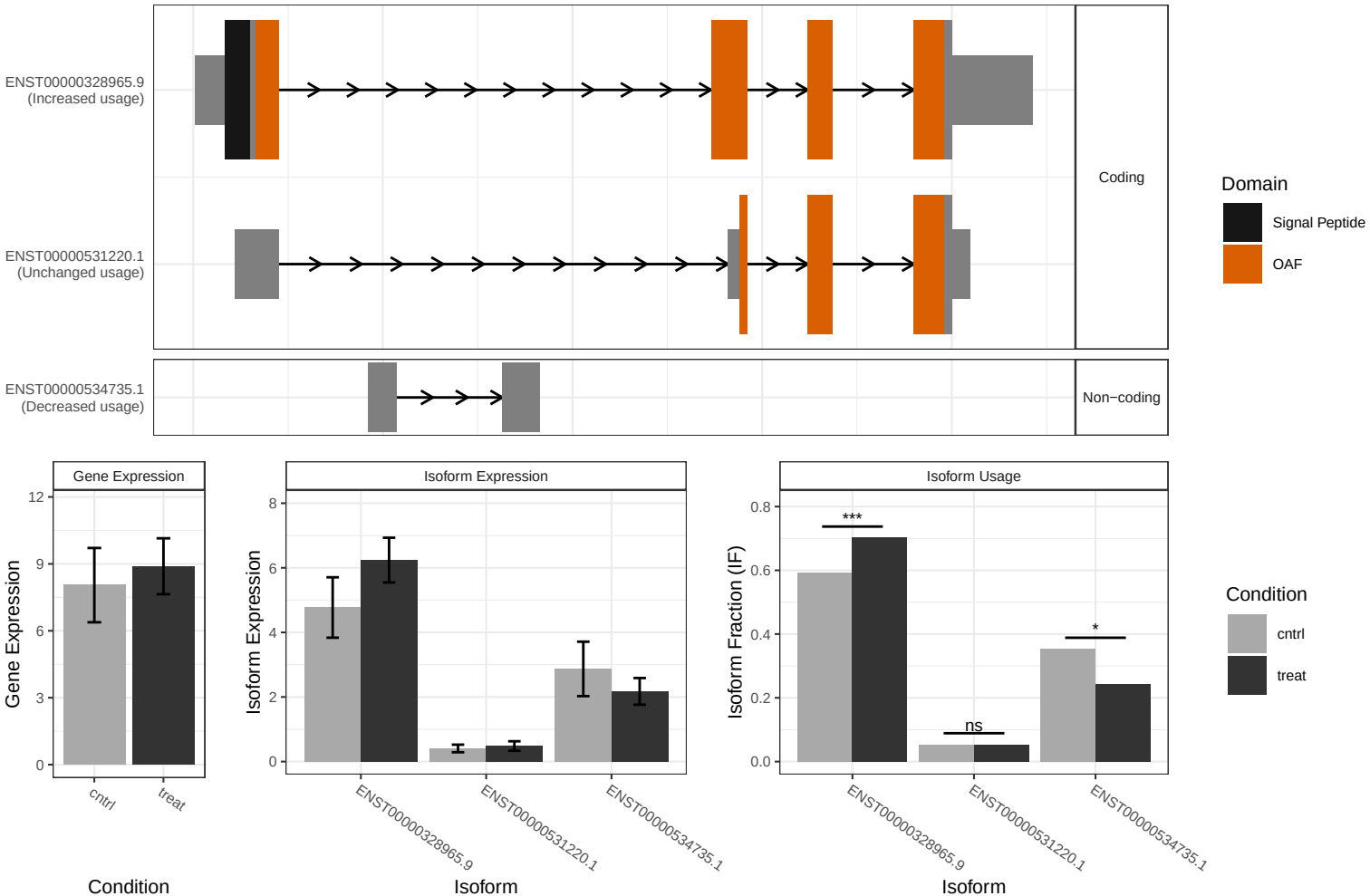




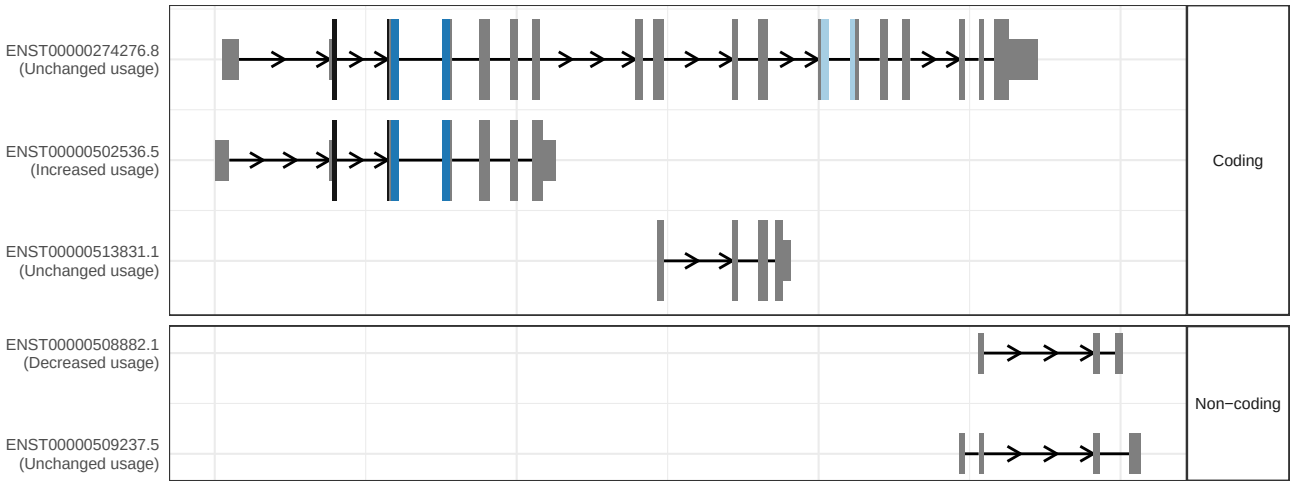


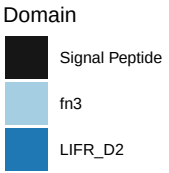


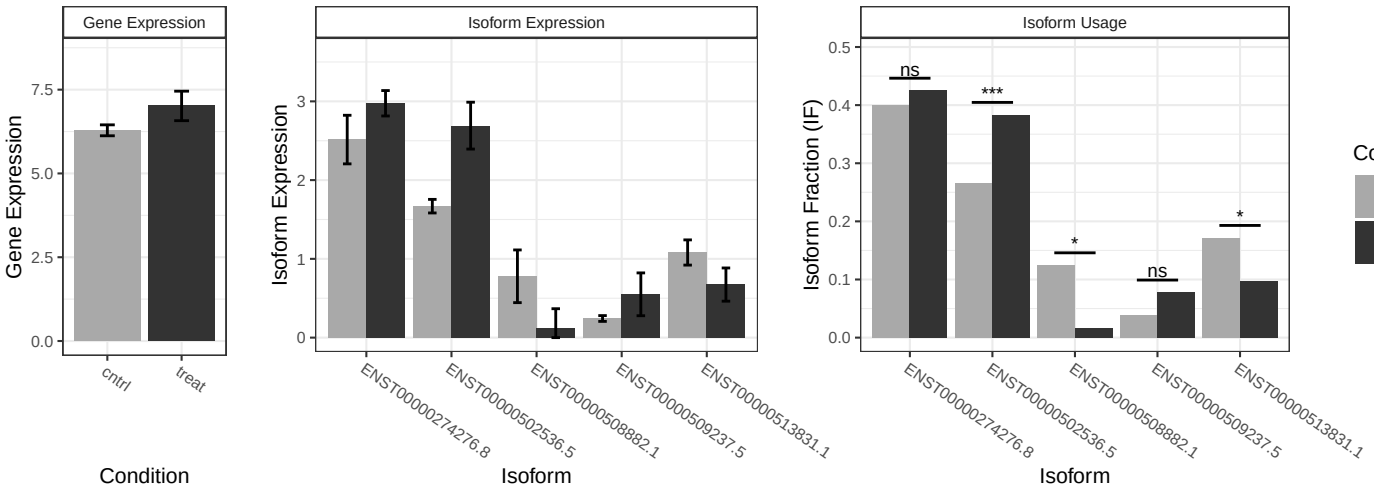
The isoform switch in OAF (cntrl vs treat)

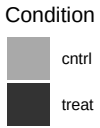


The isoform switch in OSMR (cntrl vs treat)

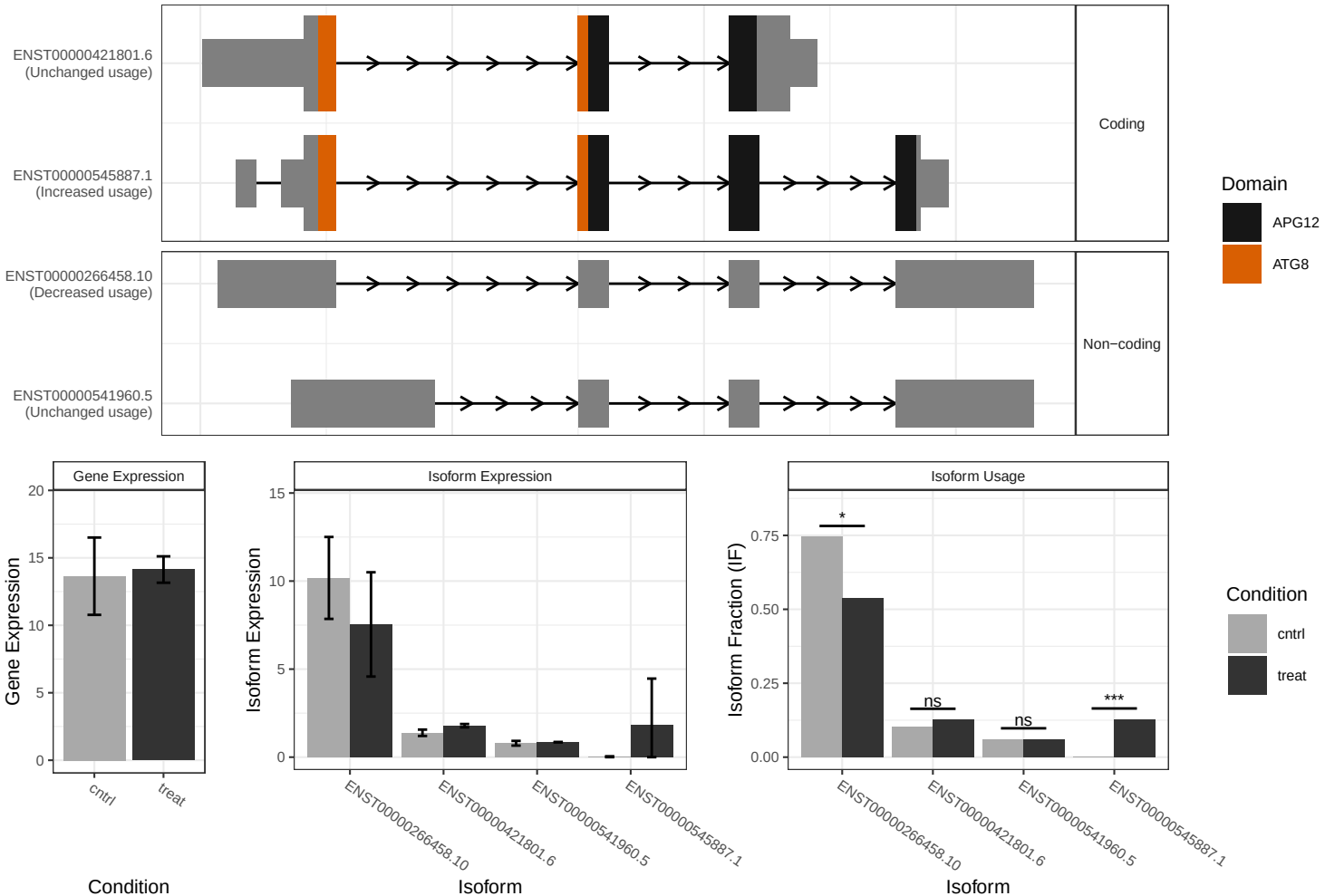




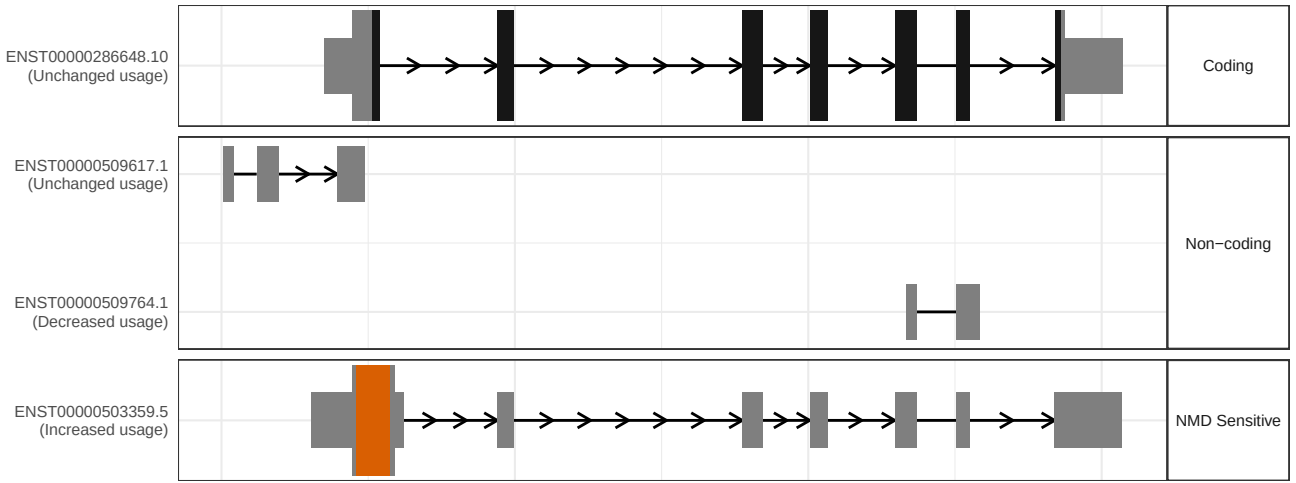


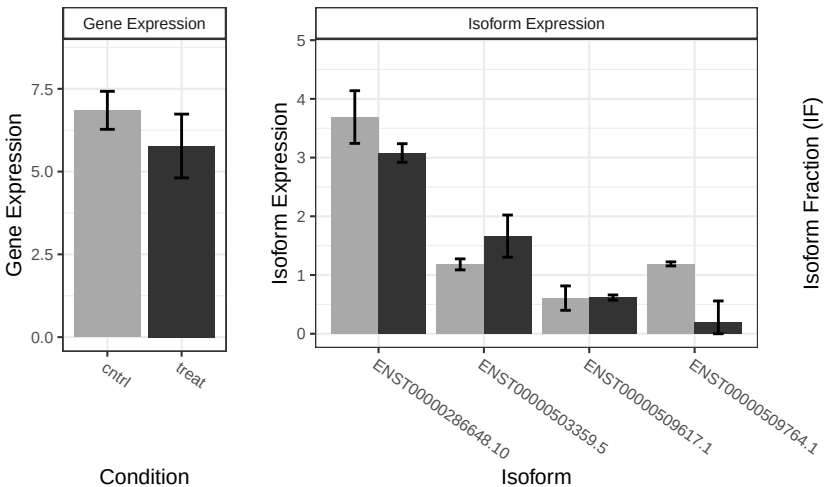


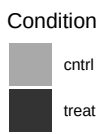
The isoform switch in GABARAPL1 (cntrl vs treat)



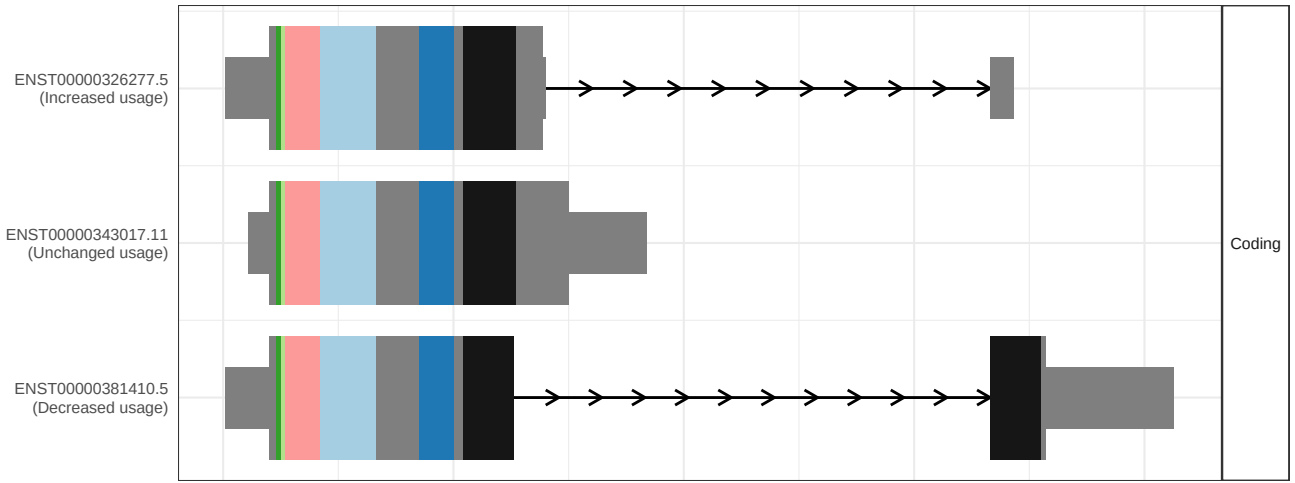
The isoform switch in DCK (cntrl vs treat)

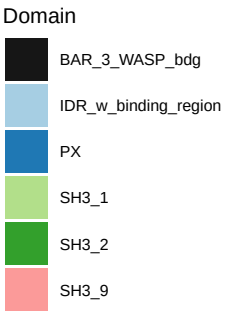


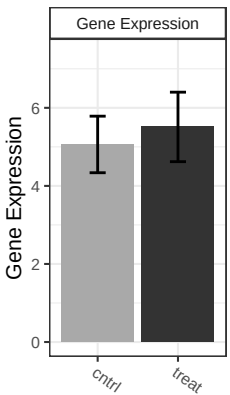


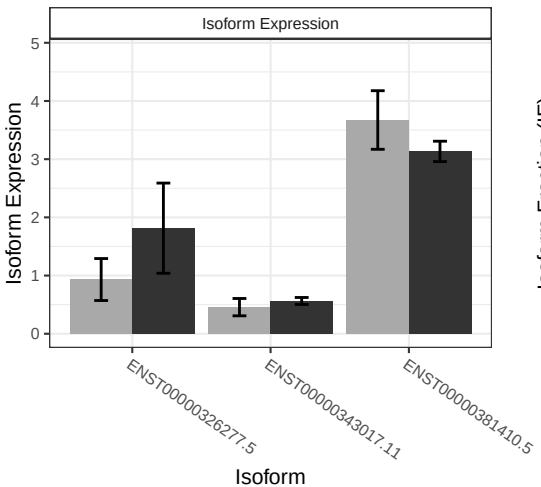


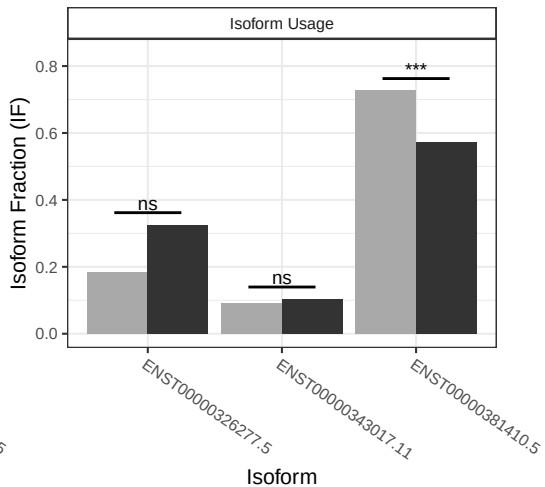
The isoform switch in SNX18 (cntrl vs treat)

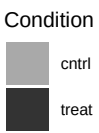




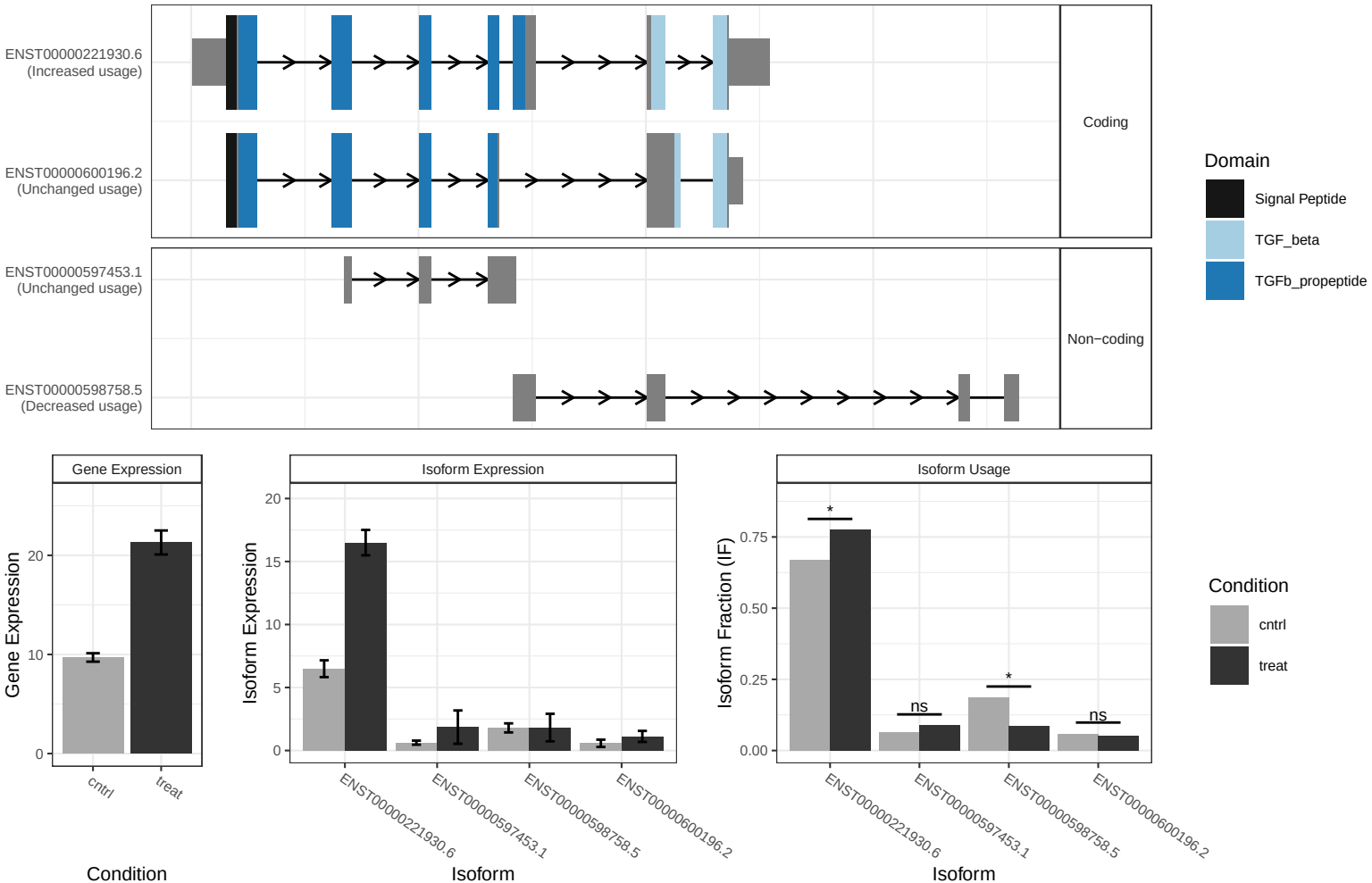




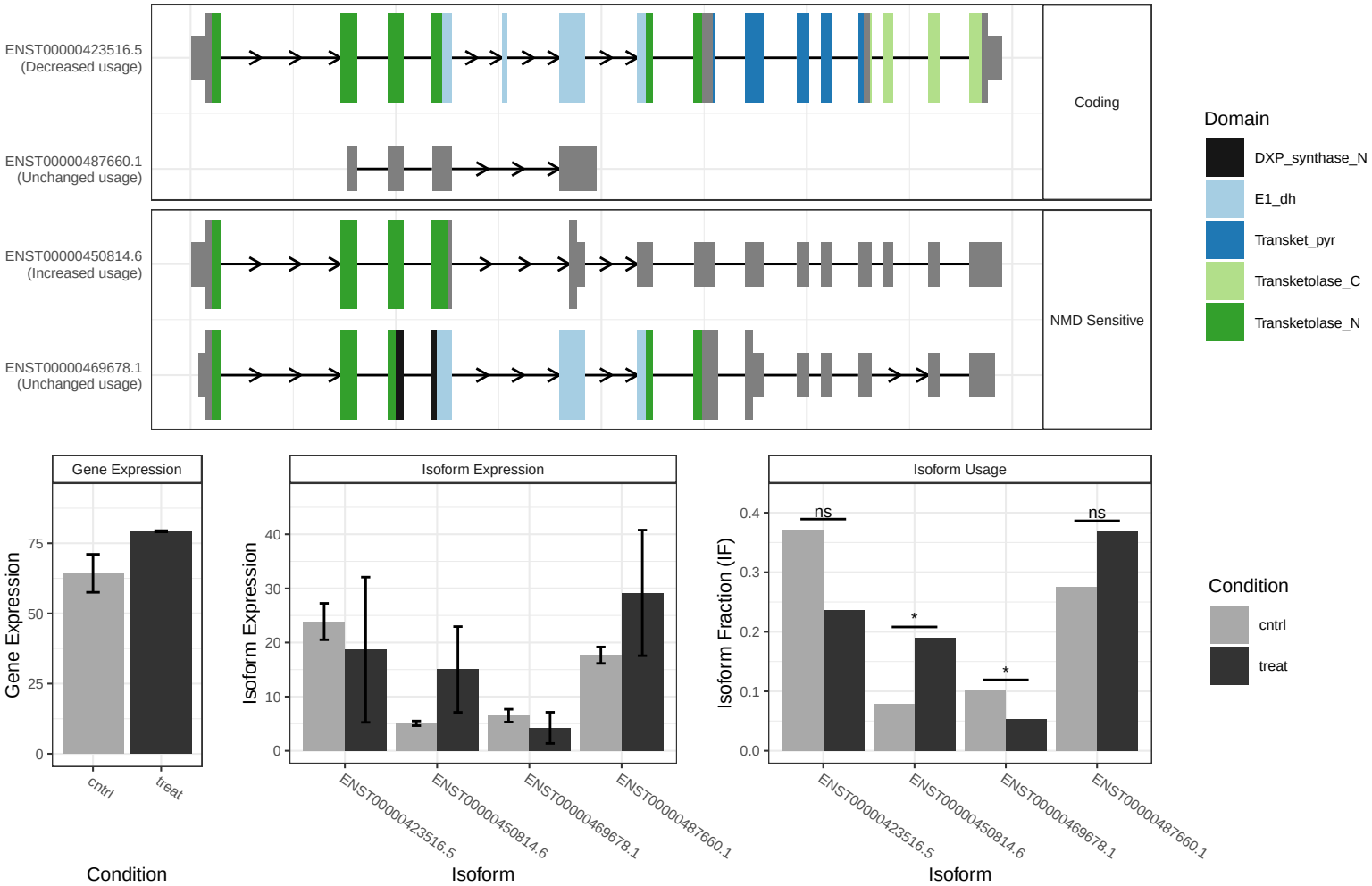




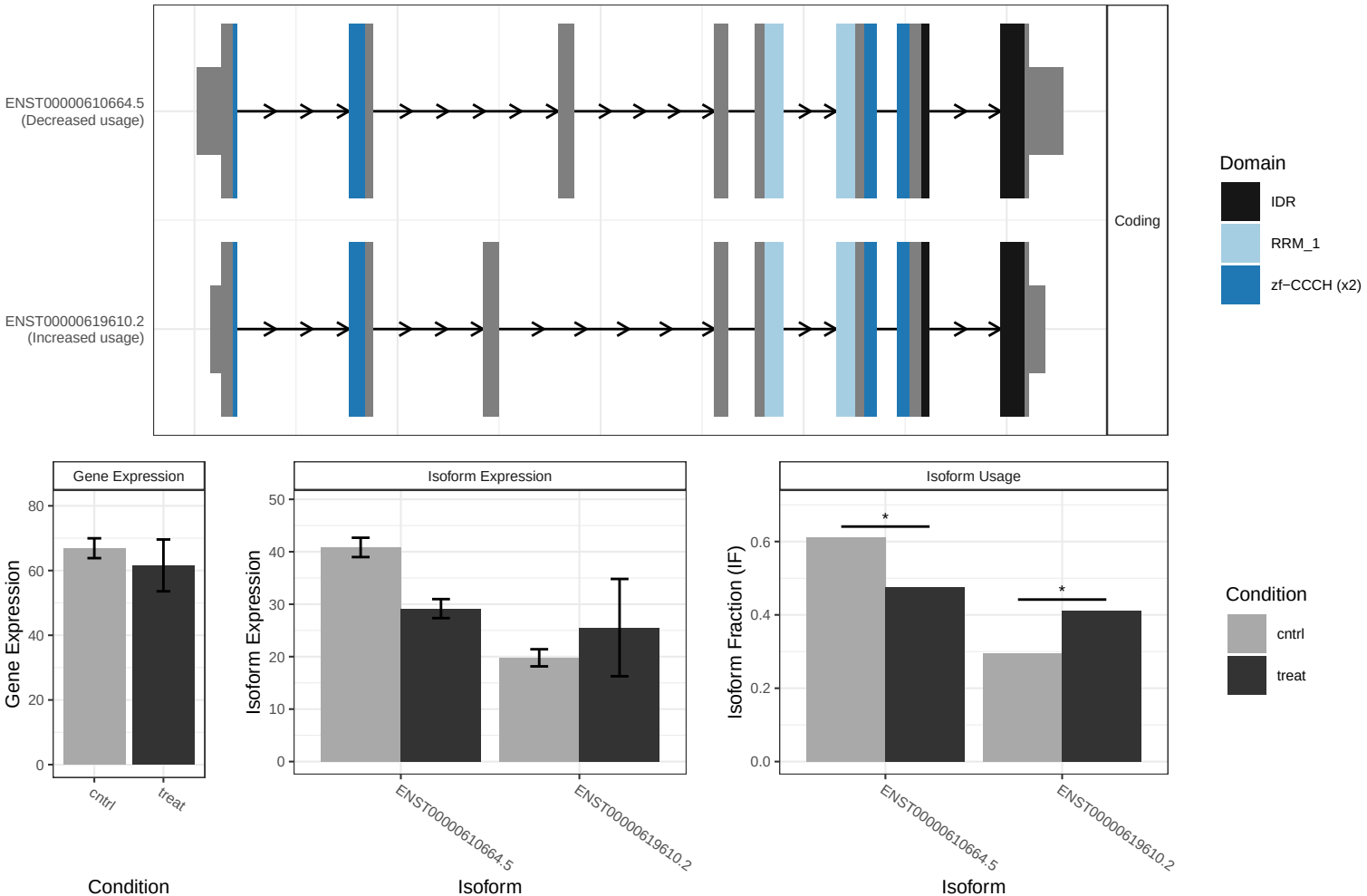
The isoform switch in TGFB1 (cntrl vs treat)



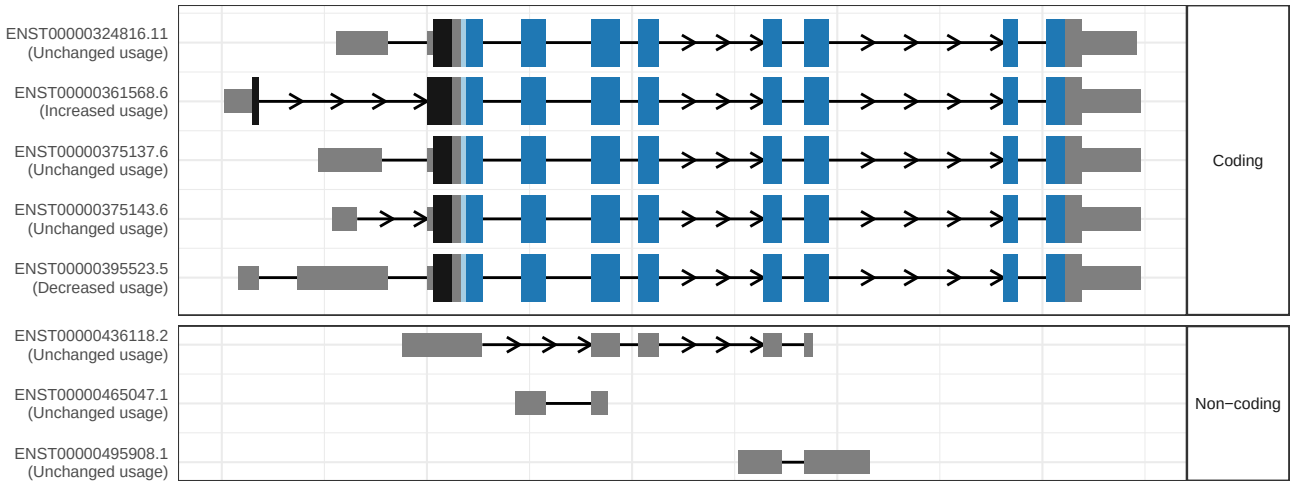
The isoform switch in TKT (cntrl vs treat)

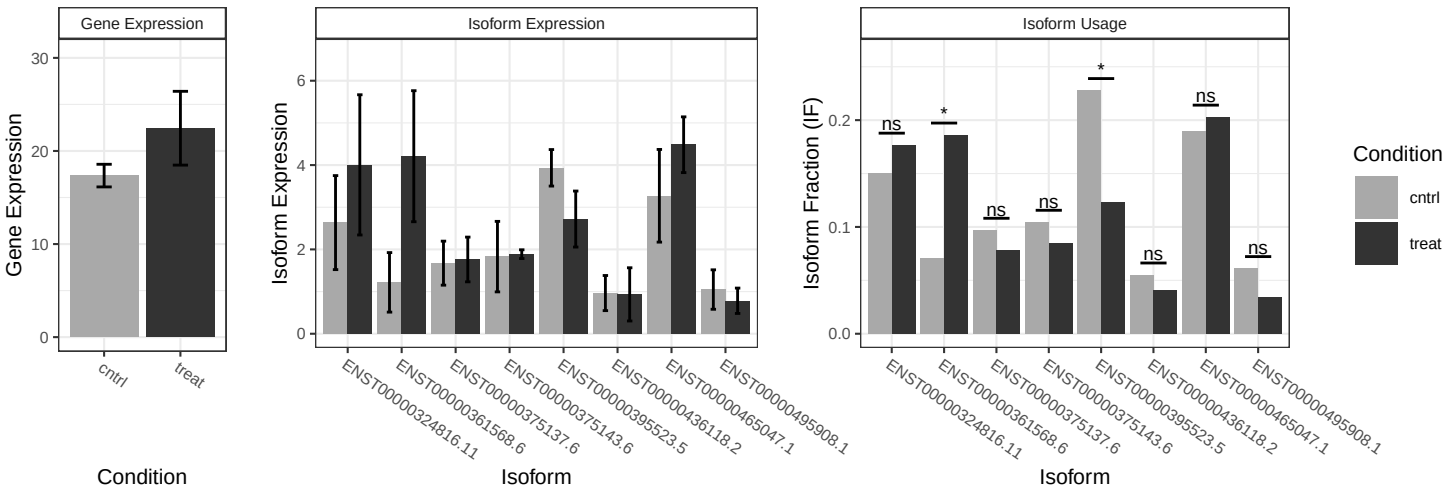


The isoform switch in U2AF1L5 (cntrl vs treat)

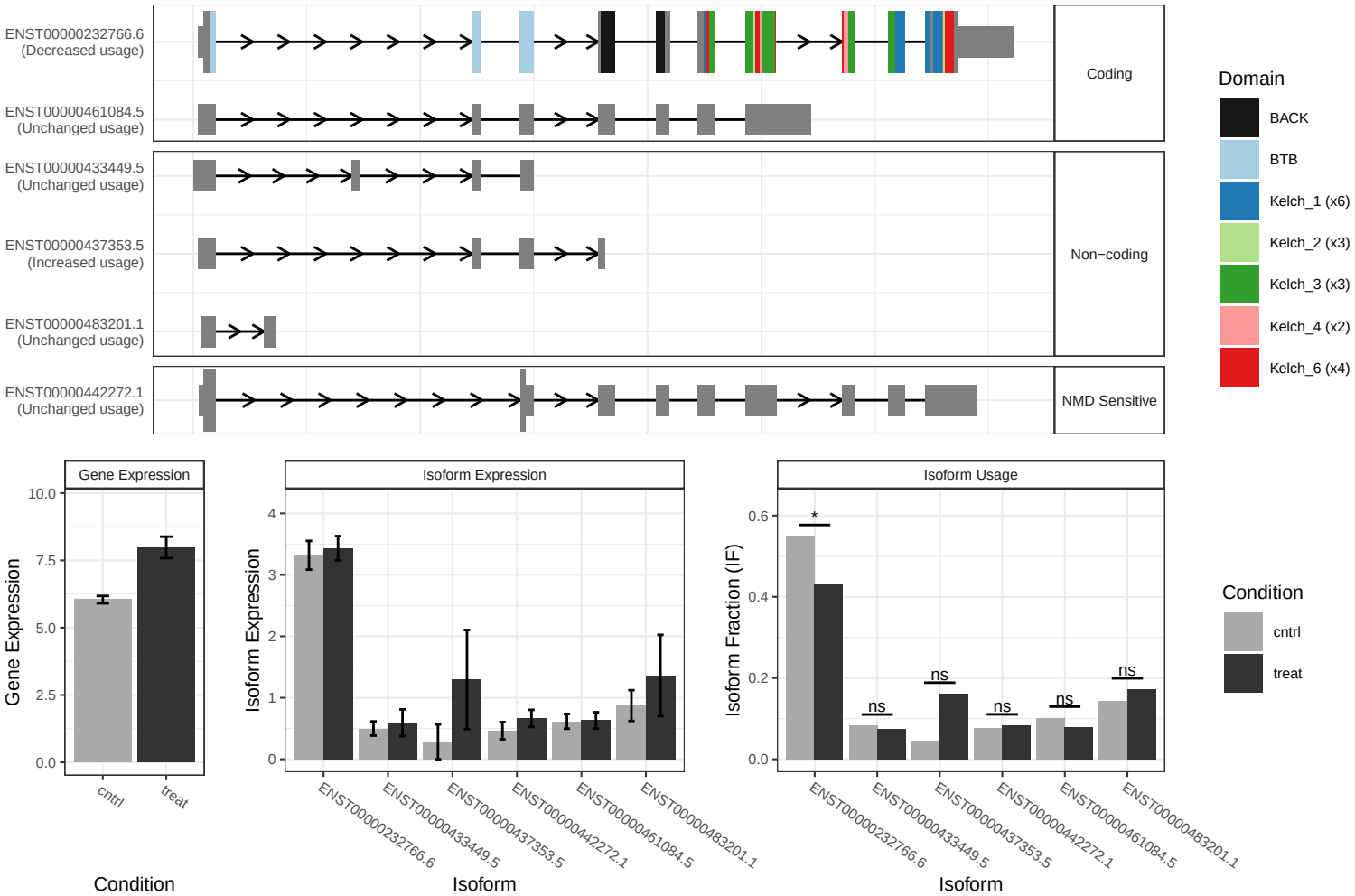


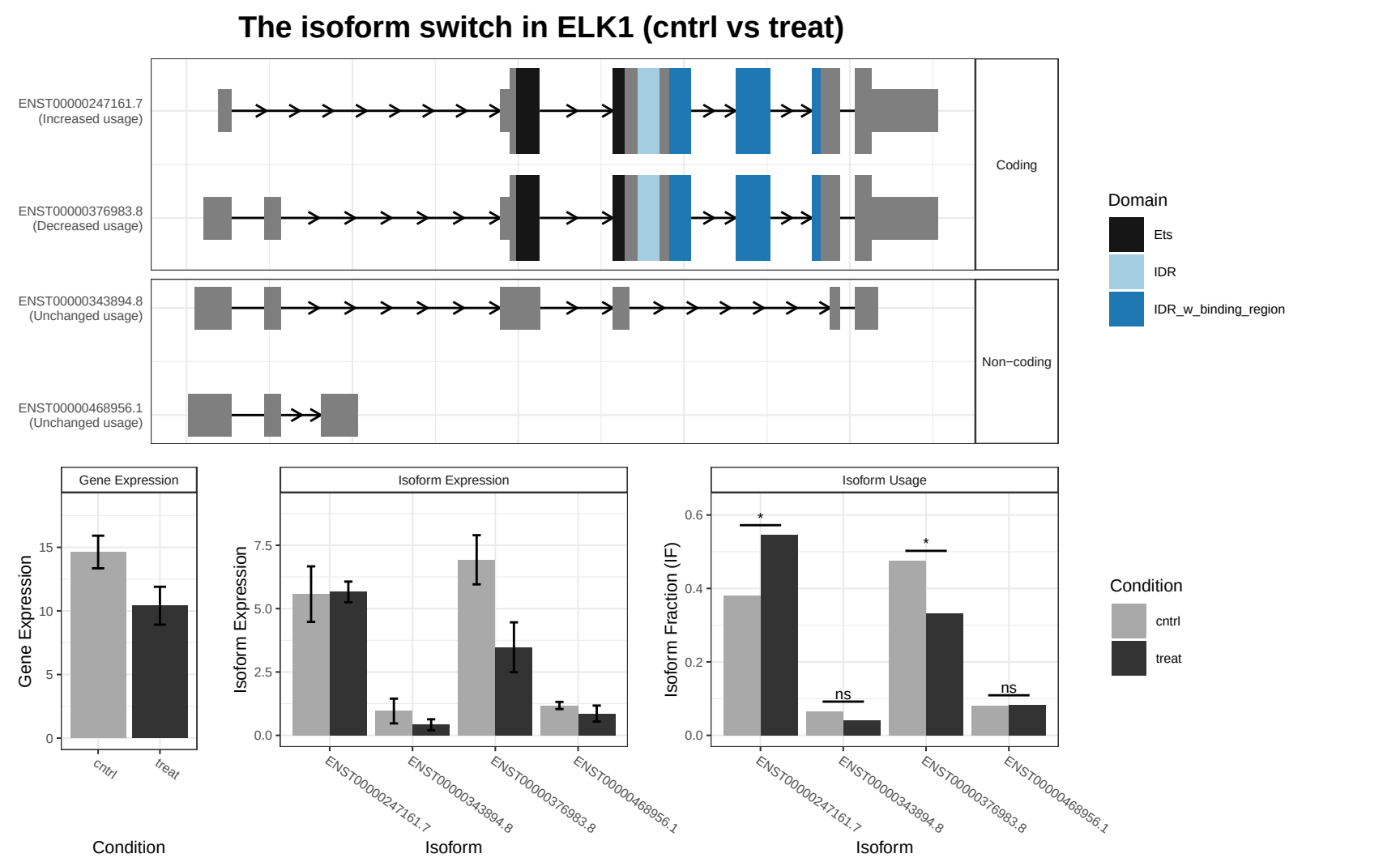
The isoform switch in PPT2 (cntrl vs treat)



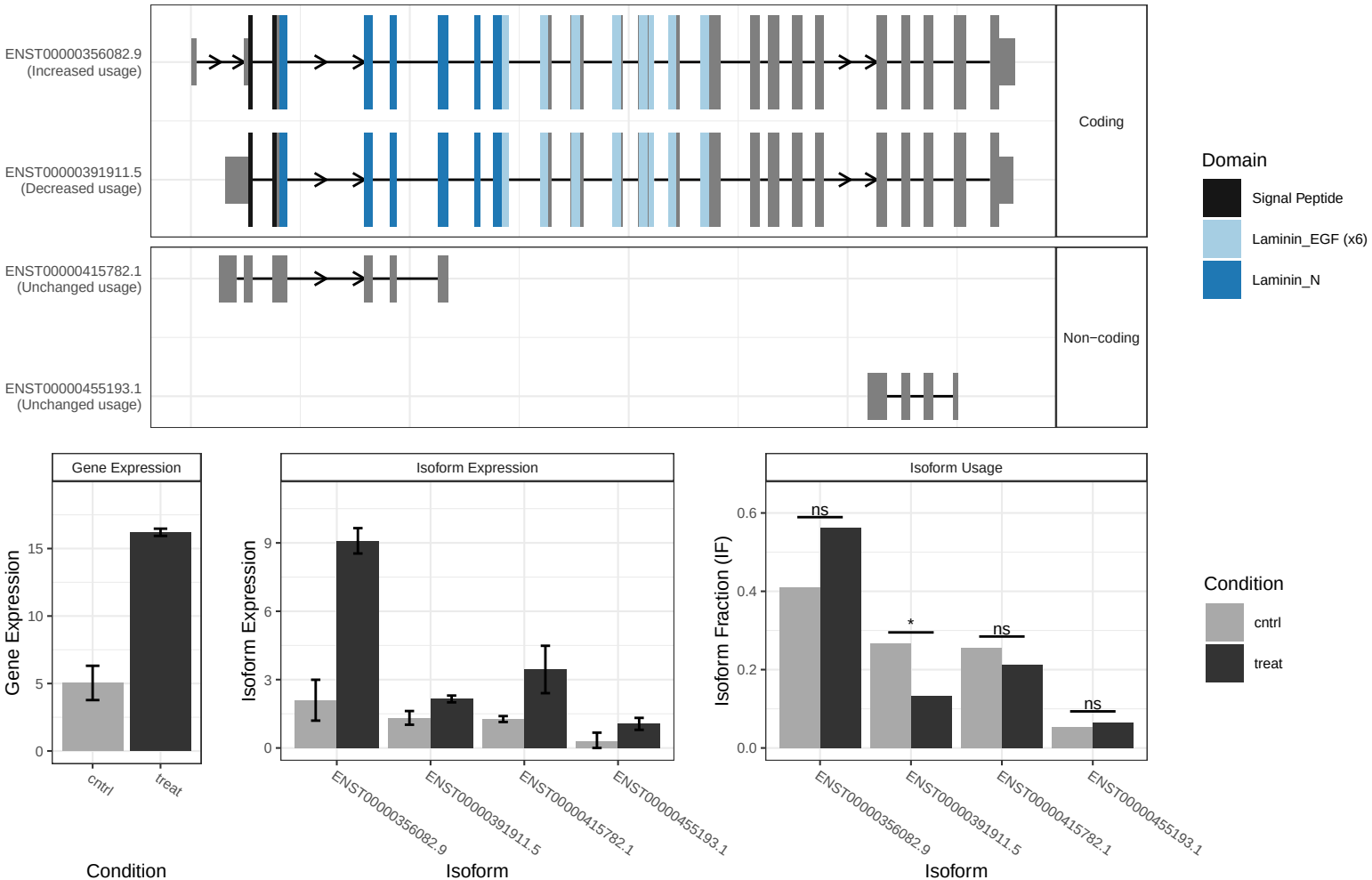


The isoform switch in KLHL18 (cntrl vs treat)

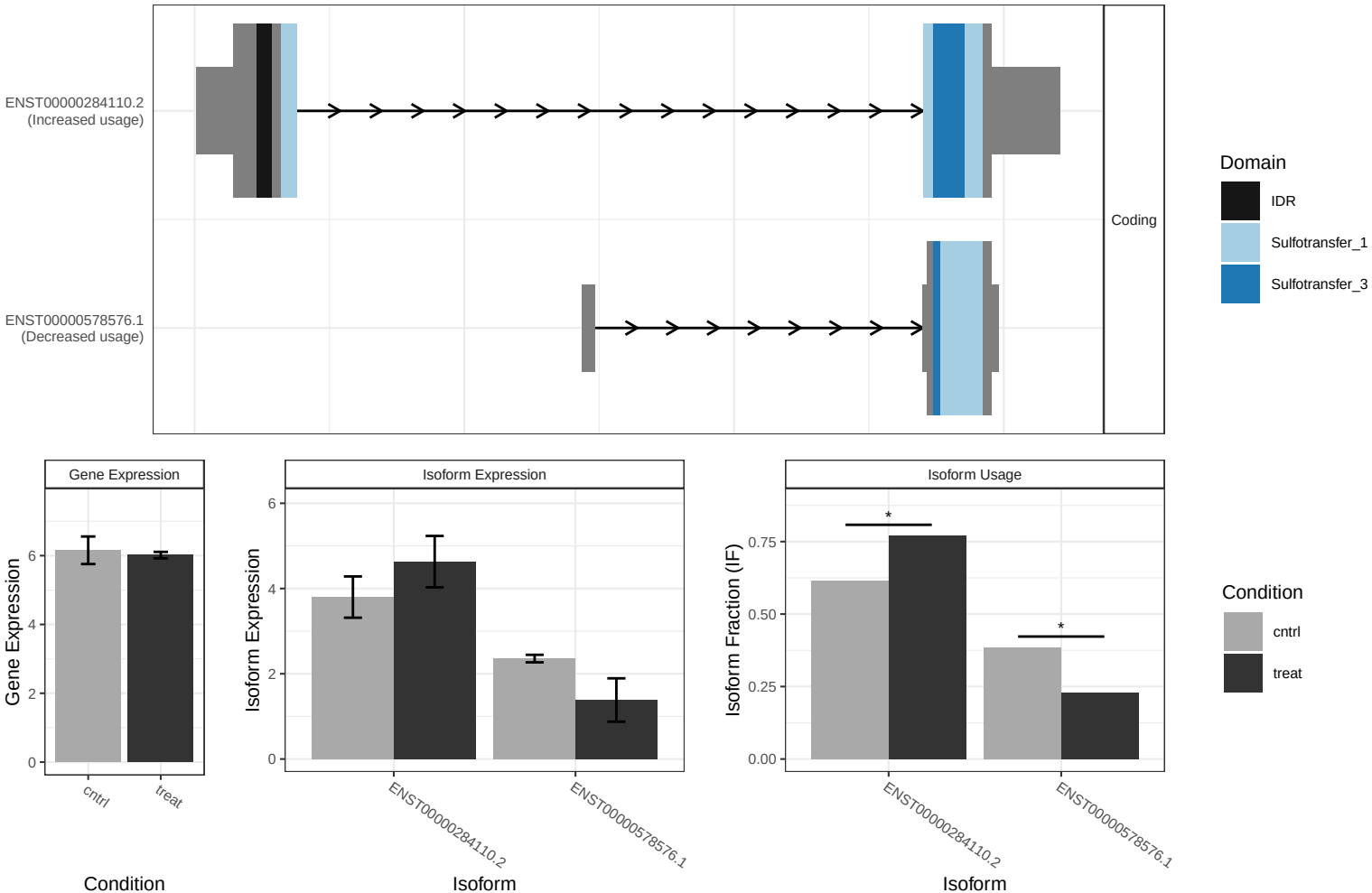



The isoform switch in ELK1 (cntrl vs treat)
ENST00000247161.7
(Increased usage)
ENST00000376983.8
(Decreased usage)
ENST00000343894.8
(Unchanged usage)
ENST00000468956.1
(Unchanged usage)
Coding
Non−coding
Domain
Ets
IDR
IDR_w_binding_region
Gene Expression
Gene Expression
15
10
5
0
cntrl
treat
Condition
Isoform Expression
Isoform Expression
7.5
5.0
2.5
0.0
ENST00000247161.7
ENST00000343894.8
ENST00000376983.8
ENST00000468956.1
Isoform
Isoform Usage
Isoform Fraction (IF)
0.6
0.4
0.2
0.0
*
ns
*
ns
ENST00000247161.7
ENST00000343894.8
ENST00000376983.8
ENST00000468956.1
Isoform
Condition
cntrl
treat

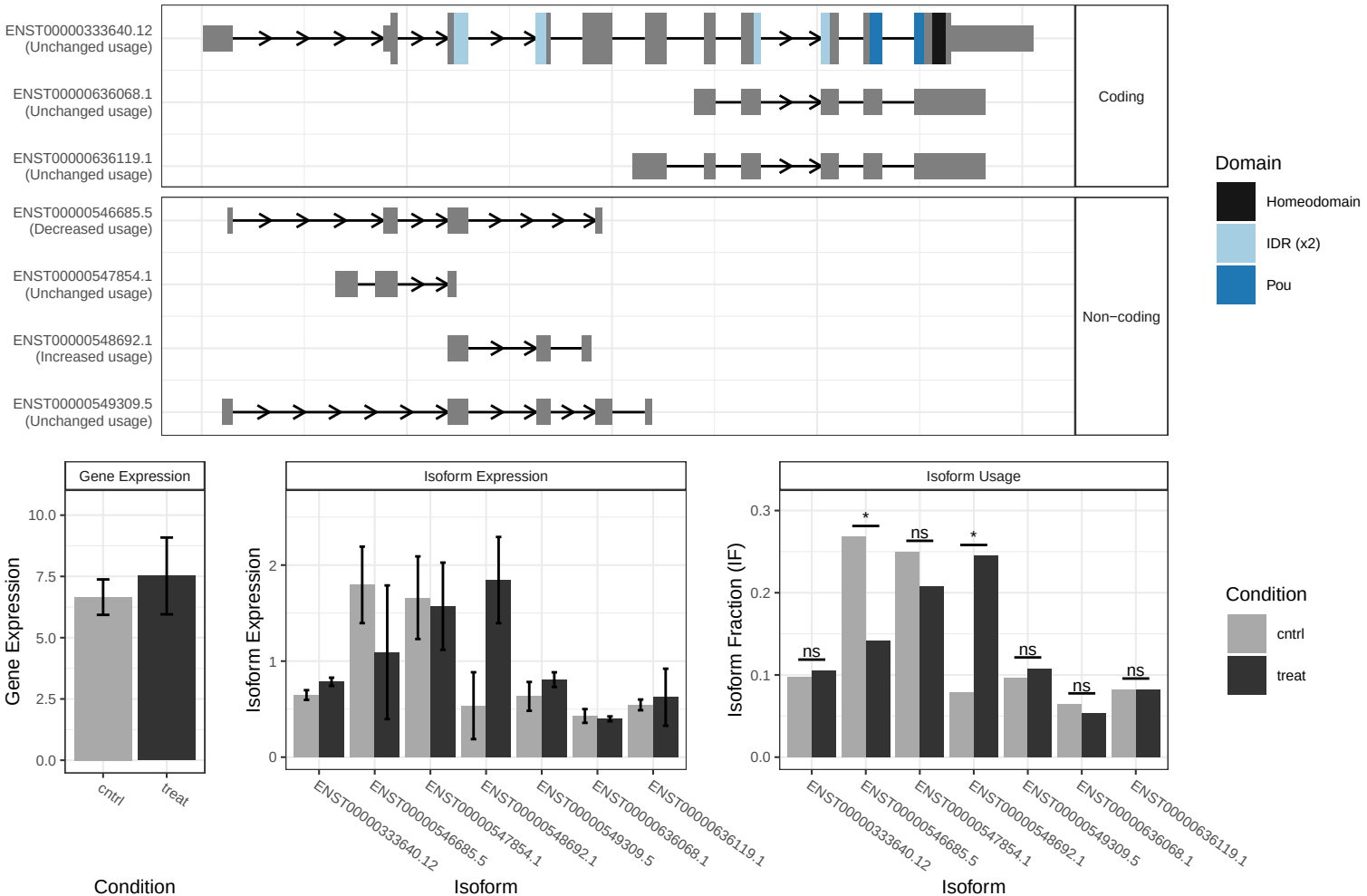
The isoform switch in LAMB3 (cntrl vs treat)



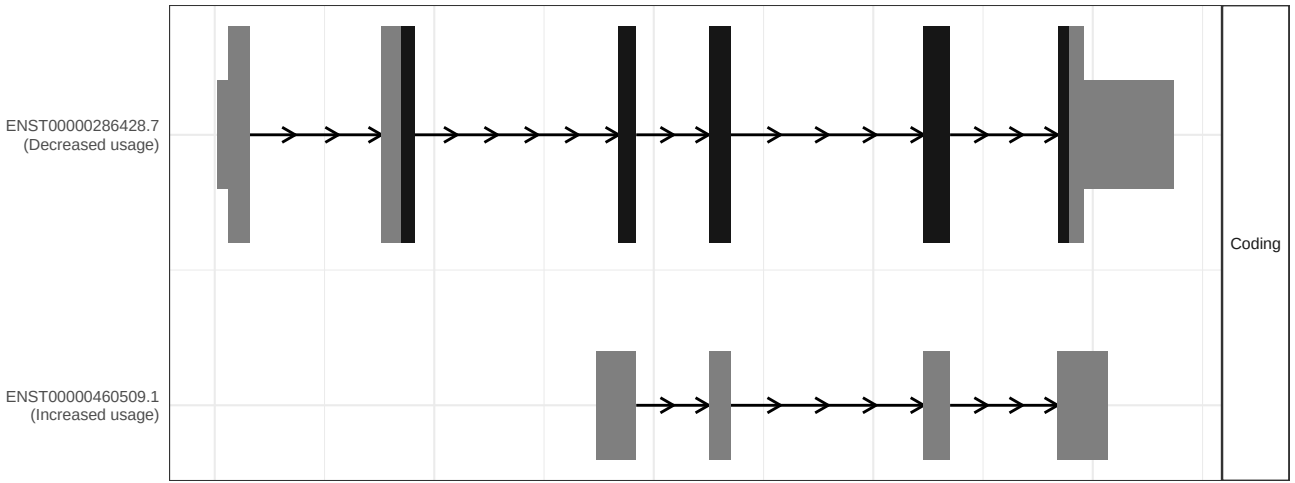
The isoform switch in HS3ST3A1 (cntrl vs treat)



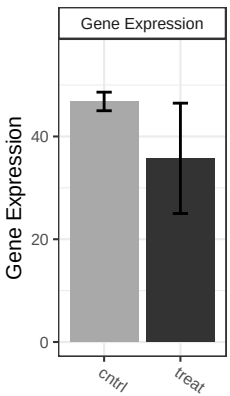
The isoform switch in POU6F1 (cntrl vs treat)

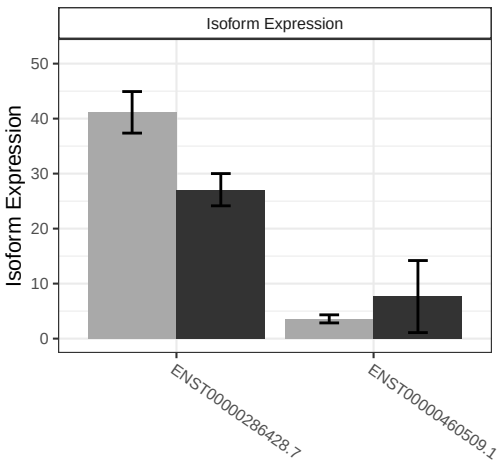


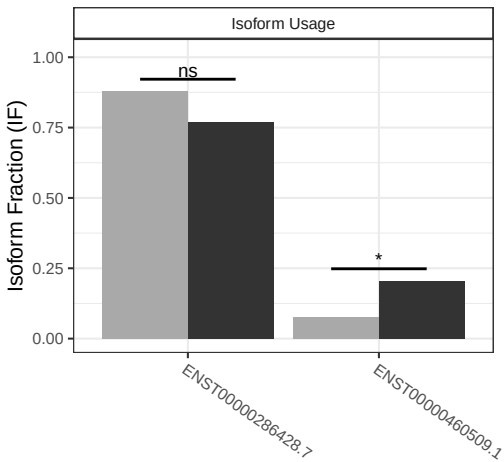
The isoform switch in VBP1 (cntrl vs treat)



Domain
Prefoldin

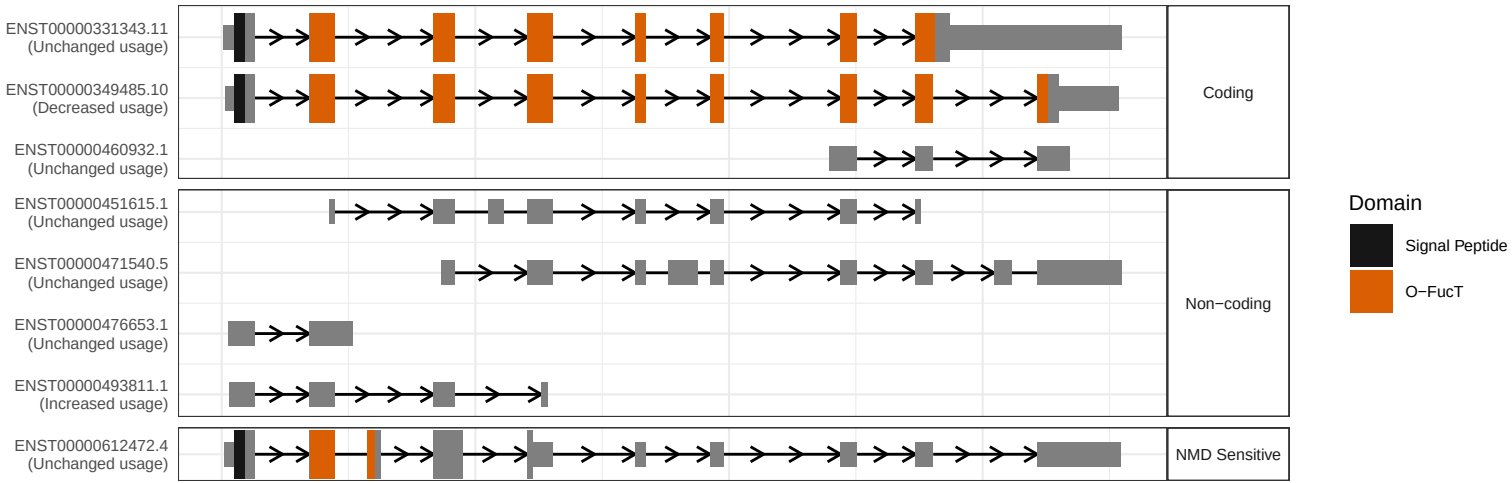


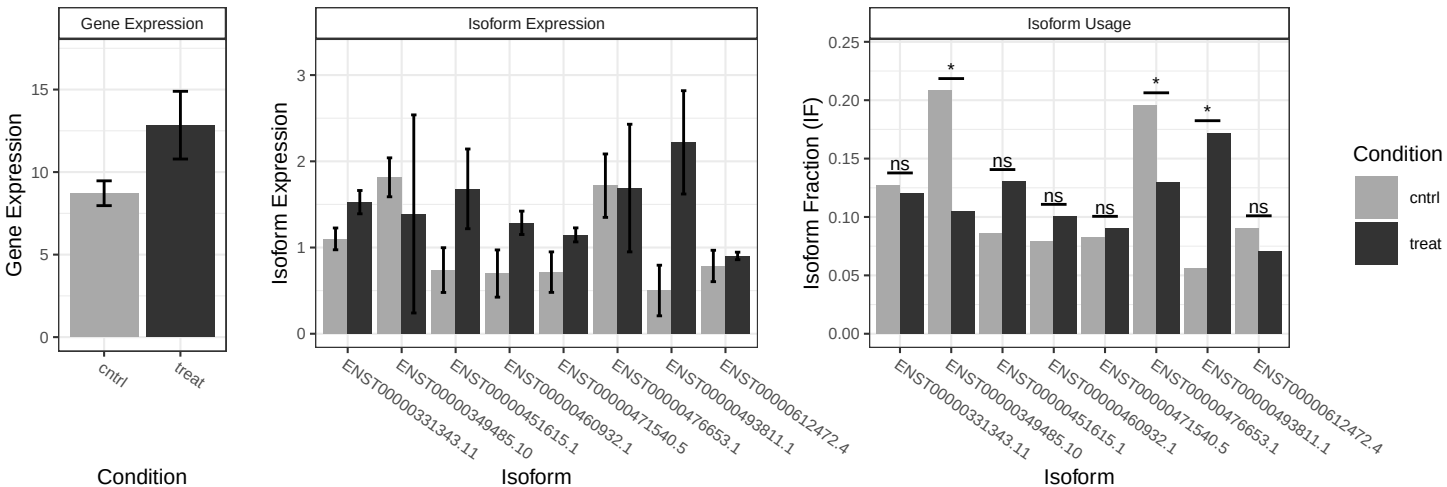




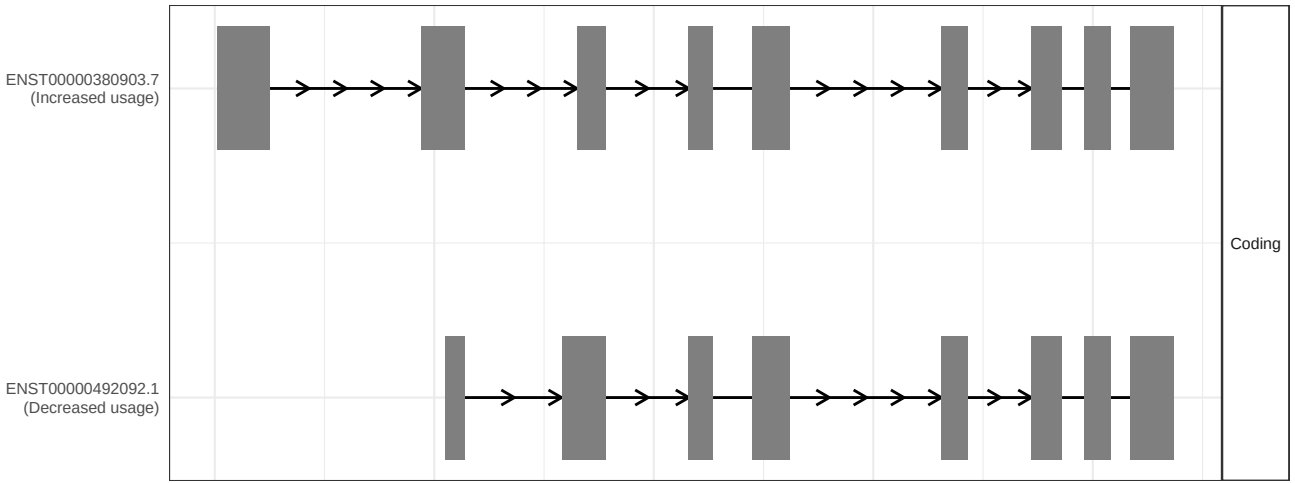
Condition
cntrl
treat

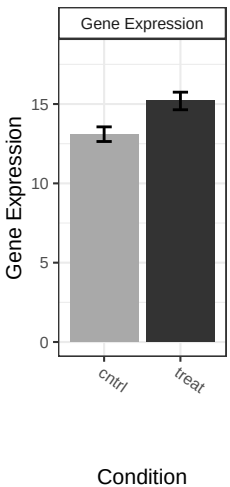
The isoform switch in POFUT2 (cntrl vs treat)

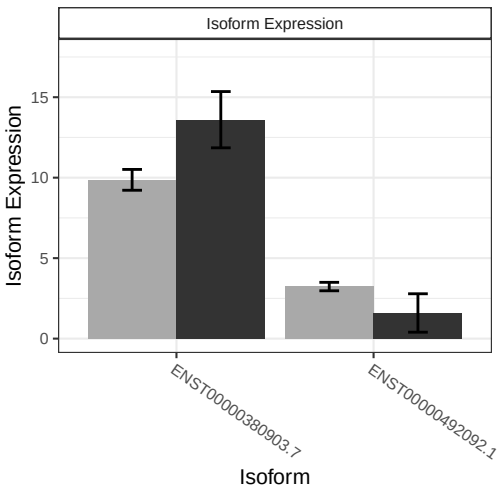


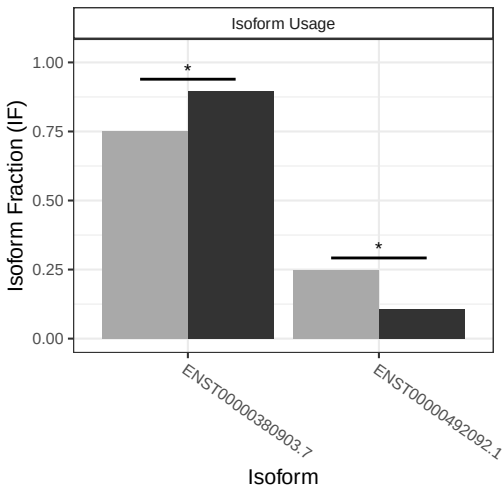


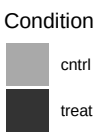
The isoform switch in SELENOO (cntrl vs treat)



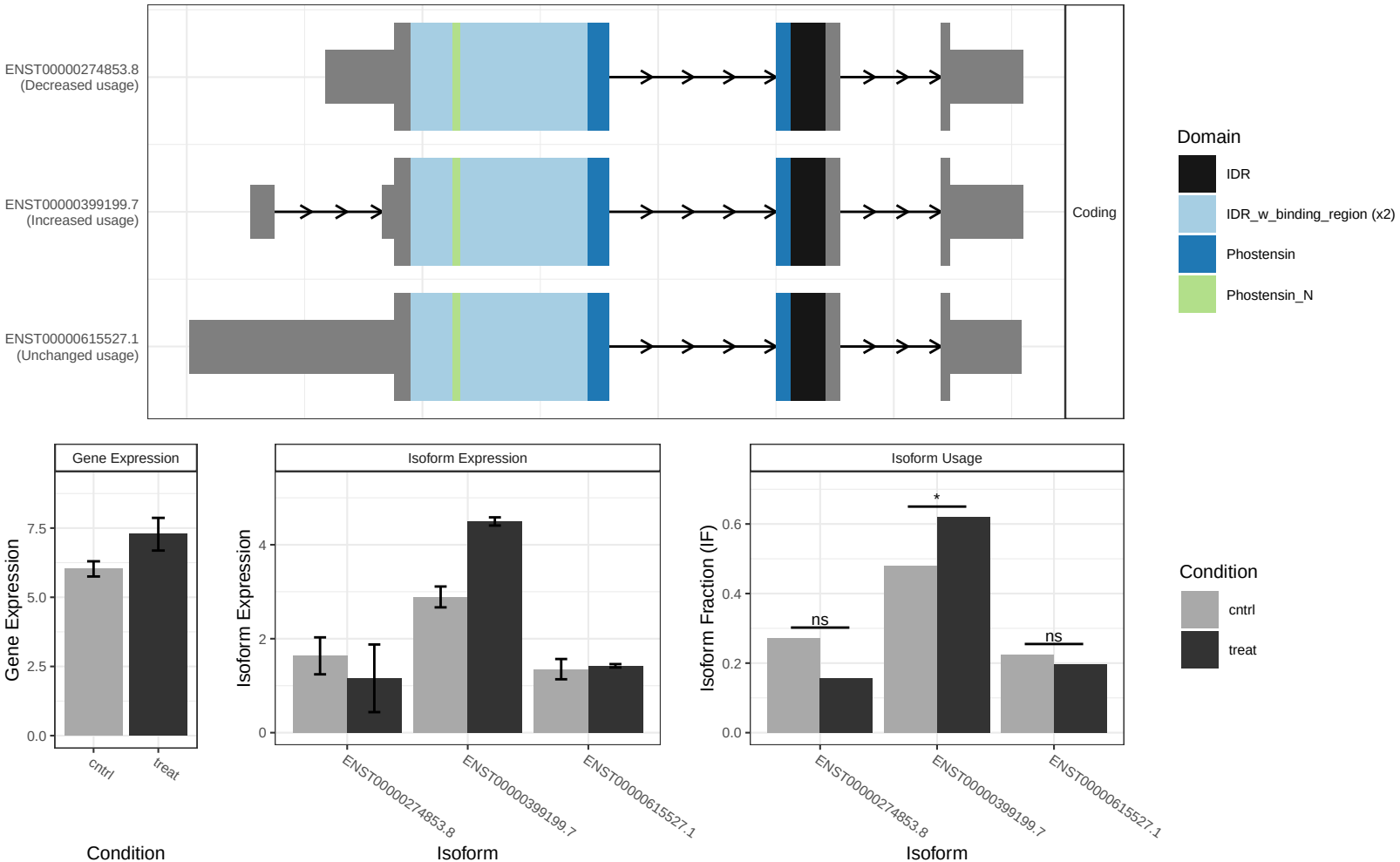








The isoform switch in PPP1R18 (cntrl vs treat)



APPENDIX B

```
#Set working directory
setwd("~/School/Spring 2022/Thesis Work/RStudio_misc/Gencode_020322")

#Load Needed Libraries
library(knitr)
library(IsoformSwitchAnalyzeR)
library(readr)
library(ggVennDiagram)
library(openxlsx)
library(patchwork)
#library(gt)

#Import Salmon data - rename file folders first
salmonQuant<-importIsoformExpression(parentDir = "SalmonOutput_Gencode/output")

#Making our data frame - gsub('_ <-underscore is where name splits
#Need to rename data files first to correct format; Ex. cntrl_07
myDesign <- data.frame(
  sampleID = colnames(salmonQuant$abundance)[-1],
  condition = gsub('_.*', '', colnames(salmonQuant$abundance)[-1])
)

#Importing Annotation and Fasta files
aSwitchList <- importRdata(
  isoformCountMatrix   = salmonQuant$counts,
  isoformRepExpression = salmonQuant$abundance,
  designMatrix         = myDesign,
  isoformExonAnnoation = "gencode.v39.chr_patch_hapl_scaff.annotation.gtf",
  isoformNtFasta       = "gencode.v39.transcripts.fa",
  fixStringTieAnnotationProblem = TRUE,
  showProgress = FALSE
)
  #55463 ( 22.74%) isoforms were removed since they were not expressed in any samples.
  #47307 genes_id were assigned their original gene_id instead of the StringTie gene_id.
  #The GUESSTIMATED number of genes with differential isoform usage are:
  #  comparison estimated_genes_with_dtu
  #1 cntrl vs treat                 26 - 43
  #The gene_ids or isoform_ids were not unique - we identified multiple instances of the
  #same gene_id/isoform_id on different chromosomes. To solve this we removed 58 gene_id.

#Filter through data to remove unwanted isoforms
aSwitchListFiltered<-
  preFilter(
    switchAnalyzeRlist = aSwitchList,
    geneExpressionCutoff = 5,
    isoformExpressionCutoff = 0,
```

```
  removeSingleIsoformGenes = TRUE
 )
 #The filtering removed 111312 ( 59.41% of ) transcripts. There is now 76042 isoforms left

#Running the DEXSeq Test on the filtered set
aSwitchListDEXSeqTest<-
 isoformSwitchTestDEXSeq(
  switchAnalyzeRlist = aSwitchListFiltered,
  reduceToSwitchingGenes = TRUE #need?
 )
 #Isoform switch analysis was performed for 8761 gene comparisons (100%).
 #Total runtime: 2.75 min

aSwitchListCombined<-
 isoformSwitchAnalysisCombined(
  switchAnalyzeRlist = aSwitchListDEXSeqTest,
  pathToGTF = "gencode.v39.chr_patch_hapl_scaff.annotation.gtf",
  n = Inf,
  pathToOutput = "Plots",
  outputPlots = FALSE  #change to output plots
 )
 #The number of isoform switches found were:
 #  Comparison nrIsoforms nrSwitches nrGenes
 #1 cntrl vs treat        116         75      65

 #The number of isoform switches with functional consequences identified were:
 #  Comparison nrIsoforms nrSwitches nrGenes
 #1 cntrl vs treat         75         50      43

 #Warning messages:
 #  1: In extractSequence(switchAnalyzeRlist = switchAnalyzeRlist,  ... :
 #     There were 17 isoforms where the amino acid sequence had a
 #     stop codon before the annotated stop codon. These was removed.

#Splicing Enrichment Analysis
#This tells me if splicing is significantly different across treatment groups
SpliceEnrichment<-extractSplicingEnrichment(
 aSwitchListCombined,
 returnResult = TRUE
)

#write.csv(SpliceEnrichment,'F_SpliceEnrichment.csv')

#For making Gene names into a Frequency Table - Filtered
F_ASAnalysis<-analyzeAlternativeSplicing(aSwitchListCombined)
IsoformFeatures<-F_ASAnalysis$isoformFeatures$gene_id
```

```
Freq_Table<-as.data.frame((table(IsoformFeatures)))
Genes_regardless<-as.data.frame(Freq_Table$IsoformFeatures)
colnames(Genes_regardless)<-c("gene_id")
#write.csv(Freq_Table, 'Freq_Table_total.csv')
#write.csv(aSwitchListCombined$isoformFeatures, "IsoformFeatures.csv")

#CuffDiff_freqTable<-
as.data.frame((table(CuffDiff_genes_sorted_by_expression_sig_OUTLIER$gene_name)))
#write.csv(CuffDiff_freqTable, 'CuffDiff_freqTable.csv')

#CuffDiff_ALL_freqTable<-as.data.frame((table(CuffDiff_ALL_2102022$gene_name)))
#write.csv(CuffDiff_ALL_freqTable, 'CuffDiff_ALL_freqTable.csv')

#Sorting through table of genes - need to import excel sheet made from .csv file
#IsoformFeatures_wConsequences<-IsoformFeatures_consequences$gene_id
#Freq_Table_wConsequences<-as.data.frame((table(IsoformFeatures_wConsequences)))
#write.csv(Freq_Table_wConsequences, 'Freq_Table_wConsequences.csv')

#Plots - Export as pdf, change to US Letter size
extractSplicingSummary(aSwitchListCombined)
SplicingGenomeWide<-extractSplicingGenomeWide(aSwitchListCombined)

#This outputs all plots, regardless of consequence
switchPlotTopSwitches(
  switchAnalyzeRlist = aSwitchListCombined,
  n = Inf,   # Set to Inf for all
  filterForConsequences = FALSE,
  fileType = "pdf",
  pathToOutput = "Total_Plots",
  splitComparison = FALSE,
  splitFunctionalConsequences = FALSE
)

#Returns list of Top Switches
TopSwitches<-extractTopSwitches(
  aSwitchListCombined,
  filterForConsequences=FALSE,
  extractGenes=TRUE,
  alpha=0.05,
  dIFcutoff = 0.1,
  n=Inf,
  inEachComparison=FALSE,
  sortByQvals=TRUE
)

#write(TopSwitchList, "TopSwitchList", sep = "\t") #writes gene list as tab deliminated file
```

```
#NoPlotList<-read.csv("Total_vs_wConsequences_ListEDIT.txt") #import txt file

#Don't Need ~ V V V V
#library(plyr)

#SNF Target Gene List (Euskirchen 2011) VS Gene List Regardless of Consequences
library(readr)
#Import SNF Target List
Euskirchen_genefreq <- read_csv("~/School/Spring 2022/Thesis
Work/Euskirchen_genefreq.csv")
SNF_Targets<-Euskirchen_genefreq[,2]
#write.csv(SNF_Targets,'SNF_Targets_List.csv')

#Trying to make fancy Venn Diagram
#if (!require(devtools)) install.packages("devtools")
#devtools::install_github("gaospecial/ggVennDiagram")
library("ggVennDiagram")
SNF_Targets_a<-as.character(SNF_Targets$gene_id)
Genes_regardless_a<-as.character(Genes_regardless$gene_id)
x<-list(SNF_Targets_a,Genes_regardless_a)

#SNF Targets vs AS Genes (1)
VennDiagram<-ggVennDiagram(
  x,
  category.names  = c(" SNF Targets","AS Genes"),
  label = c("count"),
  #set_color = c("chartreuse4","turquoise4") #Change label name colors
  ) + theme(legend.position = "none") + scale_color_brewer(palette = "Dark2")

SNF_vs_AS<-VennDiagram[["plot_env"]][["data"]]@region[["item"]][[3]] #list of genes in both
pile

#DEG vs AS Genes (2)
DEGList<-CuffDiff_ALL_freqTable[-c(1),]
DEGList<-DEGList[,2]
DEG<-as.character(DEGList$Var1)
Genes_regardless_a<-as.character(Genes_regardless$gene_id)
y<-list(DEG,Genes_regardless_a)

VennDiagram2<-ggVennDiagram(
  y,
  category.names  = c(" DEG","AS Genes"),
  label = c("count"),
  #set_color = c("chartreuse4","turquoise4") #Change label name colors
) + theme(legend.position = "none") #+ scale_color_brewer(palette = "Dark2")
```

```
DEG_vs_AS<-VennDiagram2[["plot_env"]][["data"]]@region[["item"]][[3]] #list of genes in both pile

#SNF Targets VS DEG (3)
z<-list(SNF_Targets_a,DEG)

VennDiagram3<-ggVennDiagram(
  z,
  category.names  = c(" SNF Targets","DEG"),
  label = c("count"),
) + theme(legend.position = "none")

SNF_vs_DEG<-VennDiagram3[["plot_env"]][["data"]]@region[["item"]][[3]] #list of genes in both pile

#SNF Targets VS DEGwFC (4)
#Import List of DEG with Fold Change >=2.0
DEGwFC<-read.csv('CuffDiff_freqTable.csv')
DEGwFC<-DEGwFC[-c(1),]
DEGwFC<-DEGwFC[,2]
zb<-list(SNF_Targets_a,DEGwFC)

VennDiagram4<-ggVennDiagram(
  zb,
  category.names  = c(" SNF Targets","DEG (FC>=2.0)"),
  label = c("count"),
) + theme(legend.position = "none")

SNF_vs_DEGwFC<-VennDiagram4[["plot_env"]][["data"]]@region[["item"]][[3]] #list of genes in both pile

#DEG VS DEGwFC (5)
zc<-list(DEG,DEGwFC)

VennDiagram5<-ggVennDiagram(
  zc,
  category.names  = c(" DEG","DEG (FC>=2.0)"),
  label = c("count"),
) + theme(legend.position = "none")

DEG_vs_DEGwFC<-VennDiagram5[["plot_env"]][["data"]]@region[["item"]][[3]] #list of genes in both pile

#Creating Master List of Gene Piles
GenePiles <-list(
  "SNF Targets vs AS Genes"=SNF_vs_AS,
```

```
  "DEG vs AS Genes"=DEG_vs_AS,
  "SNF Targets VS DEG"=SNF_vs_DEG,
  "SNF Targets VS DEGwFC"=SNF_vs_DEGwFC,
  "DEG VS DEGwFC"=DEG_vs_DEGwFC)

#Writing Excel Files
#install.packages("openxlsx", dependencies = TRUE)
library(openxlsx)
#Needs list of data frames
'SNF Targets vs AS Genes'<-as.data.frame(SNF_vs_AS)
'DEG vs AS Genes'<-as.data.frame(DEG_vs_AS)
'SNF Targets VS DEG'<-as.data.frame(SNF_vs_DEG)
'SNF Targets VS DEGwFC'<-as.data.frame(SNF_vs_DEGwFC)
'DEG VS DEGwFC'<-as.data.frame(DEG_vs_DEGwFC)

GenePiles_DF<-list(
  `SNF Targets vs AS Genes`,
  `DEG vs AS Genes`,
  `SNF Targets VS DEG`,
  `SNF Targets VS DEGwFC`,
  `DEG VS DEGwFC`)

write.xlsx(GenePiles_DF, file = "GenePiles.xlsx") #Excel Sheet with tabs

#Exon Info testing
test<-as.data.frame(aSwitchListCombined$exons) #outputs table with potential exon data
exons_freqTable<-as.data.frame((table(test$gene_id)))
write.xlsx(exons, file = 'exons.xlsx')
#Trying to edit switch plots to include exon sizes and/or x-axis
switchPlotTranscript(
  aSwitchListCombined,
  gene = 'HS3ST3A1', #choose a gene in the list
  plotXaxis = TRUE, #adds x axis labels in
  rescaleTranscripts = TRUE,) #rescales exons and introns to relative sizes

#Adding External Functional Analysis
#Part 1 - Filtered
aSwitchListpt1 <- isoformSwitchAnalysisPart1(
  switchAnalyzeRlist   = aSwitchListDEXSeqTest,
  pathToOutput = "Adding_Functional_Analysis",
  outputSequences      = FALSE, #change to TRUE to get sequences
  prepareForWebServers = FALSE #outputs in subsets
)

    #Step 1 of 3 : Detecting isoform switches...
    #Step 3 of 3 : Extracting (and outputting) sequences
```

```
   #The 'removeLongAAseq' and 'removeShortAAseq' arguments:
   #  Removed : 1 isoforms.
   #Trimmed : 6 isoforms (to only contain the first 1000 AA)
   #The 'alsoSplitFastaFile' caused 1 fasta files, each with a subset of the data,
   #to be created (each named X of Y).

   #The number of isoform switches found were:
   #  Comparison nrIsoforms nrSwitches nrGenes
   #1 cntrl vs treat        116         75      65

#External Analysis Servers \/ \/ \/ \/ \/
#### BATCH - Nt file: http://cpc2.gao-lab.org/batch.php  Ignore 'Genome Assembly Version"
####PFAM - aa file: https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan   Need to input email
####SignalP - aa file:http://www.cbs.dtu.dk/services/SignalP/   Download 'Prediction Summary"
####IuPRED - aa file: https://iupred2a.elte.hu   Input email

aSwitchListpt2 <- isoformSwitchAnalysisPart2(
  switchAnalyzeRlist        = aSwitchListpt1,
  #dIFcutoff                 = 0.3,   #default is 0.1
  n                         = Inf,
  removeNoncodinORFs        = TRUE,
  pathToCPC2resultFile      = "Adding_Functional_Analysis/result_cpc2.txt",
  pathToPFAMresultFile      = "Adding_Functional_Analysis/result_pfam.txt",
  pathToIUPred2AresultFile  = "Adding_Functional_Analysis/result_iupred.txt",
  pathToSignalPresultFile   = "Adding_Functional_Analysis/result_signalP.txt",
  outputPlots               = FALSE, #change to output
  pathToOutput              = "Adding_Functional_Analysis/Plots_wConsequences",
)

   #The number of isoform switches with functional consequences identified were:
   #  Comparison nrIsoforms nrSwitches nrGenes
   #1 cntrl vs treat         75         49      42
   #The switch analysis plot for each of these, as well as a plot summarizing the functional
   #consequences have been outputted to the folder specified by 'pathToOutput'.
   #There were 50 or more warnings (use warnings() to see the first 50)

switchPlotTopSwitches(
  switchAnalyzeRlist = aSwitchListpt2,
  n = Inf,   # Set to Inf for all
  filterForConsequences = FALSE,
  fileType = "pdf",
  pathToOutput = "Adding_Functional_Analysis/All_plots",
  splitComparison = FALSE,
  splitFunctionalConsequences = FALSE
)
```

```
   #Made 65 plots of genes with isoform switching

#Changing 'consequences to analyze'
aSwitchListpt2_REDO <- isoformSwitchAnalysisPart2(
  switchAnalyzeRlist       = aSwitchListpt1,
  #dIFcutoff                = 0.3,  #default is 0.1
  n                        = Inf,
  removeNoncodinORFs       = TRUE,
  pathToCPC2resultFile     = "Adding_Functional_Analysis/result_cpc2.txt",
  pathToPFAMresultFile     = "Adding_Functional_Analysis/result_pfam.txt",
  pathToIUPred2AresultFile  = "Adding_Functional_Analysis/result_iupred.txt",
  pathToSignalPresultFile   = "Adding_Functional_Analysis/result_signalP.txt",
  outputPlots              = TRUE,
  pathToOutput             =
"Adding_Functional_Analysis/Plots_wConsequences/testing_consequences",
  consequencesToAnalyze = c(
    'tss',
    'tts',
    'exon_number',
    'intron_structure',
    'intron_retention',
    'coding_potential',
    'NMD_status',
    'domains_identified',
    'IDR_identified',
    'signal_peptide_identified'
  )
)

View(aSwitchListpt2_REDO[["switchConsequence"]])
write.xlsx(aSwitchListpt2_REDO[["switchConsequence"]], file =
"REDO_switchConsequences.xlsx")

REDO_ntCutOff<-analyzeSwitchConsequences(
  switchAnalyzeRlist = aSwitchListpt2,
  consequencesToAnalyze = c(
    'tss',
    'tts',
    'exon_number',
    'intron_structure',
    'intron_retention',
    'coding_potential',
    'NMD_status',
    'domains_identified',
    'IDR_identified',
    'signal_peptide_identified'
```

```
  ),
  ntCutoff = 1 #should register any change in nucleotide length greater than 1
)

write.xlsx(REDO_ntCutOff[["switchConsequence"]], file =
"REDO_ntCutOff_switchConsequences.xlsx")

#test for all consequences - block out ones that showed zero isoforms
test<-analyzeSwitchConsequences(
  aSwitchListpt2_REDO,
  ntCutoff = 1,
  consequencesToAnalyze = c(
    'tss',
    'tts',
    'exon_number',
    #'intron_structure',
    'intron_retention',
    'coding_potential',
    'NMD_status',
    'domains_identified',
    'IDR_identified',
    'signal_peptide_identified',
    'last_exon',
    'isoform_seq_similarity',
    'isoform_length',
    'isoform_class_code',
    'ORF_seq_similarity',
    #'ORF_genomic',
    'ORF_length',
    '5_utr_seq_similarity',
    '5_utr_length',
    '3_utr_seq_similarity',
    '3_utr_length',
    'domain_length'
    #'genomic_domain_position',
    #'IDR_length',
    #'IDR_type'
  )
)
extractConsequenceSummary(test)
test_more<-extractSplicingSummary(test, returnResult = TRUE)
write.xlsx(test_more,file = 'test_more.xlsx')

write.xlsx(test[["AlternativeSplicingAnalysis"]], file = "test_ASanalysis.xlsx")
write.xlsx(test[["switchConsequence"]], file = "test_switchConsequences.xlsx")
```

```
switchPlotTopSwitches(
  test,
  n = Inf,
  filterForConsequences = TRUE,
  fileType = "pdf",
  pathToOutput = "Adding_Functional_Analysis/test",
)

write.xlsx(test$isoformFeatures, file = 'test_isoformfeatures.xlsx')

#Barplot of Switching Isoform Features
sums<-c(
  sum(S4_GenePiles$ATSS),
  sum(S4_GenePiles$ATTS),
  sum(S4_GenePiles$ES),
  sum(S4_GenePiles$IR),
  sum(S4_GenePiles$`A5'SS`),
  sum(S4_GenePiles$`A3'SS`)
  #sum(S4_GenePiles$`Domain loss`),
  #sum(S4_GenePiles$`Domain gain`)
  )

sums_labels<-c(
  "ATSS",
  "ATTS",
  "ES",
  "IR",
  "A5'SS",
  "A3'SS"
 #"Domain loss",
  #"Domain gain"
)

#barplot(
#  sums,
#  names.arg = sums_labels,
#  #legend.text = "test"
#  xlab = "Splicing Event Type",
#  ylab = "# Events",
#  #axes = TRUE
#  #xpd = TRUE
#  main = "Number of Splicing Event Types per Gene",
#  ylim=c(0,80),
#  yaxt = "n"
#  )
#axis(2, seq(0,81,10))
```

```
#Using ggplot to make it better
sums_df<-data.frame(sums)
sums_AS_SNF_df<-data.frame(sums)
windowsFonts(Times=windowsFont("TT Times New Roman"))

#the 65 pile
ggplot(
  sums_df,
  aes(x=sums_labels, y=sums, fill=sums_labels)) +
  geom_bar(stat = "identity") +
  labs(x="Consequence Type",y="Consequence Frequency")+
  ylim(0,58)+
  theme(legend.position = "none")

#SNF Target vs AS
#ggplot(
#  sums_AS_SNF_df,
#  aes(x=sums_labels, y=sums, fill=sums_labels)) +
#  geom_bar(stat = "identity") +
#  labs(x="Splicing Event Type",y="# Events",title = "Number of Splicing Event Types per
Gene")+
#  ylim(0,76)+
#  theme(legend.position = "none")

# create a dataset
#"Switch Consequence"<-c(rep("ATSS", 2),rep("ATTS", 2),rep("ES", 2),rep("IR", 2),
#           rep("A5'ss", 2),rep("A3'ss", 2),rep("Domain loss", 2),rep("Domain gain", 2) )
#"Gene List"<-rep(c("AS","AS vs SNF Targets"),8)
#"Consequence Frequency"<-c(37,58,32,48,21,36,7,9,3,6,7,12,65,76,68,74)
#data<-data.frame(`Switch Consequence`, `Gene List`,`Consequence Frequency`)
#
# Grouped
#ggplot(data, aes(fill=`Gene List`, y=`Consequence Frequency`, x=`Switch Consequence`)) +
#  geom_bar(position="dodge", stat="identity")

#Two proportion z test for AS events by abnormal exon size - from switching transcripts
(counted by hand)
#27 AS events for 141 abnormal size exons
#37 AS events for 512 average size exons
prop.test(x=c(26,37),n=c(141,474))

       #2-sample test for equality of proportions with continuity correction

      #data:  c(26, 37) out of c(141, 474)
      #X-squared = 12.234, df = 1, p-value = 0.0004694
```

```
    #alternative hypothesis: two.sided
    #95 percent confidence interval:
    #  0.03332185 0.17935433
    #sample estimates:
    #  prop 1     prop 2
    #0.18439716 0.07805907

#Figures and Tables - Final Constructions

#1 - Frequency of Different AS and TC Events in Response to SNF5 Inhibition (OUt of the 65)
  #Input data from Switch Plots (counted manually)
  AS65.counts<-c(36,9,6,12)
  TC65.counts<-c(58,48)
  AS65.labels<-c("ES","IR","A5'ss","A3'ss")
  TC65.labels<-c("ATSS","ATTS")
  AS65<-data.frame(AS65.labels,AS65.counts)
  TC65<-data.frame(TC65.labels,TC65.counts)

  #Craft AS Barplot
  AS.bar<-ggplot(AS65,aes(x=AS65.labels, y=AS65.counts, fill=AS65.labels))+
    geom_bar(stat = "identity") +
    labs(x="Splicing Event",y="Event Frequency",title = "A")+
    ylim(0,60)+
    theme(legend.position = "none")

  #Craft TC Barplot
  TC.bar<-ggplot(TC65,aes(x=TC65.labels, y=TC65.counts, fill=TC65.labels))+
    geom_bar(stat = "identity") +
    labs(x="Trancription Event",y=" ",title = "B")+
    ylim(0,60)+
    theme(legend.position = "none")

  #Combine into Single Image (Using Patchwork)
  Event.bars<-(AS.bar+TC.bar)

#R.Table1 - Events per Gene and SNF5 Target Status (Using GT)
#  #Import Data
#  R.Table1.data<-read.xlsx("R.Table1.EventsperGene.xlsx")
#
#  #Make Table
#  gt(R.Table1.data) %>% #%>% means put left side into right
#  opt_table_outline() %>% #adds outline to table

#4 - Venn Diagram of SNF5 Targets vs AS(49) Genes
  #Data Inputs
  SNF5.Targets<-as.character(Euskirchen_genefreq$gene_id)
```

```
  AS.49<-read.xlsx("AS(49).xlsx") #import data
  AS<-as.character(AS.49$gene_id)
  AS.SNF5<-list(SNF5.Targets,AS)

  #Craft Venn Diagram
  SNF5.vs.AS<-ggVennDiagram(
    AS.SNF5,
    category.names  = c(" SNF5 Targets","AS Genes"),
    label = c("count"),
  ) + theme(legend.position = "none")

#5 - Venn Diagram of SNF5 Targets vs DEG
  #Data Inputs
  DEG.file<-read.csv("CuffDiff_ALL_freqTable.csv")
  DEG.list<-DEG.file$Var1
  SNF5.DEG<-list(SNF5.Targets,DEG.list)

  #Craft Venn Diagram
  SNF5.vs.DEG<-ggVennDiagram(
    SNF5.DEG,
    category.names  = c(" SNF5 Targets","DEG"),
    label = c("count"),
  ) + theme(legend.position = "none")

#6 - Venn Diagram of AS vs DEG
  #Data Inputs
  AS.DEG<-list(AS,DEG.list)

  #Craft Venn Diagram
  AS.vs.DEG<-ggVennDiagram(
    AS.DEG,
    category.names  = c(" AS","DEG"),
    label = c("count"),
  ) + theme(legend.position = "none")

#7 - Can we do three comparisons? - Yes! AS.vs.DEG.vs.SNF5
  #Data Inputs
  SNF5.AS.DEG.list<-list(SNF5.Targets,AS,DEG.list)

  #Craft Venn Diagram
  SNF5.AS.DEG<-ggVennDiagram(
    SNF5.AS.DEG.list,
    category.names  = c("   SNF5"," AS","DEG  "),
    label = c("count"),
  ) + theme(legend.position = "none")
```

```
##8 - DAVID Sideways Barplot (fun.....?!)
#  #Data Inputs - Using Functional Clustering Groups
#  pathway<-c("zinc finger/transcription","glycoproteins",
#            "extracellular matrix","cadherins and cell adhesion",
#            "development","matrix metalloprotein","protein binding",
#            "chromatin and nucleosome structure")
#  e.score<-c(4.88,3.56,3.15,2.92,2.29,2.08,1.60,1.49)
#  DAVID.info<-data.frame(pathway,e.score)
#  DAVID<-DAVID.info[order(-e.score),] #Sort by e.score (minus sign means descending order)
#
#  DAVID.factor<-factor(DAVID$pathway, levels = DAVID$pathway)
#  #factor so ggplot doesn't reorder my groups
#
#  DAVID.finalboss<-data.frame(DAVID.factor,e.score)
#
#  #Crafting Sideways Barplot
#  DAVID.bar<-ggplot(DAVID.finalboss,aes(x=e.score, y=DAVID.factor, fill=DAVID.factor))
+
#    geom_bar(stat = "identity")+
#    labs(x="Enrichment Score",y="Pathway",title = "DAVID Pathway Summary")+
#    theme(legend.position = "none")
#  DAVID.bar

##UPDATED DAVID Sideways Barplot
  #Data Inputs - Using Functional Clustering Groups
    pathway<-c("zinc finger/transcription","glycoproteins",
               "cancer pathways","cadherins and cell adhesion",
               "axon guidance","differentiation","metallopeptidase",
               "transcription/RNA pol II binding",
               "chromatin and nucleosome structure")
    e.score<-c(4.89,2.68,2.4,2.3,1.96,1.94,1.66,1.48,1.38)
    DAVID.info<-data.frame(pathway,e.score)
    DAVID<-DAVID.info[order(-e.score),] #Sort by e.score (minus sign means descending order)

    DAVID.factor<-factor(DAVID$pathway, levels = DAVID$pathway)
    #factor so ggplot doesn't reorder my groups

    DAVID.finalboss<-data.frame(DAVID.factor,e.score)

    #Crafting Sideways Barplot
    DAVID.bar<-ggplot(DAVID.finalboss,aes(x=e.score, y=DAVID.factor, fill=DAVID.factor))
+
      geom_bar(stat = "identity")+
      labs(x=" ",y="Pathway")+
      theme(legend.position = "none")
    DAVID.bar
```