Investigating the Regulation of Nonsense Mediated Decay During the Inflammatory Response

Matthew Cattle
Department of Biomolecular Engineering
Brooks Lab
April 7, 2020
Contents

Abstract 3

Introduction 3

2.1 Project Goal ................................................. 3
2.2 Inflammatory Transcriptional Cascade .......................... 4
2.3 RNA sequencing ............................................. 5
  2.3.1 Short-read RNA sequencing ................................. 5
  2.3.2 Long-read RNA sequencing ................................ 5
  2.3.3 Frac-seq .................................................. 5
2.4 Nonsense Mediated Decay ..................................... 6
  2.4.1 NMD Regulation .......................................... 6

Aim 1: Whole-cell PTC+ Enrichment 7

3.1 Sample Descriptions .......................................... 8
3.2 RNA-seq Alignment .......................................... 8
  3.2.1 Short-read RNA-seq alignment ............................ 8
  3.2.2 Long-read RNA-seq alignment ............................ 8
3.3 Transcriptome assembly ...................................... 8
3.4 PTC+ predictions ............................................. 9
3.5 Isoform quantification .................................. 9
3.6 Differential isoform expression ............................. 10
3.7 PTC+ transcripts are enriched post-stimulation ............ 11

Aim 2 - Future work 11

4.1 Transcriptome assembly .................................. 11
4.2 TPM quantification ...................................... 11
4.3 Fraction enrichment analysis ............................ 12

Conclusion 12

References 12
Abstract

Chronic inflammatory diseases affect millions of patients across the world and occur when normal inflammatory pathways become dysregulated. In order to understand more about the regulatory mechanisms that govern innate immunity and how they may become dysregulated in the context of chronic inflammation, I am computationally investigating the regulation of nonsense-mediated decay during the mouse macrophage inflammatory response. The regulation of nonsense-mediated decay has not been studied in innate immunity despite recent observations that show nonsense-mediated decay as an important regulatory process across many different cellular systems. By analyzing whole-cell RNA sequencing data from both short-read and long-read sequences, I determined that transcripts that should normally be targeted by nonsense-mediated decay are enriched post-inflammation, a phenomenon that is unexpected if nonsense-mediated decay is active. Future studies for this project will involve analyzing polysome profiling data to understand changes in translational efficiency of transcripts and expected nonsense-mediated decay targets. This will determine if the normal nonsense-mediated decay pathway is being inhibited during the inflammatory response and allowing for translation of these isoforms. This study may work to identify a potentially important aspect of regulation in the mammalian immune response and provide direction for further studies in understanding transcriptional and translational control of inflammation.

Introduction

2.1 Project Goal

My goal for this project is to computationally identify whether or not regulation of nonsense-mediated decay exists during the mammalian inflammatory response. In this study, I will quantify the usage of transcripts with premature termination codons (PTC+) and assess the relative translational efficiency of these isoforms in inflamed mouse and human macrophages. I hypothesize that PTC+ transcripts will be enriched in the cell post-stimulation, similar to a phenomenon seen in plant immunity. My approach will use both short-read RNA-seq, for the high coverage and low error rates it provides, as well as direct RNA nanopore sequencing for the high-confidence full-length transcripts.

During previous work on characterizing alternative splicing in the mammalian inflammatory response in the Brooks Lab, we found that several important splicing genes in the SR gene family express alternative isoforms with premature stop codons. When we performed Frac-seq, a technique to assess the relative translational efficiency of extracted transcripts, we see that some of the PTC+ isoforms are found in the high polysome fraction. This implies translation of these PTC+ isoforms, despite the presence of an early stop codon that should flag the transcript for nonsense mediated decay. When looking through the literature, it appears there is other evidence that show that PTC+ isoforms can be translated. Previous studies have found evidence for the truncated SRSF3 protein from a PTC+ transcript (Kano et al., 2013), and ribosome profiling data from the Genome Wide Information on Protein Synthesis database shows ribosomes elongating over the entire PTC+ transcript (Michel et al., 2014). There is a growing body of evidence to support the translation of an Srsf3 PTC+ transcript.

This phenomenon of translating PTC+ transcripts is also seen during bacterial infection of plants in the Arabidopsis genus (Gloggnitzer et al., 2014). By downregulating nonsense mediated decay, these plant species prevent degradation of transcripts encoding for immune receptors to increase production of membrane receptor proteins (Gloggnitzer et al., 2014). This was shown to be a crucial aspect of plant immunity and to occur directly through suppression of nonsense mediated decay (Gloggnitzer et al., 2014). By coming across this study, our observation of Srsf3 translation during inflammation turned into a hypothesis that NMD is potentially being inhibited as a part of the innate immune response to allow translation to occur.

My aim is to globally quantify the expression and translation of PTC+ transcripts in lipopolysaccharide (LPS) stimulated human and mouse macrophages. This study will use a hybrid sequence analysis approach, where both long-read and short-read sequencing data will be incorporated into the pipeline, in order to perform a comprehensive global analysis of PTC+ isoform usage that might be missed by either sequencing approach alone. Long-read sequences will be used to construct high-confidence full-length isoforms, while short-read sequences provide the coverage necessary for quantification and ensure that signals of low abundance are not missed. Full-length transcripts will be assembled using StringTie2 on both the short-read
Figure 1: Cartoon diagram showing that alternative splicing can create two different products from the same gene. By combining Exon 1, Exon 2, and Exon 3, this gene normally creates Transcript A. Under a different condition, however, this gene may combine Exon 1, Exon 2, and Exon 4 to create Transcript B that codes for a slightly modified protein or other RNA product. Alternative splicing expands the coding space of genes and allows a cell to modify RNA products based on their needs.

and long-read sequences (Kovaka et al., 2019). These full isoforms will also be processed using FLAIR for identification of premature stop codons. By using kallisto (Bray et al., 2016), isoform usage can be quantified to look for enrichment of the PTC+ transcripts after LPS stimulation in the whole cell and in the individual polysome fractions.

While the regulatory mechanisms of the inflammatory response are a heavily studied system in human biology, there are still many unknowns that prevent a full understanding of how dysregulation occurs and leads to chronic inflammation. By investigating the underappreciated role of nonsense-mediated decay in the innate immune response, my goal is to provide future direction for researchers and clinicians in their search for therapeutic targets for chronic inflammatory diseases.

### 2.2 Inflammatory Transcriptional Cascade

In order to be both fast and robust, the inflammatory transcriptional cascade must incorporate tight regulatory processes in its pathways (Smale and Natoli, 2014). Stimulation of the macrophage membrane protein Toll-like receptor 4 with lipopolysaccharide (a common component of Gram-negative bacteria) induces the NFkB pathway leading to widespread gene expression changes (Smale and Natoli, 2014). These gene expression changes include gene activation, gene suppression, and more recently has been shown to induce massive alternative splicing changes (Smale and Natoli, 2014). During the alternative splicing process, different mRNA isoforms are produced from the same gene through incorporation of a combination of exons and introns (Wang et al., 2015). Figure 1 shows a cartoon diagram of the alternative splicing process, where the one gene produces two different RNA products, referred to as isoforms, under certain conditions. These gene expression changes all work to drastically change the cell’s transcriptional profile and activate an inflammatory state.

Uncontrolled activation of the inflammatory pathways has the potential to develop into highly harmful chronic inflammatory diseases including systemic lupus erythematosus, rheumatoid arthritis, chronic obstructive pulmonary disease, and a range of cardiovascular problems (Pawha et al., 2020). Chronic inflammatory diseases affect millions of patients throughout the world and are ranked as the most significant threats to human health according to the World Health Organization (Pawha et al., 2020). Developing a thorough understanding of how chronic inflammation occurs can lead to identification of therapeutic targets...
for a dysregulated immune response and could save millions of patients suffering from chronic inflammatory disorders but requires a deep analysis of the transcriptome of inflamed macrophages.

2.3 RNA sequencing

Sequencing of mRNA transcripts allows one to study the transcriptional profile of a cell with individual nucleotide resolution. RNA sequencing allows for accurate assessment of gene expression and allows one to read the entire transcriptome of a cell by determining the sequence of nearly every mature RNA product that is transcribed (Stark, Grzelak, & Hadfield, 2019). RNA sequencing protocols can be broken down into two categories: short-read RNA sequencing and long-read RNA sequencing.

2.3.1 Short-read RNA sequencing

For short-read RNA sequencing protocols, fully processed transcripts are extracted and broken up into smaller fragments (Stark et al., 2019). These fragments are each sequenced, and all the sequences make up what is referred to as a sequencing library (Stark et al., 2019). Fragments have to be aligned back to the genome to determine which gene or region they come from and assign genomic coordinates (Stark et al., 2019). Because transcripts are broken up into smaller pieces, it becomes extremely difficult to determine exactly what isoform a sequenced fragment comes from. It is relatively easy to determine which exons a sequence comes from, but the exact combination of exons that a sequence comes from is difficult or impossible to determine. To assemble a full transcriptome from short-read RNA sequences to construct novel transcripts, inferences about which transcript a read aligns to must be made from statistical models (CITE). Short-read sequencing is primarily used for its cost efficiency, low sequencing error rates, and high sequencing coverage (Stark et al., 2019).

2.3.2 Long-read RNA sequencing

For long-read RNA sequencing, full transcripts are extracted from a cell and are not broken up into smaller fragments (Stark et al., 2019). These full transcripts are placed onto a lipid membrane that has a charge differential across the membrane. Located on the membrane is a channel protein that is able to thread nucleic acids through to the other side. As a nucleic acid strand is threaded through the protein channel, the current across the membrane is slightly altered. Each base that is threaded through creates a unique differential current signature that is recorded. By analyzing the data for the charge across the membrane over time as a transcript is threaded through, the sequence of nucleic acids can be inferred. This sequencing method is incredibly useful because full nucleic acid segments can be sequenced and exactly which transcripts are being expressed can be determined and quantified (Stark et al., 2019). However, long-read sequencing is not as cost-efficient as short-read sequencing, is highly error prone, and is difficult to get deep coverage (Stark et al., 2019).

2.3.3 Frac-seq

One protocol for comparing translational efficiencies of isoforms within a transcriptome is a polysome sedimentation technique dubbed “Frac-seq”. For this sequencing protocol, ribosomes that are associated with an mRNA product at a given time are fixed to the transcript and the transcripts are then extracted (Sterne-Weiler et al., 2013). Once extracted, these polysome-associated transcripts are centrifuged through a sucrose gradient to sediment based on their molecular weights (Sterne-Weiler et al., 2013). Highly translated RNA products will be heaviest, falling deepest in the sucrose gradient (Sterne-Weiler et al., 2013). The sucrose gradient is then separated into fractions and all transcripts found within that gradient are sequenced (Sterne-Weiler et al., 2013). For the purposes of this project, the sucrose gradient is only separated into three fractions: monosome (one ribosome), polysome low, and polysome high. For each condition there is also a control input fraction that represents all transcripts found in the cytoplasm of the cell and can be thought of as a combination of the three fractions. By calculating expression values of a transcript within each fraction, it can be determined whether that transcript is highly translated or not (Sterne-Weiler et al., 2013). By performing this method between two cellular conditions, such as inflammatory stimulated macrophages, differences in translational efficiency for a transcript could be identified.
Figure 2: When a transcript does not have a premature termination codon (PTC-), ribosomes associate with the transcript and proceed with translation as normal. Upon identifying a transcript that contains a premature termination codon (PTC+), the protein UPF1 associates with the transcript and recruits other degradation factor proteins along with nucleases to destroy the transcript and prevent translation.

2.4 Nonsense Mediated Decay

Nonsense mediated decay (NMD) is a post-transcriptional process that alters the available transcriptional profile and translational efficiency through targeted degradation of mRNA (Hug et al., 2016). In the NMD pathway, cells recognize transcripts that contain a premature termination codon (PTC), a stop codon that occurs more than 50 to 55 nucleotides before the final splice junction, and degrade these transcripts before translation or other RNA activities can occur (Hug et al., 2016). Recognizing an NMD substrate begins at the splicing process, when a complex of proteins called the exon-junction complex (EJC) is deposited slightly upstream from the final splice junction (Hug et al., 2016). When a stop codon is identified upstream from the EJC, a degradation inducing protein named UPF1 (and other degradation factors) is recruited to the transcript, along with nucleases intended to degrade the nucleic acid product (Hug et al., 2016). An illustration of this post-splicing process is seen in Figure 2. Cells use NMD as a way to prevent harmful protein production, maintain protein synthesis levels, or control other aspects of RNA activity (Hug et al., 2016). Recently, the link between alternative splicing and producing NMD substrates has also become apparent. Studies across a wide variety of systems have shown that under certain conditions, splicing changes result in the inclusion of a ‘poison exon’, or an exon that contains a PTC (Lareau et al., 2007). The recent observations of both splicing-dependent NMD response and poison exon alternative splicing have made it clear that splicing and NMD are very closely linked, and that further investigation of these two processes across different regulatory systems has the potential to discover unknown mechanisms of transcriptional and translational regulation.

2.4.1 NMD Regulation

There is evidence across different cell types that shows translation of PTC+ transcripts that should be degraded by NMD. In unpublished preliminary work done in the Brooks Lab, we identified an isoform of an important splicing factor gene, Srsf3, that contains a PTC. When we performed Frac-seq, we found the specific PTC+ isoform of Srsf3 in the sucrose fraction corresponding to high polysome association, indicating that it is being translated by the cell. Figure 3 shows a pileup of RNA sequencing reads aligned to the UCSC Genome Browser (Kent et al., 2002) for the high polysome fraction of control and inflammatory induced macrophages. The height of the peak corresponds to its abundance in the high polysome fraction. The section of the isoform that contains the PTC is highlighted in red. Based on comparative read counts, it is clear that this isoform is preferentially translated in inflammatory stimulated macrophages. Two different groups of researchers, Kano et al. and Jumaa et al., have also independently confirmed the existence of the truncated SRSF3 protein resulting from the PTC+ transcript in two different human cancer cell lines, the HCT116 colon cancer cell line and K46 B cell lymphoma cell line, respectively (Kano et al., 2013; Jumaa et al., 1997). This evidence indicates that despite containing a PTC, some genes, specifically Srsf3, may
Figure 3: A screenshot from the UCSC Genome Browser showing the Srsf3 locus in the mouse mm10 genome with predicted isoforms and Frac-seq read pileups to represent translational efficiency. Orange constructs represent PTC+ isoforms, blue constructs represent protein coding transcripts, and black isoforms represent isoforms with no start codon that are likely non-coding or aberrant transcripts. The light blue read pileups are calculated from sequencing the high polysome fraction in the unstimulated condition while the yellow read pileups are calculated from sequencing the high polysome fraction in the LPS stimulated condition. The red box highlights the region that contains the premature termination codon. It can be seen that the LPS stimulated condition contained more reads that align to the transcript that contains a PTC, indicating higher translation of this isoform after LPS stimulation.

become translated under certain cellular conditions and implies that cells have a way to modulate the NMD response pathway to prevent RNA degradation.

Other researchers have shown that some cell types have the ability to specifically regulate the process of NMD in response to certain stimuli, such as HeLa cells that downregulate NMD after doxorubicin stimulation (Popp and Maquat, 2015). These studies have shown that modulating the response of the NMD pathway is a functionally necessary adaptive mechanism for many different conditions. Interestingly, plants of the Arabidopsis genus show a phenomenon of NMD suppression during bacterial infection as a way to upregulate surface receptor proteins and mount a faster and more effective response to infection (Gloggnitzer et al., 2014). While it is shown across different cell types that PTC+ transcripts can be translated and that NMD can be regulated as a cellular response, there is no work done on the regulation of NMD in mammalian macrophage cells and its role in maintaining a healthy acute immune response.

Aim 1: Whole-cell PTC+ Enrichment

The first aim of this project is to determine if PTC+ transcripts are globally enriched in LPS stimulated macrophages. That is, is the number of upregulated PTC+ transcripts different from what is expected by random chance? While this whole-cell RNA sequencing analysis approach is not able to address the translational efficiency of specific isoforms, this initial approach was used to identify patterns in the data to inform future experiments pertaining to the second aim. For example, should the counts of differentially expressed PTC+ transcripts be greater than expected, this would provide further evidence to the NMD inhibition hypothesis. However, the opposite result (fewer differentially PTC+ transcripts) also provides useful information, such as the possibility of a highly specific modulation of the NMD pathway that only acts on (or does not act on) specific transcripts or genes. Rather than looking for global inhibition of NMD, this would suggest that future analysis of isoform translational efficiency should be used to look for specific changes with possibly functional effects.

In order to address this aim, macrophage samples were extracted, LPS treated, and sequenced. Once sequenced, transcriptomes need to be assembled with PTC+ calls as well as quantified for reads mapping to
specific isoforms and then statistically tested for significant differences in expression values. This project is
concerned with the computational analysis of the transcriptome data and follows the steps post-sequencing.

3.1 Sample Descriptions
Primary mouse bone marrow-derived macrophages were isolated from the femurs of twelve different mice.
Macrophage samples were split up into two groups consisting of an untreated control group and a two hour
LPS treatment group. Total cellular mRNA was polyA extracted from treatment and control cells. In order
to capture as many isoforms as possible, this project incorporated both short-read RNA Illumina sequencing
and long-read direct RNA nanopore sequencing. Six isolates (three control, three LPS stimulated) were
subjected to short-read RNA sequencing while the other six were direct RNA long-read sequenced.

3.2 RNA-seq Alignment
3.2.1 Short-read RNA-seq alignment
Short-read sequences were aligned to the transcriptome using TopHat, an RNA-seq splice junction mapping
tool (Kim et al., 2013). In order to do a fast and accurate alignment, TopHat requires a genome index file
created from the DNA sequencing alignment predecessor tool bowtie2 (Langmead et al., 2009). Using the
bowtie2-build module and its default parameters, an index was created from the GENCODE GRCm38.p6
FASTA reference sequence. Once the index was created, both strands of the sequencing run were provided
to TopHat along with the created bowtie index and the GENCODE mm10 vM18 transcriptome reference in
GTF format. For all of the alignment parameters, library type was set to fr-firststrand, segment length was
set to 20, and coverage based search for junctions was disabled. Across all six samples, average concordant
alignment rate (percent of read pairs that map to the genome with correct insert size between them) was
53.7%, with a high of 72.3% and a low of 34.8%. Of the mapped alignment pairs, the average multimapping
rate was only 3.6%, meaning most reads only mapped to one place in the genome.

3.2.2 Long-read RNA-seq alignment
Base-called nanopore fastq files were independently aligned using the alignment tool within FLAIR (Tang et
al., 2019). The FLAIR alignment tool utilizes minimap2 for its sequence alignment (Li, 2018). FLAIR was
run using the basic alignment parameters and samples were aligned to the mm10 vM18 GENCODE reference
genome. Alignment of the long-read sequences was done by Pratibha Jagannatha, a previous undergraduate
student that worked on a related project.

3.3 Transcriptome assembly
Once aligned to the genome, both short-read and long-read sequences were assembled into a set of full tran-
scripts using StringTie2 (Kovaka et al., 2019). In the case of short-read RNA-seq data, this involves making
predictions about the transcripts to which the short sequences belong. The GENCODE mm10 vM18 refer-
cence transcriptome is used to guide the assembly process. For the short-read transcript assembly, StringTie2
was run with default parameters, along with the BAM files created from the TopHat alignment and the refer-
cence transcriptome as input. A set of isoforms was created from each sample, making six total transcript
sets for the short-read sequencing data. All six transcript sets were then merged using the StringTie2 merge
functionality to create a non-redundant set of isoforms from all samples. All default merge parameters were
used, except for the addition of the reference transcriptome to assist with the transcript merging process
and assign identifiable ENSEMBL ID’s to isoforms that are seen within the reference transcriptome.

Assembling the long-read sequencing data followed a nearly identical process, using the aligned BAM
sequences from the long-read sequencing runs and the GENCODE reference transcriptome. The one differ-
ence in running StringTie2 for the long-read assembly is the addition of the -L parameter to signal that the
provided alignments are from a long-read sequencing experiment. StringTie2 was run individually for all six
long-read samples, at which point the isoform sets were merged using StringTie2 merge with the GENCODE
reference transcriptome. After creating transcript sets for both the short and long-read sequencing samples,
both sets were merged into one final unified set of isoforms. Once again, StringTie2 merge was used, with the reference transcriptome, to create the union of the two transcript sets.

One difficulty with assembling a transcriptome from direct RNA nanopore sequencing is the relatively high abundance of clipped reads. Segments of RNA that are threaded through a membrane pore have a tendency to break off as they go through, leading to a shortened transcript at the 5’ start of the transcript. In order to deal with this technical artifact, transcripts were filtered out from the final set if there was not a chromatin accessible region at the transcript start site, based on the assumption that actively transcribed promoters are free from histone clusters. Using previously published ATAC-seq data for untreated and LPS stimulated macrophages (Tong et al., 2016), transcripts were retained in the final isoform set if they intersected with a chromatin accessible region. First, ATAC-seq reads were aligned to the genome and assembled into consensus regions referred to as “peaks”. Raw ATAC-seq fastq sequence files were pulled from the GEO accession number GSE67357. A bowtie2 index file was created from the Gencode M18 (mm10 assembly) basic annotation and the untreated and LPS-treated ATAC-seq reads were aligned using the created index file with the default bowtie2 parameters (Langmead et al., 2009). Peaks were then called separately by treatment type on untreated and treated samples using the ENCODE published ATAC-seq peak calling pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline) using the aligned reads as sequence input. Parameters that were used followed the basic JSON input file template, using an IDR threshold of 0.05. Peaks from both conditions were then merged using bedtools merge if the tail ends were less than 10bp away from each other, in order to create a set of consensus peaks from both conditions. A GFF file was created from the merged peaks, assigning a unique ID to each peak. This GFF file was provided to HTSeq-count (Anders et al., 2015) along with the aligned reads for each condition in each replicate to count reads aligning to the unique peaks. The read count matrix was provided to DESeq2 (Love et al., 2014) to call differential peaks. All peaks were considered significant if log2FC $\geq$ 0.8 and p-value $\leq$ 0.15. Intersections were determined using bedtools intersect (Quinlan & Hall, 2010). The first nucleotide of each transcript was overlapped with the consensus peaks and all transcripts that overlapped at the start site with a chromatin accessible region were kept in the final transcriptome set.

The final transcriptome consisted of 65,793 total transcripts. Of these 65,793, 41,106 of these isoforms were previously known and found in the reference annotation while 24,687 transcripts were completely novel and assembled by StringTie2.

3.4 PTC+ predictions

After full transcriptome assembly, transcripts were run through the predictProductivity.py tool within FLAIR. Using annotated start codons from the reference genome, this tool identifies premature (>50nt away from a splice junction) stop codons in the open reading frames of the provided transcripts. As input to predictProductivity.py, the assembled transcriptome in BED12 format was provided along with the GENCODE mm10 vM18 mouse genome reference sequence in FASTA format and the GENCODE mm10 vM18 transcript reference annotations in GTF format. This module compares assembled transcripts to the reference transcriptome and reference sequence, identifying transcripts that contain a stop codon more than 50nt away from its canonical stop codon and designating them as PTC. predictProductivity.py also flags transcripts as NGO if they do not contain a start codon (unlikely to be protein coding), NST if they do not contain a stop codon, and PRO, for productive, if they contain a start codon and do not have a PTC. PRO transcripts are the most likely to code for proteins, while the other transcripts are likely to be non-coding. Out of the 65,793 total transcripts assembled, 17,073 (36.8%) were predicted to be unproductive PTC+ transcripts. 36,263 (55.1%) were categorized as PRO, while 7,414 (11.2%) and 5,043 (7.6%) were classified as NGO and NST, likely representing different classes of regulatory RNAs or otherwise aberrant transcripts.

3.5 Isoform quantification

After transcripts were assembled and PTC+ prediction calls were made, isoforms were annotated for read counts and expression values in transcripts per million (TPM). While the previous steps used both long and short-read sequencing data to assemble a high-confident and comprehensive transcriptome, quantifying isoform usage was done using the Illumina short-read RNA sequences due to the higher coverage and lower error rates. kallisto, a popular software for transcript-level quantification, was used to count reads mapping
Figure 4: A volcano plot of the differentially expressed transcripts that plots the significance p-value over the log transformed TPM fold change. Significant differentially expressed PTC+ transcripts are shown as red dots.

Transcripts were tested for statistically significant differential expression between conditions by using sleuth, an R package designed as an extension of kallisto for differential isoform expression (Pimentel et al., 2017). This package employs a linear model to decouple biological variance and identify biologically relevant expression differences between conditions (Pimentel et al., 2017). In running sleuth, a table explaining the experimental design and path to kallisto counts for each sample was first loaded into a dplyr dataframe. The dplyr dataframe was then loaded into a sleuth_prep object with default arguments to preprocess the transcript counts quantified by kallisto, including smoothing read counts and filtering transcripts without enough supporting reads. The processed counts from sleuth_prep were fit to both a full and reduced model using sleuth_fit with the “full” and “reduced” arguments, separately. The likelihood ratio test for significance was performed on the fit models within the sleuth_lrt object. 2,511 transcripts were calculated as differentially expressed, 698 of which were PTC+ transcripts. Of the 698 differentially expressed PTC+ transcripts, 576 were upregulated and only 122 were downregulated PTC+ transcripts. These results are represented in the volcano plot seen in Figure 4, plotting the negative log transformed significance p-value over the log transformed TPM fold change after inflammation. Differentially expressed PTC+ transcripts are marked as red dots. The volcano plot is skewed to the right for an increased number of upregulated genes because the inflammatory response is followed by a massive increase in gene expression and most transcripts are induced post-stimulation.
### Table 1: Table representing the contingency matrix for the Chi-squared enrichment test

<table>
<thead>
<tr>
<th>Category</th>
<th>Upregulated</th>
<th>Not upregulated</th>
<th>Expected Upregulated</th>
<th>Expected Not upregulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTC</td>
<td>576</td>
<td>16497</td>
<td>489</td>
<td>16,584</td>
</tr>
<tr>
<td>PRO</td>
<td>951</td>
<td>35,312</td>
<td>1,038</td>
<td>35,225</td>
</tr>
</tbody>
</table>

Table 1: Table representing the contingency matrix for the Chi-squared enrichment test. The table on the left shows the observed values while the table on the right shows the expected values. There are more upregulated PTC transcripts than what would be expected by chance, and this difference is determined to be significant (p-value ≤ 0.05).

### 3.7 PTC+ transcripts are enriched post-stimulation

In order to test for PTC+ enrichment during LPS treatment, a Chi-squared test of independence was used to look for significant differences in expected versus observed upregulated PTC+ transcripts. The contingency matrix for the Chi-squared test is shown in Table 1. 87 more upregulated PTC+ transcripts were observed than what is expected given a random distribution and a statistically significant (p = 0.0000012313) enrichment of PTC+ transcripts in the stimulated whole-cell isolates was identified.

This observation of a statistically significant enrichment of PTC+ transcripts post-LPS stimulation supports the hypothesis for a regulated NMD response, but additional evidence is necessary to make a strong case in support of this hypothesis. In order to look at how translation of PTC+ transcripts is affected after the inflammatory response is activated in mouse macrophages, I will spend the remainder of the year analyzing differences between treatment conditions in polysome profiling Frac-seq data for each PTC+ transcript.

### Aim 2 - Future work

After PTC+ transcripts were determined to be enriched in LPS stimulated macrophages, the next goal is to look for a global increase in translational efficiency of PTC+ transcripts after treatment. In a previous experiment performed by members of the Carpenter and Sanford lab, polysome fractionation of transcripts was performed and transcripts were subjected to direct RNA long-read sequences. This aim of the project will look at differences in expression values across Frac-seq fractions between the two treatment conditions. Expression values within each fraction will be normalized to the TPM values in the cytoplasmic fraction. This analysis will computationally predict whether each identified transcript is more highly translated after LPS treatment, indicating a regulation of the NMD response for those transcripts during the immune response.

### 4.1 Transcriptome assembly

For this section of the project, the long-read Frac-seq samples have been previously aligned using FLAIR to the GENCODE mm10 vM18 genome by a previous undergraduate student in the Brooks Lab, Pratibha Jagannatha. Since sequences had already been aligned, the first step of this aim is to assemble the transcriptomes for isoform quantification. As was done before in the previous aim, the aligned BAM files will be provided to StringTie2 using the default parameters for long-read sequencing data for each replicate, condition, and Frac-seq fraction. Once assembled, StringTie2 will merge the transcript sets for corresponding replicates. Because TPM comparisons will be done within fractions, transcriptomes will then be merged with the cytoplasmic set in pairwise fashion, creating the polysome fractionation isoform sets: Cyto-Mono, Cyto-Low, Cyto-High, for each condition. A further comparison is done between treatment conditions, at which point the Cyto-Mono isoform set for the control group will be merged with the Cyto-Mono transcript set for the LPS stimulated cells. This will be done for each polysome fractionation isoform set, making three supersets of isoforms created from both control and LPS treated macrophages corresponding to Cyto-Mono, Cyto-Low, and Cyto-High.

### 4.2 TPM quantification

After transcriptome assembly, isoform usage will be quantified by condition and polysome fraction using kallisto. Once again, a kallisto index is made for each transcriptome set. Raw FASTQ short-read sequences from each sequencing sample will then be used to count reads that pseudoalign to the assembled isoforms.
generating TPM values by replicate within each condition and fraction. For each fraction (monosome, polysome low, and polysome high), the TPM value will be divided by the TPM value for that transcript in the cytoplasmic to normalize expression values.

4.3 Fraction enrichment analysis

Once TPM values are determined for all transcripts in each condition and replicate, these normalized values will be used as input to the sleuth differential transcript expression R package. The differential comparison will be done between the monosome fraction and the polysome low fraction, the monosome fraction and the polysome high fraction, as well as the polysome low and polysome high fraction. Similar to the first aim, a contingency matrix will be established in order to perform a Chi-squared test to look for upregulated PTC+ enrichment between the comparing polysome fractions.

Conclusion

Through identification of enriched SR PTC+ transcripts associated with a relatively high number of ribosomes, this work has shown that transcripts of the SR gene family exhibit increased translational efficiency during the inflammatory response despite containing an early stop codon. NMD has recently been recognized as a process to not only ensure quality control over mRNA transcripts, but also functions as a fast and robust regulatory process that controls endogenous protein levels (Hug et al., 2016). It is likely that these important functions are a part of the complicated regulatory network of the inflammatory response. The apparent modulated NMD response towards SR PTC+ transcripts could be a result of specific inhibitions of the NMD pathway preventing recognition of these mRNA products, but more work must be done to confirm the mechanism behind the observed changes in degradation substrates. Further functional studies are necessary to understand the role that NMD regulation plays in the immune response. In order to experimentally validate the computational predictions made through this project, NMD inhibition can be observed through an NMD reporter assay experiment. Developing a greater understanding of NMD and its role in the inflammatory response could provide insights to aid in discovery of new therapeutics for chronic inflammatory diseases.

References


