

**DEVELOPING A BIOINFORMATICS PIPELINE TO ASSESS THE
POTENTIAL FUNCTIONAL IMPACT OF NOVEL PROTEIN ISOFORMS**

by

Alyssa Klein

B.S. Biology (Lebanon Valley College) 2017

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

In

BIOINFORMATICS

In the

GRADUATE SCHOOL

of


HOOD COLLEGE

May 2020

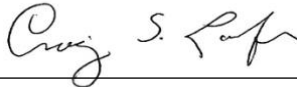
Accepted:



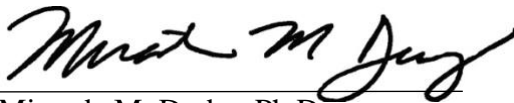
Georgette N. Jones, Ph.D.
Committee Member



April M. Boulton, Ph.D.
Dean of the Graduate School



Craig S. Laufer, Ph.D.
Committee Member



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

STATEMENT OF COPYRIGHT WAIVER

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

DEDICATION

I would like to dedicate this study to all scientists whose passion continues to drive discovery and increase the understanding of the world in which we live.

ACKNOWLEDGEMENTS

First and foremost I want to thank my family for being on my journey through school right alongside me. They were always willing to move plans around so I could finish schoolwork, or help me with making a meal because I didn't have the time. They have always been my biggest supporters and cheerleaders, and there are no words I can say that would encompass my appreciation for them.

I would also like to extend a thank you to Dr. Miranda Darby of Hood College. Without her, I would not have been able to complete my thesis. After my initial idea for my thesis imploded, she was there to help me go back to the drawing board and draft an idea that honed in on my interests. Dr. Darby has been there for countless vent sessions, and was always more than willing to listen when life didn't go as planned.

I would like to thank Dr. Craig Laufer and Dr. Georgette Jones for serving as members of my thesis committee. They were always more than willing to meet and discuss my project at various points along the way, offering suggestions to make the project successful.

Thank you also to the Hood College Biology Department, for allowing me to be a part of the department. It was a very warm and inviting department to be a member of, and everyone was always willing to lend a hand when the situation presented itself.

Finally, I wish to thank Dr. Robert Carey of Lebanon Valley College, who introduced me to the field of bioinformatics and computational biology through independent research for over half of my undergraduate career.

TABLE OF CONTENTS

	Page
ABSTRACT	vii
Supplementary Material	viii
LIST OF TABLES	xvi
LIST OF FIGURES	xvii
INTRODUCTION	1
Pipeline Development	1
Rationale	1
Overall Goals	4
Assessing Novel Transcripts and Their Importance	6
Annotation Databases: Gencode, RefSeq, and Ensembl	7
RNA-Seq	9
Putative Exons and Novel Protein Isoforms	10
Alternative Splicing	11
Frameshifts	12
Functional Protein Domains	13
Premature Stops	14
Nonsense Mediated Decay (NMD)	14
3D Protein Modeling	15
Differential Expression	17
RESULTS	18
Identification of the loss and/or gain of protein domains in the novel isoform	18

versus the current annotated form of the protein	
Analysis of proteins that have lost or gained functional domains by generating 3D models of the annotated proteins and novel protein isoforms to facilitate potential understanding of functional impact	30
Differential expression analysis of the putative exons at the RNA level to investigate disease-specificity of expression and potential changes in expression of the novel isoform in disease	41
DISCUSSION	48
REFERENCES	54
APPENDICES	58
Appendix A: Tables and text files referenced in the paper	58

ABSTRACT

While we know the sequence of the nucleotides that make up the DNA of the human genome, the process of annotating those nucleotides according to the transcripts that originate from them remains incomplete. Novel transcripts continue to be identified, and so methods must be devised to characterize these novel transcripts and prioritize them for future study by assessing potential hallmarks of function. One of the potential hallmarks of function is presence of an open reading frame with potential to produce a protein isoform that is not yet annotated. Based on the sequence of the novel protein isoform, an initial assessment of the potential functional impact of expression of the novel protein can be made:

- 1) Identification of the loss and/or gain of protein domains in the novel isoform versus the current annotated form of the protein.
- 2) Analysis of proteins that have lost and/or gained functional domains by generating 3D models of the annotated proteins and novel protein isoforms to facilitate potential understanding of functional impact.
- 3) Differential expression analysis of the putative exons at the RNA level to investigate disease-specificity of expression and potential changes in expression of the novel isoform in disease.

Supplementary Material

The following includes a list of files generated over the duration of the project, along with a description of each.

Scripts

canonical_seq_download_nov22_2019.R

R script that identifies gene names that correspond to a UniProt ID and those that do not. After identifying the UniProt IDs, the appropriate annotated sequences are downloaded from UniProt.

ERP001304_DE_FINAL.R

R script for differential expression analysis using the recount study ERP001304.

nmd_determination.R

R script that determines if a novel protein isoform is or is not a potential NMD candidate.

protein_domain_comparison.R

R script that generates files for each protein which contain the domains identified by SUPERFAMILY for both the annotated sequence and the isoform, followed by a determination as to the loss and/or gain of functional domains for each protein.

SRP033725_DE_FINAL.R

R script for differential expression analysis using the recount study SRP033725.

SRP035524_DE_FINAL.R

R script for differential expression analysis using the recount study SRP035524.

SRP043684_DE_FINAL.R

R script for differential expression analysis using the recount study SRP043684.

Files

Count tables for created for each of the recount2 studies used: ERP001304, SRP033725, SRP035524, and SRP043684

- **Initial count tables generated for regions of interest per chromosome**
 - counts_chr1_{001304, 033725, 035524, 043684}.csv
 - counts_chr10_{001304, 033725, 035524, 043684}.csv
 - counts_chr11_{001304, 033725, 035524, 043684}.csv
 - counts_chr12_{001304, 033725, 035524, 043684}.csv
 - counts_chr13_{001304, 033725, 035524, 043684}.csv
 - counts_chr14_{001304, 033725, 035524, 043684}.csv
 - counts_chr15_{001304, 033725, 035524, 043684}.csv
 - counts_chr16_{001304, 033725, 035524, 043684}.csv

- **Rounded count tables for regions of interest per chromosome generated based on read lengths**

counts_chr1_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr10_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr11_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr12_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr13_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr14_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr15_rounded_{001304, 033725, 035524, 043684}.csv
counts_chr16_rounded_{001304, 033725, 035524, 043684}.csv

- **Combined rounded count tables for regions of interest per chromosome into one matrix for each study**
counts_all_rounded_{001304, 033725, 035524, 043684}.csv
- **counts_all_rounded_for_pseudocounts.csv**
csv file that includes all regions that had at least one count in one sample (removed regions where the count was 0 for all samples) in study SRP035524
- **counts_all_rounded_pseudocount_added.csv**
csv file that includes all regions of interest with 1 added to each count in study SRP035524

canonical.html

SUPERFAMILY html file generated through the use of AWS for the canonical annotated sequences used in the batch analysis.

canonical_seqs.fasta

FASTA file containing all of the canonical annotated protein sequences that correspond to the gene IDs in SupplementaryTable_19_mod.xls.

df_canonical_initially.csv

Output csv file from the protein_domain_comparison.R script that includes the protein name and a column entitled “Canonical_Domains_Initially” to convey to the user if the canonical annotated sequence had domain assignments given to it by the SUPERFAMILY tool.

domain_information_output.txt

Output text file generated from the protein_domain_comparison.R script that includes specifics as to the name and number of domains lost and/or gained by each protein used in the batch analysis.

exon_reading_frames.xlsx

Table that includes the Gene_ID, Protein_Name, the reading frame identified for exon 1, and the reading frame identified for exon 2.

Filtering_gene_ids_for_batch_analysis.xlsx

Excel document with multiple tables that illustrate how identifiers were removed at various points in the batch analysis:

- **Initial_Gene_Ids**- List of initial fifty gene identifiers used.
- **UniProt_Gene_Ids**- List of gene identifiers and the UniProt ID associated with them.
- **Filter_Protein_Coding_Gene_Ids**- List of gene identifiers that were unique and protein coding (removal of readthroughs).
- **Determination_for_NMD_Testing**- Notation of which identifiers experienced premature stops to lead to a check for nonsense-mediated decay (NMD).
- **Remaining_After_NMD_Testing**- List of identifiers remaining after checking for NMD candidacy that were most likely to be translated multiple times.
- **Proteins_for_RaptorX** - Proteins remaining after checking for loss and/or gain of functional protein domains.

frameshift_determination_and_event_caused_by_insertion.xlsx

Table that contains information for each of the fifty gene identifiers used in the batch analysis regarding the reading frames identified for the flanking annotated exons, as well as if the putative exon insertion caused a frameshift in the annotated protein sequence, and if the insertion caused only an insertion or a premature stop.

gain_loss_domains_manually.xlsx

Table created to make sure that the output for loss/gain of domains being determined by the protein_domain_comparison.R script was correct. Table includes if there were domains initially identified in the annotated sequence, the domains gained, the domains lost, and if there was potential for further analysis.

gene_ids_for_conversion_script.txt

Text file that contains all of the initial gene identifiers from SupplementaryTable_19 provided by Darby et al.

gene_ids_no_uniprot_ids.csv

csv file containing all of the gene ids that did not identify with a UniProt ID.

gene_names_to_entrnames_orig_no_matches.csv

csv file containing all UniProt entry names for the gene ids determined to have alternative names.

gene_names_to_entrnames.csv

csv file containing all UniProt entry names for the gene ids that initially identified with a UniProt match.

gene_names_to_ids_orig_no_matches.csv

csv file containing all UniProt IDs for the gene ids determined to have alternative gene names.

gene_names_to_ids.csv

csv file containing all UniProt IDs for the gene ids that initially identified with a UniProt match.

intron_positions_for_nmd.csv

input file used in the nmd_determination.R script that includes the necessary information to determine NMD candidacy.

intron_positions_for_nmd_with_nmd_candidacy.csv

output csv file for nmd_determination.R script that includes a column highlighting if a novel protein isoform is a potential NMD candidate or not.

isoform_amino_acid_sequences.fasta

FASTA file with all of the manually generated novel protein isoform sequences for the proteins used in the batch analysis.

isoform_nuc_seqs_edit.txt

The nucleotide sequences downloaded from the UCSC Genome Browser for the flanking annotated exons and putative exon as (coordinates taken from SupplementaryTable_19).

isoform_html

SUPERFAMILY html file generated through the use of AWS for the novel isoform sequences used in the batch analysis.

nmd_information_output.txt

Output text file generated by the nmd_determination.R script that includes information as to the NMD candidacy of a novel protein isoform.

no_match_category_names.xlsx

Table generated that includes the gene IDs that did not initially identify with a UniProt ID, their alternative gene names, and the category of gene type that they fall under as notated by GeneALaCart.

nuc_and_isoform_sequences.txt

File that includes the generation of the novel protein isoform for each protein used in the batch analysis, including the frame used for translation, the nucleotide sequence of the flanking exons and the putative exon, followed by the translation of that nucleotide sequence from Expaty translate, the canonical sequence for the protein, and the entire novel protein isoform sequence.

proteins_for_RaptorX.xlsx

Table that lists the proteins that made it through the criteria used for pipeline developed that would be selected for further analysis by RaptorX.

ProteinName_SeqID_can_iso_seqs.fasta

Files generated that include the canonical annotated sequence and the novel protein isoform sequence. These files were used as input in the nmd_determination.R script. The following are the “ProteinName_SeqID” of each file:

AAMDC_Q9H7C9, ABCG4_Q9H172, BORA_Q6PGQ7, CDC16_Q13042, CE350_Q5VT06, CMTA1_Q9Y6Y1, DGKA_P23743, DIP2B_Q9P265, DNMI1L_O00429, DPP3_Q9NY33, F149B_Q96BN6, FPPS_P14324, HACD3_Q9P035, KCRU_P12532, KCTD3_Q9Y597, KLDC1_Q8N7A1, KTN1_Q86UP2,, LRFN5_Q96NI6, NEUR3_Q9UQ49, PLBL2_Q8NHP8, RBM4_Q9BWF3, RTN3_O95197, SAC2_Q9Y2H2, SNX1_Q13596, SYT11_Q9BT88, TM39B_Q9GZU3

ProteinName_SeqID_domains_canonical_isoform_compared.csv

Files generated that include the domains determined by the SUPERFAMILY tool for both the canonical annotated sequence and the novel isoform sequence. These files were used as input in the protein_domain_comparison.R script. The following are the “ProteinName_SeqID” of each file:

AAMDC_Q9H7C9, ABCG4_Q9H172, BORA_Q6PGQ7, CDC16_Q13042, CE350_Q5VT06, CMTA1_Q9Y6Y1, DGKA_P23743, DIP2B_Q9P265, DNMI1L_O00429, DPP3_Q9NY33, F149B_Q96BN6, FPPS_P14324, GDE_P35573, GL1D1_Q96MS3, HACD3_Q9P035, KCAB2_Q13303, KCRU_P12532, KCTD3_Q9Y597, KLDC1_Q8N7A1, KPBB_Q93100, KTN1_Q86UP2, LRFN5_Q96NI6, NEIL1_Q96FI4, NEUR3_Q9UQ49, PLBL2_Q8NHP8, RBM4_Q9BWF3, RTN3_O95197, SAC2_Q9Y2H2, SNX1_Q13596, SYT11_Q9BT88, TM39B_Q9GZU3, WDTC1_Q8N5D0

putative_region_info_for_GRanges.csv

csv file containing the information necessary to construct a GRanges object used in the differential expression analysis of the novel exons.

README.Rmd

Summation of the batch analysis conducted for the fifty gene identifiers used.

superfamily_domains_canonical_for_script.csv

Modified table from the SUPERFAMILY output table (canonical.html) used as one of the input files for the protein_domain_comparison.R script.

superfamily_domains_isoform_for_script.csv

Modified table from the SUPERFAMILY output table (isoform.html) used as one of the input files for the protein_domain_comparison.R script.

SupplementaryTable 19_mod.xls

Initial data from Darby et al. used for batch analysis that contains the RE putative exon region coordinates, the flanking exon coordinates, coverage, and the gene id each associated with. Also contains two additional columns entitled “ALT_GENE_IDS” and

“GENE_IDS_TO_USE” to identify all possible gene names that identify with a UniProt ID.

test_canonical_50.fasta

FASTA file containing the canonical annotated sequences that were submitted to the SUPERFAMILY tool for domain determination.

uniprot_ids_for_canonical_download.txt

Text file containing a list of the UniProt IDs to be used to download the appropriate canonical annotated sequences from the UniProt database in the canonical_seq_download_nov22_2019.R script.

unique_gene_ids_no_uniprot_ids.csv

csv file containing all UniProt IDs for unique gene ids that initially identified with a UniProt match (minimizes gene_names_to_ids.csv).

unique_gene_names_to_entrynames_orig_no_matches.csv

csv file containing all UniProt entry names for unique gene ids that initially identified with a UniProt match (minimizes gene_names_to_entrynames.csv).

unique_gene_names_to_entrynames.csv

csv file containing all UniProt entry names for unique gene ids that initially identified with a UniProt match (minimizes gene_names_to_entrynames.csv).

unique_gene_names_to_ids_orig_no_matches.csv

csv file containing all UniProt IDs for unique gene ids determined to have alternative gene names (minimizes gene_names_to_ids_orig_no_matches.csv).

unique_gene_names_to_ids.csv

csv file containing all UniProt IDs for unique gene ids that initially identified with a UniProt match (minimizes gene_names_to_ids.csv).

updated_gene_ids_for_original_no_matches.txt

Text file containing a list of alternative gene names for protein coding genes that should have a UniProt ID match (but initially did not have a match).

Folders

ERP001304

ERP001304.csv

Phenotype information converted from tsv form to csv form

ERP001304.tsv

tsv file downloaded with phenotype information for the study

rse_gene_ERP0013004.Rdata

allows for the study to be downloaded into R so that the DE analysis can be run.

SraRunTable_ERP001304.txt

Table containing information about the samples in the recount study being used for the DE analysis.

SRP033725

SRP033725.csv

Phenotype information converted from tsv form to csv form

SRP033725.tsv

tsv file downloaded with phenotype information for the study

rse_gene_SRP033725.Rdata

allows for the study to be downloaded into R so that the DE analysis can be run.

SraRunTable_SRP033725.txt

Table containing information about the samples in the recount study being used for the DE analysis.

SRP035524

SRP035524.csv

Phenotype information converted from tsv form to csv form

SRPP035524.tsv

tsv file downloaded with phenotype information for the study

rse_gene_SRP035524.Rdata

allows for the study to be downloaded into R so that the DE analysis can be run.

SraRunTable_SRP035524.txt

Table containing information about the samples in the recount study being used for the DE analysis.

SRP043684

SRP043684.csv

Phenotype information converted from tsv form to csv form

SRP043684.tsv

tsv file downloaded with phenotype information for the study

rse_gene_SRP043684.Rdata

allows for the study to be downloaded into R so that the DE analysis can be run.

SraRunTable_SRP043684.txt

Table containing information about the samples in the recount study being used for the DE analysis.

Plots**chr11/63746946-63747024_+.pdf**

Counts plot generated for putative exon at genomic coordinates chr11:63746946-63747024 in recount study SRP037725.

chr15/75007587-75007728_+.pdf

Counts plot generated for putative exon at genomic coordinates chr15:75007597-75007728 in recount study SRP037725.

chr15/75007660-75007728_+.pdf

Counts plot generated for putative exon at genomic coordinates chr15:75007660-75007728 in recount study SRP037725.

chr15/75351300-75351376_+.pdf

Counts plot generated for putative exon at genomic coordinates chr15:75351300-75351376 in recount study SRP037725.

LIST OF TABLES

Tables		Page
1	Databases and tools utilized in pipeline development.	5
2	Proteins selected for further analysis.	32

LIST OF FIGURES

Figure		Page
1	NR4A2 mRNA expression in control samples and schizophrenia samples.	4
2	Schematic of the bioinformatics pipeline being developed and its relation to other important steps of novel transcript investigation.	7
3	Possible locations of putative exons in relation to annotated genes.	10
4	Potential impact of putative exon insertions at the transcriptional and translational level.	11
5	Downstream effects of altered functionality of a protein as a consequence of the loss and/or gain of functional protein domains.	19
6	Process of determining alternative gene names and UniProt IDs.	22
7	User input for the single isoform analysis tool as determined by initial batch analysis in pipeline development.	23
8	Process followed upon user input for the single isoform analysis.	25
9	Alternative method to JuncDB to determine the final exon-exon junction.	27
10	Selecting a novel protein isoform for further analysis.	33
11	RaptorX output of NEUR3.	36

12	RaptorX output of KLDC1.	36
13	Comparison of the domain assignments generated by the SUPERFAMILY versus the RaptorX.	37
14	Example of RaptorX Output for entire protein model and summary information.	39
15	Summary information regarding the predicted 3D protein model.	39
16	Secondary structure information for predicted 3D model.	40
17	Visualization of domain-parsing in RaptorX output.	40
18	Genomic coordinate information used to construct Granges object for DE analysis.	42
19	Counts plot for putative exon: chr15:75007587-75007728.	45
20	Counts plot for putative exon: chr15:75007660-155866737.	45
21	Counts plot for putative exon: chr15:75351300-1800523216.	46
22	Counts plot for putative exon: chr11: 63746946- 6807088.	46
23	Additional user input for the single isoform analysis tool.	47
24	Integration of pipeline developed with future single isoform analysis tool.	49

INTRODUCTION

Pipeline Development

Rationale

The idea for development of a pipeline that determined potential functional impact of novel protein isoforms begins with the all-encompassing idea that the current gene annotations available are incomplete.

Only about 2% of the human genome is considered protein coding, with the remainder of the genome being considered “junk DNA” (Nowacki et al. 2009). Coined in around 1972 by Susumu Ohno, this idea of a dispensable portion of the genome began to change in the 1990s, as scientists discovered that some of the “junk DNA” is actually functionally important, for example, in the case of repetitive elements. One function of the “junk DNA” could be that it is material included in the production of novel transcripts. Some of these novel transcripts could have the potential to code for novel protein isoforms.

The Sequence Read Archive (SRA), a repository for RNA sequencing (RNA-Seq) data, has experienced a decline in new exon-exon junctions being submitted, suggesting that sequencing data now contained in the SRA is sufficient to identify the majority of existing exon-exon junctions. Therefore, the task researchers are now faced with is how to address insufficient isoform annotation through identification and analysis of novel transcripts (Darby 2019).

Furthermore, the annotation databases that currently exist, such as GENCODE, RefSeq, and Ensembl, while often helpful, are not always in sync with one another. Annotations found in one database are not guaranteed to be found in another. Because of

this, researchers may obtain different results depending on the annotation database being referenced. Thus, it is important to achieve consistency across annotation databases so that research being conducted is robust and reproducible.

In addition to the fact that current gene annotations are incomplete, many methods currently used in scientific research rely on the use of current annotations. The two main methodologies used for RNA-Seq quantification are based only on information from current annotations. However, it is also possible to identify novel transcripts by examining sequencing reads that align to the reference genome outside of annotated transcripts using an annotation-agnostic approach.

Regarding proteomics and data produced from experiments in this field, annotated transcripts are used to generate hypothetical peptides. Since these peptides are produced from annotated transcripts, they are not novel, and therefore do not allow for assessment of potential function of unannotated regions in the genome.

In Genome Wide Association Studies (GWAS), it is found that most disease-associated polymorphisms are outside of annotated coding regions, which make these polymorphisms difficult to interpret. Hence, it is important to establish methods that allow for assessment of the impact of such polymorphisms in order to learn more about the development of disease on a cellular level. While current methods exist to determine whether such polymorphisms affect expression of annotated gene, expanding our understanding of the functional significance of unannotated transcripts that might contain these polymorphisms will allow for discovery of additional therapeutic targets, bolstering research efforts to develop new treatments, and helping those living with disease.

Darby et al. identified novel mRNAs with the potential to code for novel protein

isoforms (Darby et al. 2016 Sep 20). Conducting RNA-Seq on only polyadenylated mRNAs, the group then counted sequencing reads for the repetitive elements (RE) mapping to the human genome. Since RNA-Seq was completed on polyadenylated sequences, it was expected that the RE would be mapping to only exonic regions, though this was not the case. Some of the RE were mapping to intronic and intergenic regions (Darby et al. 2016 Sep 20). This allowed for the conclusion that these RE in the intronic and intergenic regions were actually being expressed. Further investigation revealed that these regions contained splice junctions indicating that they were incorporated into novel transcript isoforms of known genes. Furthermore, it raised questions as to the impact of including these novel exons as an insertion into a mRNA transcript, as well as what the functional impact could be of a novel protein isoform produced from one of these novel mRNA transcripts.

One such example of an insertion of a novel exon into a mRNA transcript includes the NR4A2 gene. Figure 1 highlights an example of a NR4A2 mRNA transcript expression in schizophrenia samples compared to the control samples. In the schizophrenia samples, it can be seen that there is expression in the control samples of what was previously annotated as an intronic region. This leads to further questioning as to what the functional impact could be of a novel protein isoform produced from a mRNA that included this intronic region should it possess coding potential, and therefore such an isoform's implication in psychiatric disease.

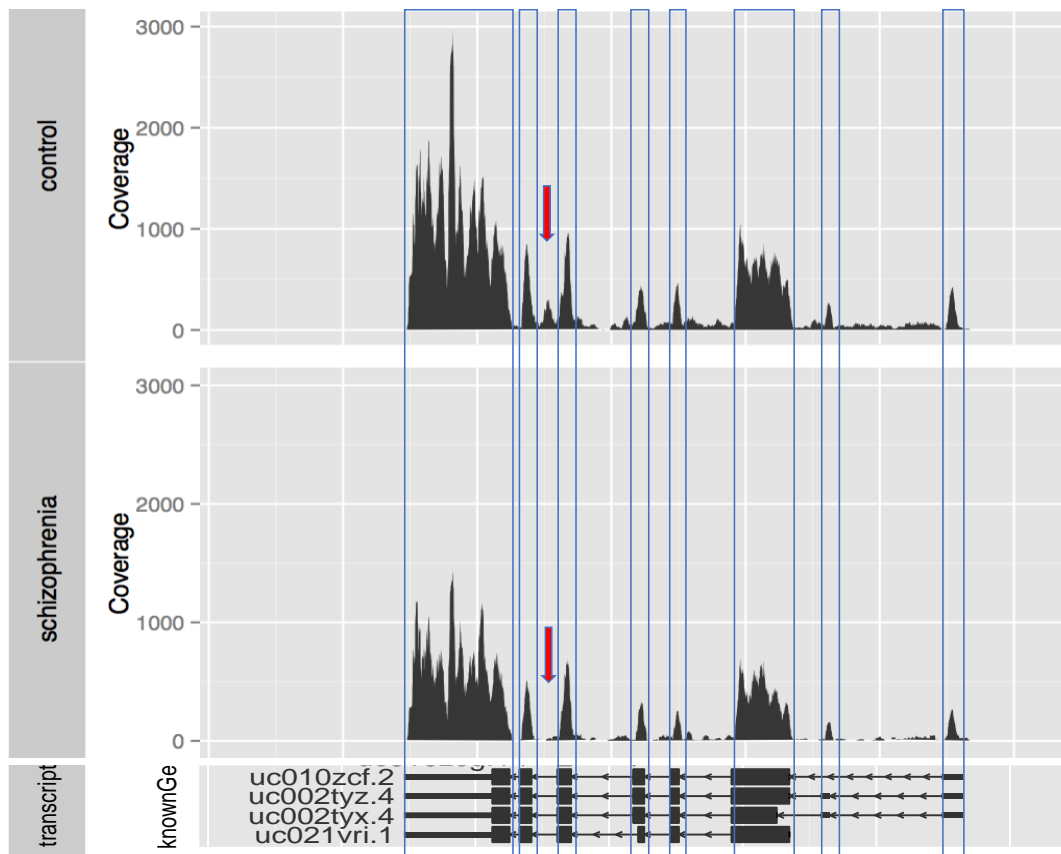


Figure 1. NR4A2 mRNA expression in control samples and schizophrenia samples. An intronic region is shown to be expressed in the control samples and not the schizophrenia samples (indicated by the red arrows), giving evidence of a potential novel mRNA transcript being expressed.

Overall Goals

There were two overall goals to be achieved in developing this pipeline:

- 1) Conduct an analysis of novel mRNAs identified by Darby et al. through a batch-type in order to assess all considerations that must be taken when implementing the pipeline into a single isoform analysis tool.
- 2) Draw conclusions from the batch-analysis that will be used in future implementation of the pipeline into a user-friendly single isoform analysis tool.

The approach behind development of this pipeline was to make use of the databases and tools currently available, with the intent to develop a new tool with the purpose of protein isoform analysis. The main databases used in developing this pipeline included GeneCards, UniProt, and JuncDB (Chorev et al. 2016; European Molecular Biology Laboratory 2019; GeneCards 2019). The main tools used in developing the pipeline included SUPERFAMILY and RaptorX (Gough et al. 2001; Källberg et al. 2012a). Table 1 contains a brief description of the database or tool, followed by its use in pipeline development.

Table 1. Databases and tools utilized in pipeline development. There were several databases and tools used in the development of the bioinformatics pipeline geared towards assessing the functional impact of novel protein isoforms.

Database/Tool	Description	Use in Pipeline
GeneCards	Database that provides information on all annotated and predicted human genes.	Identification of genes with no initial UniProt ID match as protein coding or not; determination of alternative gene names.
UniProt	Freely-accessible database of protein sequence and functional information.	Used to download annotated protein sequences.
JuncDB	Freely-accessible exon-exon junction database.	Retrieval of final exon-exon junctions for potential NMD candidates.
SUPERFAMILY	Database of structural and functional annotation of proteins based on a collection of Hidden Markov Models (HMM).	Tool used for domain assignments of both annotated proteins and novel protein isoforms.
RaptorX	Template-based modeling; predicts 3D protein structure based on sequence similarity.	Tool used to generate predicted 3D models of both the annotated proteins and novel protein isoforms.

Assessing Novel Transcripts and Their Importance

Other projects currently taking place in the Darby Lab work in conjunction with the pipeline being developed. All projects involve the assessment and investigation of novel transcripts through initial identification and determining hallmarks of potential function

The suite of tools being developed to assess and investigate novel transcripts includes the following steps. Steps that are bolded in the list below are the ones being focused on in the development of this pipeline. Figure 2 highlights the pipeline and how it relates to some of the other steps regarding novel transcript investigation.

1. Identify that a genomic region in an intron or between genes that is expressed based on RNAseq data.
2. Find splice junctions connecting the novel portion of the mRNA as an alternative or additional portion of an annotated mRNA.
- 3. Translate the new mRNA to identify the novel protein sequence.**
- 4. Identify whether the novel sequence will undergo nonsense mediated decay.**
- 5. Identify any loss and/or gain of functional protein domains in the novel isoform compared to the annotated protein.**
- 6. Generate structural models of the novel and annotated proteins.**
7. Perform expression analysis of the novel region of the mRNA to assess specificity of expression across tissue types and brain regions.
- 8. Perform expression analysis of the novel region of the mRNA to assess altered expression in disease.**

9. Search shotgun proteomics data for peptides that match the novel portion of the protein sequence.

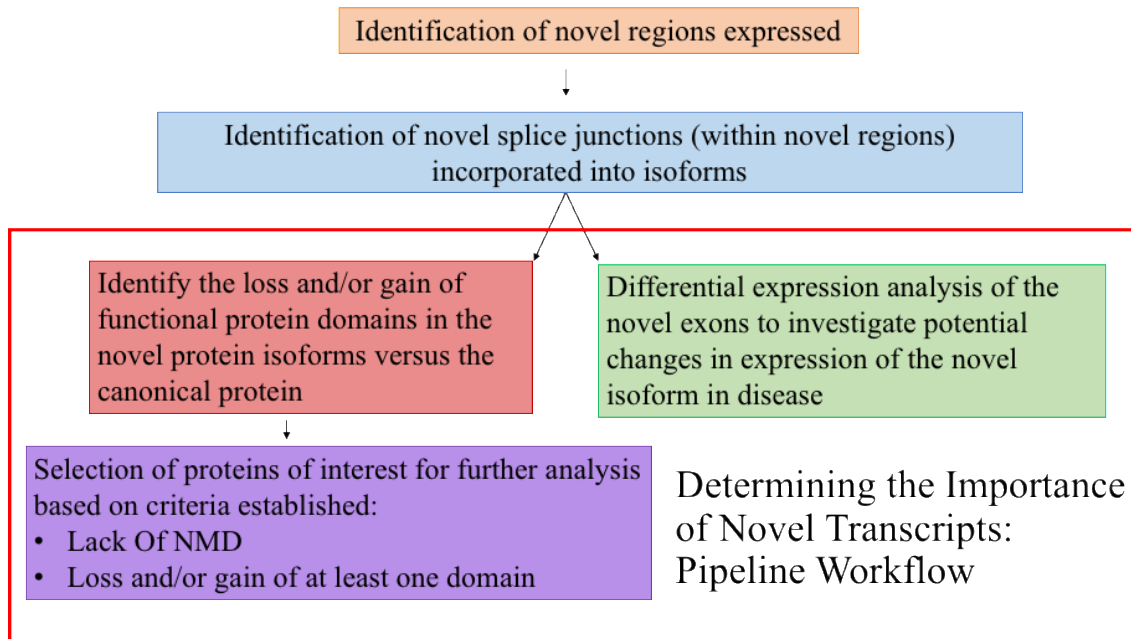


Figure 2. Schematic of the bioinformatics pipeline being developed and its relation to other important steps of novel transcript investigation. This pipeline will help to determine the potential functional impact of novel protein isoforms. Boxed in red are the processes specific to the pipeline development in this report.

Annotation Databases: GENCODE, RefSeq, and Ensembl

There are several annotation databases currently available that compile the information that is currently known about the RNAs, proteins, and DNA sequences in the genome from which they are generated. The annotation databases that are most commonly used are GENCODE (Harrow et al. 2012), RefSeq (RefSeq 2019), and Ensembl (Zhao and Zhang 2015). RefSeq, created and maintained by the United States National Center for Biotechnology Information, and Ensembl, created and maintained by European Molecular Biology Laboratory’s European Bioinformatics Institute (EMBL-EBI), both use a combination of computational pipelines and manual curation to annotate

the genome based on RNA and protein sequences that are submitted to GenBank (Frankish et al. 2015). The GENCODE and Ensembl annotations are considered to essentially be identical. The GENCODE annotation is created by merging the manual and automated Ensembl gene annotations. The GENCODE annotation is the default Ensembl gene annotation used in its genome browser (GENCODE 2019), while the RefSeq gene and protein annotations are the primary annotations used by NCBI tools such as BLAST. Currently, there are over 200,000 human transcripts available in each of these annotations, though each has a slightly different number (GENCODE 2019), and the individual entries are not identical.

While RefSeq has been determined to be very similar to the GENCODE Basic Set, it does not compare to GENCODE's Comprehensive Set, which contains greater annotations of alternative splicing, novel CDSs, novel exons and has higher genomic coverage (Frankish et al. 2015). In a previous analysis, it was found that GENCODE contained over 30,000 annotated transcripts not available in RefSeq (Harrow et al. 2012). The differences between the annotation databases are problematic both because they are frustrating to researchers who may not know which database to use and because the lack of similarity between the annotations can lead the results of gene expression studies and other analyses to vary depending on the annotation tool used. For example, while there are 21,598 common genes between RefSeq and Ensembl, it has been found that the choice of annotation has quite an effect on data analysis in relation to gene quantification and expression (Zhao and Zhang 2015). More importantly in the context of this thesis, the differences between the databases highlight the fact that none of the annotation databases are fully complete, so tools are needed to allow researchers to interpret the

likely functional impact (or lack thereof) of new, unannotated transcripts that they discover using RNA sequencing or other methods.

RNA-Seq

RNA sequencing (RNA-Seq) is the process whereby the nucleotide sequences of RNAs extracted from a sample are determined and quantified. The vast majority of RNA-Seq studies utilize RNA from the whole genome that has been purified to include only polyadenylated sequences, the majority of which are messenger RNAs (mRNAs). The two major methodologies for quantification of RNA-Seq data are: 1) to align sequencing reads to the full reference genome and to count the number of sequencing reads that overlap genomic regions of interest, or 2) to align the sequencing reads to canonical transcripts of annotated genes prior to downstream analysis and quantification (Conesa et al. 2016). While most RNA-Seq studies focus on the measurement of annotated transcripts, it is also possible to identify novel transcripts by examining sequencing reads that align to the reference genome outside of annotated transcripts or by using de novo sequence analysis tools that attempt to reconstruct sequencing reads into continuous transcripts without the guidance of a reference genome. Sequencing reads derived from polyadenylated RNAs that align to a part of the genome that is not annotated as an mRNA or noncoding transcript indicate that an unannotated transcript is likely to exist.

Putative Exons and Novel Protein Isoforms

Inclusion of putative exons in currently annotated mRNA transcripts has the ability to create novel mRNAs with novel protein coding potential. Putative exons can occur upstream, downstream, or within annotated genes. Putative exons that occur upstream of annotated genes can be included in alternative transcripts of those genes as alternative transcription start sites that may add to the five prime untranslated portion of the transcript, introduce an alternative translation site, or trigger changes in splice site usage downstream in the transcript. Those that occur downstream of annotated genes can be included as alternative transcription termination sites that could extend the coding sequence or three prime untranslated sequence of the transcript and can affect the efficiency whereby the alternative transcript is translated. Putative exons that occur in regions that are annotated as introns are likely to alter the coding sequence of the gene, but could potentially be included in the untranslated regions of the transcript (Figure 3).

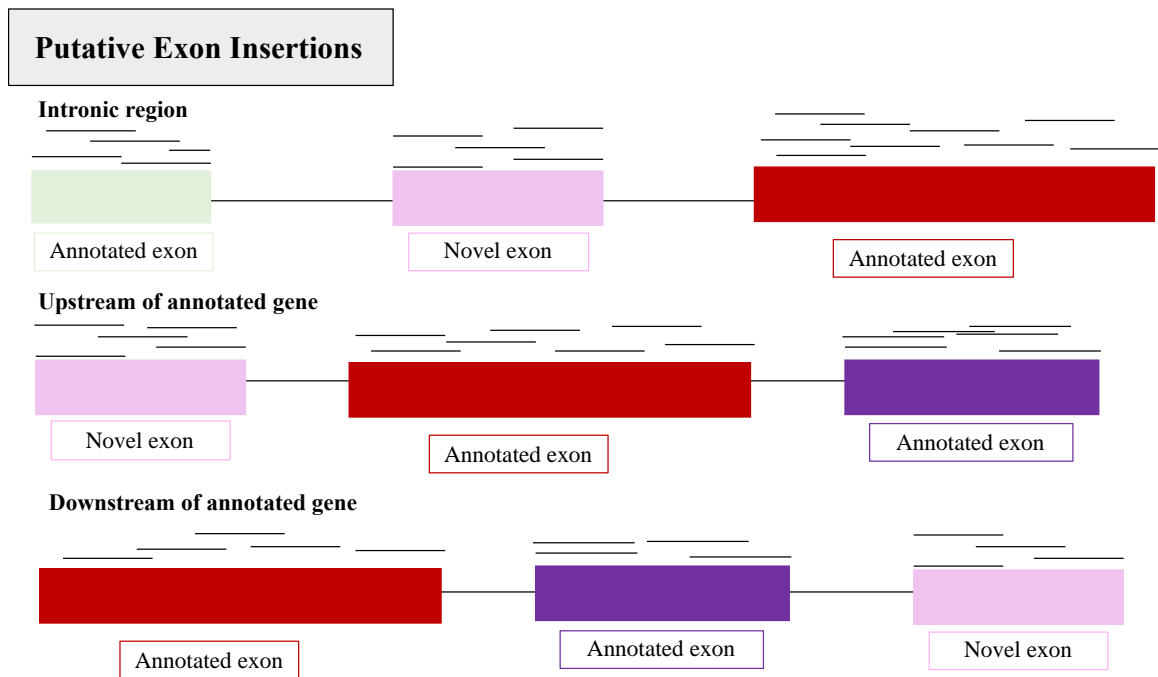


Figure 3. Possible locations of putative exons in relation to annotated genes. Putative exons can be upstream, downstream, or within introns of annotated genes. Location of the putative exon contributes to its effect on the transcripts being created.

In summary, the effect of putative exons could occur at both the transcriptional and translational level. At the transcriptional level, putative exons can add new transcriptional start sites, alter splicing patterns, or add an alternative termination site. At the translational level, putative exons can cause nonsense-mediated decay, truncated proteins, frameshifts, and the (likely) loss and/or (unlikely) gain of functional protein domains (Figure 4).

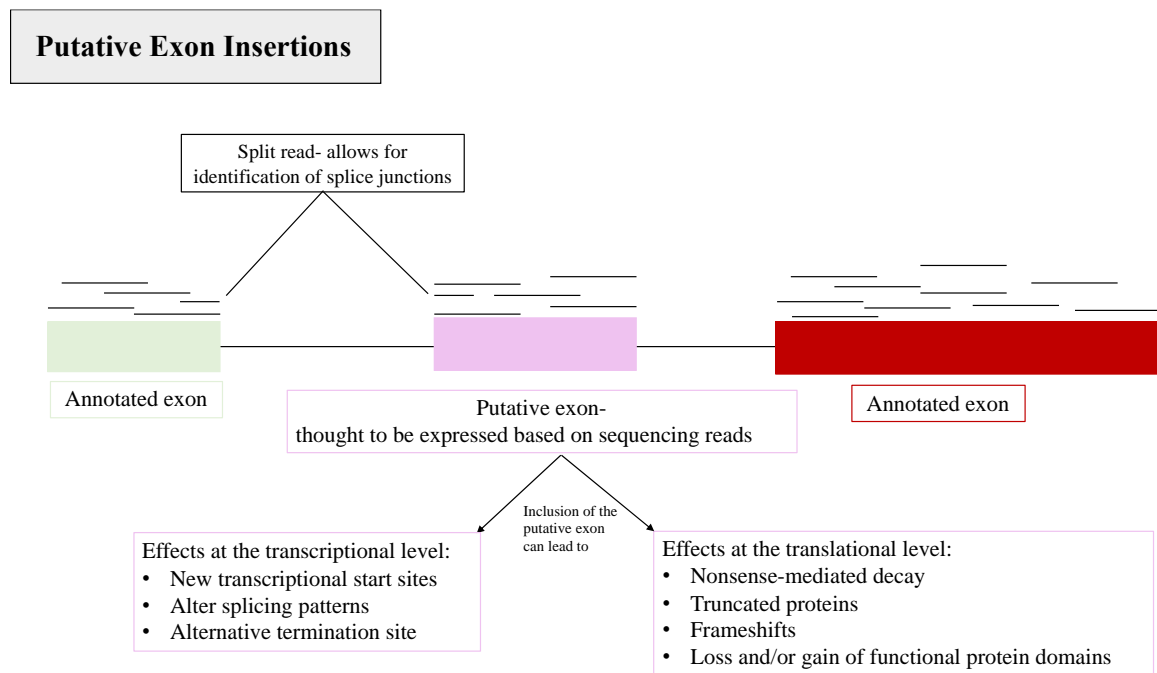


Figure 4. Potential impact of putative exon insertions at the transcriptional and translational level. Insertion of a putative exon into an annotated sequence can cause several downstream effects, including new transcriptional start sites, alter splicing patterns, alternative termination sites, nonsense-mediated decay, truncated proteins, frameshifts, and the loss and/or gain of functional protein domains.

Alternative Splicing

The underlying purpose of alternative splicing (AS) is to aid in regulation of gene expression (Yeo et al. 2004), and such regulation leads to an increased complexity in cellular functioning, specifically in eukaryotes (Barmak Modrek and Christopher Lee

2002). AS events generate tissue-specific mRNAs, and those with coding potential can then be translated into protein isoforms. Of the estimated 25,000 genes in the human genome, there are thought to be between one-third and two-thirds that undergo alternative splicing events (Yeo et al. 2004). Types of AS events include exon skipping, alternative start or termination sites, and intron retention, though exon skipping is the most prevalent form (Sultan et al. 2008).

AS events help to produce mature mRNA from pre-mRNA. Variance in the splice junctions, the genomic coordinates used to determine the start and end point of an exon, leads to the production of alternatively-spliced transcripts. These transcripts could potentially go on to produce protein isoforms that could impact overall cellular function if the alternative protein functions differently than the canonical version.

Functional Protein Domains

Proteins contain various levels of structure, including primary, secondary, tertiary and quaternary structure. These levels of structure, when broadly conserved, form what are known as functional protein domains. Domains are distinct, functional units within a protein, and are able to exist and function independently of one another (EMBL-EBI 2020). Each domain has a specific function in the overall protein, known as “functional division of responsibility” (OpenLearn University 2020). Some common families of domains are: DNA-binding domains, Regulation G-protein signaling (RGS) domains, and RNA-binding domains (EMBL-EBI 2020). Domains such as these are present in similar proteins that may serve an overall different function from one another. Each of these domains is biologically significant to the overall role of the protein in the cell, and the

loss and/or gain of a domain could result in the expression of minor isoforms whose function differs from the annotated form, affecting downstream biological pathways.

Frameshifts

Frameshifts are a mutation caused by the insertion or deletion of nucleotides in a coding sequence. As a result, the reading frame for translation of the sequence is shifted, leading to a different codon translation for all codons following the frameshift. This leads to a different amino acid sequence compared to the original protein sequence, and can lead to deleterious effects on the structure and function of the protein that is produced.

It is important to note that a frameshift will almost always cause a string of nonsensical sequence to follow downstream of the frameshift. In a previously annotated protein that experiences a frameshift, structure and function of the protein, such as its functional domains, are typically lost. In fact, it has been found that a +1 frameshift in mRNAs leads to an average protein sequence identity of just 6.2% between the wild-type protein and the protein resulting from the frameshift (Bartonek et al. 2020).

While it is often thought that polypeptides resulting from a frameshift are completely unrelated to the wild-type, there are special cases when the structure and function of the protein, or at least part of it, may be conserved even in the case of a frameshift (Bartonek et al. 2020). Simple structural domains such as protein loops caused by disulfide bonds between cysteines can be preserved if the frameshift generates new cysteines near original cysteines that were lost.

It has been noted that key physiochemical properties of the protein, such as hydrophobicity and hydrophilicity, are often retained in many cases (Bartonek et al. 2020). However, this retaining of properties does not contribute to the same globular structure of the protein forming through folding. It contributes to how the protein may be viewed by the cell. For example, the presence of several hydrophobic residues on the outside of a protein could signal that the protein is unfolded in some way, ultimately leading to degradation of the protein. This degradation process serves as part of the cell's proteostasis system.

Premature Stops

Premature stops are a mutation type that results in a stop codon before the end of a reading frame. At the transcriptional level, this premature stop can be considered an alternative termination site, one of the many effects of novel exons. At the translational level, premature stops can cause truncation of the protein, and ultimately lead to the loss of functional protein domains. This loss of domains is the result of a lack of a protein isoform's ability to fold and function as it normally would. In extreme cases, mutations that lead to premature stops can then lead to an mRNA that undergoes nonsense-mediated decay.

Nonsense-mediated decay (NMD)

Nonsense-mediated decay (NMD) is the process of mRNA degradation when a premature stop codon is located more than 50-55 nucleotides upstream of the final exon-exon junction (Hug et al. 2016). The main purpose of NMD is to degrade mRNAs that

would translate into truncated protein that could cause deleterious effects on the organism, including disease development (Kurosaki and Maquat 2016). However, it has also been determined that NMD also targets naturally occurring and ‘normal’ transcripts for the purpose of regulating cellular response when influenced by environmental stimuli (Hug et al. 2016). If NMD is triggered in either case, it will affect overall levels of a protein expression in an organism, and so it is important to determine whether a novel mRNA transcript is undergoing NMD or not.

3D Protein Modeling

3D protein modeling is a way of visualizing the globular structure of a protein. By generating this structure, a researcher can visualize any loss or gain of functional protein domains in relation to the overall structure, which then allows for a judgement to be made regarding potential functional impact of that protein. In addition, 3D modeling allows for comparisons to be made between proteins of different types, as well as between protein isoforms, allowing for an expansion in the understanding of general protein structure and function.

There are two main types of protein modeling currently used: template-based modeling and template-free modeling. Template-based modeling involves determination of related proteins to the input sequence that are then used as templates to generate structural models using the templates as a “scaffolding” to build the model. In template-free modeling, predicted structural models are generated using *ab initio* methods, building models without any sort of homology information. (Källberg et al. 2012).

Template-based modeling includes two methods used for protein modeling: 1) homology modeling and 2) protein threading. In homology modeling, the template in the alignment generated for the template and the query is treated as a sequence, and this is used to model the predicted structure. In protein threading, also known as “fold-recognition”, the template in the alignment is instead treated as a structure, and both the sequence and structure information determined from the alignment are used for generating the 3D model. Homology modeling is best to use when proteins are thought to be “easy” targets- they have homologous proteins that have known structure. Protein threading is best for “hard” targets- those that have only fold-level homology and a lack of homology to other proteins.

While protein modeling tools can provide insight into protein structure and therefore its function, there are pitfalls when it comes to 3D protein modeling. When homology is not able to be determined between an input sequence and a known structure, models could be extremely unreliable if they are able to be generated at all (Kelley et al. 2015). In this case, different approaches would then need to be used, such as de novo protein structure predicting tools. While there are over 85 million protein sequences available in protein annotation databases, there are only around 120,000 protein structures that have been experimentally determined and are available in the Protein Data Bank (PDB) (Rachel 2018). This large gap between the availability of sequences versus structure highlights the fact that 3D protein modeling is a tool that must be used with extreme caution and consideration. A user must take in to account the output generated along with the 3D model in determining its validity, including the percentage of residues modeled with a specific threshold of confidence, residues modeled as “disordered”, and

the abundance of homologous structures to the input sequence that were used as template to generate the predicted 3D structure.

Differential Expression

A differential expression analysis is completed to identify regions of the genome that are expressed differently in one group of samples compared to another. In order to perform differential expression analysis using RNA-Seq data, one counts the number of sequencing reads aligning to a region of the genome, followed by statistical modeling to observe mean differences in expression between groups of samples (i.e. tissue A vs. tissue B or individuals affected by diseases versus unaffected controls).

One database that houses RNA-Seq data that can be used in a differential expression analysis is recount2. Recount2 is a curated repository of RNA-Seq data of human tissue developed by the lab of Jeff Leek (Darby 2019). There are over 2,000 RNA-Seq studies in the database, with samples defined through categories such as psychiatric disease, stages of neurodevelopment, and patient condition (Collado-Torres et al. 2017).

The analysis tool being used to investigate differential expression is based off of the DER Finder approach. DER Finder is annotation-agnostic, determining differentially expressed regions (DERs) through identification of expression at each base. Regions of bases that have similar levels of expression are grouped together, and a statistical significance assigned to each, leading to conclusions as to the presence of differentially expressed regions (Frazee et al. 2014).

RESULTS

Identification of the loss and/or gain of protein domains in the novel isoform versus the current annotated form of the protein

Rationale

Identification of the loss or gain of functional protein domains in a novel isoform can signify a protein with altered functionality compared to the current annotated isoform. This altered functionality leads to the protein working differently than it normally would overall, leading to abnormalities occurring at the cellular level. Such abnormalities could then lead to the potential for disease development (Figure 5).

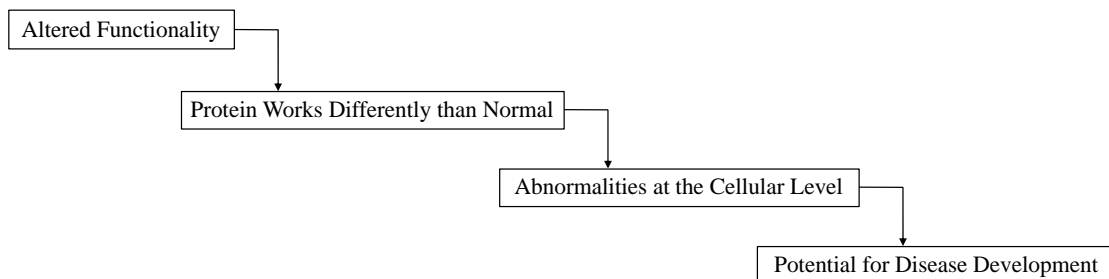


Figure 5. Downstream effects of altered functionality of a protein as a consequence of the loss and/or gain of functional protein domains. Altered functionality of a protein can lead to the potential for disease development.

Analysis Design

The following is a list of steps completed to identify the loss and/or gain of protein domains.

1. Downloaded the annotated protein sequences to give context to novel exon.
 - 1a) Fetched UniProt IDs.
 - 1aa) Determined alternative gene names.

2. Translated the novel mRNA into novel protein sequence.
 - 2a) Fetched nucleotide sequences for the flanking annotated exons and the putative exon using the UCSC Genome Browser.
 - 2b) Determined the appropriate reading frame for the flanking annotated exon that precedes the putative exon.
 - 2c) Used the reading frame identified for the first flanking annotated exon to translate the nucleotide sequence which consisted of the flanking exon, the novel exon, and the second flanking exon.
 - 2d) Inserted the translated sequence into the annotated protein sequence in order to produce the entire novel protein isoform sequence.
3. Checked the protein isoform sequence for nonsense-mediated decay (NMD).
 - 3a) Used JuncDB to identify final exon-exon junctions in proteins exhibiting a premature stop in the novel isoform sequence.
 - 3b) Checked for NMD candidates using an automated script process.
4. Checked the proteins for loss and/or gain of functional domains.
 - 4a) Selected the best tool for domain assignments based on evaluation of several available tools.
 - 4b) Identified domains using the SUPERFAMILY domain assignment tool.
 - 4c) Determined the loss and/or gain of functional domains using automated script process.

Downloaded annotated sequences to give context to novel exon

Analysis Outcome

When downloading the annotated protein sequences from UniProt, one issue was the occurrence of alternative gene names. There is no naming schema for genes in biology, and so over time this has allowed for one gene to be named using multiple identifiers. The data from Darby et al. specified gene names that were found to be associated with the genomic coordinates of the annotated exons (Darby et al. 2016 Sep 20). To download annotated protein sequences from UniProt, a connection needed to be established between the gene name and the appropriate UniProt ID for the protein being produced from a coding gene. Illustrated in Figure 6 is the process taken to download the annotated protein sequences from UniProt. An R script written determined UniProt ID matches to the gene names provided by data produced by Darby et al. Gene names that did not have a UniProt match were identified and searched using batch queries through GeneCards (GeneALaCart) to determine if the gene was protein coding. Gene names categorized as protein coding genes were matched with their alternative gene name as provided by the GeneALaCart search. This alternative gene name is what was then used to identify the UniProt ID associated with the gene. See Appendix A- “no_match_category_names.xlsx” for a portion of the table generated to notate if a gene was protein coding or not and what its alternative gene name was.

All of the UniProt IDs were then used to download the annotated protein sequences from UniProt. The annotated sequences downloaded were those considered “canonical” by the UniProt database. However, there were other protein isoforms annotated that have been caused by alternative splicing available in the UniProt database.

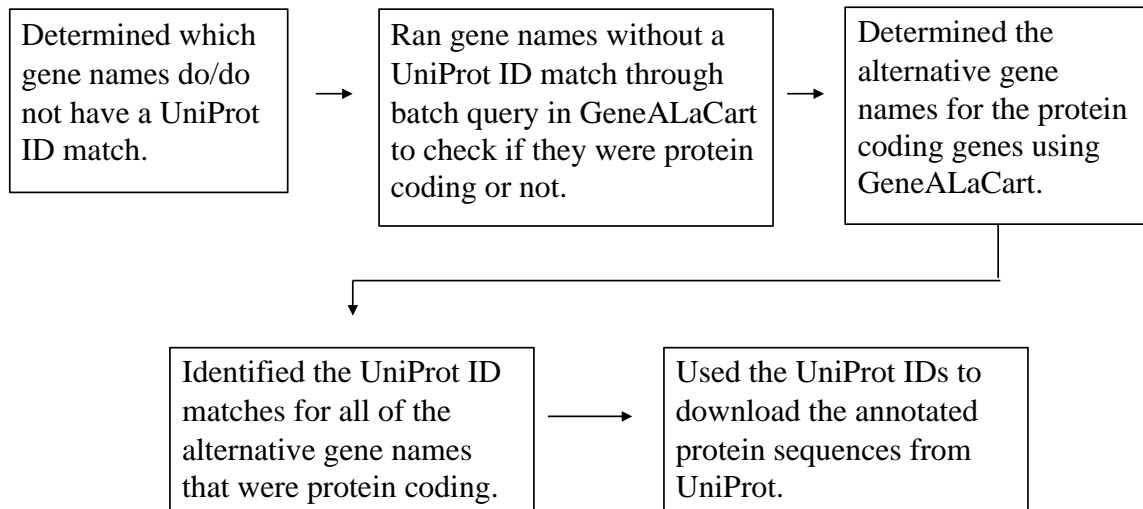


Figure 6. Process of determining alternative gene names and UniProt IDs. The flow chart illustrates the process taken to identify gene names associated with UniProt IDs. These UniProt IDs were then used to download the annotated protein sequences from UniProt.

Conclusions for Single Isoform Analysis Tool Development

Discovering that there are alternative gene names, as well as multiple protein isoform annotations a user may want to use allowed for the conclusion that there needed to be at least three inputs from the user when implementing the pipeline into a tool for single isoform analysis. These three inputs include: 1) protein isoform annotation ID (from UniProt), 2) genomic coordinates of the region to be used for insertion, and 3) gene name that the protein is coded from as seen in the UniProt database (Figure 7).

Protein_Isoform_ID ▼	Gene_Name ▼	Genomic_coordinates_for_insertion
Q13303	KCAB2	Chromosome_Number:
		1
		Start_Position:
		6145799
		End_Position:
		6145853

Figure 7. User input for the single isoform analysis tool as determined by initial batch analysis in pipeline development. This input includes the protein isoform ID, the genomic region for insertion, and the gene name (as found in UniProt).

Translated the novel mRNA into novel protein sequence

Analysis Outcome

To translate the novel mRNA into protein, the nucleotide sequences for the annotated exonic regions and the putative exonic regions were fetched from the UCSC Genome Browser. The batch analysis used only the first fifty gene identifiers in the data provided by Darby et al. This list of gene identifiers used can be found in Appendix A-“Initial_Gene_Ids.xlsx”.

Next, the translation in each frame for the flanking annotated exon (called exon 1) preceding the putative exon was determined using ExPasy Translate. This was followed by a blastp search through NCBI to determine which frame aligned with the protein annotation. Of the initial fifty gene identifiers, only those that were protein coding generated a hit from the blastp search (see Appendix A-“Filter_Protein_Coding_Gene_Ids”). The nucleotide sequences (annotated flanking exons and putative exon) were then pieced together and translated according the reading frame

identified for the translation of exon 1. This partial sequence was then inserted into the annotated sequence. The reading frame for the flanking exon following the putative exon (exon 2) was also determined. Appendix A- “exon_reading_frames.xlsx” displays the reading frames determined for the annotated exons. Also noted was if insertion of the putative exon caused a frameshift or not, and this can be found in Appendix A- “frameshift_determination_and_event_in_sequence_causes.xlsx”. In that table is also a notation as to whether the insertion caused a premature stop. Notation of a premature stop implies an isoform should be checked for NMD candidacy.

Conclusions for Single Isoform Analysis Tool Development

After generating the novel protein isoform sequences, it was concluded that in single isoform analysis, the following considerations should be made: 1) the reading frame for the appropriate flanking annotated exon that precedes the putative exon needs to be used for sequence translation, and 2) the tool needs to also report to the user if the insertion causes a frameshift, as well as if the insertion causes a premature stop.

In addition, Ensembl will be used instead of the UCSC Genome Browser in implementation due to its availability of exonic sequences for mRNA transcripts. In our case of development, we had the coordinates of the flanking exons, but in the case of the analysis tool this information will need to be determined based off of the protein isoform ID from UniProt that corresponds to a transcript in Ensembl.

Figure 8 illustrates all of the considerations mentioned above, and how to appropriately generate the novel protein isoform sequence in the single isoform analysis tool.

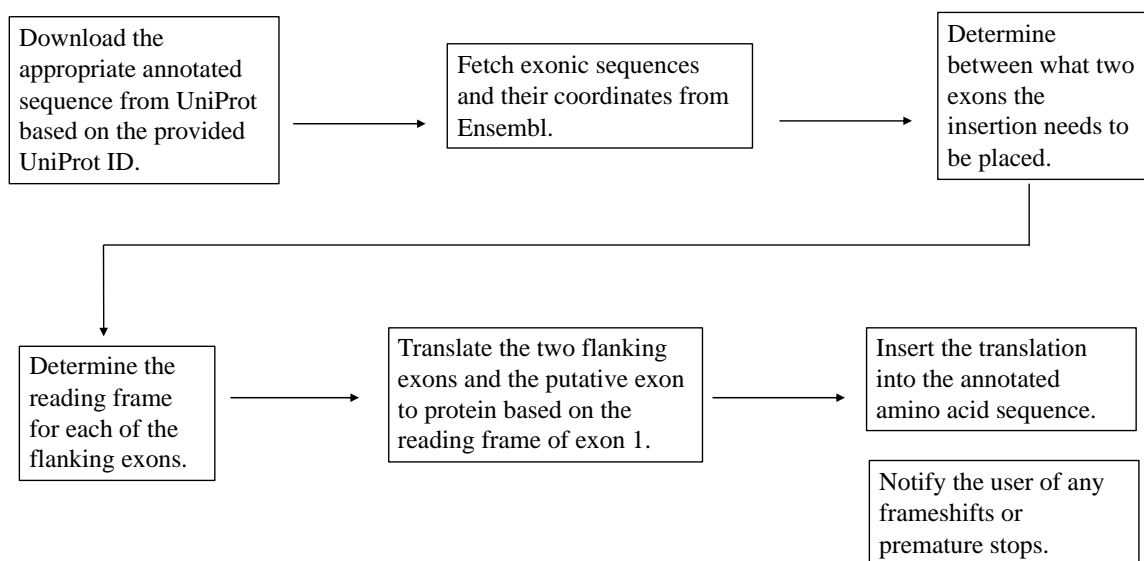


Figure 8. Process followed upon user input for the single isoform analysis. This process includes determination of the reading frames for translation of each flanking exon, followed by translation of the first flanking exon, putative exon, and second flanking exon together, and its insertion in the annotated protein sequence.

Checked the protein isoform sequence for nonsense-mediated decay (NMD)

Analysis Outcome

Determining if the novel protein isoforms were potential subjects for NMD began with the use of JuncDB, an exon-exon junction database. This database provides information using queries including gene identifiers, transcripts, or protein names. It provides intron positions across a variety of species, as well as exon lengths (Chorev et al. 2016). Using JuncDB, one is able to take the final intron position as the final exon-exon junction, which is then used in determination of an mRNA in undergoing NMD. Gene identifiers or protein names are the two query types used to search for the junction information. Queries that did not return a hit using the gene identifier or protein name

were notated in the csv file entitled “intron_positions_for_nmd.csv”, found in Appendix A. These data entries were omitted for the remainder of the project, due to the desire to be consistent and use only one database for the exon-exon junction information in the batch analysis.

After determining the final exon-exon junction for each of the proteins, a table was generated that would be used as the input file for an R script that would determine NMD candidacy (this file can be found in Appendix A- “intron_position_for_nmd.csv”). Two files are output for the user by the script: 1) a table that tells the user if the isoform is an NMD candidate or not (Appendix A- “intron_positions_for_nmd_with_nmd_candidacy.csv”) and 2) a text file that lists each file input, how far away the premature stop codon is from the final exon-exon junction, and if the isoform is an NMD candidate or not (Appendix A- “nmd_information_output.txt”).

A table named “Determination_for_NMD_Testing.xlsx” provides the proteins to be checked for NMD potential based on the finding of a premature stop in the isoform sequence.

Conclusions for Single Isoform Analysis Tool Development

While JuncDB was sufficient to use in the batch analysis, someone using the single isoform analysis tool may not use the canonical protein sequence as the annotated sequence. If this is the case, and an isoform different from the canonical is being used, it will have different intron positions, and so the JuncDB information would not be sufficient. An alternative method to determining the final exon-exon junction position

instead of using JuncDB would be to take the final two exon coordinates from Ensembl, determine the final intron position, and use this as the final exon-exon boundary in helping to determine NMD candidacy in a given isoform (Figure 9).

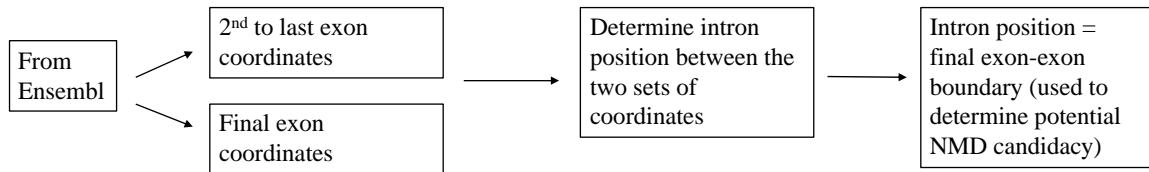


Figure 9. Alternative method to JuncDB to determine the final exon-exon junction. The user may be using an isoform as the annotated sequence rather than the canonical, and so an alternative to JuncDB would be to take the information from the Ensembl database.

Checked the proteins for loss and/ or gain of functional domains

Analysis Outcome

There are several tools currently available to determine functional domain assignments for the amino acid sequence of a given protein. Some of these tools include SUPERFAMILY, Motif Scan, SMART, RaptorX, and Pfam. These tools are similar in that they each take protein sequences as input, and generate domain assignments for each sequence. However, the method used to obtain the domain assignments differs between the tools. SUPERFAMILY utilizes Hidden Markov Models (HMM) to determine different domain groups based on the Structural Classification of Proteins (SCOP) database, using these to generate domain assignments for a given query (Gough et al. 2001). Motif Scan outputs domain assignments by extracting information from multiple annotation databases, such as Pfam, and creates an output of assignments from those different databases. SMART determines domain assignments through the use of manually curated Hidden Markov Models, based off of the integration of the combined

UniProt and Ensembl databases (Letunic and Bork 2018). Pfam also utilizes Hidden Markov Models, in addition to multiple sequence alignments, to generate domain assignments. Pfam entries are created using an alignment that consists of members within the same family, generating all detectable protein sequences that belong to the family (Sonnhammer 1998). These are used to determine domain assignments for a given protein query. RaptorX assigns domains through parsing based also off of Hidden Markov Models, and these models are generated from the Pfam database (Källberg et al. 2012).

There was a lot of research done to determine the best tool to use for identifying the protein domains of both the annotated proteins and the novel protein isoforms. Several tools were used in this determination, and included RaptorX, SMART, Motif Scan and SUPERFAMILY. SUPERFAMILY was the most-user friendly, and was also capable of domain identification for batches of sequences, including those not yet annotated. SUPERFAMILY utilizes Amazon Web Services (AWS), which allowed for ease of submitting sequences for domain assignments, and it also generated an easy-to-read and use file output. Below is a brief rationale for the choosing of the SUPERFAMILY domain assignment tool over the others under consideration:

- RaptorX did not generate the names of the domains. For the pipeline it was desired to highlight specific domain names that could be determined as being lost or gained.
- SMART would not allow for a batch search when non-annotated sequences were being passed into a query.

- Motif Scan generated domains by extracting information from a series of domain assignment tools- it wasn't generating domain assignments as part of the tool itself. This output included too many possibilities that would be overwhelming for a user.

The proteins used to develop the pipeline were run through a script written to determine the loss and/or gain of functional protein domains. This script used slightly modified files output from AWS, and split them up according to protein name, placing the annotated domains and the isoform domains into the same file for each individual protein. The second portion of the script determined the loss and/or gain of domains. There were two files output for the user to view: 1) a table that notifies the user if there were domains identified in the annotated sequence to begin with (Appendix A- "df_canonical_initially.csv") and 2) a text file that highlights the name of the file being reviewed, domains lost and/or gained, and the number of each domain lost and/or gained (Appendix A- "domain_information_output.txt"). The output files include the potential NMD candidates, due to the determination of the loss and/or gain of domains before checking for NMD. This is important since not every NMD candidate will actually undergo NMD.

A table named "Remaining_After_NMD_Testing.xlsx" contains the proteins remaining after evaluating NMD candidacy (Appendix A). These are the proteins that are most likely to be translated multiple times.

Conclusions for Single Isoform Analysis Tool Development

After determining what proteins in the batch analysis experienced a loss and/or gain of functional protein domains, the following conclusions were drawn when implementing the pipeline into the single isoform analysis tool:

- 1) The user will be made aware if the annotated sequence received domain assignments or not. This is important because if no domains were identified in the annotated sequence, it would not be legitimate to convey that any domains were lost and/or gained in the novel protein isoform.
- 2) In addition to the file(s) generated by the script, the following should be viewable for the user:
 - Table comparing the domains in both the annotated protein and the novel isoform should be included
 - Output statement(s) of what domains were lost and/or gained (and the number of each)

Analysis of proteins that have lost and/or gained functional domains by generating 3D models of the annotated proteins and novel protein isoforms to facilitate potential understanding of functional impact

Rationale

The expression of novel protein isoforms that have experienced a loss and/or gain of functional domains could alter a biological pathway in several ways. Expression could cause a change in the overall function and activity of the domains, causing changes in characteristics of the protein, such as binding affinity. Therefore, it is important to model predicted 3D models of the novel protein isoforms to facilitate an understanding of its functional impact. Through observation of a 3D model, a researcher is able to: 1) visualize any loss and/or gain of domains relative to the globular structure of the protein, allowing the user to make a judgement as to potential functional impact of this isoform and 2) expand potential understanding of protein isoforms in relation to an annotated sequence.

Analysis Design

The following is a list of steps completed to produce 3D models of both the annotated protein and the novel protein isoform.

1. Selected the proteins that have passed through the user-specified criteria. Criteria used for pipeline development included:
 - Lack of NMD
 - Loss and/or gain of at least one functional protein domain
2. Generated 3D models of the annotated proteins and the novel protein isoforms.

- 2a) Determined the best tool to use for 3D model visualization.
- 2b) Used the tool determined in (2a) to generate the 3D models for the proteins that passed through the specified criteria.

Select the proteins that have passed through the user-specified criteria

Analysis Outcome

The criteria used to develop the pipeline included a lack of NMD, as well as the loss or gain of at least one functional protein domain. After determining which proteins produced from the initial fifty gene identifiers passed these criteria, two proteins were to move forward in further analysis (Table 2).

Table 2. Proteins selected for further analysis. The two proteins to be used in further analysis that involved 3D modeling, identified by their gene ID and protein name, included NEUR3 and KLDC1.

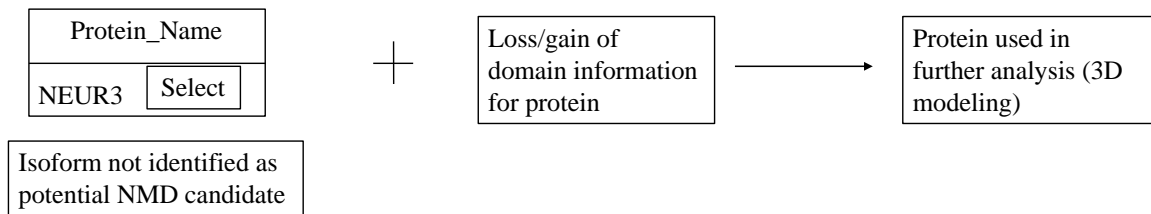
Gene_ID	Protein_Name
NEU3	NEUR3
KLHDC1	KLDC1

Conclusions for Single Isoform Analysis Tool Development

Determining our own criteria for pipeline development allowed for the conclusion that when implementing the pipeline for use in single isoform analysis, the user needs to be responsible for determining what their criteria is for selecting proteins to move

forward in analysis. A user would select the protein for further analysis if it passed their criteria regarding the loss and/or gain of functional domains (Figure 10).

The lack of NMD should be used as one of the default criteria, as the potential of the isoform to undergo NMD is important in determining if the protein is likely to be produced more than once per transcript. For example, if the novel transcript undergoes NMD and the annotated one does not, it could lead to decreased functioning protein concentration. If there were 100% of a type of annotated transcript in normal tissue and 50% annotated transcripts/50% novel transcripts (that undergo NMD) in samples of psychiatric disease, there would be only 50% of the normal functioning protein that comes from the mRNA.



the user in generating the model. We wanted the models provided to the user to be generated with ease and convenience, not making the user feel as though they needed to be experts in 3D modeling. In addition, Pymol required a paid membership after a tier of services. The pipeline is not being developed to use something that someone would have to pay for. One point of the pipeline is to demonstrate that availability of information already at one's disposal within several databases that can be used to achieve a common goal- in this case, the potential functional impact of novel protein isoforms.

SERpServer (Surface Entropy Reduction) predicts residues/groups of residues that may be subject to mutation within the protein. The output does not generate any sort of 3D model- it is simply a series of residue ranges, the residues, and scores in reference to how likely the mutation is to occur.

RaptorX and Phyre2 are two of the main template-based modeling tools used in research today. They both use remote homology recognition techniques, and so are able to still generate models when those based only on homology modeling are not.

Phyre2 carries out its homology detection method using alignments of Hidden Markov Models (HMM) through HHSearch, which is a software that generates HMM-HMM alignments in order to perform protein sequence searching (Kelley et al. 2015).

RaptorX utilizes the homology detection method, MRFAAlign, which is able to model long-range residue interaction for a given protein (Ma et al. 2014). This method is characterized as being able to be more sensitive to homology detection than HMM-HMM alignments, generating greater alignment accuracy and remote homology detection.

RaptorX also developed a nonlinear scoring function, which combines homologous information determined with structural information to determine the most appropriate 3D

model (Källberg et al. 2012). For example, those sequences that have few homologs have a “sparse sequence profile”, and so more weight will be placed on structural information when determining the most accurate alignment between the target sequence and sequences of the template structures that have been determined (Peng and Xu 2011).

Both Phyre2 and RaptorX conduct multiple-template protein threading, which is the prediction of a protein structure using more than one template structure. However, in Critical Assessment of Structure Prediction (CASP) experiments, which are community-wide experiments used to compare and continue the advancement of protein modeling, RaptorX was found to outperform Phyre2 in multiple-template threading (Peng and Xu 2011). RaptorX was ranked second overall in the CASP9 experiments, which evaluated several individual homology modeling/threading programs, and was ranked first in alignments generated for “hard” targets (Källberg et al. 2012). In more recent evaluation of the best modeling tools, Continuous Automated Model EvaluatiOn (CAMEO) has tested ~40 web servers, including RaptorX and Phyre2 (Xu 2018). CAMEO has shown that RaptorX outperformed Phyre2 in the case of easy, medium, and hard protein model targets (Haas et al. 2013; Haas et al. 2018; Haas et al. 2019)

In addition, in the results output for both of these tools, the domain information provided by RaptorX is of greater quality than that of Phyre2. RaptorX carries out domain parsing, which then allows for the visualization of each domain identified, in addition to the numbering of the domains identified, what residues aligned to the domain assignment, and the E-value for the domain assignment. Phyre2’s output contains a domain analysis, but only includes which amino acid residues at the domain coincides with, and does not always clearly state the domain identified. Since we are heavily

interested in the loss and/or gain of functional protein domains to drive further analysis and potential functional impact, we want as much information regarding the domains and their modeling in relation to the overall model generated as possible.

Based on the result of the research completed above, RaptorX was chosen as the tool of choice to generate potential 3D models of the novel protein isoform and its respective annotated sequence to which it was being compared.

The following figures are the RaptorX 3D models generated for the two proteins (NEUR3 and KLDC1) that passed both of the criteria being used for pipeline development (Figure 11 and Figure 12).



Figure 11. RaptorX output of NEUR3. The 3D model of the NEUR3 annotated protein and the novel protein isoform of NEUR3 predicted using RaptorX.

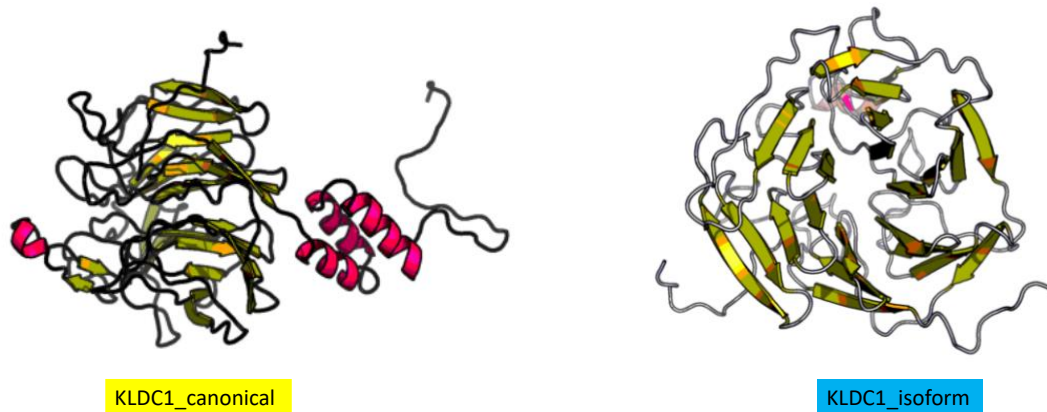


Figure 12. RaptorX output of KLDC1. The 3D model of the KLDC1 annotated protein and the novel protein isoform of KLDC1 predicted using RaptorX.

In both instances of the 3D models generated, it can be observed that there are visual differences in the annotated models and the isoform models. These visual distinctions between the proteins could be made by the user, and from there they could conclude that expression of the isoform could have potential functional impact.

It is important to note that the pipeline is not intended to provide the user with the specific functional impact of the novel protein isoform. The generation of 3D models is intended to provide the user with a visual as to the isoform versus the annotated protein that they are comparing it to, keeping in mind the domain information that was previously provided. It is up to the user to draw conclusions as to the specific functional impact that may be a result of the isoform being created.

However, there is discrepancy sometimes when RaptorX identifies domains and generates visualizations for them, compared to the domain assignments generated using the SUPERFAMILY domain assignment tool. While RaptorX is a good visualization tool, the domains identified by RaptorX differ for both proteins KLDC1 and NEUR3 compared to the domains identified by SUPERFAMILY. Figure 13 illustrates the

differences in the domain assignments given by the SUPERFAMILY tool, versus the RaptorX tool for both proteins.

KLDC1 Domains					NEUR3 Domains					
	Protein_name	Domains	Match_region	E.value		Protein_name	Domains	Match_region	E.value	
SUPERFAMILY	1	KLDC1_canonical	Kelch motif	13-163	2.09E-27	1	NEUR3_canonical	Sialidases	13-289	5.05E-98
	2	KLDC1_canonical	Kelch motif	168-333	3.53E-33	2	NEUR3_canonical	Sialidases	319-407	5.05E-98
	3	KLDC1_isoform	Galactose oxidase, central domain	10-234	7.32E-34	3	NEUR3_isoform	Sialidases	13-65	1.37E-08
RaptorX	1	KLDC1_canonical	1	1-340	5.3e-14	1	NEUR3_canonical	1	13-289	5.42E-15
	2	KLDC1_canonical	2	341-406	5.44e-03	2	NEUR3_isoform	2	1-109	9.30E-08
	3	KLDC1_isoform	1	1-348	1.83e-15	3	NEUR3_isoform	2	110-163	NA

Figure 13. Comparison of the domain assignments generated by the SUPERFAMILY versus the RaptorX. It can be observed that at times the domain assignments are not the same, leading to confusion as to what domain assignments are more appropriate.

Conclusions for Single Isoform Analysis Tool Development

The discrepancy between the domains assigned and modeled by RaptorX versus the domains determined by the SUPERFAMILY tool could lead to much confusion for the user. An alternative to SUPERFAMILY in determining domains and their loss or gain is the Pfam database. Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models- similar to the SUPERFAMILY tool (Sonnhammer 1998). RaptorX uses the Pfam database to carry out domain-parsing and further 3D modeling (Källberg et al. 2012). The use of Pfam may allow for better consistency between domain identification and visualization of the domains, and so it will be used in pipeline implementation in the single isoform analysis tool.

While it was originally thought that the 3D models generated by RaptorX would be sufficient for model prediction, it was determined that in the single isoform analysis tool, additional information regarding the models would be necessary. RaptorX generates information about 2D structure, as well as the percentage of residues modeled with structure, and the percentage of residues modeled with disorder. All of this information could be helpful to the user in determining the biological functional impact of the novel protein isoform, and so in the single isoform analysis tool the job URL generated by RaptorX will be provided to the user so that they may use the information provided to them at their leisure. Figures 14- 17 illustrate the example output from RaptorX.

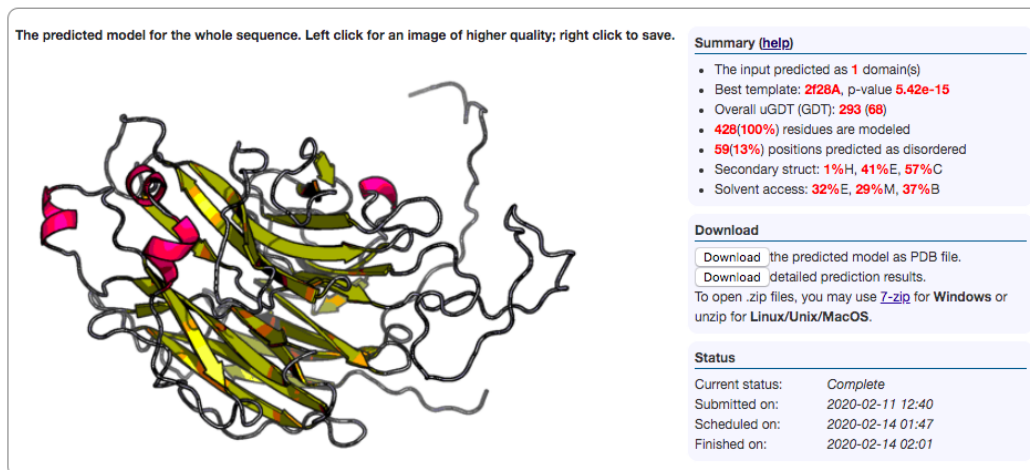


Figure 14. Example of RaptorX Output for entire protein model and summary information. At the top of the job page, the predicted 3D model of the protein is provided, in addition to a short summary of information regarding the model.

Summary [\(help\)](#)

- The input predicted as **1** domain(s)
- Best template: **2f28A**, p-value **5.42e-15**
- Overall uGDT (GDT): **293 (68)**
- **428(100%)** residues are modeled
- **59(13%)** positions predicted as disordered
- Secondary struct: **1%H, 41%E, 57%C**
- Solvent access: **32%E, 29%M, 37%B**

Figure 15. Summary information regarding the predicted 3D protein model. The summary of information includes the template best suited to help generate the model, residues modeled, and overall percentages of secondary structure.

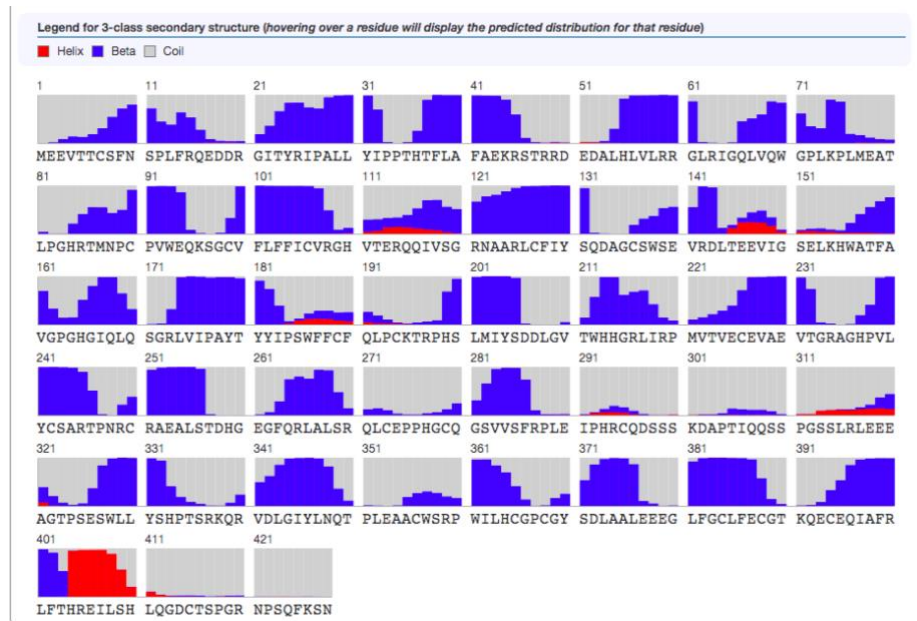


Figure 16. Secondary structure information for predicted 3D model. Example output from RaptorX that includes the secondary structure assignment for each residue in the submitted sequence.

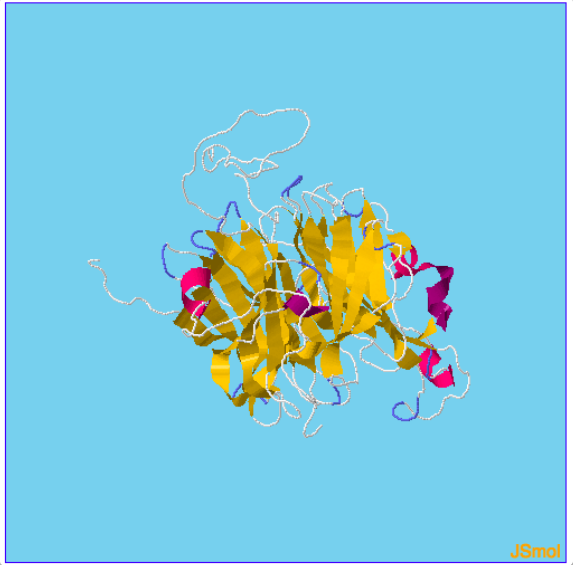
[-] Click to view 3D model(s) for domain 1 [1, 428] P-value:5.42e-15

View Alternative Model Template: [2f28:A](#)

Rank: 1 **P-value:** 5.42e-15 **uGDT(GDT):** 294(69) **uSeqId(SeqId):** 153(36) **Score:** 267

Jsmol viewer quick guide

- Left-click+drag to rotate the structure.
- Use the middle-scroller to zoom.
- Right-click the structure for more options.
- Hover over a target residue in the sequence alignment box to see it highlighted in the structure.
- Visit the [Jmol mouse manual wiki](#)



Rotation

spin on/off

fast

slow

Coloring

color AAs

color SSs

Display

side chain

back bone

cartoon

Zoom

Show Quality

low

high

Save

Figure 17. Visualization of domain-parsing in RaptorX output. Each of the individual domains identified in the 3D model predicted are able to be viewed and manipulated by the user.

Differential expression analysis of the putative exons at the RNA level to investigate disease-specificity of expression and potential changes in expression of the novel isoform in disease

Rationale

Completing a differential expression (DE) analysis of the putative exons could allow for implication of an isoform in a specific disease. Such an implication could lead to a broadened understanding of diseases found in the human population.

Analysis Design

The following process was devised to achieve the completion of a differential expression (DE) analysis:

- 1) Determined the best way to include only regions of interest (putative exons) in the DE analysis.
- 2) Conducted DE analysis using each study in the recount2 database that contained samples of interest.

Determined the best way to include only regions of interest in the DE analysis

Analysis Outcome

To make the pipeline most efficient, it was deemed necessary to create a GRanges object that would then be used as input for the differential expression analysis. A GRanges object stores genomic information that is easy to access for different purposes of analysis. In our case, the object allows for specification as to the exact

coordinates of the putative exons that wish to be searched for differential expression. The information necessary to create a GRanges object includes seqnames (i.e. chromosome number), strand (+/-), and ranges for the coordinates (start/end position).

Conclusions for Single Isoform Analysis Tool Development

As previously mentioned, the user input for the tool would include the gene name (as seen in UniProt), the UniProt ID, and the genomic region for insertion. The information input for the genomic insertion is what will be used to construct the GRanges object in the implemented tool for single isoform analysis (Figure 18).

Protein_Isoform_ID ▼	Gene_Name ▼	Genomic_coordinates_for_insertion	
Q13303	KCAB2		Chromosome_Number:
			1
		Start_Position:	
		6145799	
		End_Position:	
		6145853	

Necessary information to construct GRanges object:

- Seqnames- determined by chromosome number entered
- Ranges (start/end)- determined by Start_Position and End_Position
- Strand- determined by if start or end position is greater
 - Start < End: + strand
 - End < Start: - strand

Figure 18. Genomic coordinate information used to construct Granges object for DE analysis. The user input of the genomic coordinates will be used to construct the necessary information for creating a GRanges object: 1) Seqnames, 2) Ranges, and 3) Strand.

Conducted DE analysis using each study in the recount2 database that contained samples of interest

Conducting the differential expression analysis required the use of the individual BigWig files for each sample in each of the recount2 studies. A BigWig file contains the output from the Rail align process, where the sequencing reads are aligned to the genome. Recount2 studies contain a mean BigWig for all samples, as well as individual BigWig files each sample in a study. Because we wanted to investigate differential expression between the control and disease groups, we needed to obtain counts for the regions of interest for each sample. Using the mean BigWig file would have meant using a combined mean of the counts for the regions of the genome that included both the control and disease groups.

The following studies from the recount2 database were used in the differential expression analysis:

- SRP035524- This study includes 20 schizophrenia samples, 24 control samples, and 28 bipolar disorder samples. Obtained from the Southwest Brain Bank, samples originate from the frontal cortex and the anterior cingulate gyrus of the brain (Xiao et al. 2014).
- ERP001304- This study includes 9 control samples and 9 schizophrenia samples. Samples originate from the superior temporal gyrus of the brain, and come from only male subjects (Wu et al. 2012).
- SRP033725- This study includes 31 bipolar samples, and 31 control samples. Samples are from the dorsolateral prefrontal cortex of subjects, with matching of age and gender between the disease samples and the controls (Akula et al. 2014).

- SRP043684- This study includes 11 bipolar samples and 8 control samples obtained from the Stanley Medical Research Institute. Samples are from the hippocampal dentate gyrus (DG) granule neurons. Subjects of the study were all white males who were patients in clinical trials for lithium ion monotherapy at the University of California, San Diego, and the Pharmacogenomics of Bipolar Disorder Study (The Pharmacogenomics of Bipolar Disorder Study et al. 2015).

Analysis Outcome

Upon completion of the differential expression analysis using these four studies, differential expression of a series of putative exons was found in study SRP033725, a study involving bipolar disorder. The following are a series of counts plots generated illustrating the differences in expression of the putative region in control samples versus bipolar samples (Figures 19-22).

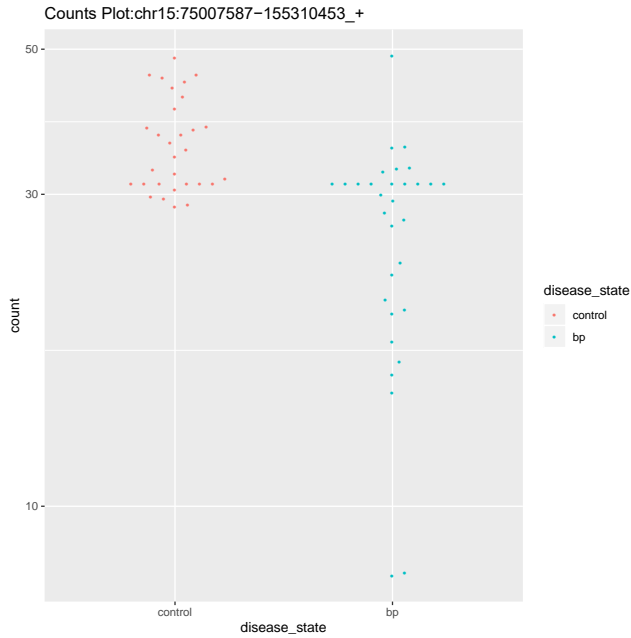


Figure 19. Counts plot for putative exon: chr15:75007587-75007728.

It can be seen that there are bipolar disorder samples from the dorsolateral prefrontal cortex exhibiting lower counts of expression compared to the normal samples. This putative exon associates with the gene SCAMP5/protein SCAM5.

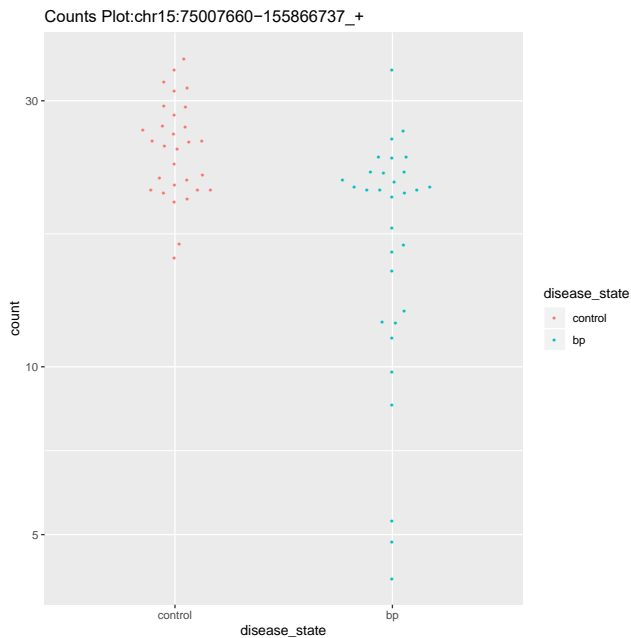


Figure 20. Counts plot for putative exon: chr15:75007660-155866737.

It can be seen that there are bipolar disorder samples from the dorsolateral prefrontal cortex exhibiting lower counts of expression compared to the normal samples. This putative exon associates with the gene SCAMP5/protein SCAM5.

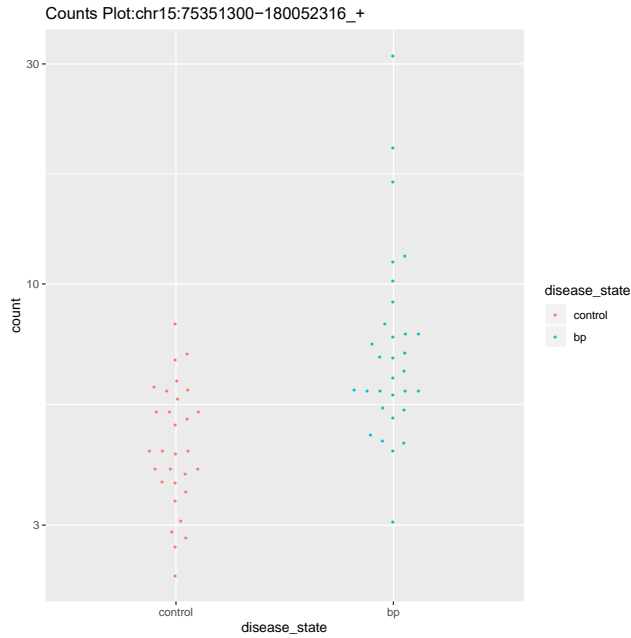


Figure 21. Counts plot for putative exon: chr15:75351300-1800523216. It can be seen that there are bipolar disorder samples from the dorsolateral prefrontal cortex exhibiting higher counts of expression compared to the normal samples. This putative exon associates with the gene NEIL1/protein NEIL1.

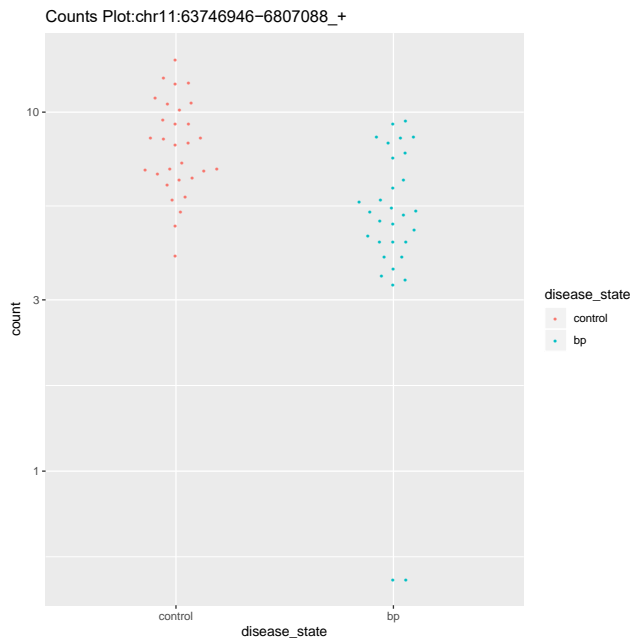


Figure 22. Counts plot for putative exon: chr11: 63746946- 6807088. It can be seen that there are bipolar disorder samples from the dorsolateral prefrontal cortex exhibiting lower counts of expression compared to the normal samples. This putative exon associates with the gene RTN3/protein RTN3.

Differences in expression of the novel putative exons could imply several things in the case of disease development in an individual. When the putative exon displays lower counts in some of the disease samples, it may lead to the thought that the isoform needed for a normal state is not being generated in the abundance that it needs to be. In contrast, when the putative exon displays higher counts in some of the disease samples, it could imply that an isoform implicated in disease state is being generated in an amount that leads to a diseased state/disease development. This sort of information that can be gained from carrying out a differential expression analysis of these putative exons is very vital to the continued expansion of knowledge regarding disease development and its relation to protein isoform expression.

Conclusions for Single Isoform Analysis Tool Development

Carrying out the DE analysis allowed for the determination that there needs to be a fourth input from the user when implementing this pipeline into a single isoform analysis tool. This includes the recount2 project ID for the study they desire to use for the DE analysis (Figure 23). An additional output to the user from running the analysis through the tool will also be notifying them of a lack of differential expression found (should that be the case).

Protein_Isoform_ID▼ Q13303	Gene_Name ▼ KCAB2	Genomic_coordinates_for_insertion Chromosome_Number: 1 Start_Position: 6145799 End_Position: 6145853
Recount2_project_ID▼ SRP033725		

Figure 23. Additional user input for the single isoform analysis tool. An additional user input for the pipeline tool upon its implementation to a single isoform analysis tool includes recount2 project ID. This is necessary in order to run the differential expression analysis.

DISCUSSION

While there is currently data in the scientific world identifying novel expressed regions in the human genome, there is not an efficient way to go about determining the impact of their expression. The purpose for the development of the bioinformatics pipeline described was to allow for the assessment of potential functional impact of novel protein isoforms generated from novel transcripts that showed evidence of coding potential. This is an important step in helping to aid in the evaluation of the impact of novel expressed regions that generate novel mRNAs.

Developing the pipeline consisted of two overall goals: 1) Conduct an analysis of novel mRNAs identified by Darby et al. through a batch-type in order to assess all considerations that must be taken when implementing the pipeline into a single isoform analysis tool. 2) Draw conclusions from the batch-analysis that will be used in future implementation of the pipeline into a user-friendly single isoform analysis tool. Both of these goals allowed for the purpose of the pipeline states above to be achieved. Figure 24 highlights the integration between the pipeline that was developed with the future single isoform analysis tool, and the output that will be generated for the user.

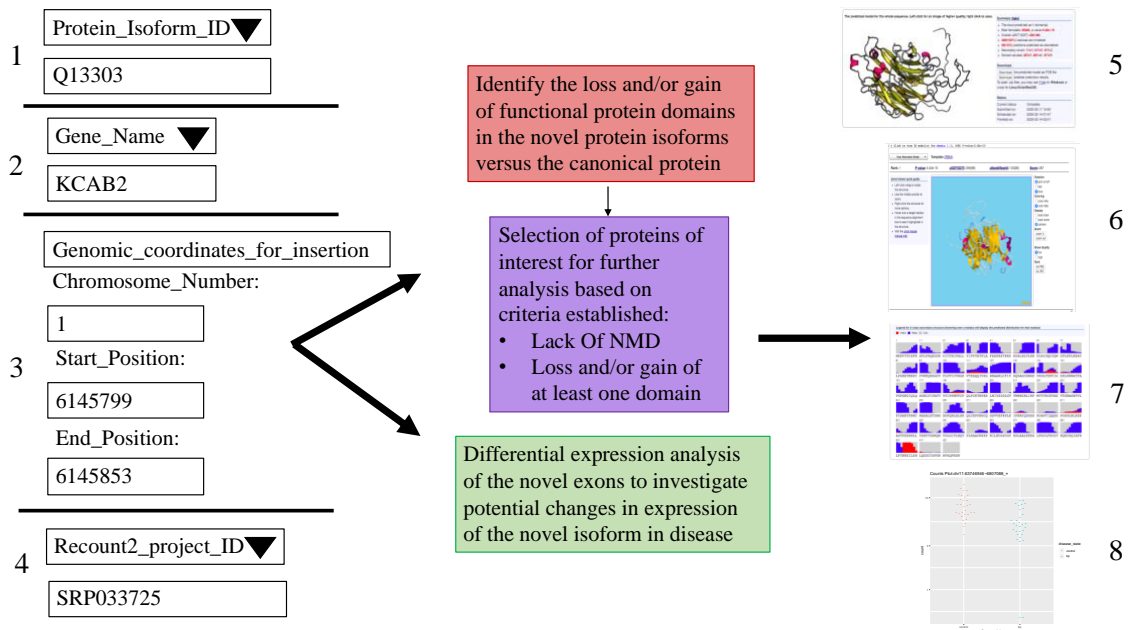


Figure 24. Integration of pipeline developed with future single isoform analysis tool. The user input for the single isoform analysis tool takes in four inputs (noted 1, 2, 3 and 4). These four inputs are then used in executing the pipeline that was developed. Execution of the pipeline leads to output from RaptorX (denoted by 5, 6, and 7), as well as counts plots for regions differentially expressed that are significant (8).

Achieving these two goals provided much insight as to current limitations in the field of biological research. For starters, there is not a consistent naming schema for genes in biology. Alternative gene names can cause confusion, as well as create potential to miss an important result due to a difference in names being used in an experimental process. This naming inconsistency demonstrated the need for specificity in user input required.

Another limitation is multiple isoform annotations for a given protein. While it is beneficial to have multiple isoforms annotations available, it requires greater consideration when determining which isoform to select for an analysis or experiment at the bench. Therefore, in the case of pipeline development and considerations to be taken

for the single isoform analysis tool, it was determined that the user would need to specify the unique UniProt entry ID for the isoform of interest.

The domain assignment tools and 3D visualization tools do not always agree in their output, generating a discrepancy between the two that can lead to confusion and a lack of validity regarding the isoform and its functional protein domains or overall structure. Each tool, whether it be for domain assignment or 3D modeling, uses different methods and algorithms that often generate slightly different output. In this case, a different method was being used to determine the domains present in the given protein in SUPERFAMILY compared to RaptorX. This inconsistency illustrates the need for careful consideration in tool selection, including its relation to other tools similar to it or performing a downstream action.

3D protein modeling is a finicky process, for it is very difficult to predict the structure of a protein from only other modeled proteins. While there are several modeling tools available and they continue to improve over time, one limitation in the field of research is the lack of confidence regarding predicted 3D protein structures. Such a lack in confidence caused the conclusion to be drawn that the most extensive information provided with the output 3D structure of a given protein should be included in the single isoform analysis tool. Providing this information would allow for an increase in confidence of the 3D structure generated, or the providing of information to the user that could be evaluated in a way that they deemed appropriate.

While the limitations on the broader level of research were highlighted through the development of this pipeline, the pipeline itself also displayed its own limitations. The biggest limitation of the pipeline includes the fact that a protein removed from

further analysis when it does not meet the user-specified criteria could still be significant in some way cellularly. Its removal from further analysis does not constitute it being deemed insignificant. The point of developing the pipeline is to help the user rationalize and/or prioritize what protein they might pursue in further research, specifically at the lab bench. There were examples of this case in developing the pipeline using the first fifty gene identifiers in the work of Darby et al. The first includes gene identifiers KCNAB2 (protein KCAB2) and WDTC1 (protein WDTC1), which lacked NMD, but did not lose or gain at least one functional protein domain. It is important to note that the expression of these isoforms is something that could be further analyzed in research, even though they did not meet the user-specified criteria in all capacities. These isoforms are being produced for a reason, though unknown at this time. They are still different from the annotated protein that they are being compared to due to the putative exon. This difference could be a shift in residues of the protein that were once on the outside of or inside of the globular structure. Another difference could be that a domain that was once on the outside of the globular structure is now more inward. Such shifts then cause downstream effects regarding binding of molecules to the protein (i.e. if it is a receptor), or interaction of the protein with other proteins.

Examples of gene identifiers with proteins that did lose and /or gain at least one domain but were determined to be potential NMD candidates include CMTA1 (protein CMTA1) and SYT11 (protein SYT11). Both of these particular novel isoforms lost at least two functional protein domains when being compared to the annotated protein. This loss of domains could still be significant at the cellular level, because even in the case of

NMD the protein would still be made once, and there could still be an impact even in generating this protein with a loss-of-function even one time.

An additional limitation to the pipeline includes the domain assignments of the annotated protein sequence. If there are no functional protein domains identified in the annotated sequence, there is not reported to be any loss and/or gain of domains in the novel protein isoform, since there were no domains identified in the initial protein. The tool being used for the domain assignments could be missing a domain that is in fact present in the annotated protein, therefore causing the identification of a loss and/or gain of functional protein domains to be missed when comparing the annotated protein to the novel isoform. As a result, this could lead to a protein not being selected for further analysis involving 3D modeling.

The approach taken to develop the pipeline considered the putative exon to be an insertion between two annotated, flanking exons. It would be important to also determine the potential functional impact of putative exons upstream or downstream of annotated genes. This could be an addition made to the single isoform tool in the future, where the user denotes the location of the putative exon insertion (upstream, downstream, or intronic) to an annotated gene.

In the future, additional batch analyses using the pipeline could be conducted. These analyses would use novel expressed regions with evidence of splice junctions as input. This information is currently being determined in research by Ms. Bianca Hoch.

An additional future direction would include extension of the differential expression analysis to include tissue-specificity. At this point the pipeline was developed to determine differential expression on the level of disease-specificity. Extension to

tissue-specificity would allow for implications as to if the putative exon is being regulated (due to differences in expression among different tissues), strengthening the idea of it being functionally important.

The main point of future direction regarding this pipeline is its implementation into a user-friendly single isoform analysis tool. This tool will be created through a collaboration between myself, Dr. Darby, Ms. Bianca Hoch, and Mr. Conor Jenkins, and allow for the beginnings of a way to determine the potential functional impact of novel protein isoforms.

REFERENCES

- Akula N, Barb J, Jiang X, Wendland JR, Choi KH, Sen SK, Hou L, Chen DTW, Laje G, Johnson K, et al. 2014. RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. *Mol Psychiatry*. 19(11):1179–1185. doi:10.1038/mp.2013.170.
- Barmak Modrek, Christopher Lee. 2002. A genomic view of alternative splicing. *Nat Genet*. 30:13–19.
- Bartonek L, Braun D, Zagrovic B. 2020. Frameshifting preserves key physicochemical properties of proteins. *Proc Natl Acad Sci*. 117(11):5907–5912. doi:10.1073/pnas.1911203117.
- Chorev M, Guy L, Carmel L. 2016. JuncDB: an exon–exon junction database. *Nucleic Acids Res*. 44(D1):D101–D109. doi:10.1093/nar/gkv1142.
- Collado-Torres L, Nellore A, Kammers K, Ellis S, Taub M, Hansen K, Jaffe A, Langmead B, Leek J. 2017. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol*. 35:319–321. doi:https://doi.org/10.1038/nbt.3838.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 17(1):13. doi:10.1186/s13059-016-0881-8.
- Darby M. 2019. NSF CAREER Project Description.
- Darby MM, Leek JT, Langmead B, Yolken RH, Sabunciyar S. 2016 Sep 20. Widespread splicing of repetitive element loci into coding regions of gene transcripts. *Hum Mol Genet*.:ddw321. doi:10.1093/hmg/ddw321.
- EMBL-EBI. 2020. Domains | EMBL-EBI Train online. [accessed 2020 Mar 31]. <https://www.ebi.ac.uk/training/online/course/biomacromolecular-structures-introduction-ebi-reso/proteins/domains>.
- EMBL-EBI. 2020. What are protein domains? | EMBL-EBI Train online. [accessed 2020 Mar 31]. <https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification/what-are-protein-domains>.
- European Molecular Biology Laboratory. 2019. UniProt. [accessed 2020 Mar 6]. <https://www.uniprot.org/>.
- Frankish A, Uszczyńska B, Ritchie GR, Gonzalez JM, Pervouchine D, Petryszak R, Mudge JM, Fonseca N, Brazma A, Guigo R, et al. 2015. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*. 16(S8):S2. doi:10.1186/1471-2164-16-S8-S2.

- Frazer AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT. 2014. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics*. 15(3):413–426. doi:10.1093/biostatistics/kxt053.
- GENCODE. 2019. GENCODE - Human Release Statistics. Stat Curr GENCODE Release Version 33. [accessed 2020 Mar 13]. <https://www.encodegenes.org/human/stats.html>.
- GeneCards. 2019. GeneCards - Human Genes | Gene Database | Gene Search. [accessed 2020 Mar 6]. <https://www.genecards.org/>.
- Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure - ScienceDirect. *J Mol Biol*. 31(4):903–919. doi:<https://doi.org/10.1006/jmbi.2001.5080>.
- Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, Mostaguir K, Gumienny R, Schwede T. 2018. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*. 86:387–398. doi:DOI: 10.1002/prot.25431].
- Haas J, Gumienny R, Barbato A, Ackermann F, Tauriello G, Bertoni M, Studer G, Smolinski A, Schwede T. 2019. Introducing “best single template” models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins*. 86(S1):1378–1387. doi:DOI: 10.1002/prot.25815.
- Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T. 2013. The Protein Model Portal - a comprehensive resource for protein structure and model information. Database.
- Harrow J, Frankish A, Gonzalez J, Tapanari E, Diekhans M, Kokocinski F, Aken B, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 22(9):1760–74. doi:10.1101/gr.135350.111.
- Hug N, Longman D, Cáceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res*. 44(4):1483–1495. doi:10.1093/nar/gkw010.
- Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc*. 7(8):1511–1522. doi:10.1038/nprot.2012.085.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 10(6):845–858. doi:10.1038/nprot.2015.053.
- Kurosaki T, Maquat LE. 2016. Nonsense-mediated mRNA decay in humans at a glance. *J Cell Sci*. 129(3):461–467. doi:10.1242/jcs.181008.

- Letunic I, Bork P. 2018. 20 Years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46. doi:10.1093/nar/gkx922.
- Ma J, Wang S, Wang Z, Xu J. 2014. MRAlign: Protein Homology Detection through Alignment of Markov Random Fields. Lengauer T, editor. *PLoS Comput Biol.* 10(3):e1003500. doi:10.1371/journal.pcbi.1003500.
- Nowacki M, Higgins B, Maquilan G, Swart E, Doak T, Landweber L. 2009. A Functional Role for Transposases in a Large Eukaryotic genome. *Science.* 324(5929):935. doi:10.1126/science.1170023.
- OpenLearn University. 2020. Proteins: 1.4.3 Protein domains - OpenLearn - Open University - S377_2. [accessed 2019 Nov 5]. <https://www.open.edu/openlearn/science-maths-technology/science/biology/proteins/content-section-1.4.3>.
- Peng J, Xu J. 2011. Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins Struct Funct Bioinforma.* 79(S10):161–171. doi:10.1002/prot.23175.
- Rachel D. 2018. Protein structure prediction. *Int J Mod Phys B.* 32:17. doi:10.1142/S021797921840009X.
- RefSeq. 2019. RefSeq growth statistics. [accessed 2020 Mar 11]. <https://www.ncbi.nlm.nih.gov/refseq/statistics/>.
- Sonnhammer E. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26(1):320–322. doi:10.1093/nar/26.1.320.
- Sultan M, Schulz MH, Klingenhof A, Scherf M, Richard H, Magen A. 2008. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science.* 321(5891):956–960. doi:10.1126/science.1160342.
- The Pharmacogenomics of Bipolar Disorder Study, Mertens J, Wang Q-W, Kim Y, Yu DX, Pham S, Yang B, Zheng Y, Diffenderfer KE, Zhang J, et al. 2015. Differential responses to lithium in hyperexcitable neurons from patients with bipolar disorder. *Nature.* 527(7576):95–99. doi:10.1038/nature15526.
- Wu JQ, Wang X, Beveridge NJ, Tooney PA, Scott RJ, Carr VJ, Cairns MJ. 2012. Transcriptome Sequencing Revealed Significant Alteration of Cortical Promoter Usage and Splicing in Schizophrenia. Zhang XY, editor. *PLoS ONE.* 7(4):e36351. doi:10.1371/journal.pone.0036351.
- Xiao Y, Camarillo C, Ping Y, Arana TB, Zhao H, Thompson PM, Xu Chaohan, Su BB, Fan H, Ordonez J, et al. 2014. The DNA Methylome and Transcriptome of Different Brain Regions in Schizophrenia and Bipolar Disorder. Liu C, editor. *PLoS ONE.* 9(4):e95875. doi:10.1371/journal.pone.0095875.

Xu J. 2018. Distance-based Protein Folding Powered by Deep Learning. Toyota Technol Inst Chic.:16.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across human tissues. *Genome Biol.*:15.

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*. 16(1):97. doi:10.1186/s12864-015-1308-8.

APPENDICES

Appendix A: Tables and text files referenced in the paper

The following appendix contains each table and/or text file referenced throughout the paper.

Determination_for_NMD_Testing.xlsx

Gene_ID	Protein_Name	Exon_1_frame	Exon_2_frame	putative_exon_causes_frameshift	event_caused_by_inclusion_of_putative_exon
KCNAB2	KCAB2		2	3 frameshift	insertion
CMTA1	CMTA1		1	2 no_frameshift	premature stop
WDTC1	WDTC1		1	2 no_frameshift	insertion
TMEM39B	TM39B		1	3 frameshift	premature stop
AGL	GDE		1	3 frameshift	insertion
FDP5	FPP5		2	2 no_frameshift	premature stop
SYT11	SYT11		3	1 no_frameshift	premature stop
CEP350	CE350		2	1 frameshift	premature stop
KCTD3	KCTD3		3	3 no_frameshift	premature stop
FAM149B1	F149B		2	1 no_frameshift	premature stop
INPP5F	SAC2		3	1 frameshift	premature stop
LHPP	LHPP			2	
TKFC	TKFC			2	
RTN3	RTN3		3	1 frameshift	premature stop
DPP3	DPP3		3	2 frameshift	premature stop
RBM4	RBM4		1	1 no_frameshift	premature stop
NEU3	NEUR3		3	2 frameshift	premature stop
AAMDC	AAMDC		1	1 no_frameshift	premature stop
ABCG4	ABCG4		1	1 no_frameshift	premature stop
DNM1L	DNM1L		1	2 frameshift	premature stop
DIP2B	DIP2B		1	3 frameshift	premature stop
DGKA	DGKA		3	1 frameshift	premature stop
SYT1	STY1				
PLBD2	PLBL2		2	2 frameshift	premature stop
GLT1D1	GL1D1		1	2 no_frameshift	insertion
BORA	BORA		1	1 frameshift	premature stop
CDC16	CDC16		1	1 no_frameshift	premature stop
LRFN5	LRFN5		3	3 frameshift	premature stop
KLHDC1	KLDC1		1	3 no_frameshift	premature stop
KTN1	KTN1		1	2 frameshift	premature stop
KLC1	KLC1				
CKMT1A	KCRU		2	2 no_frameshift	premature stop
GALK2	GALK2_A			3	
GALK2	GALK2_B			3	
SNX1	SNX1		1	1 no_frameshift	premature stop
HACD3	HACD3		3	2 frameshift	premature stop
SCAMP5	SCAM5_A			2	
SCAMP5	SCAM5_B			2	
NEIL1	NEIL1		2	3 no_frameshift	insertion
PHKB	KPBB		3	2 no_frameshift	premature stop

*only protein sequences (bolded) that showed a premature stop were tested for NMD- the table above includes all proteins
**additional proteins were removed when they did not obtain a hit for the flanking exons through NCBI blast (UTR regions)- highlighted

df_canonical_initially.csv

	Protein_name	Canonical_Domains_Initially
1	KCAB2	yes
2	CMTA1	yes
3	WDTC1	yes
4	TM39B	no
5	GDE	yes
6	FPPS	yes
7	SYT11	yes
8	CE350	yes
9	KCTD3	yes
10	F149B	no
11	SAC2	no
12	RTN3	no
13	DPP3	no
14	RBM4	yes
15	NEUR3	yes
16	AAMDC	yes
17	ABCG4	yes
18	DNM1L	yes
19	DIP2B	yes
20	DGKA	yes
21	PLBL2	no
22	GL1D1	yes
23	BORA	no
24	CDC16	yes
25	LRFN5	yes
26	KLDC1	yes
27	KTN1	no
28	KCRU	yes
29	SNX1	yes
30	HACD3	yes
31	NEIL1	yes
32	KPBB	yes

domain_information_output.txt (screenshot of part of output file)

```

domain_information_output.txt
KCAB2_Q13303_domains_canonical_isoform_compared.csv
No domains have been gained or lost in this novel isoform.
CMTA1_Q9Y6Y1_domains_canonical_isoform_compared.csv
Domains lost: Ankyrin repeat 1
Domains lost: E set domains 1
Domains lost: P-loop containing nucleoside triphosphate hydrolases 1
Domains gained:
WDTC1_Q8N5D0_domains_canonical_isoform_compared.csv
No domains have been gained or lost in this novel isoform.
TM39B_Q9GZU3_domains_canonical_isoform_compared.csv
There were no domains identified in the canonical sequence by the SUPERFAMILY tool.
No domains have been gained or lost in this novel isoform.
GDE_P35573_domains_canonical_isoform_compared.csv
No domains have been gained or lost in this novel isoform.
FPPS_P14324_domains_canonical_isoform_compared.csv
Domains lost: Terpenoid synthases 1
Domains gained:
SYT11_Q9BT88_domains_canonical_isoform_compared.csv
Domains lost: C2 domain (Calcium/lipid-binding domain, CaLB) 2
Domains gained:
CE350_Q5VT06_domains_canonical_isoform_compared.csv
Domains lost: Cap-Gly domain 1
Domains gained:
KCTD3_Q9Y597_domains_canonical_isoform_compared.csv
Domains lost: WD40 repeat-like 1
Domains gained:
F149B_Q96BN6_domains_canonical_isoform_compared.csv
There were no domains identified in the canonical sequence by the SUPERFAMILY tool.
No domains have been gained or lost in this novel isoform.
SAC2_Q9Y2H2_domains_canonical_isoform_compared.csv

```

exon_reading_frames.xlsx

Gene_ID	Protein_Name	Exon_1_frame	Exon_2_frame
KCNAB2	KCAB2	2	3
CMTA1	CMTA1	1	2
WDTC1	WDTC1	1	2
TMEM39B	TM39B	1	3
AGL	GDE	1	3
FDPS	FPPS	2	2
SYT11	SYT11	3	1
CEP350	CE350	2	1
KCTD3	KCTD3	3	3
FAM149B1	F149B	2	1
INPP5F	SAC2	3	1
LHPP	LHPP		2
TKFC	TKFC		2
RTN3	RTN3	3	1
DPP3	DPP3	3	2
RBM4	RBM4	1	1
NEU3	NEUR3	3	2
AAMDC	AAMDC	1	1
ABCG4	ABCG4	1	1
DNM1L	DNM1L	1	2
DIP2B	DIP2B	1	3
DGKA	DGKA	3	1
SYT1	STY1		
PLBD2	PLBL2	2	2
GLT1D1	GL1D1	1	2
BORA	BORA	1	1
CDC16	CDC16	1	1
LRFN5	LRFN5	3	3
KLHDC1	KLDC1	1	3
KTN1	KTN1	1	2
KLC1	KLC1		
CKMT1A	KCRU	2	2
GALK2	GALK2_A		3
GALK2	GALK2_B		3
SNX1	SNX1	1	1
HACD3	HACD3	3	2
SCAMP5	SCAM5_A		2
SCAMP5	SCAM5_B		2
NEIL1	NEIL1	2	3
PHKB	KPBB	3	2

frameshift_determination_and_event_in_sequence_causes.xlsx

Gene_ID	Protein_Name	Exon_1_frame	Exon_2_frame	putative_exon_causes_frameshift	event_caused_by_inclusion_of_putative_exon
KCNAB2	KCAB2		2	3 frameshift	insertion
CMTA1	CMTA1		1	2 no_frameshift	premature stop
WDTC1	WDTC1		1	2 no_frameshift	insertion
TMEM39B	TM39B		1	3 frameshift	premature stop
AGL	GDE		1	3 frameshift	insertion
FDPS	FPPS		2	2 no_frameshift	premature stop
SYT11	SYT11		3	1 no_frameshift	premature stop
CEP350	CE350		2	1 frameshift	premature stop
KCTD3	KCTD3		3	3 no_frameshift	premature stop
FAM149B1	F149B		2	1 no_frameshift	premature stop
INPP5F	SAC2		3	1 frameshift	premature stop
LHPP	LHPP			2	
TKFC	TKFC			2	
RTN3	RTN3		3	1 frameshift	premature stop
DPP3	DPP3		3	2 frameshift	premature stop
RBM4	RBM4		1	1 no_frameshift	premature stop
NEU3	NEUR3		3	2 frameshift	premature stop
AAMDC	AAMDC		1	1 no_frameshift	premature stop
ABCG4	ABCG4		1	1 no_frameshift	premature stop
DNM1L	DNM1L		1	2 frameshift	premature stop
DIP2B	DIP2B		1	3 frameshift	premature stop
DGKA	DGKA		3	1 frameshift	premature stop
SYT1	STY1				
PLBD2	PLBL2		2	2 frameshift	premature stop
GLT1D1	GL1D1		1	2 no_frameshift	insertion
BORA	BORA		1	1 frameshift	premature stop
CDC16	CDC16		1	1 no_frameshift	premature stop
LRFN5	LRFN5		3	3 frameshift	premature stop
KLHDC1	KLDC1		1	3 no_frameshift	premature stop
KTN1	KTN1		1	2 frameshift	premature stop
KLC1	KLC1				
CKMT1A	KCRU		2	2 no_frameshift	premature stop
GALK2	GALK2_A			3	
GALK2	GALK2_B			3	
SNX1	SNX1		1	1 no_frameshift	premature stop
HACD3	HACD3		3	2 frameshift	premature stop
SCAMP5	SCAM5_A			2	
SCAMP5	SCAM5_B			2	
NEIL1	NEIL1		2	3 no_frameshift	insertion
PHKB	KPBB		3	2 no_frameshift	premature stop

Initial_Gene_Ids.xlsx

GENE_ID
KCNAB2
CAMTA1
WDTC1
TMEM39B
MROH7-TTC4
FPGT-TNNI3K
AGL
FDPS
SYT11
CEP350
KCTD3
FAM149B1
MFSD13A
INPP5F
LHPP
TKFC
RTN3
DPP3
RBM4
NEU3
AAMDC
ABCG4
FAM66C
DNM1L
DIP2B
DGKA
SYT1
PLBD2
LINC00507
GLT1D1
TPT1-AS1
BORA
CDC16
LRFN5
KLHDC1
KTN1
KLC1
HERC2P9
LINC02256
DPH6-DT
CKMT1A
GALK2
GALK2
SNX1
HACD3
SCAMP5
SCAMP5
NEIL1
SLX1B-SULT1A4
PHKB

intron_positions_for_nmd_with_nmd_candidacy.csv

	Gene_ID	Seq_ID	Protein_Name	Query_match_in_JuncDB	Final_intron_position	Strand	NMD_candidate
1	CAMTA1	Q9Y6Y1	CMTA1	yes	6780	+	yes
2	TMEM39B	Q9GZU3	TM39B	yes	1398	+	yes
3	FDPS	P14324	FPPS	yes	1254	+	yes
4	SYT11	Q9BT88	SYT11	yes	1195	+	yes
5	CEP350	Q5VT06	CE350	yes	9678	+	yes
6	KCTD3	Q9Y597	KCTD3	yes	2111	+	yes
7	FAM149B1	Q96BN6	F149B	yes	1756	+	yes
8	INPP5F	Q9Y2H2	SAC2	yes	3341	+	yes
9	RTN3	O95197	RTN3	yes	3149	+	yes
10	DPP3	Q9NY33	DPP3	Not_in_JuncDB	NA	+	NA
11	RBM4	Q9BWF3	RBM4	yes	436	+	no
12	NEU3	Q9UQ49	NEUR3	yes	357	+	no
13	AAMDC	Q9H7C9	AAMDC	yes	261	+	no
14	ABCG4	Q9H172	ABCG4	yes	2573	+	yes
15	DNM1L	O00429	DNM1L	Not_in_JuncDB	NA	+	NA
16	DIP2B	Q9P265	DIP2B	yes	7082	+	yes
17	DGKA	P23743	DGKA	yes	3945	+	yes
18	PLBD2	Q8NHP8	PLBL2	yes	1890	+	yes
19	BORA	Q6PGQ7	BORA	yes	2022	+	yes
20	CDC16	Q13042	CDC16	yes	3787	+	yes
21	LRFN5	Q96NI6	LRFN5	yes	2214	+	yes
22	KLHDC1	Q8N7A1	KLDC1	yes	1061	+	no
23	KTN1	Q86UP2	KTN1	Not_in_JuncDB	NA	+	NA
24	CKMT1A	P12532	KCRU	Not_in_JuncDB	NA	+	NA
25	SNX1	Q13596	SNX1	Not_in_JuncDB	NA	+	NA
26	HACD3	Q9P035	HACD3	yes	1507	+	yes

intron_positions_for_nmd.csv

Gene_ID	Seq_ID	Protein_Name	Query_match_in_JuncDB	Final_intron_position	Strand
CAMTA1	Q9Y6Y1	CMTA1	yes	6780	+
TMEM39B	Q9GZU3	TM39B	yes	1398	+
FDPS	P14324	FPPS	yes	1254	+
SYT11	Q9BT88	SYT11	yes	1195	+
CEP350	Q5VT06	CE350	yes	9678	+
KCTD3	Q9Y597	KCTD3	yes	2111	+
FAM149B1	Q96BN6	F149B	yes	1756	+
INPP5F	Q9Y2H2	SAC2	yes	3341	+
RTN3	O95197	RTN3	yes	3149	+
DPP3	Q9NY33	DPP3	Not_in_JuncDB	NA	+
RBM4	Q9BWF3	RBM4	yes	436	+
NEU3	Q9UQ49	NEUR3	yes	357	+
AAMDC	Q9H7C9	AAMDC	yes	261	+
ABCG4	Q9H172	ABCG4	yes	2573	+
DNM1L	O00429	DNM1L	Not_in_JuncDB	NA	+
DIP2B	Q9P265	DIP2B	yes	7082	+
DGKA	P23743	DGKA	yes	3945	+
PLBD2	Q8NHP8	PLBL2	yes	1890	+
BORA	Q6PGQ7	BORA	yes	2022	+
CDC16	Q13042	CDC16	yes	3787	+
LRFN5	Q96NI6	LRFN5	yes	2214	+
KLHDC1	Q8N7A1	KLDC1	yes	1061	+
KTN1	Q86UP2	KTN1	Not_in_JuncDB	NA	+
CKMT1A	P12532	KCRU	Not_in_JuncDB	NA	+
SNX1	Q13596	SNX1	Not_in_JuncDB	NA	+
HACD3	Q9P035	HACD3	yes	1507	+

nmd_information_output.txt (screenshot of part of output file)

```

CMTA1_Q9Y6Y1_can_iso_seqs.fasta
CMTA1 Intron position: 6780 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 6659 nucleotides away, and therefore is a potential NMD
candidate.
TM39B_Q9GZU3_can_iso_seqs.fasta
TM39B Intron position: 1398 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 239 nucleotides away, and therefore is a potential NMD
candidate.
FPPS_P14324_can_iso_seqs.fasta
FPPS Intron position: 1254 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 836 nucleotides away, and therefore is a potential NMD
candidate.
SYT11_Q9BT88_can_iso_seqs.fasta
SYT11 Intron position: 1195 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 1134 nucleotides away, and therefore is a potential NMD
candidate.
CE350_Q5VT06_can_iso_seqs.fasta
CE350 Intron position: 9678 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 4760 nucleotides away, and therefore is a potential NMD
candidate.
KCTD3_Q9Y597_can_iso_seqs.fasta
KCTD3 Intron position: 2111 Strand: +
The premature stop codon (PTC) is greater than 55 nucleotides away from the final exon-
exon junction. The PTC is 523 nucleotides away, and therefore is a potential NMD
candidate.

```


no_match_category_names.xlsx (only a portion of the table)

InputTerm	Symbol	Category
TMEM180	MFSD13A	Protein Coding
FAM66C	FAM66C	RNA Gene
LINC00507	LINC00507	RNA Gene
TPT1-AS1	TPT1-AS1	RNA Gene
HERC2P9	HERC2P9	Pseudogene
LOC100996255	LINC02256	RNA Gene
DPH6-AS1	DPH6-DT	RNA Gene
SLX1B-SULT1A4	SLX1B-SULT1A4	RNA Gene
MIR548N	MIR548N	RNA Gene
SELO	SELENOO	Protein Coding
SELT	SELENOT	Protein Coding
HRASLS	PLAAT1	Protein Coding
TMEM161B-AS1	TMEM161B-AS1	RNA Gene
LINC00265	LINC00265	RNA Gene
LINC01000	LINC01000	RNA Gene
PVT1	PVT1	RNA Gene
FRG1HP	FRG1HP	Pseudogene
MINOS1-NBL1	MICOS10-NBL1	Protein Coding
LOC728730	MAP4K3-DT	RNA Gene
LINC00634	LINC00634	RNA Gene
RPL23AP82	RPL23AP82	Pseudogene
GTF2H2B	GTF2H2B	Pseudogene
TRAF3IP2-AS1	TRAF3IP2-AS2	RNA Gene
ARMCX5-GPRASP	ARMCX5	Protein Coding

Remaining_After_NMD_Testing.xlsx

Gene_ID	Protein_Name	Exon_1_frame	Exon_2_frame	putative_exon_causes_frameshift	event_caused_by_inclusion_of_putative_exon
KCNAB2	KCAB2	2	3	frameshift	insertion
WDTC1	WDTC1	1	2	no_frameshift	insertion
AGL	GDE	1	3	frameshift	insertion
NEU3	NEUR3	3	2	frameshift	premature stop
GLT1D1	GL1D1	1	2	no_frameshift	insertion
KLHDC1	KLDC1	1	3	no_frameshift	premature stop
NEIL1	NEIL1	2	3	no_frameshift	insertion