Global Characterization of the Direct Substrates of Nonsense-Mediated mRNA Decay in *Caenorhabditis elegans*

by

Virginia Lamb

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Genetics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2015

Date of final oral examination: 12/7/2015

The dissertation is approved by the following members of the Final Oral Committee:
Philip Anderson, Professor, Genetics
Audrey Gasch, Associate Professor, Genetics
Scott Kennedy, Professor, Genetics
Kate O'Connor-Giles, Assistant Professor, Genetics
Marvin Wickens, Professor, Molecular Biology and Biochemistry

This thesis is dedicated to my fiancée, Alison Muir.

Acknowledgements

A number of incredible people have assisted and supported me through my graduate career. The following is a paltry attempt to recognize major contributions; I can only hope that those mentioned realize how much I have valued their help.

Phil has always allowed me to work independently. Meetings with my mentor consistently renewed my excitement in the project, and Phil always helped me focus on the big picture when I was tangled in minutiae.

My committee members – Audrey Gasch, Scott Kennedy, Kate O'Connor-Giles, and Marv Wickens – provided excellent guidance and insightful suggestions. I'd particularly like to thank Scott and his lab for tips regarding various wet-lab protocols. Audrey also provided invaluable recommendations and support for bioinformatics analyses.

The other folks in the Anderson lab were a joy to work with. Lisa laid the groundwork for my project; Ben and Leah helped me through gazillions of technical problems; and Bonnie both kept the lab running smoothly and provided much-needed assistance toward the end of my project. Though I may still make some of the worst snowflakes in lab, I appreciated all of my lab mates' guidance.

My Genetics classmates have kept me sane for several years, now. While the camping trips and Library runs made Madison a spectacular place to live, I also owe a debt to the number of classmates who provided troubleshooting suggestions, coding assistance, and occasional reagents.

Lastly, my family has encouraged and supported me through good times and bad. My parents, sister, and extended family listened patiently and reassured me whenever I ran into problems with my project. Without Alison, however, this document would not exist. Ali has been my touchstone since the second year of grad school. She has comforted, guided, and inspired me, and I will never be able to thank her enough for the patience and proofreading skill she has demonstrated over the past few months.

Table of Contents

List of Tables and Figures

Abstract

Nonsense-mediated mRNA decay (NMD) is a highly conserved pathway that regulates gene expression in eukaryotes. Messenger RNAs containing premature termination codons (PTCs) are identified and rapidly degraded by NMD. NMD requires seven core components, *smg-1* through *smg-7,* in *C. elegans*. A significant portion of the transcriptome exhibits altered expression in NMD mutants; however, transcripts that change in abundance may either be direct substrates of NMD or secondary effects resulting from the stabilization of directly targeted mRNAs.

Here, I present the first genome-wide investigation of direct substrates in a multicellular organism to use true NMD null alleles, allowing for more sensitive detection of targets. SMG-2, the central effector of NMD, preferentially associates with mRNAs containing PTCs. By coupling SMG-2 immunoprecipitations in an NMD(-) background with high-throughput sequencing, I identified 585 candidate substrates. I defined an additional 95 high-confidence substrates that included sequence from annotated introns in NMD-targeted transcripts. This indicates that NMD targets and degrades specific spliced isoforms with such efficiency that they have not been detected in wild-type cells and have not been correctly annotated. Other classes of NMD substrates defined by this study include pseudogenes and genes belonging to recently expanded gene families. My results identify novel NMD substrates and provide context for understanding the role NMD plays in regulation of gene expression.

List of Abbreviations

NMD: nonsense-mediated mRNA decay

PTC: premature termination codon

IP: immunoprecipitation

RIP-Seq: RNA immunoprecipitation coupled sequencing

DE: differential expression

UPF: up-frameshift

SMG: suppressor with morphogenetic effects on genitalia

EJC: exon junction complex

uORF: upstream open reading frame

UTR: untranslated region

PABPC1: poly(A)-binding protein, cytoplasmic 1

GO term: gene ontology term

bp: base pair

PCR: polymerase chain reaction

RT-PCR: reverse transcription polymerase chain reaction

qRT-PCR: quantitative reverse transcription polymerase chain reaction

Chapter 1: An Introduction to Nonsense-Mediated mRNA Decay (NMD) and its Natural Targets

Overview

Organisms require highly sophisticated systems to regulate gene expression. Such systems ensure that gene products are available in precise quantities, in particular locations, and at exact times. This regulation allows organisms to function both efficiently and responsively. Nonsense-mediated mRNA decay (NMD) is a highly conserved pathway of regulating gene expression in eukaryotes. NMD is responsible for recognition and degradation of transcripts that contain premature termination codons (PTCs), and key components of the system have been identified in yeast [1, 2], nematodes [3, 4], *Drosophila* [5], *Arabidopsis* [6], zebrafish[7] , mice [8], and humans [9]. Such conservation illustrates NMD's fundamental importance in regulation of gene expression.

Aberrant, PTC-containing mRNAs are identified and destroyed by the NMD system. Frame shifts, nonsense mutations, and certain other mutations often produce NMD substrates. In budding yeast, a +1 frameshift mutation near the beginning of *his4* creates an early stop codon and a His^- phenotype [10]. mRNA stability is significantly decreased for the frameshift allele in a WT strain; however, the message is stabilized and the His^- phenotype is reversed in an NMD-deficient organism [1]. In humans, the W1282X nonsense mutation in the *CFTR* gene is responsible for a decrease in transcript abundance. When NMD factors are depleted through RNAi, mutant mRNAs are stabilized and exhibit a 1.7- to 3.2-fold increase relative to mock-treated samples [11]. In nematodes, an allele of the myosin heavy chain gene *unc-54* creates an extended 3' UTR that causes the normal stop codon to be perceived as premature [12]. This

allele is targeted by NMD, and degradation of the message causes paralysis. When NMD is abrogated, the mRNA is stabilized, and paralysis is suppressed [3]. Though extremely rare, messages that have been faultily transcribed, poorly spliced, or prematurely exported from the nucleus are also likely substrates for NMD [13, 14].

NMD was first identified as a quality-control/surveillance pathway that prevents nonsense mutations and their encoded truncated proteins from causing damage to cells [15, 16]. When the NMD system is not functioning, PTC-containing transcripts are stabilized, and these transcripts can encode malformed dominant-negative or gain-of-function proteins which are harmful to their host organisms [4]. The NMD pathway protects cells from more than just nonsense mutations. Natural targets of NMD include pseudogenes [17], transposons [18], and viruses [19, 20] – all of which may be harmful to the cell.

In addition to preventing expression of deleterious proteins, NMD serves to fine tune gene expression (reviewed [21, 22]). 4-25% of the transcriptome exhibits altered expression in the absence of NMD, though only a subset of these changes are caused by direct interaction with the NMD system (discussed below) [18, 23-29]. Certain mRNAs contain features that cause them to be substrates of NMD. Transcripts with long 3' untranslated regions (UTRs) [30] or upstream open reading frames (uORFs) [18] often trigger NMD. While the cell-wide repercussions of this targeting are unknown, it is plausible that NMD degradation of these transcripts allows for an added layer of gene regulation. Many alternatively spliced mRNAs are substrates of NMD. Up to 35% of human pre-mRNAs may produce NMD substrates via alternative splicing [31]. These targeted messages do not simply result from mutations or errors in transcription; most of these mRNAs seem to be deliberately spliced to include PTCs.

Cells may use alternative splicing-coupled NMD as a mechanism to more precisely regulate gene expression. Interestingly, while NMD mutants have relatively mild phenotypes in *S. cerevisiae* [1] and *C. elegans* [4], NMD is essential in higher animals [7, 8, 27].

NMD Factors

The first components of NMD were discovered in *S. cerevisiae* with the identification of the *up frameshift (upf)* genes [1, 2]. UPF1, 2, and 3 form the core NMD machinery in yeast; however, NMD relies on 7 genes in *C. elegans* (Table 1.1). Interfering with any one of these genes leads to the stabilization of PTC-containing messages. The core seven *suppressor with morphogenetic effects on genitalia* (*smg)* genes were first identified in *C. elegans* via mutagenesis screens [3, 4, 32]. SMG-2, -3, and -4 are homologues of UPF1 -3 respectively, while SMG-1, -5, -6, and -7 have no known homologues in *S. cerevisiae* [3, 4, 32]. RNAi screens in *C. elegans* have since identified seven other genes important for NMD: *smgl-1/NBAS, smgl-2/DHX23, ngp-1/GNL2, npp-20/SEC13, aex-6, pbs-2,* and *noah-2* [33, 34]. Two additional genes, *smg-8* and *smg-9*, were identified as members of the mammalian SMG-1 complex [35]. Very mild NMD defects are observed when homologs of *smg-8* and *smg-9* are RNAi depleted in nematodes [35, 36]. The functions of these factors are less well understood than the SMG proteins. A majority of the genes identified through RNAi screens are essential for viability in nematodes, suggesting that they provide additional functions independent of NMD. All seven of the *smg* genes and four of the remaining NMD-involved genes are conserved in most metazoans. An exception, *smg-7,* is absent in *D. melanogaster* [5].

Table 1.1: NMD Factors

C. elegans Gene	Homologs	Conserved in	Function in NMD Pathway
smg-1	*smg-1*	Nematodes through humans	SMG-2 kinase
smg-2	*upf-1*	Yeast through humans	Central effector, marks PTC-containing transcripts
smg-3	*upf-2*	Yeast through humans	Interacts with SMG-2, enhances marking of PTC-containing transcripts
smg-4	*upf-3*	Yeast through humans	Interacts with SMG-2 via SMG-3, enhances marking of PTC-containing transcripts
smg-5	*smg-5*	Nematodes through humans	Involved in recruitment of RNA decay factors and SMG-2 phosphatase PP2A
smg-6	*smg-6*	Nematodes through humans	Endonucleolytic cleavage of target mRNAs
smg-7	*smg-7*	Nematodes, zebrafish through humans	Involved in recruitment of RNA decay factors and SMG-2 phosphatase PP2A
smg-8	*smg-8*	Nematodes through humans	Found in SMG-1 complex in mammals, represses SMG-1 kinase activity; does not play a role in *C. elegans* NMD
smg-9	*smg-9*	Nematodes through humans	Found in SMG-1 complex in mammals, represses SMG-1 kinase activity
smgl-1	*NBAS*	Nematodes through humans	Unknown
smgl-2	*DHX23*	Nematodes through humans	Promotes SURF complex remodelling and SMG-2:SMG-3 interaction
ngp-1	*GNL2*	Yeast through humans	Unknown
npp-20	*SEC13*	Yeast through humans	Unknown
aex-6	*RAB27A and B*	Nematodes through humans	Unknown
pbs-2	*PSMB7 and 10*	Yeast through humans	Unknown
noah-2	*nompA*	Nematodes and flies	Unknown

Nonsense mediated mRNA decay requires the sequential action of each SMG protein to identify and degrade PTC-containing mRNAs. SMG-2 is the central effector of the pathway. It specifically marks PTC-containing mRNAs, a process that requires SMG-2 interaction with SMG-3 and SMG-4 [37]. NMD also requires cycles of SMG-2 phosphorylation and dephosphorylation in metazoans [38]. Such phosphorylation cycles likely contribute to formation and recycling of protein complexes required for NMD [39, 40]. SMG-1 is a phosphoinositide 3-kinase (PI3K) family member and directly phosphorylates SMG-2 [41] [42]. SMG-5 and SMG-7 interact with both SMG-2 and its dephosphorylating enzyme, protein phosphatase 2A (PP2A) [38-40, 43]. In several organisms, SMG-6 has endoribonuclease activity [44, 45]. Whether this function is conserved in nematodes has yet to be determined. A more thorough description of NMD's molecular mechanism is given below.

SMG Proteins Act in Non-NMD Pathways

Several *smg* genes have been implicated in important cellular processes in addition to NMD, including DNA replication, DNA repair, telomere maintenance, and the cellular response to stress (reviewed [46]). UPF1/SMG-2 is important for Staufen1-mediated mRNA decay, a distinct RNA decay pathway that may partially compete with NMD [47, 48] and for degradation of histone mRNAs [49].

Molecular Mechanism of NMD

Overview

The mechanism of NMD has been well studied in yeast, nematodes, and mammalian cells. Premature translation termination triggers recognition of PTC-containing transcripts,

after which SMG-2 and an associated "SURF complex" associate with PTC-containing mRNAs. Such mRNPs are specifically "marked" for degradation by SMG-2 and are degraded by one of three pathways. Finally, NMD factors and the translational machinery are recycled for later cycles of NMD and translation.

PTC Recognition

PTC recognition depends upon translation. PTC-containing messages are stabilized when chemical inhibitors such as cycloheximide or inhibitory stem loops block translation [50, 51]. The context of translation termination determines whether a given termination event is deemed premature (Figure 1.1). Normal termination is proximal to the 3' UTR, poly(A) tail, and the associated protein complexes. During normal termination events, poly(A) binding protein, cytoplasmic 1 (PABPC1) enhances activity of release factors eRF1 and eRF3, and both the nascent polypeptide and the ribosome are efficiently released [52, 53]. Premature termination occurs at a greater distance from the normal 3' UTR, poly(A) tail and associated complexes, creating a "faux" or false 3' UTR downstream of the early stop codon. Such a faux-UTR prevents interaction between the terminating ribosome and the PABP complexes, causing inefficient release of the ribosome and recruitment of the SMG proteins [54].

In many organisms, this faux-UTR model of NMD seems to sufficiently define PTC-recognition. Generally, the further a stop codon is from the 3'UTR, the more likely it is to be recognized as premature [4, 55, 56]. Transcripts with PTCs in the final exon are often not substrates of NMD [51, 57]. PTC recognition in mammals involves function of the exon junction complex (EJC). EJCs downstream of a stop codon enhance NMD in mammals [54, 58]. While downstream EJCs are not strictly required for NMD, they are strong enhances of NMD when the

3’ UTR of an mRNA is short [59]. Downstream EJCs influence NMD according to the “50/55-bp rule”. Translocating ribosomes displace EJCs, and if a PTC occurs within 55nt of the final intron, the final EJC will be displaced and, thus, not enhance NMD [57].

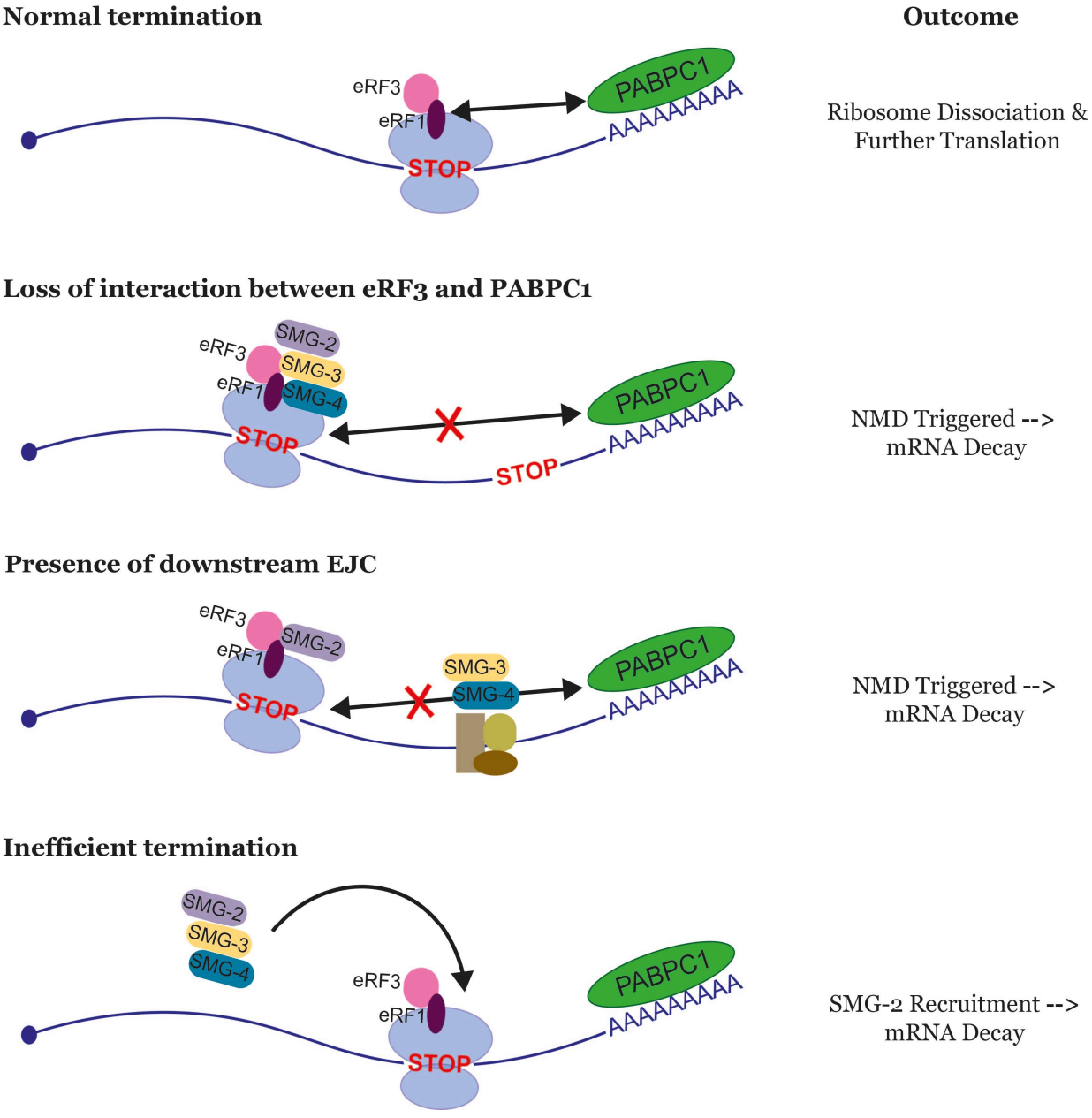
SMG-2 Specifically Marks PTC-containing Transcripts

SMG-2, the central effector of the NMD, associates strongly with PTC-containing transcripts [37]. It acts as a surveillance protein, briefly or weakly associating with most mRNAs but persistently or strongly associating with those containing PTCs. SMG-2 associates with the translation termination factors eRF3 and eRF1, bringing UPF1/SMG-2 proximal to both WT and aberrant mRNAs [60, 61]. RNA can outcompete eRF3 for UPF1/SMG-2 interaction [61]. *In vitro* work in both yeast and human cells indicates that UPF1/SMG-2 binds RNA directly and exhibits RNA-dependent ATPase and RNA helicase enzymatic activities [62, 63]. After SMG-2, eRF1, and eRF3 initially associate with PTC-containing mRNAs, recruitment of SMG-1, SMG-8, and SMG-9 occurs to form the SMG-1-Upf1-eRF1-eRF3 (SURF) complex, which is essential for NMD function [58].

Figure 1.1: PTC Recognition

During normal translation termination, the ribosome and associated release factors are proximal to the poly-A tail, PABPC1, and other 3' UTR-associated complexes (first panel). Interactions between eRFs and PABPC1 stimulate efficient dissociation of the ribosome. SMG proteins may associate with prematurely terminating ribosomes due to the disruption of such interactions (second panel). Atypical mRNP context, such as the presence of downstream EJCs can also trigger SMG-2 marking of PTC-containing mRNAs (panel three). Inefficient termination can leave the ribosome in an atypical conformation that allows association with SMG-2 (panel four). Such abnormal termination contexts lead to SMG-2 marking and mRNA degradation via NMD.

Figure 1.1


Normal termination
Outcome
eRF3
eRF1
STOP
PABPC1
AAAAAAAAA
Ribosome Dissociation & Further Translation
Loss of interaction between eRF3 and PABPC1
SMG-2
SMG-3
SMG-4
NMD Triggered --> mRNA Decay
Presence of downstream EJC
NMD Triggered --> mRNA Decay
Inefficient termination
SMG-2 Recruitment --> mRNA Decay

SMG-3 and SMG-4 enhance SMG-2 dependent marking of PTC-containing mRNAs [37]. Several different models hypothesize how the SMG-3:SMG-4 complex is brought into contact with SMG-2 on NMD-target transcripts. In one model, UPF3b/SMG-4 associates with newly formed EJCs in mammalian nuclei and is thought to traffic with nascent mRNPs to the cytoplasm [64]. EJCs are removed from mRNAs during translation, but EJCs located downstream of a PTC would remain bound. Certain substrates of NMD targets do not contain introns downstream of a PTC [54, 59, 65, 66], but ~50% of mammalian EJCs map to non-canonical positions (i.e. not at exon-exon boundaries). Thus, EJCs may still occur downstream of such PTCs [67, 68]. When downstream EJCs are absent, SMG-3 and SMG-4 can still associate with SMG-2, eRF1, and eRF3 as part of a surveillance complex [60].

Phosphorylation of SMG-2 by SMG-1, which is required for NMD [38], likely facilitates SMG-2 interactions with SMG-5, SMG-6, SMG-7 and PNRC-2 [39, 69]. This phosphorylation occurs in the SURF complex [58]. In mammals, association of SMG-8 and SMG-9 with the SMG-1 complex represses SMG-1 kinase activity [35]. However, UPF2/SMG-3 can disrupt SMG-8:SMG-9 interactions, thus activating the SMG-1 kinase [70]. In addition to interacting with SMG-1, SMG-3 interacts with SMG-2 and induces a conformational shift in SMG-2, thereby activating its helicase and ATPase activities [71, 72]. (Figure 1.2)

Modes of Target Degradation

Following UPF1/SMG-2 phosphorylation in metazoans, the substrates of NMD are degraded by one or more of four distinct pathways: direct endonucleolytic cleavage of target mRNAs by SMG-6, deadenylation- and decapping-dependent degradation mediated by SMG-7 and executed by the 5' exoribonuclease XRN1, decapping-dependent degradation following

activation by the proline-rich nuclear receptor coactivator PNRC2, and independent degradation by XRN1 and the 3’ to 5’ exosome (reviewed in [31]). These pathways can act either independently or in concert. While the specific contributions of each pathway and the activating determinants for each are unclear, the four mechanisms likely act in a partially redundant manner.

SMG-6 is a telomerase-binding protein with a C-terminal PilT N terminus (PIN) domain that cleaves single stranded RNA both *in vitro* and *in vivo* [45, 73]. Following cleavage of mRNAs by SMG-6, the resulting 5’ and 3’ fragments are rapidly degraded by sequence-nonspecific exoribonucleases, including the 5’-3’ exonuclease XRN1 and the 3'-5' exosome [74]. SMG-6-mediated cleavage of target mRNAs requires UPF1/SMG-2 [73]. Interactions between mammalian SMG-6 and UPF3B bound to EJCs may facilitate recruitment of SMG-6 to substrate mRNPs and their subsequent degradation [75].

SMG-5 and SMG-7 form a complex that interacts with UPF1/SMG-2 and promotes interactions with other mRNA decay factors. SMG-7 elicits decay when tethered to a mRNA without requiring other NMD factors [76]. The unstructured C-terminus of SMG-7 is essential for such decay, but the key interacting proteins are presently uncertain. β-globin reporter mRNAs containing PTCs are degraded via a two-phase deadenylation process that occurs concurrently with decapping [77, 78]. The PAN2-PAN3 deadenylase shortens poly(A) tails to approximately 110nt during an initial phase, followed by a second phase of deadenylation by the CCR4-CAF1 complex, which may be recruited to mRNPs by its interaction with SMG7 [79]. Both of these deadenylation complexes interact with PABPC1. eRF3 would normally outcompete PAN3 and TOP (in the CCR4-CAF1 complex) for interaction with PABPC1, however,

in an aberrant translation termination context, UPF1/SMG-2 interacts with eRF3 instead [80]. This rearrangement may explain why deadenylation occurs so rapidly on NMD-target mRNAs.

NMD factors may additionally trigger decapping and degradation of target mRNAs by interacting with PNRC2, with interacts both with the mRNA decapping enzyme (DCP1A) and UPF1/SMG-2. PNRC2 copurifies with SMG-5 but not with SMG-7, and downregulation of PNRC2 disrupts the link between SMG-5 and DCP1A. PNRC2 preferentially interacts with hyperphosphorylated UPF1/SMG-2; this interaction can facilitate both decapping of NMD substrates and relocation of the PTC-containing RNPs to P bodies, where many mRNA decay factors are found [81, 82].

In *S. cerevisiae*, which does not have homologs of SMG-5, SMG-6, and SMG-7, PTC-containing messages are decapped and then are primarily degraded 5' to 3' by XRN1 [83]. The 3' to 5' exosome does provide a redundant mechanism for degradation in yeast; its contribution is most obvious when the 5' to 3' decay pathway is blocked.

NMD acts during translation to prevent the further generation of potentially deleterious truncated proteins. However, the peptide formed during this first round may or may not be rapidly degraded. In yeast, UPF1 promotes polypeptide degradation in a manner that may rely upon inherent E3 ubiquitin ligase activity [84, 85]. This general destabilization of peptides from NMD-targeted mRNAs has not been observed in human cell lines, and it is unclear what factors affect the stability of these truncated proteins.

NMD Factor and Ribosomal Recycling

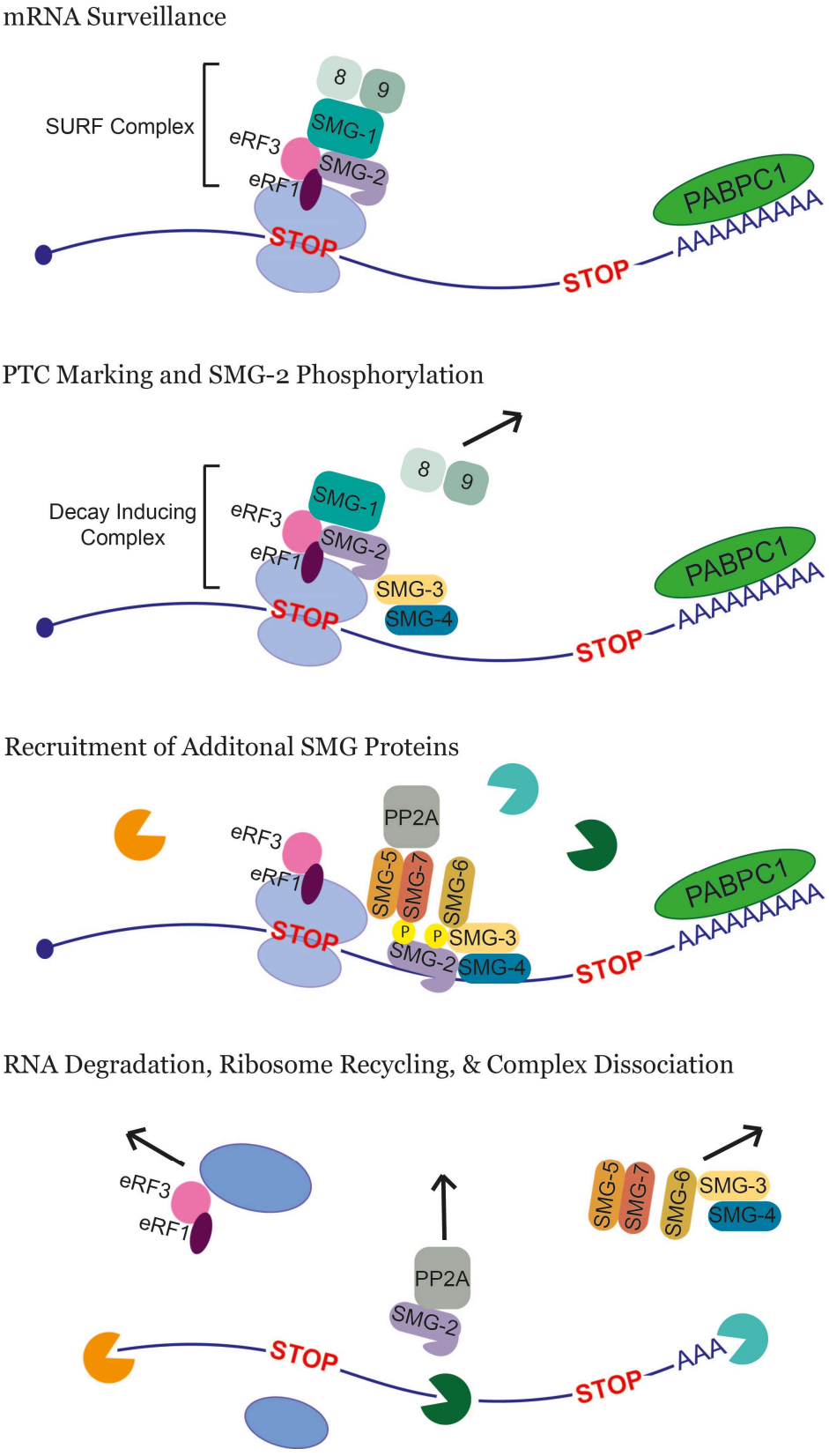
Following degradation of individual mRNAs, complexes of NMD factors must dissociate and SMG-2 must be dephosphorylated for subsequent cycles of NMD to occur. The SMG-2

phosphatase PP2A is recruited by the SMG-5:SMG-7 complex [43]. This complex ensures the dissociation of not only NMD factors but also translation equipment, including the ribosome. Translation termination at a PTC is less efficient than normal termination, and a ribosome can become stalled at an early stop [86, 87]. Ghosh, et al., demonstrated that a functional NMD system is required for efficient ribosome recycling and ongoing translation in the presence of PTC-containing mRNAs [88]. Therefore, while the degradation of aberrant transcripts and truncated proteins is an important result of NMD, NMD factors are also vital for recycling "stuck" ribosomes (reviewed [89]).

Figure 1.2: Generalized Model of NMD

SMG-2 associates with eRF1, eRF3, SMG-1, and, in mammals, SMG-8 and SMG-9 to form the SURF complex. Interaction between SMG-2, -3, and -4 on a prematurely terminating mRNA leads to the formation of a decay inducing complex, the phosphorylation of SMG-2, and preferential marking by SMG-2. Phosphorylated SMG-2 interacts with SMG-6, an endonuclease, and the SMG-5:SMG-7 complex to promote degradation of PTC-containing mRNA. PP2A, the phosphatase that interacts with SMG-5:SMG-7 aids in the recycling of NMD complexes.

Figure 1.2


mRNA Surveillance
SURF Complex
8
9
SMG-1
eRF3
SMG-2
eRF1
STOP
PABPC1
AAAAAAAAA
STOP
PTC Marking and SMG-2 Phosphorylation
Decay Inducing Complex
8
9
SMG-1
eRF3
SMG-2
eRF1
SMG-3
SMG-4
STOP
PABPC1
AAAAAAAAA
STOP
Recruitment of Additonal SMG Proteins
PP2A
eRF3
eRF1
SMG-5
SMG-7
SMG-6
P
P
SMG-3
SMG-2
SMG-4
STOP
PABPC1
AAAAAAAAA
STOP
RNA Degradation, Ribosome Recycling, & Complex Dissociation
eRF3
eRF1
SMG-5
SMG-7
SMG-6
SMG-3
SMG-4
PP2A
SMG-2
STOP
STOP
AAA

Identifying Targets of NMD

In order to better understand the biological significance of NMD, numerous studies have measured transcriptome-wide changes in gene expression when NMD is inactivated (Table 1.2) [18, 23, 25, 27, 90, 91]. NMD is thought to regulate approximately 10% of the genome from yeast through humans. Genes which show elevated levels of expression in an NMD-deficient background are involved in an incredible variety of cellular processes including RNA-processing [25], cellular transport [23], amino acid metabolism [18], DNA repair [91], and cell surface dynamics [90].

Many of these microarray and RNA-Sequencing (RNA-Seq) experiments have failed to separate direct substrates of NMD from indirect effects of the pathway. Direct substrates are those mRNAs associated with UPF1/SMG-2 and degraded by the NMD pathway. Destruction of NMD substrates can lead to further fluctuations in the transcriptome. For example, if NMD targets a transcription factor mRNA for degradation, the genes regulated by that transcription factor will exhibit altered expression in the absence of NMD. Such changes are classified as indirect effects of NMD. SMG-2 marking of PTC-containing mRNAs provides a means to separate indirectly affected genes from direct substrates.

A smaller subset of the literature has attempted to separate direct substrates of NMD from the indirect effects of NMD-target degradation [90, 92-97]. These experiments largely employ one of three strategies to determine whether a transcript is subjected to NMD: (1) isolating RNAs that co-precipitate with UPF1/SMG-2 [94, 95, 97], (2) tracking the rate of an mRNA's degradation [90, 92, 93], or (3) measuring mRNA abundance following NMD reactivation [94, 96, 98]. These studies show that only 20-40% of the transcripts that exhibit

increased expression in an NMD-deficient background are direct substrates of NMD [90, 92, 97]. Studies of direct substrates also identify large subsets of mRNAs which meet the additional requirements used to define NMD direct substrates (i.e. more rapid degradation in an NMD-competent setting) but which do not demonstrate a significant change in expression between wild-type and mutant samples.

While some classes of NMD targets, such as pseudogenes and transposons, are relatively simple to identify in genome-wide studies, identification of targeted PTC-containing alternate splice isoforms can be confounded both by the existence of multiple isoforms, by poor annotation, and by NMD's regulation of splicing factors [99, 100]. Ramani, et al., combined tiling arrays and sequencing in *C. elegans* to identify genome-wide expression changes between wild-type, *smg-1(-)* and *smg-5(-)* animals [25]. The study found roughly 300 genes that exhibit a splicing change in NMD mutants but which do not differ in overall expression between mutant and wild-type strains. While NMD may target and degrade specific splice isoforms, the total abundance of a gene's transcripts may not change. In addition to the changing abundance of annotated exons, Ramani, et al., also found that 7% of genes produce transcripts with retained introns in an NMD-deficient background. When insufficient data from NMD-deficient lines is used to annotate an organism's transcriptome, PTC-containing exons are poorly defined.

Table 1.2: Summary of Previous NMD Target Studies

Citation	Organism	Identifies direct substrates?	Fold change threshold	% Transcriptome with increased abundance in an NMD-deficient strain	Approach	% of Altered Transcripts that are Direct Substrates
He, 2003	*S. cerivisiae*	no	2-fold enrichment	10-12%	Microarray	-
Ramani, 2009	*C. elegans*	no	1.5-fold enrichment	10.0%	Microarray	-
Metzstein, 2006	*D. melanogaster*	no	2-fold enrichment	1.5%	Microarray	-
Rehwinkel, 2005	*D. melanogaster*	no	1.5-fold enrichment	3.4%	Microarray	-
Wittmann, 2006	*H. sapiens*	no	2-fold enrichment	1.5%	Microarray	-
Mendell, 2004	*H. sapiens*	no	1.9-fold enrichment	4.9%	Microarray	-
Yepiskoposyan, 2011	*H. sapiens*	no	1.5-fold enrichment	4.0%	Microarray	-
Johansson, 2007	*S. cerivisiae*	yes	2-fold enrichment	12.3%	UPF1 IP - Microarray	45%
Guan, 2006	*S. cerivisiae*	yes	1.5-fold enrichment	9.2%	mRNA Decay Rates - Microarray	46%
Matia-Gonzales, 2013	*S. pombe*	yes	1.5-fold enrichment	4.2%	UPF1 IP - Microarray	29%
Chapin, 2014	*D. melanogaster*	yes	1.8-fold enrichment	4.0%	NMD Reactivation - Seq	16%
Hurt, 2013	*M. musculus*	yes	1.1-fold enrichment	7.2%	UPF1 IP - Seq	4%
Tani, 2012	*H. sapiens*	yes	2-fold enrichment	3.7%	mRNA Stabilization - Seq	22%
Schmidt, 2015	*H. sapiens*	yes	1.5-fold enrichment	6.1%	mRNA Stabilization & SMG6 cleavage – Seq	2%

As NMD targets have become better-defined, several questions regarding the conservation of NMD direct substrates and their physiological roles have arisen. Genome-wide studies that did not discriminate between direct substrates and indirect effects of NMD showed little conservation of specific targets [27, 91]. Interestingly, recent direct substrate studies have demonstrated that NMD direct substrates may be conserved across some phyla (e.g. between mouse and human tissues [95]) but not others [97].

Introduction to Future Chapters

In the following chapters of this dissertation, I present my work to identify and characterize direct substrates of the NMD pathway in *C. elegans*. Chapter two describes the definition and characterization of NMD substrate mRNAs from a high-throughput sequencing dataset. In chapter three, I delve into both alternative analyses of the data and characteristics that are shared among NMD-targeted transcripts. In chapter 4, I review the results and conclusions of this dissertation and suggest future experiments.

Chapter 2: Global Identification of NMD Substrates in *C. elegans*

Abstract

Nonsense-mediated mRNA decay (NMD) modulates gene expression by triggering the rapid degradation of premature termination codon (PTC)-containing transcripts. While PTC-containing transcripts are the direct targets of NMD, stabilization of these transcripts has widespread secondary effects on the transcriptome. While multiple studies have investigated NMD's effect on the transcriptome, a relatively small number have determined direct substrates of NMD. In this chapter, I focus on the identification of *C. elegans* NMD direct substrates via the deep-sequencing of RNAs that co-purify with SMG-2. This is the first genome-wide study of NMD in *C. elegans* that has separated direct substrates from the indirect effects of NMD. I identify 773 genes whose RNA expression is effected by the loss of NMD (5.5% expressed genes). 585 of these genes (75.7%) appear to be direct targets of NMD.

Introduction

Gene expression is a tightly controlled process which is influenced by many different regulatory mechanisms. Nonsense-mediated mRNA decay (NMD) regulates gene expression in eukaryotes by recognizing and rapidly degrading transcripts that contain premature termination codons (PTCs). NMD was originally considered a eukaryotic RNA surveillance mechanism that prevents the production and accumulation of truncated proteins through the recognition and degradation of aberrant PTC-containing mRNA transcripts [101, 102]. NMD plays a crucial role in preventing potentially toxic dominant-negative effects of the buildup of

truncated, nonfunctional proteins in the cell. However, the role of NMD is now believed to be two fold. In addition to removing aberrant transcripts, NMD regulates gene expression via the downregulation of physiological transcripts. A number of classes of natural targets of NMD have been identified, including transcripts with upstream open reading frames (uORFs), certain cases of alternative splicing, and transcripts arising from retroviruses and transposons [18]. Clearly, NMD is a major regulator of gene expression. However, further research is needed to separate the direct substrates of NMD from indirectly affected transcripts.

When NMD is reduced or eliminated, expression of both direct substrates and indirectly affected transcripts is altered. PTC-containing messages are stabilized, but NMD substrates cannot be identified by this altered abundance alone, as there is also an indirect upregulation (or downregulation) of many transcripts. I chose to exploit SMG-2's selective and preferential association with NMD substrates in order to clarify which transcripts are true substrates. Since SMG-2 plays a role in multiple RNA degradation pathways, it was important not only that I ensure that SMG-2 associated RNAs were stabilized in an NMD(-) strain but also that SMG-2 immunoprecipitations were performed in a substrate-rich environment. When SMG-2 is immunoprecipitated in *smg-1(null)* nematodes, substrate mRNAs are strongly enriched in the IP pellet [37]. SMG-2 "marks" PTC-containing messages under such conditions.

RNA-sequencing data was generated for NMD(+) and NMD(-) nematodes, and SMG-2 RIP-Seq was used to discriminate true NMD substrates from indirectly affected transcripts. Here, I describe analysis of the resulting data using multiple statistical packages, identification of NMD substrates, and characterization of NMD-regulated genes via functional enrichment.

Results

Sample Preparation & Immunoprecipitation of SMG-2

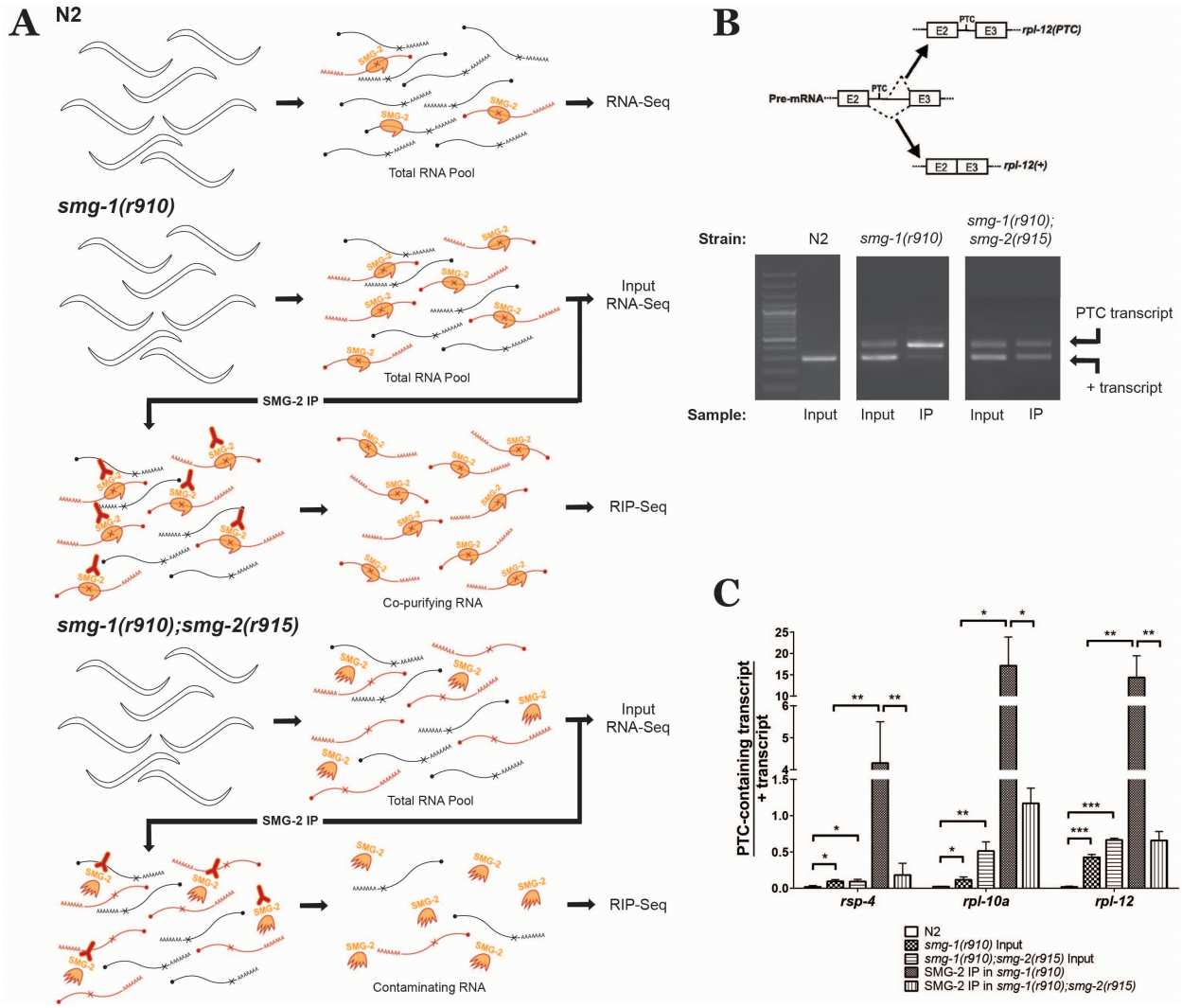
To analyze gene expression in the absence of NMD, I performed deep sequencing on RNA extracted from staged N2(CGC), smg-*1(r910),* and *smg-1(r910) smg-2(r915)* L4 larvae. *Smg-1(r910)* is a null allele created by a transposon insertion in *smg-1*. SMG-1 protein is not detected on a western blot and these strains are strongly NMD defective (L. Johns and P. Anderson, unpublished). *Smg-2(r915)* is a 1,064 bp deletion within *smg-2* [38]. *Smg-2(r915)* expresses a truncated SMG-2 protein that is nonfunctional, as demonstrated by the strong NMD defect of *smg-2(r915)*, which is similar to that of the null allele *smg-2(r908)* (L. Johns and P. Anderson unpublished data). *Smg-2(r915)* was used because *smg-2(r908)* also deletes a large portion of an upstream gene. Total RNA samples were DNase treated to reduce gDNA contamination of the sequencing sample.

To identify which transcripts the NMD pathway directly targets, I isolated RNA that co-purifies with SMG-2 (Figure 2.1A). I performed RNA immunoprecipitation coupled with sequencing (RIP-Seq) in *smg-1(r910)* mutants, in which expression levels of both direct and indirect targets should be affected, and in which direct substrates are marked by association with SMG-2 [37]. Negative control RIP-Seq was performed in *smg-1(r910) smg-2(r915)* animals, in which SMG-2 cannot be immunoprecipitated. *Smg-1(r910) smg-2(r915)* was used to minimize any variability between my experimental and control samples. RNA sequenced from these negative control samples was used to set a threshold for RIP contamination. IPs were not performed in wild-type extracts due to the rapid degradation of NMD substrates and their

Figure 2.1: SMG-2 preferentially associates with PTC-containing mRNAs

(A) Schematic diagram of SMG-2 IPs performed in this study. (B) mRNPs containing SMG-2 were immunoprecipitated from wild-type (N2) and *smg(-)* extracts. RNA from the IP pellets and a small portion of the input samples was subjected to RT-PCR for validation of known substrates of NMD prior to sequencing of the samples. Gene schematic and PCR products for *rpl-12* shown. Both the schematic and gel are representative of other tested positive controls. (C) Semi-quantitative PCR used to validate IP efficacy prior to high-throughput sequencing. Columns show the measured ratio of PTC-containing transcript to "+" (non-PTC-containing) transcripts for each gene. Error bars indicate standard error of the mean. P-values generated via Student's t-test. Asterisks indicate significance: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Figure 2.1


A
N2
RNA-Seq
Total RNA Pool
smg-1(r910)
Input
RNA-Seq
Total RNA Pool
SMG-2 IP
RIP-Seq
Co-purifying RNA
smg-1(r910);smg-2(r915)
Input
RNA-Seq
Total RNA Pool
SMG-2 IP
RIP-Seq
Contaminating RNA
B
E2
PTC
E3
rpl-12(PTC)
Pre-mRNA
rpl-12(+)
Strain:
N2
smg-1(r910)
smg-1(r910); smg-2(r915)
PTC transcript
+ transcript
Sample:
Input
Input
IP
Input
IP
C
PTC-containing transcript / + transcript
rsp-4
rpl-10a
rpl-12
N2
smg-1(r910) Input
smg-1(r910);smg-2(r915) Input
SMG-2 IP in smg-1(r910)
SMG-2 IP in smg-1(r910);smg-2(r915)

resulting low abundance. Prior to high-throughput sequencing, I verified the efficacy of the IP using previously described NMD targets *rpl-12*, *rsp-4*, and *rpl-10a* [37]. Each of these genes produces a larger, PTC-containing, spliced mRNA that is stabilized in an NMD(-) background and is enriched in a SMG-2 IP, as seen in my validation (Figure 2.1B & C).

High-Throughput Sequencing

Library preparation and sequencing were performed by the UW Biotechnology Center DNA Sequencing Facility. RNA samples were poly(A) selected to reduce ribosomal RNA contamination and amplified by 15 cycles of PCR. While RT-PCR validation showed moderate levels of contamination in the negative control [*smg-1(r910) smg-2(r915)* strain] IP, quantification of my submitted samples demonstrated approximately a 4-fold reduction in the amount of RNA in the negative control IP compared to the experimental IP. Input RNA samples were not markedly different in abundance between the two strains. Following poly(A) selection and library preparation, the negative control IP samples contained roughly one-eighth the amount of cDNA that the experimental IPs contained. Together these results indicate that my SMG-2 IP preferentially pulls down SMG-2 bound transcripts. Pre-sequencing quantification for the three biological replicates is provided in Table 2.1.

One-hundred base pair, single-end reads were generated on the Illumina HiSeq 2000 platform. Approximately 80-140 million reads were generated for each input sample and 65-87 million reads for each IP sample. Following general quality checks in FastQC, a gentle sliding-window quality trim was performed using Trimmomatic [103]. Bases within the first or last 10nt with quality-scores below 20 were clipped; the sliding window trim cut the sequence when four consecutive bases' quality score averaged below 20; and remaining reads with less than

25bp were discarded. 86-98% of reads per sample survived the trim. These reads were aligned to the WBcel235 genome release (available through Ensembl) using Tophat2, a splice-aware aligner [104]. I largely used the Tophat2 default settings, but the Ensembl gene annotation GTF file downloaded from Illumina's iGenomes was used with the program. 3-12% of the trimmed reads for each sample could not be mapped, and 3-14% of aligned reads mapped to multiple locations. 69-116 million reads per input sample and 39-73 million reads per IP sample were uniquely aligned to the genome. A chart outlining the reads that survived each processing step is provided in Table 2.2.

In preparation for differential expression analysis, I then used HTSeq-count to count how many reads aligned to each gene [105]. Reads were counted as aligning to genes rather than transcripts based on the designer's recommendation; otherwise, when multiple transcripts share sequence (e.g. a shared exon), reads that align to that sequence are counted as ambiguous and discarded. After counting reads that aligned to genes using htseq-count, I found reads representing 14,777 genomic features that met an expression threshold of three counts per million.

Table 2.1: Sample Quantification

Sample	Pre-Library Prep Concentrations (Agilent)	Amount of RNA Used (Total IP material)	Post-Library Prep Concentrations (PicoGreen)	Material Used for Sequencing
N2 Input 1	412.60ng/uL	500ng	6.7ng/uL	200.7ng
N2 Input 2	284.90ng/uL	500ng	5.1ng/uL	153.6ng
N2 Input 3	243.00ng/uL	500ng	20.8ng/uL	23.1ng
smg-1(r910) Input 1	346.60ng/uL	500ng	12.5ng/uL	373.5ng
smg-1(r910) Input 2	235.00ng/uL	500ng	13.0ng/uL	388.5ng
smg-1(r910) Input 3	252.10ng/uL	500ng	24.3ng/uL	729.9ng
smg-1(r910) smg-2(r915) Input 1	255.80ng/uL	500ng	9.0ng/uL	268.8ng
smg-1(r910) smg-2(r915) Input 2	259.80ng/uL	500ng	13.8ng/uL	412.8ng
smg-1(r910) smg-2(r915) Input 3	292.50ng/uL	500ng	24.3ng/uL	729.9ng
smg-1(r910) IP 1	1.9ng/uL	37.75 ng	25.2ng/uL	756.0ng
smg-1(r910) IP 2	1.5ng/uL	262.5 ng	15.9ng/uL	475.8ng
smg-1(r910) IP 3	5.4ng/uL	945.0 ng	19.4ng/uL	581.1ng
smg-1(r910) smg-2(r915) IP 1	0.6ng/uL	102.9 ng	2.1ng/uL	61.5ng
smg-1(r910) smg-2(r915) IP 2	0.3ng/uL	45.85 ng	1.6ng/uL	47.4ng
smg-1(r910) smg-2(r915) IP 3	1.3ng/uL	227.5 ng	7.7ng/uL	229.8ng

Table 2.2: Read counts throughout the bioinformatic pipeline

Sample	Reads	Survive Trim	% Survived Trim	Mapped	% Mapped from Trim	Multiple Mapped	% of mapped	Single Mapped Reads	% Single Mapped from Original # Reads
N2 Input 1	101,381,500	98,644,238	97.3%	92,639,538	93.9%	5,650,580	6.1%	86,988,958	85.8%
N2 Input 2	140,139,474	136,769,327	97.6%	129,270,813	94.5%	13,592,157	10.5%	115,678,656	82.5%
N2 Input 3	91,413,733	84,992,320	93.0%	82,100,030	96.6%	3,942,348	4.8%	78,157,682	85.5%
smg-1(r910) Input 1	144,071,357	140,388,023	97.4%	135,348,943	96.4%	19,238,873	14.2%	116,110,070	80.6%
smg-1(r910) Input 2	118,543,991	115,717,113	97.6%	110,806,234	95.8%	5,778,892	5.2%	105,027,342	88.6%
smg-1(r910) Input 3	115,346,635	112,185,383	97.3%	108,989,609	97.2%	4,579,186	4.2%	104,410,423	90.5%
smg-1(r910) IP 1	87,490,899	78,591,519	89.8%	75,351,712	95.9%	2,497,642	3.3%	72,854,070	83.3%
smg-1(r910) IP 2	69,282,528	62,371,278	90.0%	59,247,329	95.0%	2,728,763	4.6%	56,518,566	81.6%
smg-1(r910) IP 3	64,956,756	58,218,307	89.6%	55,552,558	95.4%	1,675,345	3.0%	53,877,213	82.9%
smg-1(r910) smg-2(r915) Input 1	110,486,068	107,464,092	97.3%	101,955,898	94.9%	7,778,283	7.6%	94,177,615	85.2%
smg-1(r910) smg-2(r915) Input 2	78,910,455	77,181,471	97.8%	74,102,768	96.0%	5,395,695	7.3%	68,707,073	87.1%
smg-1(r910) smg-2(r915) Input 3	119,575,155	108,057,269	90.4%	104,998,554	97.2%	4,896,267	4.7%	100,102,287	83.7%
smg-1(r910) smg-2(r915) IP 1	53,619,833	46,423,323	86.6%	42,282,153	91.1%	2,857,803	6.8%	39,424,350	73.5%
smg-1(r910) smg-2(r915) IP 2	65,015,012	58,305,297	89.7%	51,489,478	88.3%	5,485,073	10.7%	46,004,405	70.8%
smg-1(r910) smg-2(r915) IP 3	58,187,815	50,189,291	86.3%	47,327,048	94.3%	1,864,645	3.9%	45,462,403	78.1%

Differential Expression Analysis

To identify both direct substrates of NMD as well as indirectly affected transcripts I applied three criteria: (1) Transcripts whose expression is increased in the absence of NMD, (2) Transcripts that are enriched in the SMG-2 IP relative to the matched *smg-1(r910)* input sample, and (3) Transcripts that are enriched in the SMG-2 IP relative to the negative control [*smg-1(r910) smg-2(r915)* double mutant] IP. I classified NMD substrates as those that are stabilized in an NMD(-) background and reliably enriched in a SMG-2 IP. I classified transcripts indirectly affected by NMD as those that are up- or down-regulated in an NMD(-) background but are not enriched in a SMG-2 IP.

To make these comparisons, I tested three common software packages that calculate differential expression: baySeq, DESeq2, and edgeR [106-108]. All three of these programs assume a negative binomial distribution of the sequencing data. However, edgeR and DESeq2 perform much more similar statistical analyses to one another than they do to baySeq. Seyednasrollah, et al, provide a brief overview and comparison of these packages [109]. Differential expression was calculated for each of the three comparisons listed above using each of the three Bioconductor software packages. Six known NMD substrates were used to determine which of the three packages best identified NMD substrates*: rsp-2, rsp-4, rsp-6, rpl-7a, rpl-10a,* and *rpl-12* (L. Johns and P. Anderson, unpublished data & [37]).

baySeq

The baySeq program uses an empirical Bayesian method to calculate estimated posterior likelihoods of differential expression. Basic differential expression analysis, based on the vignettes and reference manual available through Bioconductor, was performed to

calculate significant changes in the following comparisons: (1) NMD(+) vs NMD(-) RNA-Seq, (2) *smg-1(r910)* RIP-Seq vs *smg-1(r910)* input RNA-Seq, and (3) *smg-1(r910)* RIP-Seq vs *smg-1(r910) smg-2(r915)* negative control RIP-Seq. These three comparisons are outlined in Figure 2.2. Practically speaking, this package was the most computationally demanding and provided the smallest NMD substrate list.

Differences in wild-type and NMD-mutant transcriptomes

Wild-type RNA-Seq samples were compared to mutant input samples. Using baySeq to call differential expression, 529 of the 14,777 expressed genomic features (3.6%) are differentially upregulated in an NMD(-) background with a false discovery rate (FDR) < 0.05 and a fold-change greater than 1.5. The maximum log2 upregulation was 6.9-fold in *smg-1(r910)* and 7.9-fold in *smg-1(r910) smg-2(r915)* compared with control samples. A total of 84 genes were downregulated in NMD(-) strains. The maximum log2 downregulation was 12.2-fold in *smg-1(r910)* and 10-fold in *smg-1(r910) smg-2(r915)* compared with control samples. No known substrates were differentially expressed between NMD(+) and NMD(-) samples, according to baySeq.

Identification of SMG-2 associated RNAs

I defined SMG-2-associated RNAs as those enriched in the SMG-2 IP in a *smg-1(r910)* strain compared to the *smg-1(r910)* input sample and those enriched in the SMG-2 IP in a *smg-1(r910)* strain compared to the negative control IP. SMG-2-associated RNAs include both true direct substrates of NMD and RNAs that may either be targeted by SMG-2 for other pathways or briefly associated with SMG-2.

After comparing 1,370 genes with expression 2-fold enriched in the SMG-2 IPs relative to the negative control IPs with 1,961 genes with expression 2-fold enriched in the SMG-2 IPs relative to input samples, I identified 1,132 RNAs that are reliably associated with SMG-2. Among the SMG-2-associated RNAs, I found two out of six previously-established direct NMD substrates: *rsp-2* and *rsp-6*.

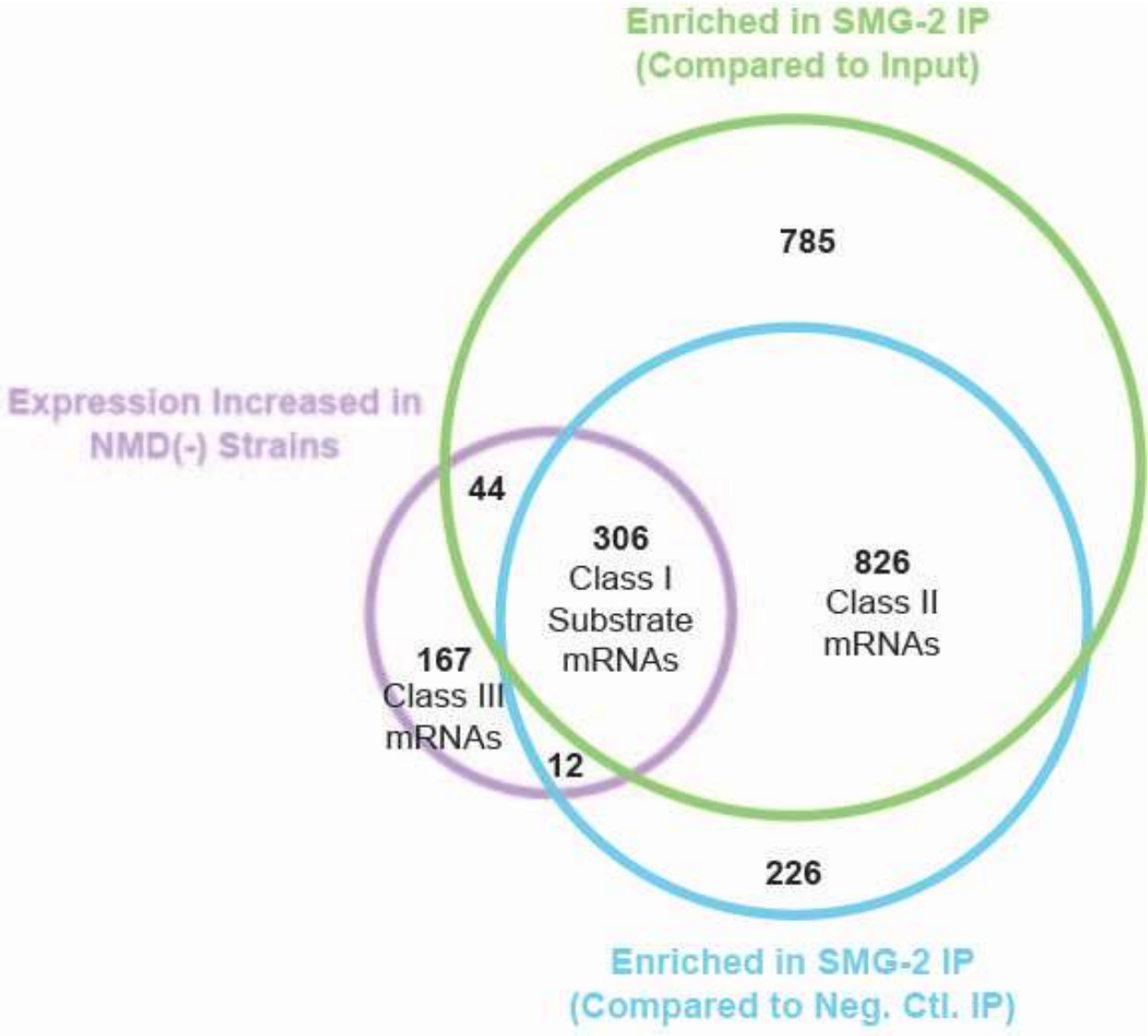
Comparison of mutant expression changes and SMG-2 associated RNAs

After combining the expression data from my input RNA-Seq experiment with my RIP-Seq experiment, I analyzed genes belonging to one of three classes. Class I genomic features, exhibit increased expression in NMD(-) strains and have transcripts that are enriched in a SMG-2 IP. These 306 RNAs (comprising 27% of SMG-2 associated RNAs, 57.8% of RNAs with increased expression in a *smg(-)* strain, and 2.1% of expressed features) are direct substrates of NMD. Notably, baySeq did not identify any previously-established NMD substrates in its Class I list. However, all NMD substrates identified by baySeq were corroborated by the other two DE-calling packages. Class II genes, including *rsp-2* and *rsp-6*, have transcripts that co-precipitate with SMG-2 but that do not demonstrate significantly increased expression in NMD(-) strains. This set of 826 features likely includes some true substrates of NMD that exhibit a shift in the abundance of specific transcripts but not an overall increase in gene expression. 167 class III genes exhibit increased expression in NMD(-) strains but are never enriched in a SMG-2 IP. Lastly, 81 class IV RNAs exhibit decreased expression in NMD(-) strains and are never enriched in a SMG-2 IP. Taken together, class III and IV genes represent the indirect effects of the NMD pathway.

Figure 2.2: Determination of NMD direct substrates and indirect effects by baySeq

Venn diagram detailing overlap between genes exhibiting at least 1.5-fold increased expression in *smg-1(r910)* and *smg-1(r910) smg-2(r915)* samples compared to N2, genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to its input sample, and genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to negative control IPs. Genes that meet all three conditions are classified as Class I NMD Substrates. Genes with transcripts that are reliably SMG-2 associated but that do not exhibit increased expression in a NMD(-) animal are deemed Class II, and genes with elevated expression in a NMD(-) sample but which never create mRNAs that are enriched in a SMG-2 IP are deemed Class III.

Figure 2.2


Enriched in SMG-2 IP
(Compared to Input)
785
Expression Increased in
NMD(-) Strains
44
306
Class I
Substrate
mRNAs
826
Class II
mRNAs
167
Class III
mRNAs
12
226
Enriched in SMG-2 IP
(Compared to Neg. Ctl. IP)

DESeq2

DESeq2 determines differential expression using the negative binomial distribution and empirical Bayes shrinkage for both dispersion and fold-change estimation to better manage small replicate numbers. The multi-factor design described in the vignettes was leveraged to account for paired samples. This package is the only one to attempt to identify and account for outliers in the sample data and created the greatest range of graphical output. DESeq2 analysis is summarized in Figure 2.3

Differences in wild-type and NMD-mutant transcriptomes

Multifactor-designed DESeq2 identified 1,513 of the 14,777 expressed genes (10.2%) as differentially upregulated in an NMD(-) background with FDR < 0.05 and a fold-change greater than 1.5. The maximum log2 upregulation was 5.4-fold in *smg-1(r910)* and 5.5-fold in *smg-1(r910) smg-2(r915)* compared with control samples. Genes demonstrating increased expression in NMD-deficient strains included four of the six known NMD substrates: *rsp-2, rsp-6, rpl-7a,* and *rpl-12*. A total of 368 genes were downregulated in NMD(-) strains. The maximum log2 downregulation was 9.3-fold in *smg-1(r910)* and 6.8-fold in *smg-1(r910) smg-2(r915)* compared with control samples. (Figure 2.3B)

Identification of SMG-2 associated RNAs

3,428 genes had RNAs at least 2-fold enriched in SMG-2 IPs performed in *smg-1(r910)* extracts relative to negative control IPs (Figure 2.3D), and 4,582 genes had RNAs at least 2-fold enriched in SMG-2 IPs relative to input samples (Figure 2.3C). 3,203 genes were shared

between these two IP-enriched lists, including positive control genes *rsp-2, rsp-4, rsp-6,* and *rpl-7a*.

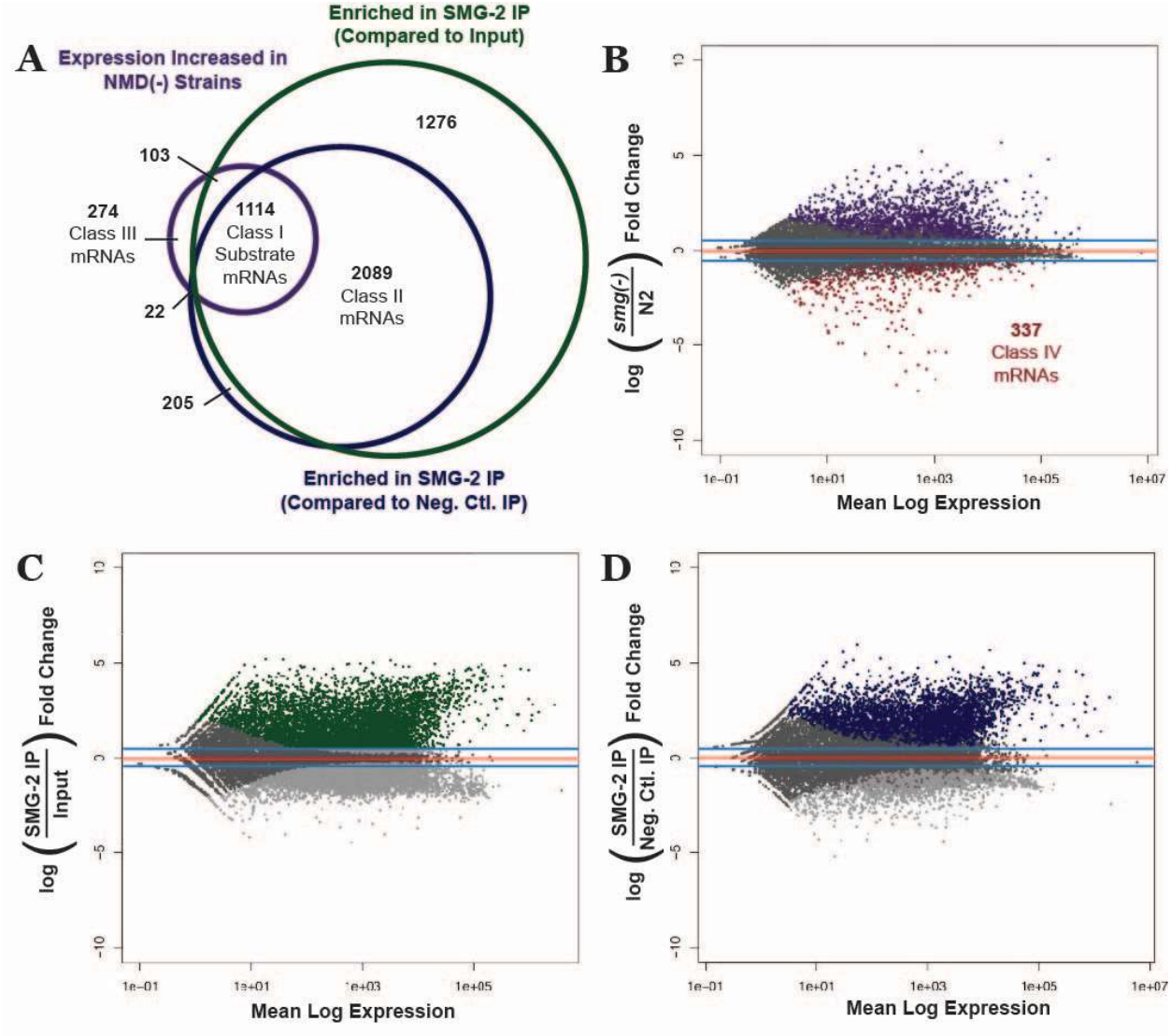
Comparison of mutant expression changes and SMG-2 associated RNAs

Input and IP comparisons were overlapped to establish expression categories, as above (Figure 2.3A). The class I category, with increased expression in NMD(-) strains and enrichment in a SMG-2 IP, included 1114 genes following DESeq2 analysis. Class I substrates comprised 34.8% of SMG-2 associated RNAs, 73.6% of RNAs upregulated in a *smg(-)* strain, and 7.5% of expressed features. DESeq2 identified *rsp-2, rsp-6*, and *rpl-7a* as direct substrates of NMD. Class II genomic features, including *rsp-4*, have transcripts that co-precipitate with SMG-2 but that do not demonstrate significantly increased expression in NMD(-) strains. DESeq2 analysis defined 765 Class II genes. Indirectly affected mRNAs, which are not enriched in the SMG-2 IPs, mapped to 274 class III genes with increased expression in NMD(-) strains and 337 class IV genes with decreased expression in NMD(-) strains.

Figure 2.3: Determination of NMD direct substrates and indirect effects by DESeq2

(A) Venn diagram detailing overlap between genes exhibiting at least 1.5-fold increased expression in *smg-1(r910)* and *smg-1(r910) smg-2(r915)* samples compared to N2, genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to its input sample, and genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to negative control IPs. Genes that meet all three conditions are classified as Class I NMD Substrates. Genes with transcripts that are reliably SMG-2 associated but that do not exhibit increased expression in a NMD(-) animal are deemed Class II, and genes with elevated expression in a NMD(-) sample but which never create mRNAs that are enriched in a SMG-2 IP are deemed Class III. (B) Scatterplot of all genes (dots) with their average expression level on the x-axis and their relative expression change in NMD(-) mutants on the y-axis. Significantly upregulated and downregulated genes (FDR<0.05) are represented by purple and red dots, respectively. Blue lines mark the 1.5-fold change threshold used in analysis of these input samples. (C&D) Scatterplot of all genes (dots) with their average expression level on the x-axes and their relative expression change between SMG-2 IPs performed in *smg-1(r910)* nematodes and the *smg-1(r910)* input samples or the SMG-2 negative control IPs performed in *smg-1(r910) smg-2(r915)* nematodes on the y-axes, respectively. Significantly enriched genes (FDR<0.05) that meet the 2-fold change threshold (blue lines) are shown in green (C) and blue (D) dots.

Figure 2.3


A
Expression Increased in NMD(-) Strains
Enriched in SMG-2 IP (Compared to Input)
Enriched in SMG-2 IP (Compared to Neg. Ctl. IP)
1276
103
274 Class III mRNAs
1114 Class I Substrate mRNAs
2089 Class II mRNAs
22
205
B
log (smg(-) / N2) Fold Change
337 Class IV mRNAs
Mean Log Expression
C
log (SMG-2 IP / Input) Fold Change
Mean Log Expression
D
log (SMG-2 IP / Neg. Ctl. IP) Fold Change
Mean Log Expression

edgeR

The edgeR package utilizes a number of statistical methods, including empirical Bayes methods and exact tests, to determine differential expression. I used the classic/pairwise approach described in the user guide and in the protocol presented by Anders, et al (2013). edgeR was the fastest-running of the three programs and was the simplest to use with paired samples (i.e. IP and input samples from the same biological replicate).

Differences in wild-type and NMD-mutant transcriptomes

Using edgeR to call differential expression, 699 of the 14,777 expressed genomic features (4.7%) are differentially upregulated in an NMD(-) background with FDR < 0.05 and a fold-change greater than 1.5. The maximum log2 upregulation was 6.3-fold in *smg-1(r910)* and 6.1-fold in *smg-1(r910) smg-2(r915)* compared with control samples. *Rsp-2, rsp-6,* and *rpl-7a* exhibited increased expression in NMD(-) strains in this analysis. A total of 74 genes were downregulated in NMD(-) strains. The maximum log2 downregulation was 12.3-fold in *smg-1(r910)* and 9-fold in *smg-1(r910) smg-2(r915)* compared with control samples. (Figure 2.4B)

Identification of SMG-2 associated RNAs

After comparing 3440 genes with expression 2-fold enriched in the SMG-2 IPs relative to the negative control IPs with 2700 genes with expression 2-fold enriched in the SMG-2 IPs relative to input samples, I identified 2588 RNAs that are reliably associated with SMG-2 (Figure 2.4C & D). Among the SMG-2-associated RNAs, I found five out of six previously-established direct NMD substrates, including *rpl-7a*, *rpl-12*, *rsp-2*, *rsp-4*, and *rsp-6*.

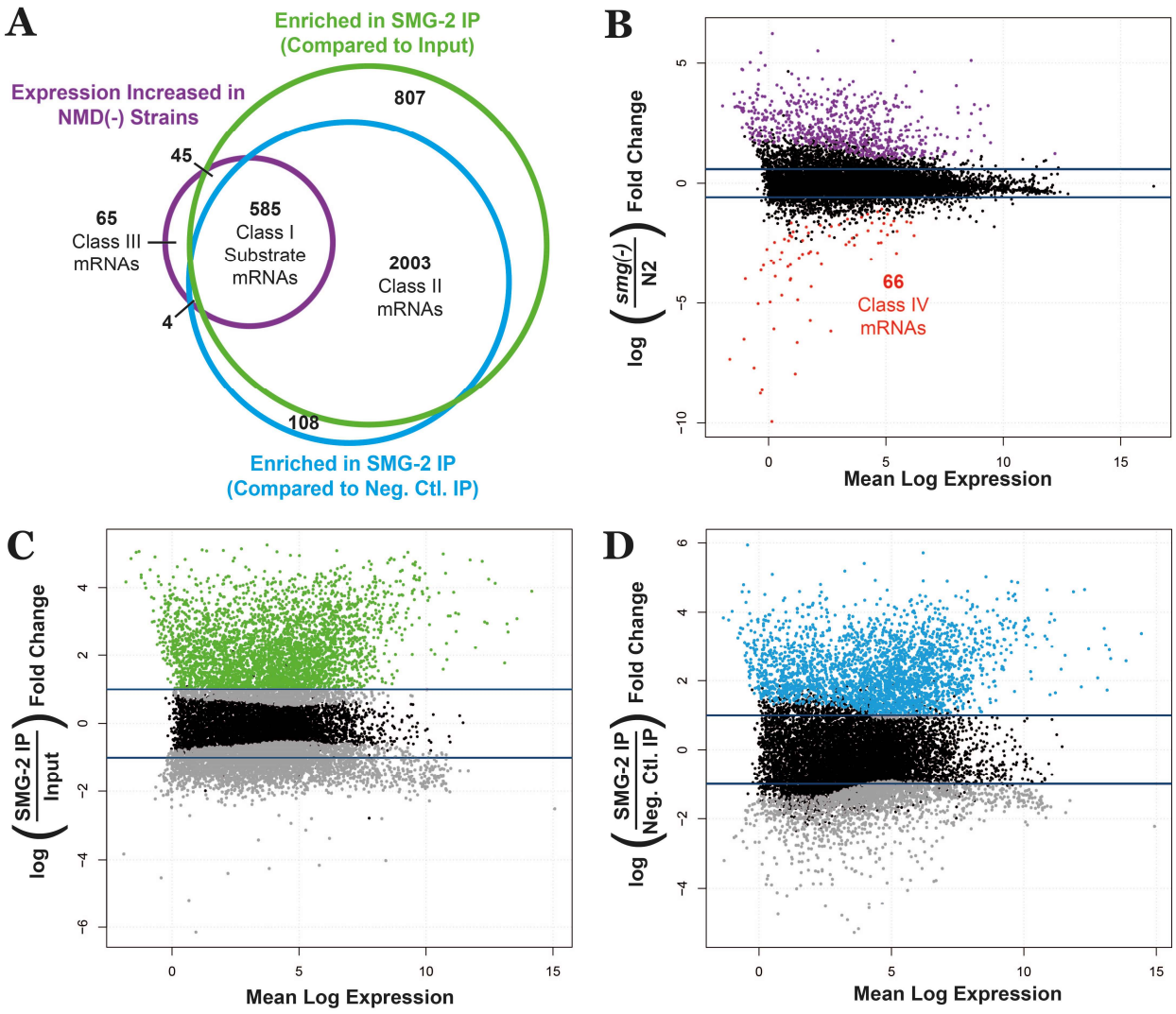
Comparison of mutant expression changes and SMG-2 associated RNAs

After combining the expression data from my input RNA-Seq experiment with my RIP-Seq experiment, I once again classified genes with specific expression patterns (Figure 2.4A). Class I RNAs, including *rsp-2*, *rsp-6*, and *rpl-7a*, exhibit increased expression in NMD(-) strains and have transcripts that are enriched in a SMG-2 IP. These 585 features (comprising 22.5% of SMG-2 associated RNAs, 83.8% of RNAs changing in expression in a *smg(-)* strain, and 2.4% of expressed features) are high-confidence direct substrates of NMD. Class II transcripts, including *rsp-4* (a previously known substrate of NMD), have transcripts that co-precipitate with SMG-2 but that do not demonstrate significantly increased expression in NMD(-) strains. This set of 2003 features likely includes some true substrates of NMD that exhibit a shift in transcript abundance but not an overall increase in gene expression. Sixty-five class III genes exhibit increased expression in NMD(-) strains but are never enriched in a SMG-2 IP. Lastly, 66 class IV transcripts exhibit decreased expression in NMD(-) strains and are never enriched in a SMG-2 IP.

Figure 2.4: Determination of NMD direct substrates and indirect effects by edgeR

(A) Venn diagram detailing overlap between genes exhibiting at least 1.5-fold increased expression in *smg-1(r910)* and *smg-1(r910) smg-2(r915)* samples compared to N2, genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to its input sample, and genes with transcripts enriched at least 2-fold in a SMG-2 IP relative to negative control IPs. Genes that meet all three conditions are classified as Class I NMD Substrates. Genes with transcripts that are reliably SMG-2 associated but that do not exhibit increased expression in a NMD(-) animal are deemed Class II, and genes with elevated expression in a NMD(-) sample but which never create mRNAs that are enriched in a SMG-2 IP are deemed Class III. (B) Scatterplot of all genes (dots) with their average expression level on the x-axis and their relative expression change in NMD(-) mutants on the y-axis. Significantly upregulated and downregulated genes (FDR<0.05) are represented by purple and red dots, respectively. Blue lines mark the 1.5-fold change threshold used in analysis of these input samples. (C&D) Scatterplot of all genes (dots) with their average expression level on the x-axes and their relative expression change between SMG-2 IPs performed in *smg-1(r910)* nematodes and the *smg-1(r910)* input samples or the SMG-2 negative control IPs performed in *smg-1(r910) smg-2(r915)* nematodes on the y-axes, respectively. Significantly enriched genes (FDR<0.05) that meet the 2-fold change threshold (blue lines) are shown in green (C) and blue (D) dots.

Figure 2.4


A
Enriched in SMG-2 IP
(Compared to Input)
Expression Increased in
NMD(-) Strains
807
45
65
Class III
mRNAs
585
Class I
Substrate
mRNAs
2003
Class II
mRNAs
4
108
Enriched in SMG-2 IP
(Compared to Neg. Ctl. IP)
B
log (smg(-)/N2) Fold Change
66
Class IV
mRNAs
Mean Log Expression
C
log (SMG-2 IP/Input) Fold Change
Mean Log Expression
D
log (SMG-2 IP/Neg. Ctl. IP) Fold Change
Mean Log Expression

Defining & Validating Direct Substrates of NMD

In the interest of generating a high-confidence list of substrates, I chose to use edgeR results for validation and all subsequent analysis. Using baySeq I was unable to identify any of my six previously identified NMD substrates. Both edgeR and DESeq2 identified *rsp-2*, *rsp-6*, and *rpl-7a* as direct substrates of NMD. However, DESeq2's target list was nearly twice the size of edgeR's. While the wealth of information provided by DESeq2 was appealing, the stringency provided by edgeR analysis was preferable. A comparison of substrates identified by each program is provided in Figure 2.5.

To validate expression changes observed in the sequencing data, quantitative RT-PCR was performed using cDNA generated for IP quality control tests (see above). I confirmed alterations in RNA abundance for 10 Class I transcripts (*B0495.8, F45D11.1, fbxa-33, linc-9, nhr-109, pqn-70, rpl-7a, rsp-2, rsp-6,* and *tdp-1* – Figure 2.6A & B), two Class II transcripts (*fib-1* and *pho-11* – Figure 2.6C), two Class III transcripts (*F53B2.6* and *Y39B6A.21* – Figure 2.6D), and three Class IV transcripts (*R08E5.2, Y51A2D.13,* and *ZK970.7* – Figure 2.6E). All 17 sets of qPCR results mimicked the trends that emerged from sequencing analysis. All Class I and III genes had increased mRNA levels in NMD(-) extracts compared to wild-type. Class II genes showed no significant elevation in expression in NMD-deficient strains, and Class IV genes exhibited a marked decrease in expression in the absence of NMD. Lastly, Class I and II transcripts were enriched in SMG-2 IPs performed in *smg-1(r910)* extracts compared to both *smg-1(r910)* input samples and negative control IPs.

Figure 2.5: Comparison of differential expression programs

Venn diagram detailing overlap between Class I NMD substrates as identified by baySeq (light blue), DESeq2 (purple), and edgeR (medium blue) statistical packages.

Figure 2.5

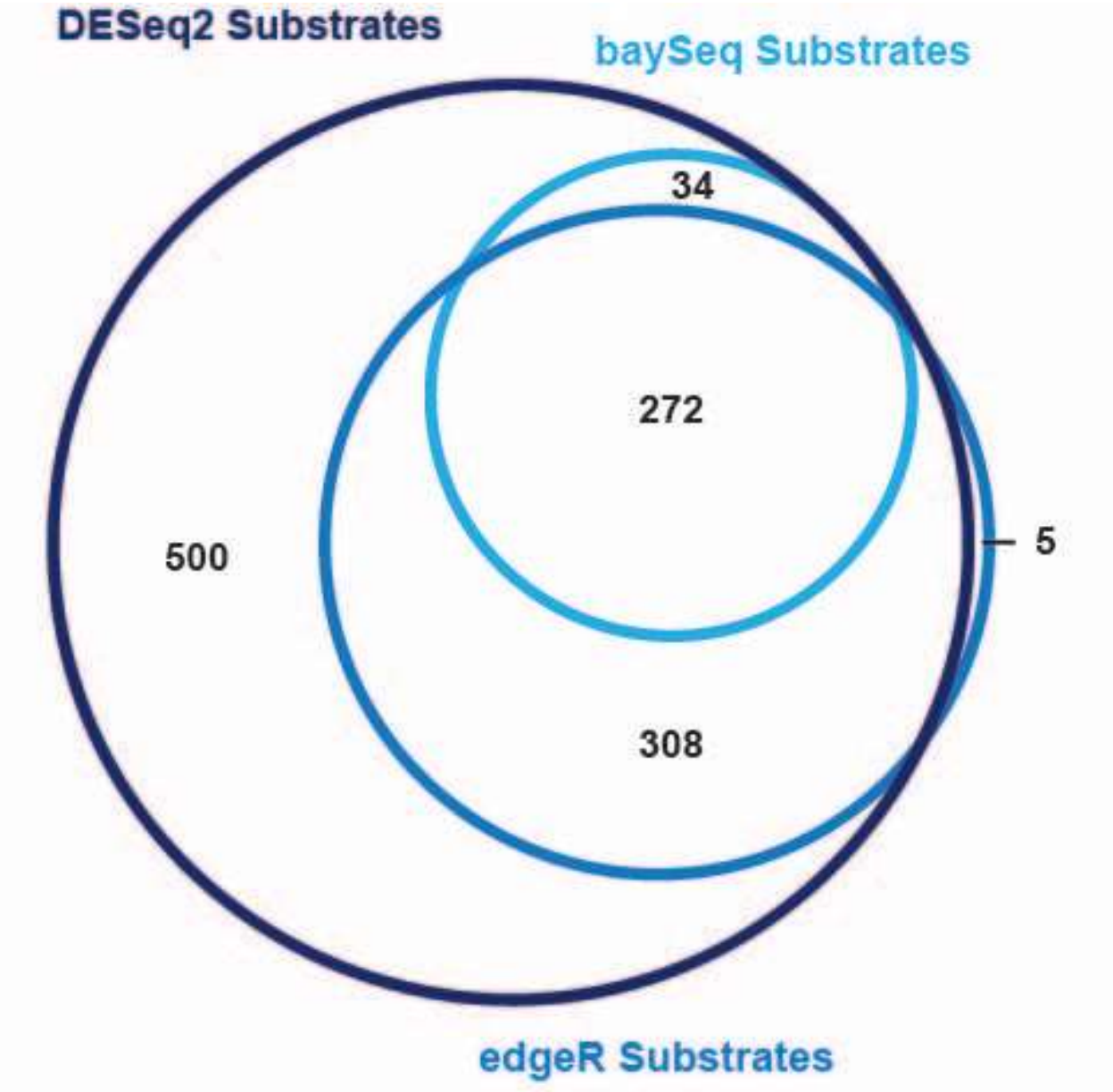
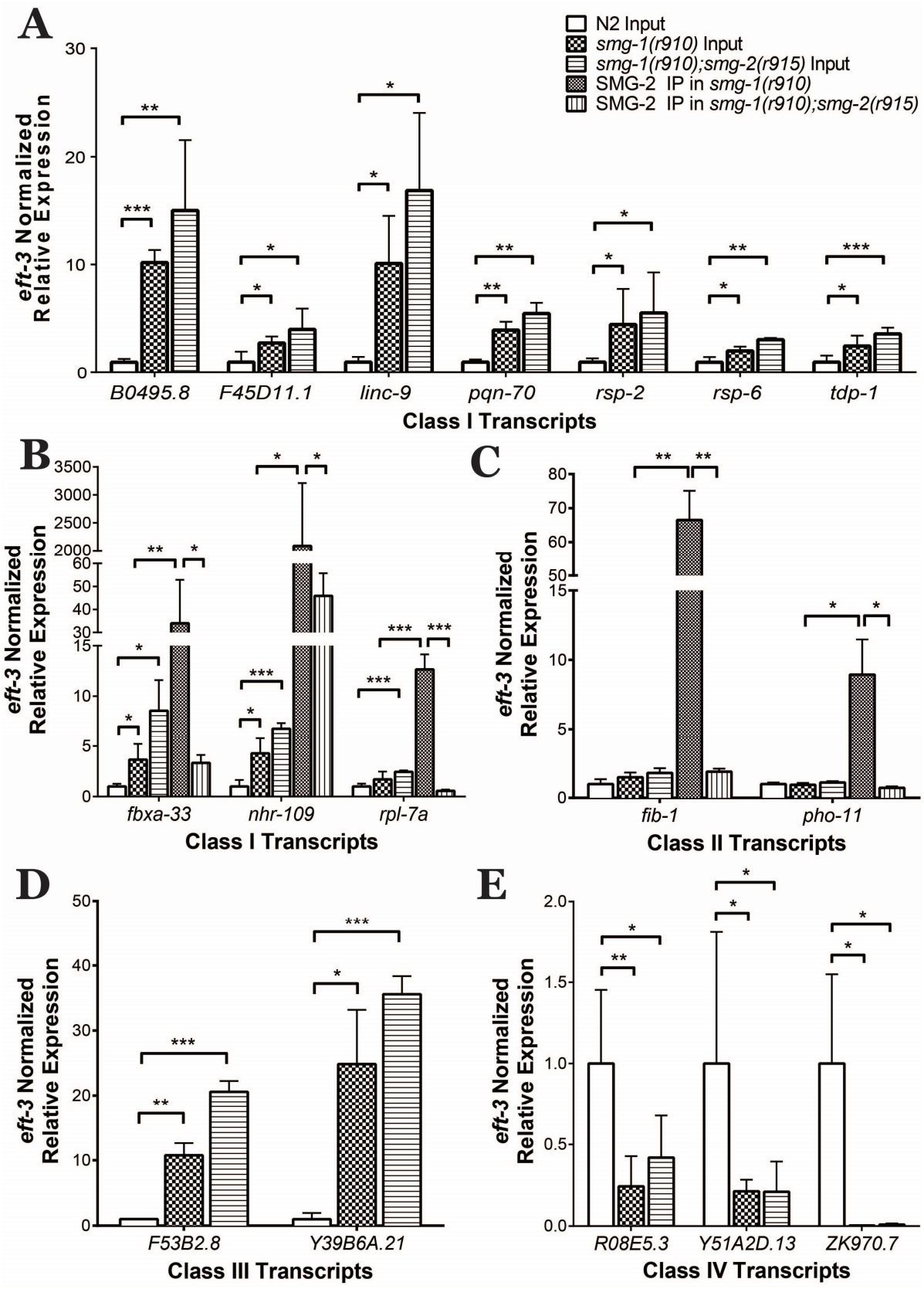

DESeq2 Substrates
baySeq Substrates
34
272
500
5
308
edgeR Substrates

Figure 2.6: Validation of sequencing data.

(A-E) qRT-PCR validation of expression changes and IP enrichments seen in high-throughput sequencing. Expression of (A&B) Class I, (C) Class II, (D) Class III , and (E) Class IV transcripts in input and IP samples (B & C only) was quantified using qRT-PCR. All expression values were normalized to *eft-3* mRNA levels. Each column indicates the expression level relative to wild-type samples. Error bars indicate standard error of the mean, and statistical significance was calculated using Student's t-test. Asterisks indicate significance: * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Figure 2.6


A
N2 Input
smg-1(r910) Input
smg-1(r910);smg-2(r915) Input
SMG-2 IP in smg-1(r910)
SMG-2 IP in smg-1(r910);smg-2(r915)
eft-3 Normalized Relative Expression
B0495.8
F45D11.1
linc-9
pqn-70
rsp-2
rsp-6
tdp-1
Class I Transcripts
B
fbxa-33
nhr-109
rpl-7a
Class I Transcripts
C
fib-1
pho-11
Class II Transcripts
D
F53B2.8
Y39B6A.21
Class III Transcripts
E
R08E5.3
Y51A2D.13
ZK970.7
Class IV Transcripts

Functional Enrichment of NMD-Regulated Genes

To better understand the physiological role of NMD, all NMD-responsive gene classes were tested for functional enrichment and statistical overrepresentation using GOrilla and PANTHER. I tested terms originating from all three gene ontology (GO) domains: cellular component, molecular function, and biological process. Most of the significantly enriched GO terms described below are members of the latter two domains. Gene lists were compared to a list of the 14,777 expressed genes in order to minimize false enrichments due to, for example, the samples' poly(A) selection (e.g. when comparing to the genome) or developmental staging (e.g. when comparing to protein-coding genes). As an example, I initially identified transcripts from operons as a significantly enriched subset of Class I substrates when compared to the genome. However, transcripts originating from operons are almost entirely protein coding, while roughly half of annotated genomic features are small or non-coding RNAs. Since my poly(A)-selected samples largely included protein-coding features, I skewed enrichment scoring by initially comparing to the genome rather than to expressed features. Similarly, sequencing was performed on L4 larval materials; thus, using a background list that included all larval stages would, by default, create a bias in the analysis.

GOrilla's functional annotation enrichment shows that Class I NMD substrates are enriched for mRNA processing, splicing, polyamine metabolism, D-amino acid metabolic processes, and nucleic acid binding (Figure 2.7 & 2.8). PANTHER found that these additional GO terms were overrepresented: RNA binding, heterocycle metabolic process, nucleobase-containing compound metabolic process, cellular aromatic compound metabolic process,

organic cyclic compound metabolic process, RNA binding, and nucleobase-containing compound metabolic process.

Class II transcripts, which are reliably SMG-2 associated but are not significantly more abundant in NMD(-) samples, show enrichment for RNA metabolic factors and also for factors involved in a number of biosynthetic and metabolic processes. The five most significantly enriched annotations according to GOrilla included nucleic acid binding, heterocyclic compound binding/metabolic process, organic cyclic compound binding/metabolic process, cellular aromatic compound metabolic process, and transcription factor activity/sequence-specific DNA binding (Figure 2.9 & 2.10). Top GO terms in PANTHER analysis only added RNA metabolic process to this list. Many overrepresented annotations were shared between Class I and Class II lists, suggesting that many Class II transcripts may be directly targeted by NMD machinery.

Class III RNAs, which exhibit increased expression in the absence of NMD but which are not pulled down in a SMG-2 IP, are enriched for the defense response GO term, the cysteine-type endopeptidase inhibitor activity term, and a number of immune-related annotations in GOrilla analysis (Figure 2.11). Class IV RNAs, which exhibit decreased expression in the absence of NMD and which do not co-precipitate with SMG-2, are similarly enriched for immune-related processes and the defense response; Class IV factors also show enrichment for response to stress and response to stimulus GO terms (Figure 2.12). While these two individual lists were too small for PANTHER to find significant overrepresentation, pooled indirectly affected genes from Class III and IV demonstrated overrepresentation for factors involved in the immune, defense, and stress response systems.

Figure 2.7: GO Biological Process enrichment for Class I substrates

GOrilla output comparing Class I substrates to all expressed genes. GO Biological Process terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}.

Figure 2.7

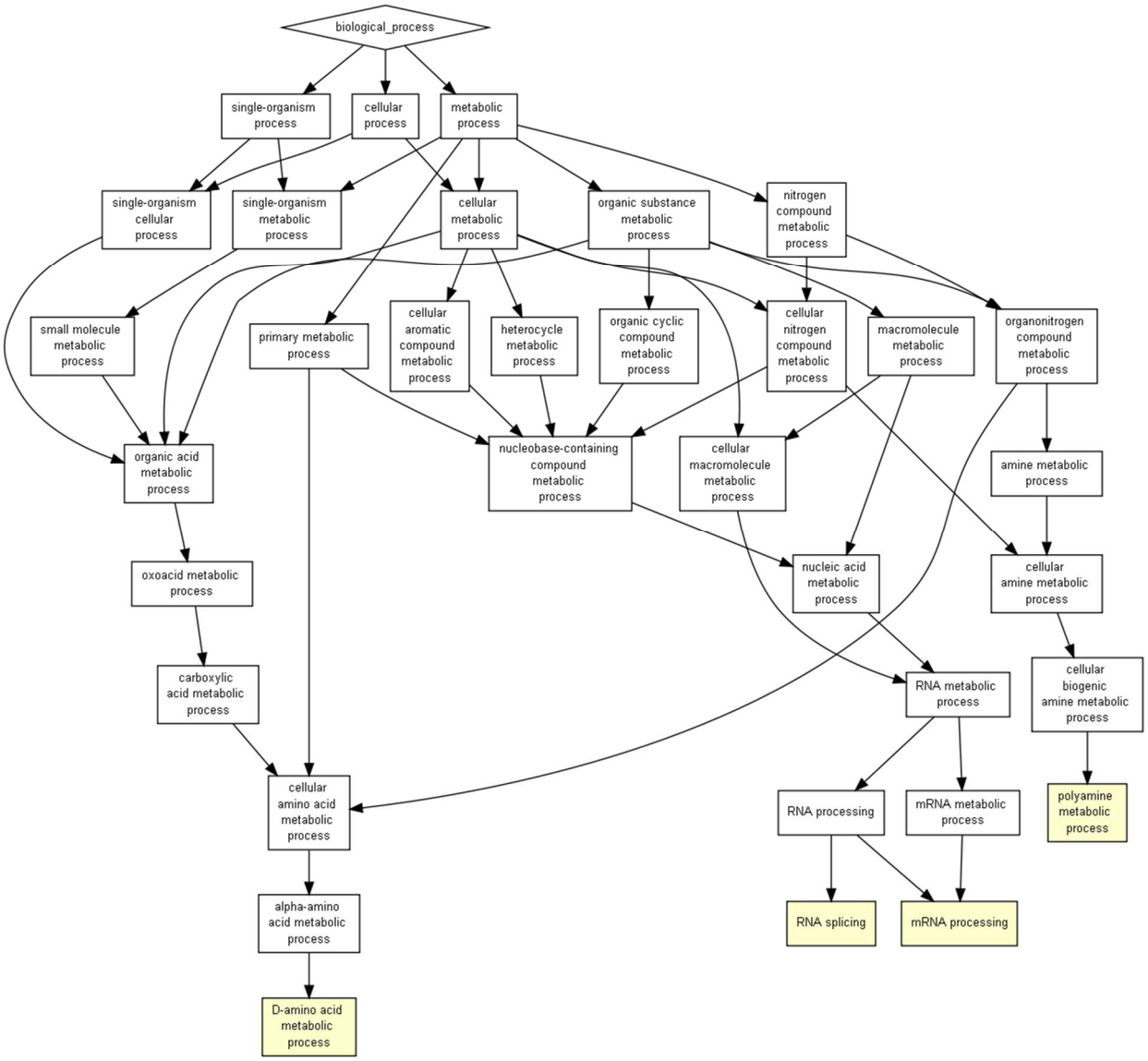

biological_process
single-organism process
cellular process
metabolic process
single-organism cellular process
single-organism metabolic process
cellular metabolic process
organic substance metabolic process
nitrogen compound metabolic process
small molecule metabolic process
primary metabolic process
cellular aromatic compound metabolic process
heterocycle metabolic process
organic cyclic compound metabolic process
cellular nitrogen compound metabolic process
macromolecule metabolic process
organonitrogen compound metabolic process
organic acid metabolic process
nucleobase-containing compound metabolic process
cellular macromolecule metabolic process
amine metabolic process
oxoacid metabolic process
nucleic acid metabolic process
cellular amine metabolic process
carboxylic acid metabolic process
RNA metabolic process
cellular biogenic amine metabolic process
cellular amino acid metabolic process
RNA processing
mRNA metabolic process
polyamine metabolic process
alpha-amino acid metabolic process
RNA splicing
mRNA processing
D-amino acid metabolic process

Figure 2.8: GO Molecular Function enrichment for Class I substrates

GOrilla output comparing Class I substrates to all expressed genes. GO Molecular Function terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}.

Figure 2.8

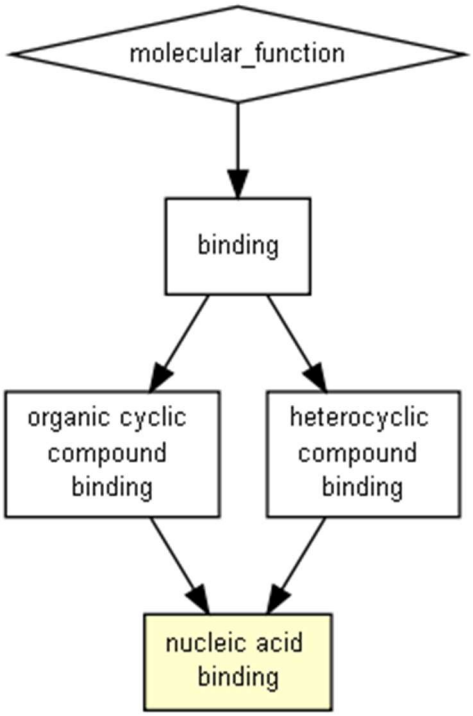

molecular_function
binding
organic cyclic compound binding
heterocyclic compound binding
nucleic acid binding

Figure 2.9: GO Biological Process enrichment for Class II transcripts

GOrilla output comparing Class II mRNAs to all expressed genes. GO Biological Process terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}, light orange indicates p between 10^{-5} to 10^{-7}, dark orange indicates p from 10^{-7} to 10^{-9}, and red indicates p $<10^{-9}$.

Figure 2.9

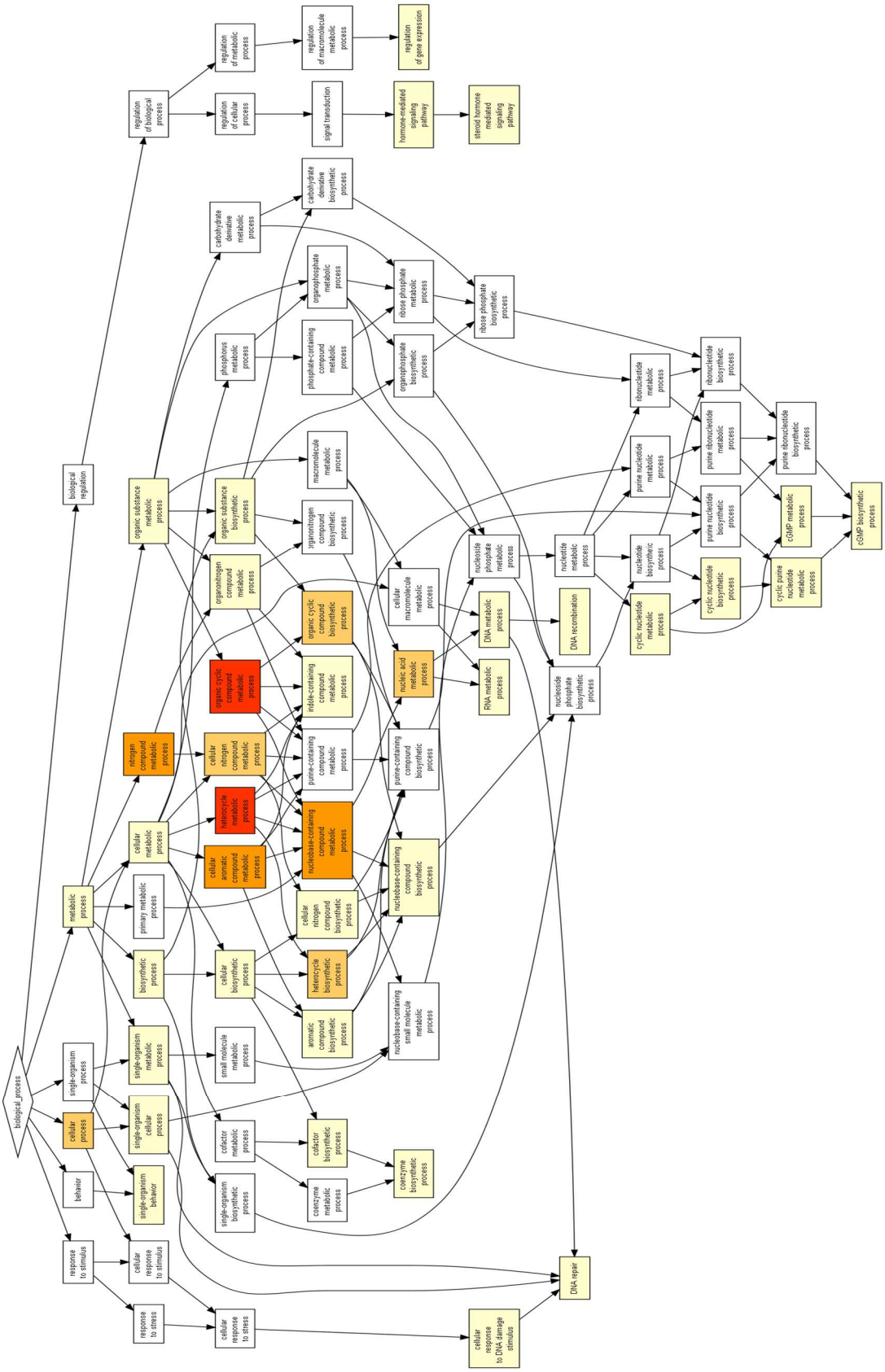

biological_process
metabolic process
cellular process
single-organism process
behavior
response to stimulus
response to stress
biological regulation
regulation of biological process
regulation of cellular process
regulation of metabolic process
regulation of macromolecule metabolic process
regulation of gene expression
signal transduction
hormone-mediated signaling pathway
steroid hormone mediated signaling pathway
primary metabolic process
biosynthetic process
single-organism metabolic process
single-organism cellular process
single-organism behavior
cellular response to stimulus
cellular response to stress
cellular metabolic process
organic substance metabolic process
nitrogen compound metabolic process
cellular nitrogen compound metabolic process
cellular aromatic compound metabolic process
heterocycle metabolic process
nucleobase-containing compound metabolic process
organic cyclic compound metabolic process
indole-containing compound metabolic process
purine-containing compound metabolic process
organonitrogen compound metabolic process
phosphorus metabolic process
organophosphate metabolic process
phosphate-containing compound metabolic process
carbohydrate derivative metabolic process
carbohydrate derivative biosynthetic process
macromolecule metabolic process
cellular macromolecule metabolic process
nucleic acid metabolic process
DNA metabolic process
RNA metabolic process
DNA recombination
DNA repair
cellular response to DNA damage stimulus
cellular cellular nitrogen compound biosynthetic process
heterocycle biosynthetic process
aromatic compound biosynthetic process
nucleobase-containing compound biosynthetic process
organic substance biosynthetic process
organonitrogen compound biosynthetic process
organophosphate biosynthetic process
ribose phosphate biosynthetic process
nucleobase-containing small molecule metabolic process
small molecule metabolic process
single-organism biosynthetic process
cofactor metabolic process
cofactor biosynthetic process
coenzyme metabolic process
coenzyme biosynthetic process
nucleoside phosphate metabolic process
nucleoside phosphate biosynthetic process
nucleotide metabolic process
nucleotide biosynthetic process
cyclic nucleotide metabolic process
cyclic nucleotide biosynthetic process
ribonucleotide metabolic process
ribonucleotide biosynthetic process
purine nucleotide metabolic process
purine nucleotide biosynthetic process
purine ribonucleotide metabolic process
purine ribonucleotide biosynthetic process
cyclic purine nucleotide metabolic process
cGMP metabolic process
cGMP biosynthetic process

Figure 2.10: GO Molecular Function enrichment for Class II transcripts

GOrilla output comparing Class II mRNAs to all expressed genes. GO Molecular Function terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}, light orange indicates p between 10^{-5} to 10^{-7}, dark orange indicates p from 10^{-7} to 10^{-9}, and red indicates p $<10^{-9}$.

Figure 2.10

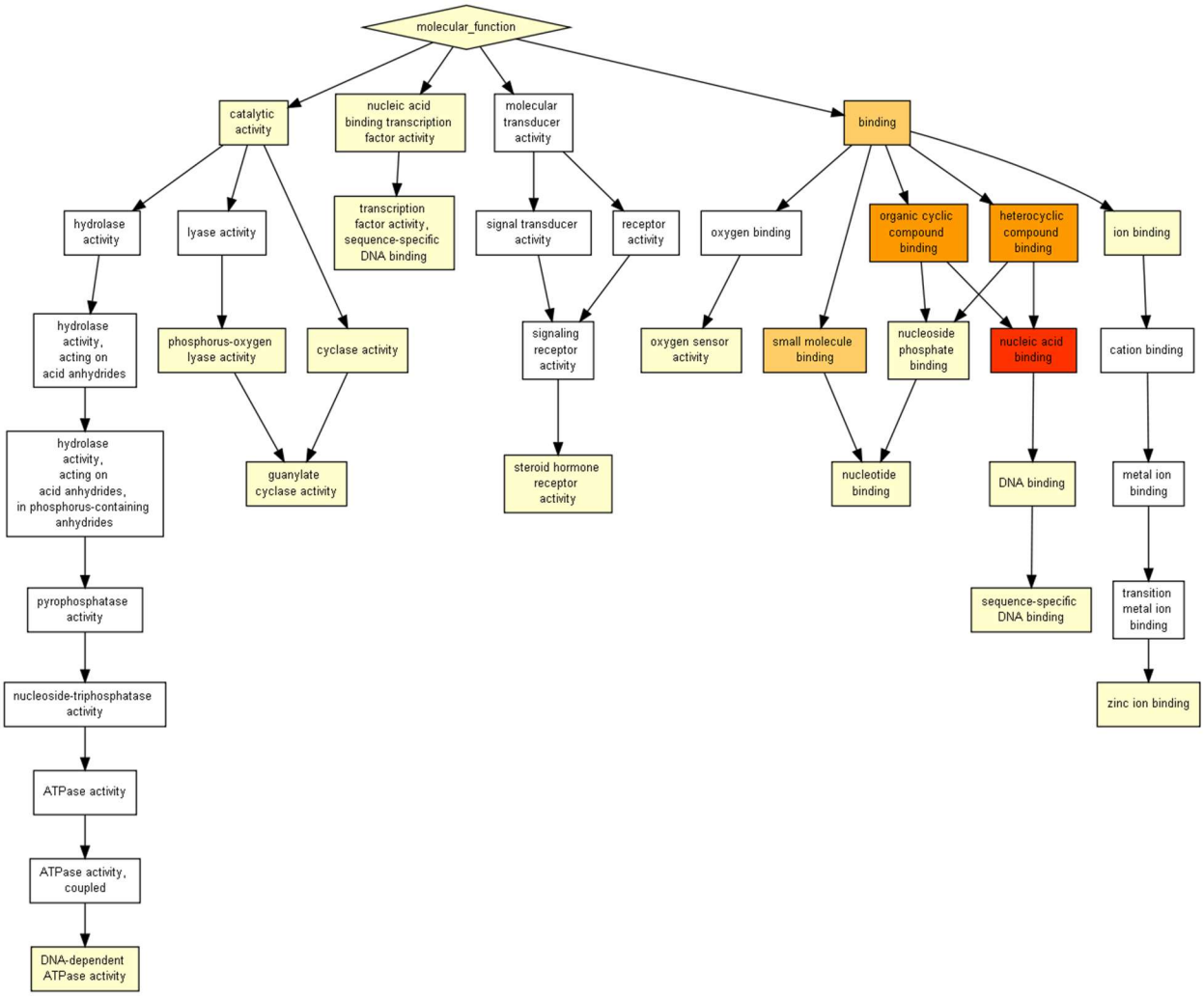

molecular_function
catalytic activity
nucleic acid binding transcription factor activity
molecular transducer activity
binding
hydrolase activity
lyase activity
transcription factor activity, sequence-specific DNA binding
signal transducer activity
receptor activity
oxygen binding
organic cyclic compound binding
heterocyclic compound binding
ion binding
hydrolase activity, acting on acid anhydrides
phosphorus-oxygen lyase activity
cyclase activity
signaling receptor activity
oxygen sensor activity
small molecule binding
nucleoside phosphate binding
nucleic acid binding
cation binding
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
guanylate cyclase activity
steroid hormone receptor activity
nucleotide binding
DNA binding
metal ion binding
pyrophosphatase activity
sequence-specific DNA binding
transition metal ion binding
nucleoside-triphosphatase activity
zinc ion binding
ATPase activity
ATPase activity, coupled
DNA-dependent ATPase activity

Figure 2.11: GO Biological Process enrichment for Class III transcripts

GOrilla output comparing Class III mRNAs to all expressed genes. GO Biological Process terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}, light orange indicates p between 10^{-5} to 10^{-7}, dark orange indicates p from 10^{-7} to 10^{-9}, and red indicates p $<10^{-9}$.

Figure 2.11

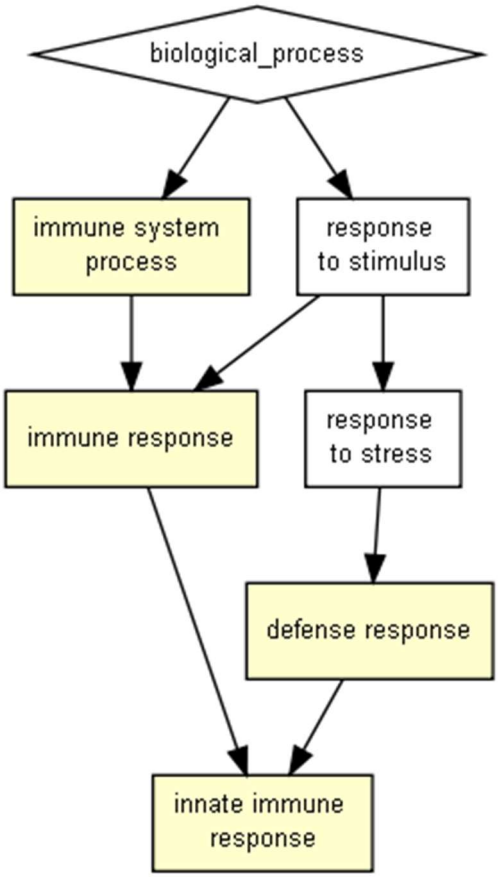
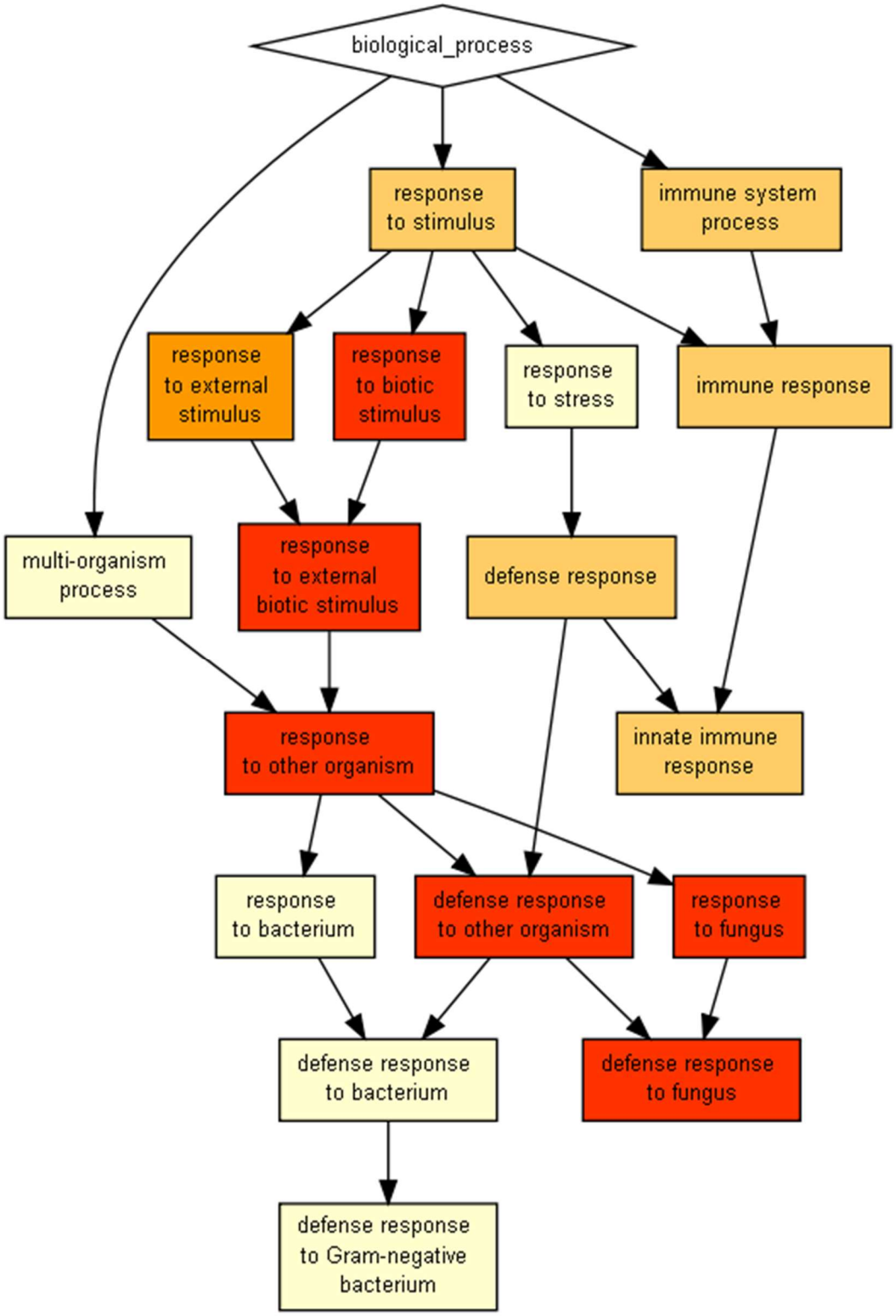

biological_process
immune system process
response to stimulus
immune response
response to stress
defense response
innate immune response

Figure 2.12: GO Biological Process enrichment for Class IV transcripts

GOrilla output comparing Class IV mRNAs to all expressed genes. GO Biological Process terms are shown. Light yellow boxes indicate a significant enrichment where p falls between 10^{-3} and 10^{-5}, light orange indicates p between 10^{-5} to 10^{-7}, dark orange indicates p from 10^{-7} to 10^{-9}, and red indicates p $<10^{-9}$.

Figure 2.12


biological_process
response to stimulus
immune system process
response to external stimulus
response to biotic stimulus
response to stress
immune response
multi-organism process
response to external biotic stimulus
defense response
response to other organism
innate immune response
response to bacterium
defense response to other organism
response to fungus
defense response to bacterium
defense response to fungus
defense response to Gram-negative bacterium

Discussion

NMD regulates gene expression in eukaryotes through the identification and degradation of PTC-containing transcripts. A relatively small number of studies have attempted to separate directly targeted substrates from the indirect effects of NMD, and only one previous study has done this *in vivo* in a multicellular organism [96]. Many of the previous studies characterizing NMD substrates relied on partial knockdown of NMD rather than true null alleles, potentially shortening the list of identified targets. In this study, I identified direct substrates of NMD by comparing the abundance of mRNAs that co-immunoprecipitate with SMG-2 to those of input samples in both wild-type and NMD(-) nematodes. I identified 585 Class I substrates of NMD; however, some of the remaining 2003 Class II genes are also likely targets. This number may further underestimate the actual number of NMD substrates since this study was only performed at one developmental stage.

I tested three common statistical packages to determine differential expression in my samples: baySeq, DESeq2, and edgeR. baySeq, the most stringent of the three programs, identified 306 NMD substrates. While baySeq did not identify any of the positive control genes (six known NMD substrates), it exhibited remarkable overlap with the other two programs. 100% of the baySeq targets were also identified by DESeq2, and 272/306 of the baySeq substrates (88.9%) were also identified by edgeR. DESeq2, the least stringent of the three programs, identified 1,114 direct substrates. 614 of these targets (55.1%) were corroborated by the other programs. DESeq2 also correctly identified 3/6 known substrates as Class I targets: *rsp-2, rsp-6*, and *rpl-7a*. edgeR, the program I used to determine which mRNAs are directly targeted for degradation by the NMD pathway, identified 585 substrates. 272 (46.5%) of these

were shared with the baySeq substrate list, and 580 (99.1%) were shared with DESeq2's target list, including all three known substrates. One known substrate, *rsp-4*, was identified as a Class II transcript by both edgeR and DESeq2.

I found that a relatively small portion of the genome is differentially expressed between NMD-competent and NMD-deficient strains compared to a previous *C. elegans* study. While I found that 4.7% of expressed genes (699/14,777) are upregulated in NMD(-) strains, while Ramani et al. estimate that 10% of genes are upregulated in NMD(-) strains following genome-wide tiling experiments [25]. This may, in part, be due to the differential expression software I chose. DESeq2 found that 10.2% of genes had elevated expression in a *smg(-)* line. However, NMD regulation of 4.7% of the transcriptome falls well within a range established by previous studies in other organisms [18, 27, 28, 92, 96, 97]. Genes with reduced expression in NMD(-) compared to NMD(+) strains are much more rare than genes with increased expression. This result is also corroborated by previous work in other organisms.

I was surprised to find that a large subset of genes (83.8%) with increased expression in an NMD(-) strain were enriched in a SMG-2 IP and identified as direct substrates. Similar work to determine high-confidence substrates has never found that more than half of transcripts with increased expression in NMD null strains are direct substrates, though Tani et al. suggested that approximately 700 transcripts are directly targeted by UPF1 [90, 92-94, 96, 97]. My high percentage implies that NMD in *C. elegans* may indirectly affect far less of the transcriptome than in other organisms.

In accordance with previous direct target studies, I find a large class of genes that satisfy other criteria for being direct substrates (for example, co-purification with SMG-2) but that do not exhibit increased expression in an NMD(-) strain [92-94, 96, 97]. 2,588 mRNAs were reliably associated with SMG-2. Of these, 585 are high confidence NMD substrates and 2,003 are Class II transcripts. While other pathways that involve SMG-2 may target many of these factors and transient associations with SMG-2 likely account for others, I find it probable that many of these transcripts are subjected to degradation by NMD.

Functional enrichment analysis of NMD-regulated genes shows that RNA processing factors, genes related to cyclic/aromatic compounds, and a variety of metabolic processes are frequently affected by SMG-2. NMD is known to self-regulate, and many of its previously identified targets encode components of the splicing machinery [28, 110, 111]. Thus, functional enrichment of RNA-processing and RNA-metabolic factors was expected. While a fair number of my functional enrichments are seen in other organisms, I did not observe the enrichment for cellular transport, telomerase maintenance, peroxisomal function, DNA repair, or cytoskeleton organization described by earlier studies [18, 23, 91].

Several GO terms related to stress response were enriched in the list of factors indirectly affected by NMD. Many other reports have linked the NMD and stress response pathways in humans, plants, nematodes, and yeast [97, 112-118]. However, SMG-1 also functions independently to regulate genotoxic stress in human cells and oxidative stress in *C. elegans* [119, 120]. As my analyses uses *smg-1(-)* animals, I cannot separate the contributions of NMD from any non-NMD functions of SMG-1.

The gene-level analyses outlined in this chapter do have limitations. Out of many transcripts for a given gene, only one may be targeted for degradation by NMD. If abundance of the other transcripts changes in a compensating manner, gene-level expression would remain stable, and the substrate would not be identified. As discussed above, analysis of differentially expressed transcripts becomes difficult and a fair portion of data is discarded when sequences are shared between transcripts. In the following chapter, I describe a method that circumvents these obstacles and further characterizes NMD substrates.

Chapter 3: Characterization of High-Confidence NMD Substrates

Abstract

The substrates of nonsense-mediated mRNA decay (NMD) commonly fall into one of several different classes. For example, specific transcripts can be alternatively spliced to include or introduce a premature termination codon (PTC) that will be recognized by NMD. Such PTCs derive from pre-mRNA sequences that are introns in other spliced products, which often causes such PTC-containing isoforms to be poorly annotated. Structural features of mRNAs, such as the presence of upstream open reading frames (uORFs) or long 3' untranslated regions (UTRs), can also elicit NMD. Lastly, certain classes of genes, such as pseudogenes and transposons, are frequently targeted for degradation by the NMD pathway. In this chapter, I identify 95 previously unrecognized direct substrates of NMD that include portions of introns. After adding these mRNAs to the list of Class I substrates identified in Chapter 2, I analyze the list of high-confidence direct targets to find common features of their biogenesis. I describe the genome-wide regulation of pseudogenes by NMD and discover that two recently expanded gene families encode several NMD substrates.

Introduction

NMD destabilizes mRNAs that contain premature termination stop codons (PTCs). Many PTC-containing transcripts are expressed in wild-type cells, which causes approximately 4-25% of the transcriptome to exhibit altered expression in the absence of NMD [18, 23-29]. Up to half of altered transcripts are likely the direct substrates of NMD, although the exact percentage varies between studies [90, 92-97]. Messenger RNAs whose expression is altered in

the absence of NMD but are not substrates of NMD likely represent the indirect effects caused by stabilization of direct targets.

Premature stop codons can arise in a number of ways. For example, genetic mutations and errors in transcription or splicing potentially create mRNAs that are substrates of NMD [4, 102]. Upstream open reading frames (uORFs), “leaky” ribosome scanning, where translation initiates downstream of the most 5' located start codon, disrupted open reading frames, the presence of selenocysteine codons, and long 3’ UTRs can all elicit NMD [18, 27, 30, 34, 56, 90, 94, 121]. Certain classes of genes often create NMD substrates; such classes include pseudogenes, transposons, and retroviruses [17, 18]. Unprocessed pseudogenes, which are the more common type of pseudogene in *C. elegans*, often contain frameshifts, insertions, premature stops, and truncations relative to their functional homologs [122]. Lastly, regulated alternative splicing can create PTC-containing isoforms that are subsequently degraded by NMD [123-126].

Genes that encode many *C. elegans* splicing factors create PTC-containing mRNAs by alternative (unproductive) splicing. Several *rpl* and *rsp* transcripts include portions of introns which either encode or induce (via frameshift) a PTC. One model suggests that unproductively spliced mRNAs are produced as a means to finely modulate gene expression [110, 127, 128]. When the encoded splicing factor is overly abundant, splicing could shift to increase the amount of PTC-containing mRNA relative to fully translatable mRNA. Conversely, when abundance of the splicing factor is insufficient, splicing could shift could to increase the amount of fully translatable mRNA relative to PTC-containing mRNA. Indeed, the proportion of *rpl-12(PTC)* is increased when RPL-12 is overexpressed [123].

Unproductively spliced transcripts are a major class of NMD substrates, yet they can be difficult to identify. Up to one third of alternative human transcripts contain PTCs [127], although the proportion of these that are direct substrates of NMD is likely lower [25, 129]. When NMD triggers degradation of specific spliced isoforms, the total mRNA abundance for a gene may not change. In *C. elegans*, for example, 300 genes exhibit a shift in splicing patterns in NMD-deficient samples without demonstrating an overall change in abundance [25]. Exon annotation can also cause difficulties in an investigation of NMD substrates. If annotations are based on wild-type samples, rapidly degraded PTC-containing transcripts are typically missed. Thus, PTC-containing or PTC-creating exons are often annotated as introns.

In this chapter, I investigate features that may create PTCs and features that are shared amongst NMD substrates. I identify alternatively included cassettes from annotated introns which likely trigger NMD. Further, I examine physical characteristics of transcripts directly targeted by NMD and describe the gene families that encode many NMD substrates.

Results

Alternatively Spliced NMD Substrates

Sequencing results described in Chapter 2 were re-analyzed in order to identify specific alternatively-spliced transcripts that are subjected to NMD. I initially used DEXSeq, a Bioconductor package in the DESeq family designed to determine differential exon usage [130], in this analysis. However, after failing to identify any known NMD substrates and very few NMD targets identified in Chapter 2, I chose to utilize edgeR instead. While edgeR is best known for its ability to test changes in gene-level expression, this program can easily be

adapted to test at the exon, transcript, or sequence tag levels [107]. I adapted edgeR analysis to identify changes in intron expression.

To identify introns and exons that were (i) differentially expressed (DE) between NMD(+) and NMD(-) samples and (ii) enriched in SMG-2 IPs, sequencing reads first had to be counted against these features. The intron/exon-level analysis utilized the aligned reads generated in Chapter 2. While exon coordinates are provided in the GTF annotation files available through Ensembl, overlapping exons and exons that are shared between transcripts can create confounding issues in HTSeq-count [105]. Introns are not defined in standard GTF files, and overlapping exons from multiple transcripts can further hinder the unambiguous definition of intronic sequence. To address these technical issues, I first collapsed exons by gene. Shared or overlapping exons were merged across transcripts to create a list of known coding sequence coordinates. Intron sequence coordinates were determined from the merged exon file, and two proper GTF files were created. HTSeq-count collated sequence reads aligning to exons and introns, and the resulting count files were combined and used in edgeR to determine differential expression.

The same three pairwise comparisons and criteria for definition of NMD substrates described in Chapter 2 were made for intron/exon counts: (1) Features whose abundance increased in a *smg(-)* sample compared to wild type; (2) features that are enriched in the SMG-2 IPs of *smg-1(r910)* samples compared to the input samples; and (3) features that are enriched in the SMG-2 IPs of *smg-1(r910)* samples compared to the SMG-2 IPs of *smg-1(r910) smg-2(r915)* negative control samples. As in Chapter 2, I required at least a 1.5-fold change for the

input (first) comparison and a 2-fold change for IP enrichment. The false discovery rate had to fall below 5%.

37,579 features were expressed with at least three counts per million in the sequencing data: 35,212 exons and 2,367 introns. 1,823 features exhibited significantly altered expression in the *smg(-)* strains compared to wild-type. 1,426 exons (4%) and 247 (10.4%) introns were upregulated in the absence of NMD; together, these features represent 809 genes. The maximum log2 upregulation in an NMD-deficient background was 7.3-fold. The maximum log2 downregulation was 10-fold. 8,652 exons and 827 introns, representing 3,397 genes, were enriched in the SMG-2 IP samples compared to their respective input samples. 7,938 exons and 774 introns from 3,128 genes were enriched in the experimental SMG-2 IPs compared to the negative control IPs. When combined, 7,400 exons and 737 introns, representing 2,866 genes, were enriched in both IP comparisons.

Introns and exons were classified by their expression patterns as in the previous chapter (Figure 3.1). 1204 exons and 191 introns were categorized as Class I NMD substrates; these features exhibited increased abundance in the absence of NMD and were consistently associated with SMG-2 in IP experiments. 661 genes were represented by these Class I introns and exons; 499 (75.5%) of these were also identified as Class I NMD substrates in the gene-level analysis presented in Chapter 2. All three of the known NMD substrate control genes which were identified as Class I targets by the gene-level analysis contained Class I substrate introns or exons. Among the newly identified genes, 94 were originally found in the Class II group via gene-level analysis. This subset includes *rsp-4*, a known substrate of NMD. 68 genes were

completely novel in the intron/exon-level analysis, and the vast majority of the differentially expressed features, 82.4%, were expressed introns.

A selection of differentially expressed Class I exons was subjected to further study. A few of the investigated exons demonstrated expression typical of an NMD substrate when examined in Broad's Integrative Genomics Viewer (IGV); however, many of the exons I investigated did not clearly change in abundance. Relative coverage for exon three of *C49C3.3* and exon two of *Y41E3.6* demonstrated a shift in *smg(-)* input and IP samples; however, the expression patterns shown for *stip-1* and *fbxa-88* were much more common (Figure 3.2). Though the absolute number of reads aligning to these genes was elevated in *smg(-)* samples, I did not observe a sharp change in exon usage in IGV. Differentially expressed exons were also far less informative than differentially expressed introns, since there were typically multiple DE exons per transcript but usually only one DE intron. The RIP-Seq procedure isolates whole RNAs that are bound to SMG-2; thus, one would generally expect all exons in a substrate transcript to exhibit increased expression. Such a trend might not be observed if the total RNA abundance for a gene remains stable while there is a shift in the predominant splice isoform. Further analysis of DE exons could perhaps locate PTCs in the transcript. However, due to the confounding results from IGV, I chose to focus on DE introns (see below).

Figure 3.1: Differential expression of introns and exons indicates NMD regulation

(A&C) Venn diagram detailing overlap between both exons and introns (A) or introns only (C) exhibiting at least 1.5-fold increased expression in *smg-1(r910)* and *smg-1(r910) smg-2(r915)* samples compared to N2, introns enriched at least 2-fold in a SMG-2 IP relative to its input sample, and introns enriched at least 2-fold in a SMG-2 IP relative to negative control IPs. Class I factors meet all three conditions. Class II factors are reliably SMG-2 associated but do not exhibit increased expression in a NMD(-) animal, and Class III factors exhibit elevated expression in a NMD(-) sample but are not enriched in a SMG-2. (B&D) Venn diagram comparing genes containing Class I introns and exons (B) or introns alone (D) with Class I and Class II mRNAs identified in the gene-wide analysis (Chapter 2). Genes containing Class I Introns but not originally identified as Class I mRNAs were added to the list of high-confidence NMD substrate.

Figure 3.1

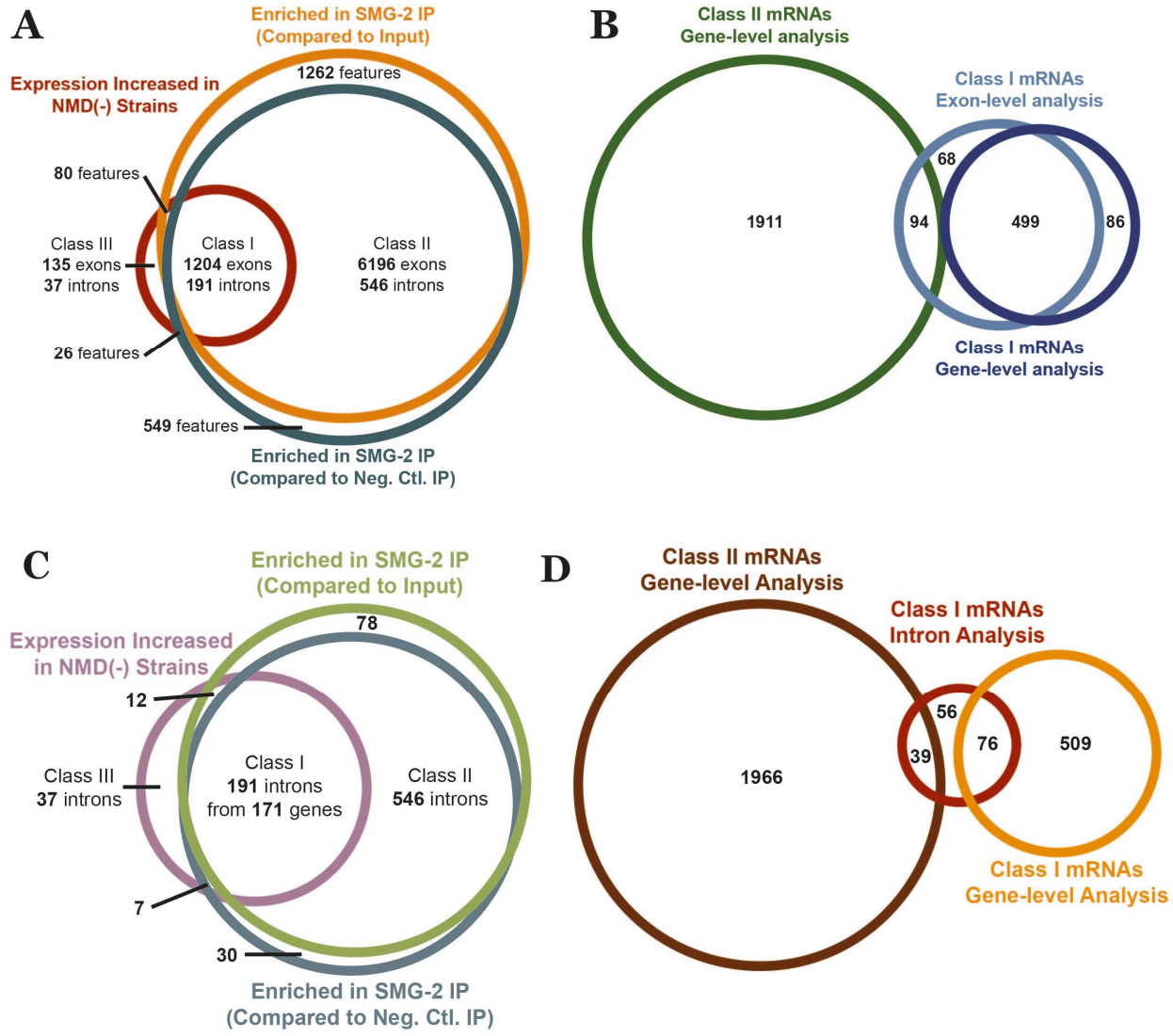
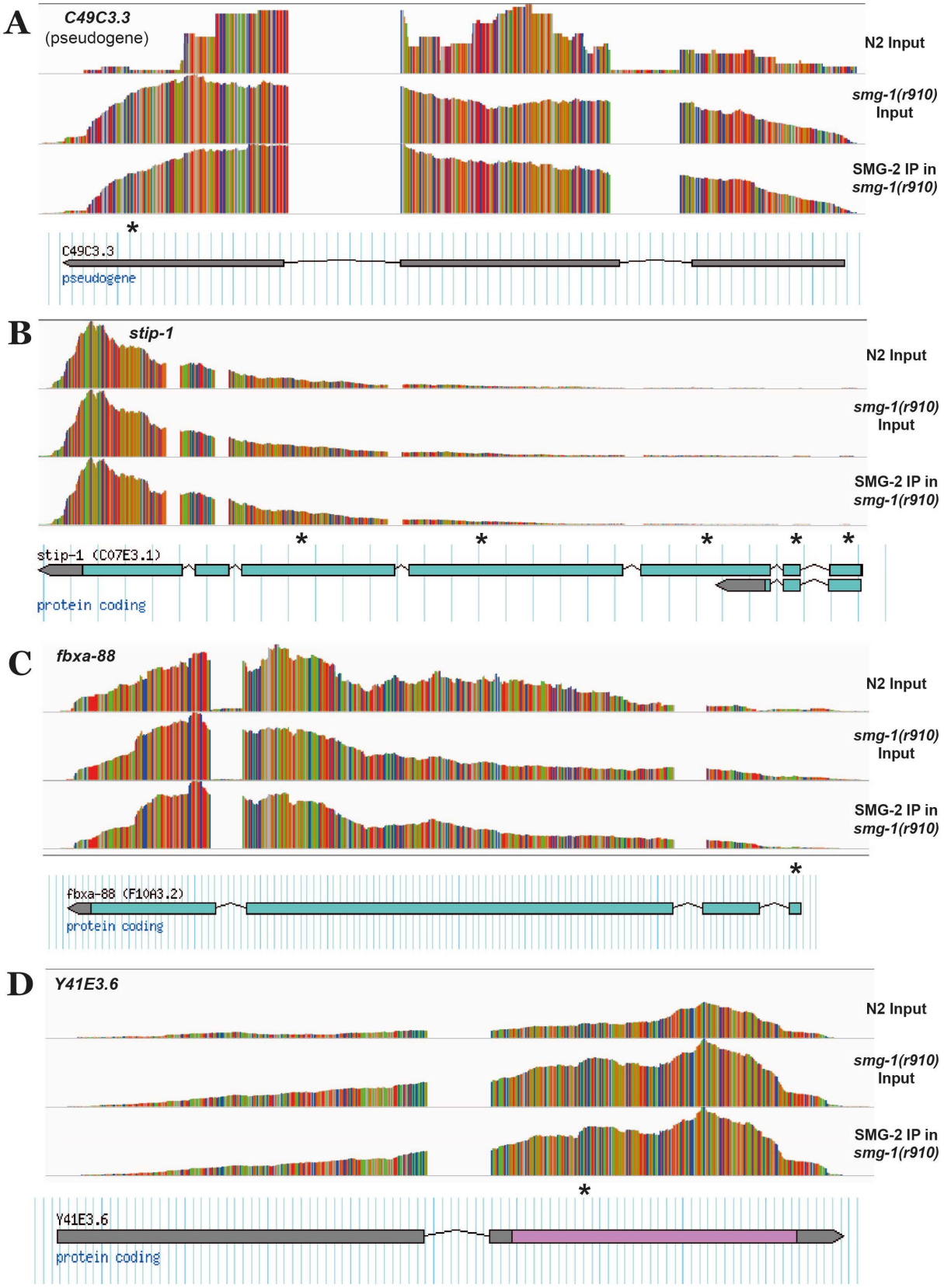

A
Enriched in SMG-2 IP
(Compared to Input)
Expression Increased in
NMD(-) Strains
1262 features
80 features
Class III
135 exons
37 introns
Class I
1204 exons
191 introns
Class II
6196 exons
546 introns
26 features
549 features
Enriched in SMG-2 IP
(Compared to Neg. Ctl. IP)
B
Class II mRNAs
Gene-level analysis
Class I mRNAs
Exon-level analysis
68
1911
94
499
86
Class I mRNAs
Gene-level analysis
C
Enriched in SMG-2 IP
(Compared to Input)
78
Expression Increased
in NMD(-) Strains
12
Class III
37 introns
Class I
191 introns
from 171 genes
Class II
546 introns
7
30
Enriched in SMG-2 IP
(Compared to Neg. Ctl. IP)
D
Class II mRNAs
Gene-level Analysis
Class I mRNAs
Intron Analysis
56
1966
39
76
509
Class I mRNAs
Gene-level Analysis

Figure 3.2: Relative coverage plots of differentially expressed exons

Relative sequencing coverage shown by IGV plots for four genes containing Class I Exons: *C49C3.3, stip-1, fbxa-88*, and *Y41E3.6*. Gene diagrams are shown below IGV plots. Colored bar height indicates relative sequencing coverage. Asterisks indicate regions that demonstrate increased expression in NMD(-) samples and enrichment in the SMG-2 IP.

Figure 3.2


A
C49C3.3
(pseudogene)
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
C49C3.3
pseudogene
B
stip-1
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
stip-1 (C07E3.1)
protein coding
C
fbxa-88
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
fbxa-88 (F10A3.2)
protein coding
D
Y41E3.6
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
Y41E3.6
protein coding

A subset of five NMD-targeted introns were chosen for further analysis. I initially examined Class I introns in IGV (Figure 3.3). Regions of identified introns that satisfied the criteria for Class I NMD substrates were unambiguous. All five of these included introns were validated via PCR (Figure 3.4). *Dpy-7* has a seven-nucleotide addition to the end of exon one that becomes detectable in a *smg(-)* strain and is particularly enriched in a SMG-2 IP (Figure 3.3A & 3.4A). Similarly, a splice variant of *R06C1.4* that includes a portion of the first intron is also stabilized in a *smg(-)* strain and enriched in a SMG-2 IP (Figure 3.3C & 3.4C). The included segment of *phy-2* was barely perceptible in a *smg(-)* strain but readily detectable in the SMG-2 IP (Figure 3.3D & 3.4D). A putative novel transcript was identified in the region between exons 14 and 15 of *deg-1* in a *smg(-)* strain. Sequencing reads aligning to four distinct areas of the intron are far more abundant than reads aligning to *deg-1* exons (Figure 3.3B), and while primers flanking the intron do not detect multiple isoforms, primers designed to the enriched region detect independent, NMD-responsive isoforms (Figure 3.4B). PCR for *ttr-46* partially validated the pattern observed in sequencing data; the upper, intron-containing transcript was enriched in a SMG-2 IP but did not exhibit an increase in abundance for the *smg-1(-)* input sample relative to the wild-type input sample (Figure 3.3E & 3.4E). However, the upper band was much more abundant in the input samples than one would expect based on the relative coverage shown in IGV. In total, I validated expressed intronic sequences in smg(-) transcripts in 5 out of 5 tested NMD-targeted introns. Due to this reproducibility, I added all mRNAs identified as Class I substrates by intron-level analysis to the set of mRNAs identified as Class I substrates via gene-level analysis. 171 genes contain Class I introns, 76 of which were also identified as Class I substrates via gene-level analysis and 39 of which were identified as Class II

Figure 3.3: Relative coverage plots of differentially expressed introns

Relative sequencing coverage shown by IGV plots for five genes containing Class I introns: (A) *dpy-7,* (B) *deg-1,* (C) *R06C1.4,* (D) *phy-2,* and (E) *ttr-46*. Gene diagrams are shown below IGV plots. Colored bar height indicates relative sequencing coverage. Asterisks indicate intronic regions that demonstrate increased expression in NMD(-) samples and enrichment in the SMG-2 IP.

Figure 3.3

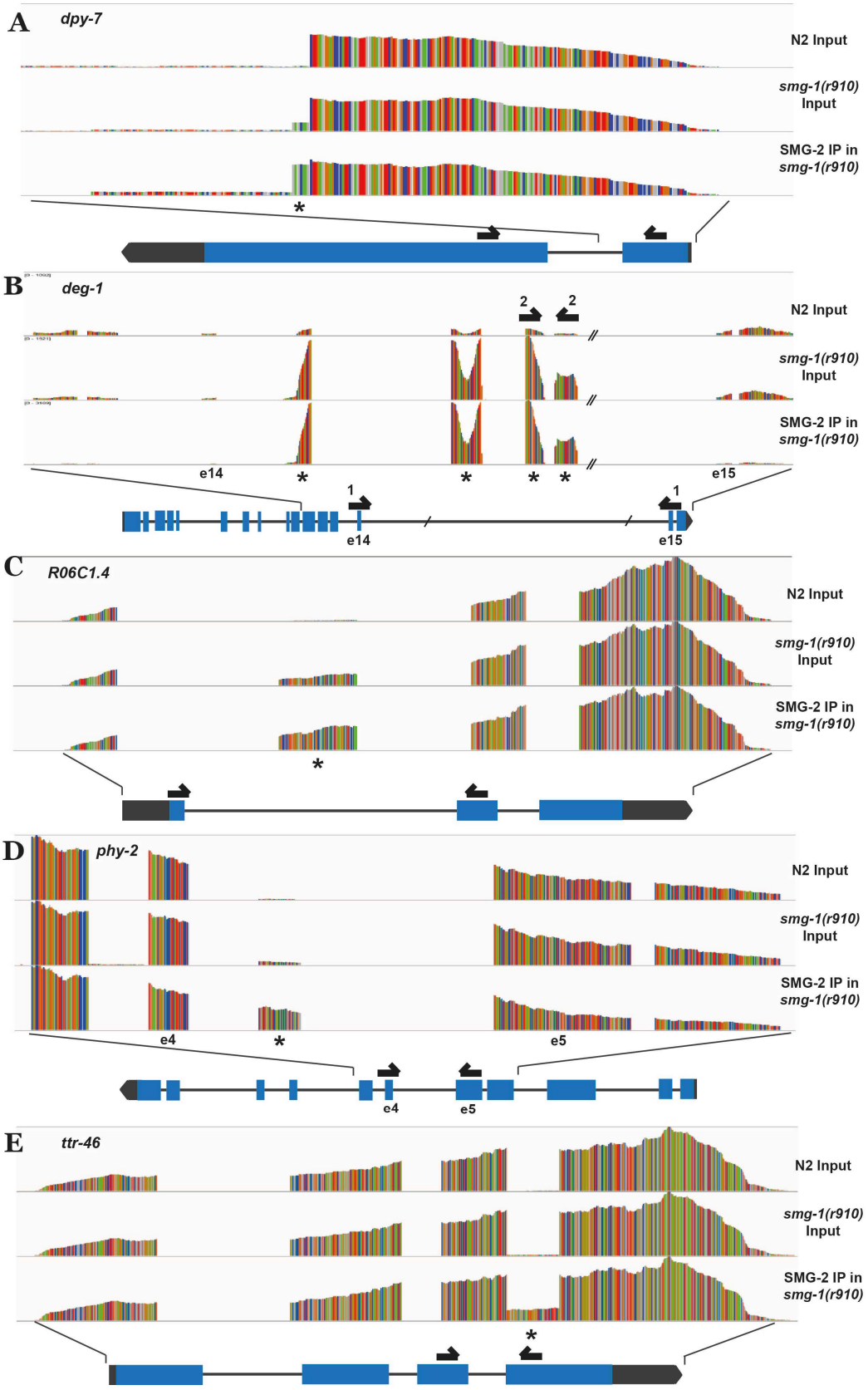
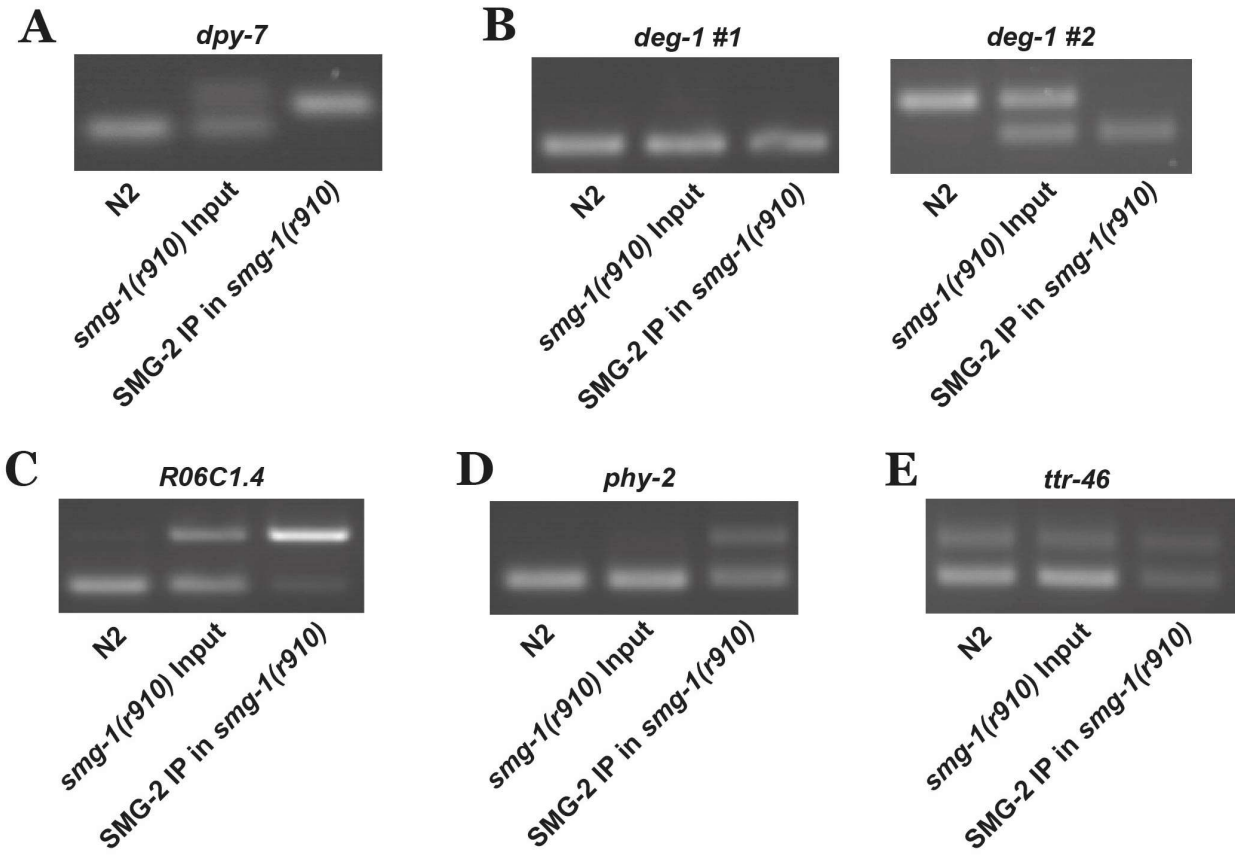

A
dpy-7
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
B
deg-1
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
e14
e15
C
R06C1.4
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
D
phy-2
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
e4
e5
E
ttr-46
N2 Input
smg-1(r910) Input
SMG-2 IP in smg-1(r910)

Figure 3.4: RT-PCR Validation of Class I Introns

Primers, noted on gene diagrams in Figure 3.3, were designed to span expressed introns for (A) *dpy-7*, (B) *deg-1* (primer set #1), *R06C1.4,* (C) *R06C1.4,* (D) *phy-2,* and (E) *ttr-46*. Upper bands for these genes indicate that a portion of the intron was included in the mature transcripts present in the *smg-1(r910)* input sample and enriched in the SMG-2 IP. *deg-1* #2 primers span the third and fourth expressed regions in the intron of *deg-1*. The smaller putative isoform shown by the lower band in this PCR co-precipitates with SMG-2 and is present in an NMD(-) sample.

Figure 3.4


A
dpy-7
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
B
deg-1 #1
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
deg-1 #2
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
C
R06C1.4
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
D
phy-2
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)
E
ttr-46
N2
smg-1(r910) Input
SMG-2 IP in smg-1(r910)

substrates via gene-level analysis. By merging the two Class I lists, a high-confidence NMD direct substrate list representing 680 genes was created.

Gene Types Directly Targeted by NMD

Of the 680 high-confidence NMD substrates identified, the vast majority (532) are protein-coding mRNAs (Figure 4A). Twelve class I features are non-coding RNAs (five lincRNAs, four ncRNAs, one ancRNA, one snoRNA, and one miRNA), four align to snoRNA hosts, and one aligns to a transposon.

Pseudogenes generate a known class of NMD substrates[17]. 423 of the 1658 annotated pseudogenes in *C. elegans* are detectable in the sequencing dataset. 135/680 (19.9%) of the direct substrates of NMD are expressed pseudogenes, a significant enrichment compared to all expressed genes (Bonferroni-corrected hypergeometric test p = 5.65e-78). Enrichment for pseudogenes can also be found among class II features (162/2003, p = 5.31e-38) and class III indirect effects (7/65, p = 2.57e-2) (Figure 3.5A).

I further identified two gene families that are frequently targeted by NMD: nuclear hormone receptors and F-box A protein encoding genes (Figure 3.5A). Both of these families have undergone remarkable expansion in *C. elegans*. *C. elegans* contains 284 *nhr* genes, compared to 21 in *Drosophila* and 48 in humans [131, 132]. Expansion of the *nhr* gene family is a fairly recent evolutionary event. An expanded set of *nhr* genes is also observed in related *Caenorhabditis* species, but proliferation and diversification of the family has likely continued since the divergence of *C. elegans* from *C. briggsae* ~40-180 million years ago [133, 134]. Fifteen *nhr* genes are found on the list of high-confidence substrates of NMD (p = 5.21e-2), and another 47 create class II transcripts (p = 2.93e-5). 222 *fbxa* genes (and 390 total F-box

proteins) are found in *C. elegans*, compared to 43 F-box proteins in *Drosophila* and 71 in humans. Thirty-eight *fbxa* genes generate high-confidence substrates of NMD (p = 1.05e-18).

Functional Enrichment & Conservation of High-Confidence Substrates

Once again, I utilized the GOrilla functional enrichment and PANTHER statistical overrepresentation tools to better characterize the physiological role of NMD. Functional enrichment was determined in the 680 high-confidence NMD substrates using the 14,777 expressed genes as the dataset for analysis. Nucleic acid binding, polyamine metabolic processes, and numerous GO terms related to RNA metabolic processes were enriched in GOrilla's analysis of high-confidence substrates. Enriched "child" terms for RNA metabolic processes included two RNA processing classifications and three RNA splicing terms. PANTHER's list of overrepresented GO terms included most of those listed by GOrilla, and linked 16 additional terms to the substrate list. Terms overrepresented in the high-confidence substrate list included structural constituent of ribosome, translation, gene expression, hydrolase activity, and a number of additional metabolic processes. Enrichment analysis of Class II factors did not change notably after excluding the 39 transcripts containing Class I introns. Selected GO term enrichment for all classes of NMD-affected RNAs, including the high-confidence substrate list, is shown in Figure 3.5 B-D.

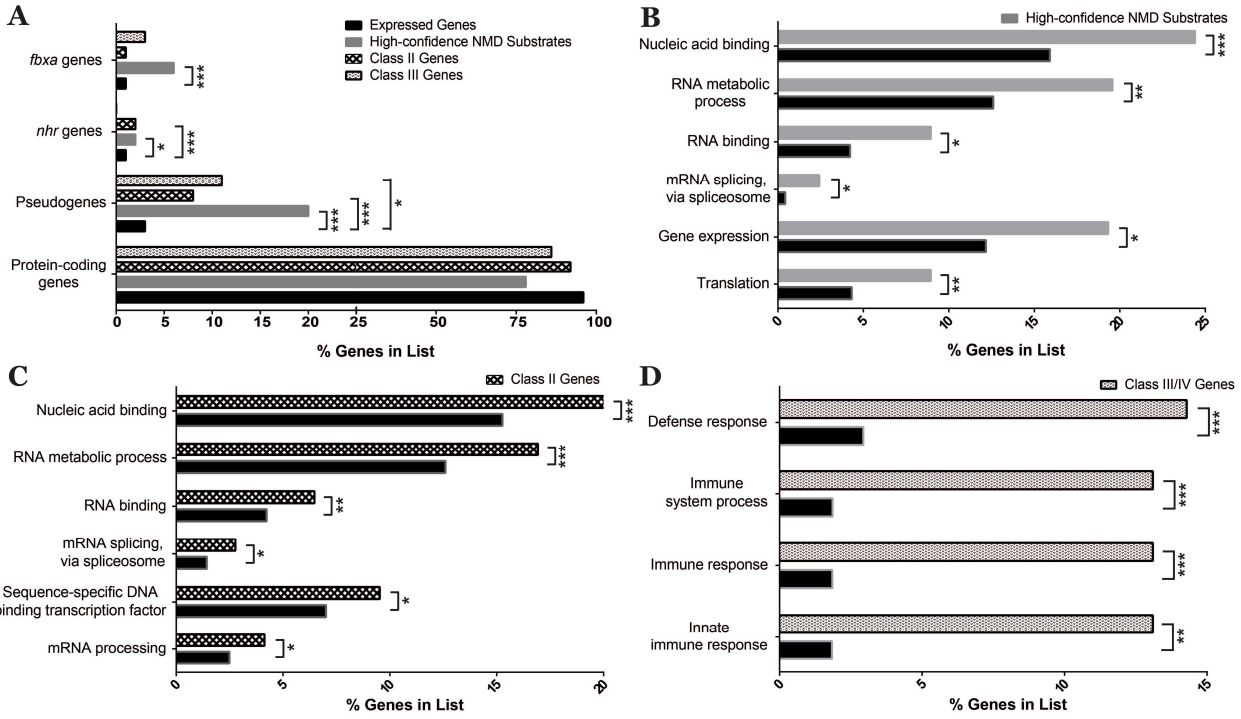
Of the 680 genes that produce high-confidence NMD substrates, 105 have orthologs in yeast, 165 have orthologs in *Drosophila*, 178 have orthologs in mouse, and 179 have orthologs in humans. I compared my direct target list to those of yeast [94], flies [96], and mice [95]. Only 2-3 genes were NMD substrates shared between *C. elegans* and each of the other species. Messenger RNA of a single gene, *aard-21* is targeted by NMD in nematodes, flies, and mice.

NMD is a highly conserved process, and classes of genes targeted by NMD, such as those with long upstream ORFs, tend to be conserved as NMD substrates. However, individual genes are not typically conserved as NMD substrates outside of Mammalia [91, 135]. Thus, the lack of conservation between these four direct target studies is not surprising.

Figure 3.5: Statistical overrepresentation of gene classes among genes affected by NMD

(A) Abundance of gene classes and gene types represented by Class I, II, and III features compared to all expressed genes. P-values are derived from Bonferroni-corrected hypergeometric tests. (B-D) Abundance of GO terms found on (B) Class I, (C) Class II, and (D) Indirect Effect (Class III & IV) lists, compared to expressed genes. P-values are derived from Bonferroni-corrected binomial tests, as performed by PANTHER's Statistical Overrepresentation function. Asterisks indicate significance: * $p<0.05$, ** $p<0.01$, ***$p<0.001$

Figure 3.5


A
Expressed Genes
High-confidence NMD Substrates
Class II Genes
Class III Genes
fbxa genes
nhr genes
Pseudogenes
Protein-coding genes
0 5 10 15 20 25 50 75 100
% Genes in List
B
High-confidence NMD Substrates
Nucleic acid binding
RNA metabolic process
RNA binding
mRNA splicing, via spliceosome
Gene expression
Translation
0 5 10 15 20 25
% Genes in List
C
Class II Genes
Nucleic acid binding
RNA metabolic process
RNA binding
mRNA splicing, via spliceosome
Sequence-specific DNA binding transcription factor
mRNA processing
0 5 10 15 20
% Genes in List
D
Class III/IV Genes
Defense response
Immune system process
Immune response
Innate immune response
0 5 10 15
% Genes in List

Physical Characteristics of NMD Substrates

A number of transcript characteristics, such as long 3' UTRs and upstream ORFs, have been correlated with NMD substrates in the past [18, 30]. I tested to see whether transcript length, the number of isoforms for a given gene, the number of exons per transcript, or UTR length were correlated with the high-confidence substrate list (Figure 3.6).

Class I mRNAs were generally shorter than the median length of all expressed transcripts (Figure 3.5B). A histograms plotting all transcript lengths demonstrate slight shifts to the left for Class I targets relative to expressed genes. The median annotated transcript length for the 14,777 genes with >3 cpm expression is 1,388bp. High-confidence NMD substrates are significantly enriched in the two transcript length bins directly below that figure: 250-750bp and 750-1250bp (Bonferroni adjusted p-values < 0.001). Short translation length is a feature of NMD substrates in yeast [136]. However, NMD-responsive genes in *Drosophila* have longer transcripts on average [27]. In this regard, nematodes are more closely aligned with *S. cerevisiae*.

I then examined the number of transcripts and the number of exons per gene for my gene lists of interest. These two analyses were meant to serve as a proxy for splicing activity. No measurable difference in the number of transcripts per gene was observed (Figure 3.6A), but both Class I and Class II RNAs seemed to have a slightly higher number of exons per gene compared to all expressed genes (Figure 3.6C). While the mean number of exons per transcript is not increased for Class I substrates compared to all expressed genes, the number of transcripts with only one exon is significantly depleted (Bonferroni adjusted p = 1.3e-3) and the numbers of transcripts containing three to six exons are significantly enriched (Bonferroni

adjusted p <0.1). I conclude that transcripts with multiple exons are more likely to be targeted by NMD.

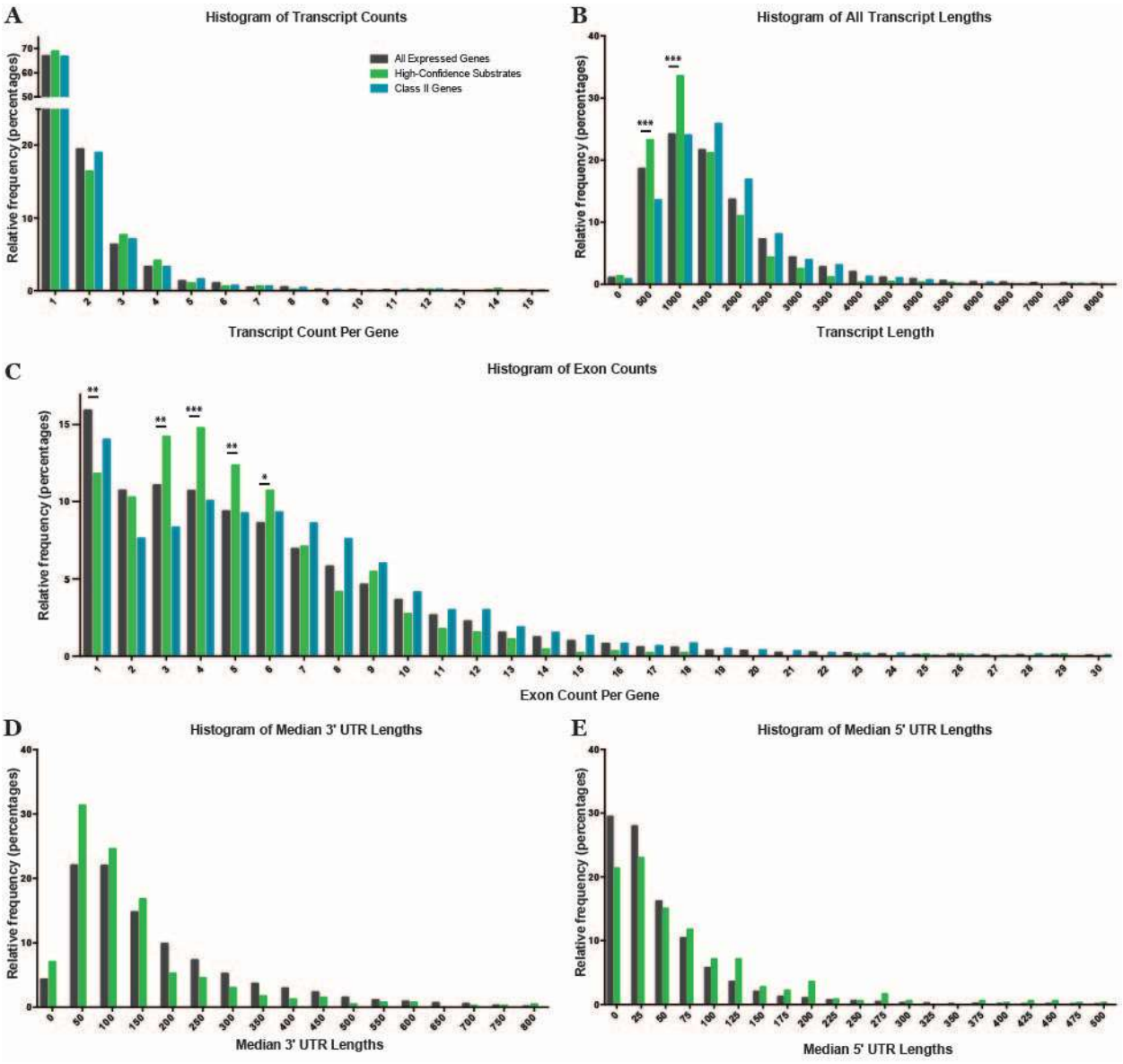
High-confidence NMD substrates do not appear to have longer-than-average 3' UTRs in this study (Figure 3.6D). Both the average and median annotated 3' UTR lengths for NMD substrates (147 and 108, respectively) are shorter than those possible for all expressed genes (204 and 148). 3' UTR lengths are only annotated for ~60% of Class I transcripts. Using the subset of high-confidence targets with annotated UTR lengths, 3' UTRs shorter than 150bp are enriched and 3'UTRs longer than 150bp are depleted (Bonferroni-corrected p = 6.2e-12) compared to expressed transcripts. While this observation runs counter to much of the published literature, which states that mRNAs containing long 3' UTRs are frequently targeted by NMD, it is also based upon available annotations rather than experimental data. Further examination of the data may reveal that targeted mRNAs include the longest possible 3' UTRs.

Though the same caveats regarding UTR annotation apply, 5' UTR length does seem to increase among high-confidence substrates (Figure 3.6E). Average 5' UTR length expands from 64bp in the general population of expressed transcripts to 86bp among Class I substrates. 5' UTR lengths were available for 366 Class I transcripts (54%). For this subset, transcripts with median 5' UTR lengths greater than 75bp were significantly enriched in the high-confidence target list (Bonferroni-corrected p-value = 3.7e-11). No previous studies have established a correlation between 5' UTR length and NMD targeting.

Figure 3.6: Histograms Testing Physical Characteristics of NMD Substrates

Histograms comparing (A) the number of transcripts per gene, (B) transcript length, (C) the number of exons per transcripts, (D) median 3' UTR length per transcript, and (E) median 5' UTR length per transcript for all expressed genes, genes encoding high confidence NMD substrates, and (for A-C) Class II transcripts. Asterisks indicate Bonferroni-corrected significance: * $p<0.05$, ** $p<0.01$, ***$p<0.001$

Figure 3.6


A
Histogram of Transcript Counts
All Expressed Genes
High-Confidence Substrates
Class II Genes
Relative frequency (percentages)
Transcript Count Per Gene
B
Histogram of All Transcript Lengths
Relative frequency (percentages)
Transcript Length
C
Histogram of Exon Counts
Relative frequency (percentages)
Exon Count Per Gene
D
Histogram of Median 3' UTR Lengths
Relative frequency (percentages)
Median 3' UTR Lengths
E
Histogram of Median 5' UTR Lengths
Relative frequency (percentages)
Median 5' UTR Lengths

I did not systematically characterize NMD substrates containing upstream ORFs. Two uORF-containing genes are readily found in the *C. elegans* literature: *zip-2* and *gna-2* [137, 138]. Notably, neither *zip-2* nor *gna-2* met the thresholds for Class I, Class II, or Class III status. These genes were only IP-enriched relative to the input sample.

Discussion

Analysis of differential expression of annotated introns significantly improved identification of alternatively spliced NMD substrates. This method presents a novel way to analyze datasets for NMD substrates, particularly when gene-level analysis proves insufficient. My analysis of alternatively included sequences revealed several previously unidentified splice variants and one putative gene. Three out of four validated alternate exons would cause a frameshift in the resulting transcript, and the one exception, *phy-2,* encodes four stop codons within its 75bp alternatively included segment. I identified four potential exons located in an intron of *deg-1* that are expressed at a much higher level than the remainder of *deg-1*. These sequences are highly expressed in NMD(-) samples compared to N2. Nucleotide BLAST found strong homology between these sequences and both *linc-20* and *linc-9,* two genes on my high-confidence substrate list. This analysis, therefore, likely identified a novel lincRNA that is a direct target of NMD. With additional intron data, I increased the size of my high-confidence target list by 16%.

Though a small number of individual transposons and pseudogenes have been identified as NMD substrates in previous studies [17], I investigated the genome-wide regulation of these features by NMD. Three expressed genes were identified as transposon in origin: *Y48G1BL.4*,

K03D7.1, and *madf-7*. *Y48G1BL.4* was identified as a direct NMD substrate, and *K03D7.1* was enriched in the SMG-2 IP compared to its input sample. 206 pseudogenes and one transposon were expressed with at least three counts per million in my wild-type samples. An additional 217 pseudogenes and two more transposons were expressed in NMD deficient samples. Of these 423 expressed pseudogenes, 135 were identified as high-confidence substrates of NMD, 162 were in the list of Class II RNAs, and 7 were identified as Class III transcripts. Altogether, 71.9% of the expressed pseudogenes demonstrate some level of NMD-related control.

Analysis of shared structural features of NMD substrates was less informative. I observe a slight preference for short ORFs among NMD substrates. Though a similar result has been described for *S. cerevisiae*, no clear models explain this phenomenon [136]. I observed no marked increase in the number of transcripts per gene for high-confidence substrates and a minimal shift in the number of exons per gene. Without extracting UTR-aligned reads from the sequencing data or having a more refined, UTR-focused methodology, my conclusions that long 3' UTRs are not common among NMD substrates and that long 5' UTRs are frequent substrate features are poorly supported.

The discovery that expanded gene classes are enriched among the substrates of NMD is provocative, particularly since many of the *nhr* and *fbxa* class genes are pseudogenes. Of the 222 *fbxa* genes in the genome, 34 are annotated pseudogenes. Likewise, 6 annotated pseudogenes are among the 283 *nhr* genes. After demonstrating the regulation of pseudogenes by NMD, it is not surprising that a significant portion of pseudogenes belonging to the *nhr* and *fbxa* families are targeted by NMD. 18/38 *fbxa* genes and 2/15 *nhr* genes on the high-confidence substrate list are pseudogenes, and the remaining *fbxa* and *nhr* NMD

substrates may be unannotated pseudogenes. NMD may be responsible for an additional layer of regulation for expression of *nhr* and *fbxa* genes. Expansion of these two gene families is an attractive explanation for the overrepresentation of these classes of genes on the NMD substrate list. Following rapid expansion, duplicated genes are more likely to collect inactivating mutations, including nonsense mutations, without detrimentally effecting the family's functionality. To my knowledge, this is the first report of expanded gene families generating large numbers of NMD direct targets**.**

Chapter 4: Conclusions and Future Directions

Nonsense-mediated mRNA decay regulates gene expression post-transcriptionally by facilitating degradation of PTC-containing transcripts. This highly conserved pathway relies on seven SMG proteins in *C. elegans*, including the central effector, SMG-2, to efficiently target a variety of PTC-containing messages. While the NMD system can identify and destroy aberrant or mutant mRNAs, it also regulates a significant fraction of the transcriptome in wild-type cells. 4-25% of genes exhibit altered expression in the absence of NMD, though only a subset of these changes are caused by direct interaction between transcripts and the NMD system [18, 23-29]. Common classes of NMD substrates include alternatively spliced mRNAs containing PTCs, transcripts with unusual physical characteristics such as upstream ORFs, and genomic "noise", including expressed pseudogenes [18] [123, 124] [17].

I defined direct NMD substrates using deep sequencing of RNAs that co-precipitate with SMG-2 and characterized transcripts affected by NMD. Four classes of messages were identified based on expression patterns and IP enrichment. Class I substrates are more abundant in the absence of NMD. I identified 585 genes that produce Class I transcripts and an additional 95 genes that produce high-confidence NMD substrates containing portions of introns. Class II mRNAs are consistently enriched in SMG-2 IPs, but do not demonstrate a significant increase in expression in the absence of NMD. 2005 mRNAs fall into this second category, a portion of which are likely true substrates of NMD. Class III and IV messages represent indirect effects that occur in the absence of NMD. Expression of Class III transcripts is elevated in *smg* mutant nematodes, while expression of Class IV transcripts is decreased under such conditions. Neither Class III nor Class IV mRNAs copurify with SMG-2. Expression of 131

genes was indirectly affected by the loss of NMD; 65 were categorized as Class III and 66 as Class IV.

Through intron-level analysis, I identified a possible targeting rationale for 171 of the 680 high-confidence substrates. Few of the included intron sequences preserved the reading frame, and one expressed sequence that incorporated full triplets introduced four premature stop codons. 136 high-confidence substrates are expressed pseudogenes and transposons. Together, these two factors offer an explanation for 285 of the 680 high-confidence genes are targeted by NMD. I additionally observed that transcripts with short ORFs, mature mRNAs that include more than one exon, and mRNAs that are encoded by two recently expanded gene families are more likely to be targeted for degradation.

Functional enrichment was used to better characterize the role of NMD. Nucleic acid binding, RNA metabolism, RNA processing, gene expression, and translation GO terms were all overrepresented in the high-confidence substrate list. Class II factors had strikingly similar enrichments. However, more functional annotations were enriched in Class II mRNAs, including transcription factors. Indirectly affected transcripts, which changed in abundance in NMD(-) samples but were not directly targeted by SMG-2, were often associated with defense and immune responses.

Future Directions

Why are the 395 unaccounted for mRNAs subjected to NMD?

The RIP-Seq data presented in this thesis was subjected to relatively basic analysis, and certain questions remain. For example, further data mining could determine whether differentially expressed exons contain PTCs or introduce early stops via translational

frameshifting. Detailed analysis could also determine which exons are mutually exclusive with PTC-causing exons. I investigated whether mean and median lengths of annotated UTRs correlate with whether the transcript is a substrate of NMD. Nearly half of high-confidence substrates did not have annotated 3' UTR information. Though I did not specifically design experiments to gather UTR sequences, UTR lengths could be determined using the available data. RNAs containing long 3' UTRs are targeted by NMD in many other systems [30, 139]. 106 of the 608 high-confidence substrates in *C. elegans* have annotated 3' UTRs that are longer than the median expressed mRNA's UTR. The contribution of 3' UTR length to degradation of these substrates could be measured either by looking for differing UTR lengths in wild-type and mutant samples or by testing constructs of these genes with shorter 3' UTRs.

How does NMD regulation overlap with other regulatory pathways?

Cells often utilize multiple layers of overlapping regulatory systems. Although bioinformatic searches for conserved motifs in other organisms do not identify specific or conserved SMG-2 binding sites [95, 140], such analyses could reveal co-regulation of NMD targets and other silencing or degradation pathways (for example, endogenous RNAi). Motif analysis could also potentially identify the Class II mRNAs that are substrates of other regulatory pathways that rely on SMG-2 as an effector protein.

How many exons (or genes) remain undiscovered due to NMD targeting?

My intron-level analysis identified novel alternative splice sites and putative cassette exons that are located in currently annotated introns and one potential new lincRNA gene. Alternatively spliced isoforms that are stabilized in *smg*-mutant nematodes were previously characterized by microarray analysis [25, 99], but my sequencing approach identified important

new NMD substrates. Intron-level analyses of existing data from other NMD-deficient organisms would likely provide a similar advances in our understanding of the physiological substrates of NMD.

What regulates the productive/unproductive splicing of substrate mRNAs?

The regulated unproductive splicing and turnover model posits that alternative splicing generates PTC-containing mRNAs as a means to finely regulate gene expression [110, 127, 128]. At its simplest, the mechanism relies on a protein (such as splicing factor RPL-3) to regulate its own alternative splicing [141]. However, mechanisms to regulate alternative splicing of pre-mRNAs that do not encode splicing factors are unclear. To partially elucidate such mechanisms, pre-mRNAs whose splicing is regulated by splicing factors that generate known NMD substrate isoforms could be identified via RIP or RNA-seq.

Are processed pseudogenes also heavily targeted by NMD?

Relative to mammalian genomes, the *C. elegans* genome contains a relative dearth of processed pseudogenes, which are generated via reverse transcription of mRNAs followed by integration into the genome [122]. Unprocessed or duplicated pseudogenes often include mutations, insertions, deletions, and frameshifts that are not found in their paralogs. Though processed pseudogenes can also accumulate mutations during evolution, I cannot estimate how frequently processed pseudogenes are directly targeted by NMD due to the dearth of such pseudogenes in *C. elegans*. Investigating expressed pseudogenes that are not NMD substrates would likely be informative. Such transcripts may have been misclassified as pseudogenes, or they may have functions as non-coding RNAs (for example, acting as a miRNA sponges or lincRNAs). Though *C. elegans* pseudogenes have homologs in other *Caenorhabditis* species, I

was unable to identify pseudogene-pseudogene conservation as a measure of potential functionality. WormBase and Ensembl did not list any pseudogenes in related species as homologs of *C. elegans* pseudogenes; direct comparison of the pseudogenes in *Caenorhabditis* may be more fruitful.

Do expanded gene families generate NMD substrates in other organisms?

The observation that transcripts encoding nuclear hormone receptors and F-box A proteins are more frequently targeted by NMD than can be expected by random chance alone is the first description that recently expanded gene families as a group are more likely to be targeted by NMD. Lynch and Conery described a trajectory where recently duplicated genes experienced relaxed selection but where nearly all gene duplicates are silenced within a few million years [142]. Substrates belonging to these families may be in the process of accumulating disabling mutations. The rate of gene duplication in *C. elegans* is not high [143], but testing whether duplicated genes are often targeted by NMD in other organisms could solidify my observations as a general principles during genome evolution. If expanded gene families often generate NMD substrates, the accumulation of mutations, the regulation of family expression via NMD targeting, and the eventual silencing of a set of genes could be tracked during evolutionary divergence.

Materials & Methods

Nematode strains and growth

The following strains were used in this study: N2 (CGC), TR2602 *smg-1(r910)*, and TR2605 *smg-1(r910) smg-2(r915)*. *Smg-1(r910)* is a null allele created by a transposon insertion in *smg-1*. SMG-1 protein is not detected on a western blot and these strains are strongly NMD defective (L. Johns and P. Anderson, unpublished). *Smg-2(r915)* is a 1,064 bp deletion within *smg-2* [38]. *Smg-2(r915)* expresses a truncated SMG-2 protein that is nonfunctional, as demonstrated by the strong NMD defect of *smg-2(r915)*, which is similar to that of the null allele *smg-2(r908)* (L. Johns and P. Anderson unpublished data). I generated synchronized strains by treating gravid adults with a 5.25% hypochlorite solution. Embryos were grown to L4 stage in S-media at 20°C.

SMG-2 Immunoprecipitation

Staged L4 animals were frozen in immunoprecipitation (IP) lysis buffer (20mM MOPS pH 7.2, 100mM NaCl, 0.01% NP-40, protease inhibitor cocktail (Sigma), pepstatin A (Sigma), RNase OUT, and PMSF), sonicated with 7x5s pulses, and centrifuged at 14,000 x g for 25 min, retaining the supernatant fraction. Bradford assays were used to determine the extract concentration. Extracts were diluted to 2mg/mL in IP lysis buffer. One aliquot of this total RNA was preserved as the input sample while a second aliquot was incubated with SMG-2 antibody coupled to Dynabead Protein G (Life Technologies) for the IP sample. Precipitated proteins were collected on the magnet and washed six times with IP wash buffer (20mM MOPS pH 7.2, 100mM NaCl, 0.02% NP-40). RNA was extracted from IP pellets and inputs using TRIzol (Life Technologies). Input samples were DNase treated. IP samples showed no gDNA contamination via PCR and

were not subjected to DNase treatment. This IP protocol was modified from Johns, et al [37]. All IP and input samples were submitted for sequencing in biological triplicate.

RNA-Sequencing

Poly(A) selection, cDNA library construction, and next-generation sequencing using the Illumina HiSeq 2000 platform was performed by the University of Wisconsin Biotechnology Center DNA Sequence Facility. Approximately 500ng of RNA was used for input sample library preparation; all of the submitted RNA (45-228ng per negative control IP sample, 263-945ng per experimental IP) was used for IP sample library preparation. Single-end reads were subjected to a mild quality trim using Trimmomatic [103] and were aligned to the WBcel235 genome using Tophat2 [104].

Bioinformatic Analysis

Following alignment, both gene counts and intron counts were compiled using htseq-count [105]. The reference intron annotation file was generated by merging each gene's intersecting exons, as defined by the Ensembl GTF available through iGenomes, and using the intervening coordinates as intron coordinates. Differential expression was determined using EdgeR as previously described [107, 144]. Sequence data was visualized using Broad's IGV [145]. GO term statistical overrepresentation and pathway analysis were performed using PANTHER and Gorilla [146, 147].

RT-PCR and qRT-PCR Validation

First-strand synthesis was performed on aliquots of all input and IP samples using random hexamers and SuperScript III reverse transcriptase (Life Technologies). Prior to high-throughput sequencing, IP quality was tested using primers on either side of a PTC-containing

alternatively spliced intron for rpl-12. Products were separated on a 2% ethidium bromide-stained agarose gel and quantified using ImageJ.

Following high-throughput sequence analysis, qPCR validation was performed on input samples using primers for B0495.8, F45D11.1, F53B2.8, fib-1, fbxa-33, linc-9, nhr-109, pho-11, pqn-70, tdp-1, R08E5.3, rpl-7a, rsp-2, rsp-6, Y39B6A.21, Y51A2D.13, and ZK970.7. Validation of SMG-2 association was performed via RIP-coupled qRT-PCR using primers for fib-1, fbxa-33, nhr-109, pho-11, and rpl-7a. Included introns discovered via differential intron expression were validated via RT-PCR for dpy-7, deg-1, phy-2, and R06C1.4. Primer sequences provided below.

Primer Sequences

Gene	Validation of?	Forward Primer	Reverse Primer
deg-1 #1	Intron inclusion	GCTACTGACCTTGGAGACTG	CCGTATGCCTCTGATTCTTG
deg-1 #2	Intron inclusion	CGCTAATCGCTCAATTTTCTG	AAGATTGCCGAACAGTCAC
dpy-7	Intron inclusion	GCTCACGAACGTCTTCAA	GTCATGCCAAATGTTGCG
R06C1.4	Intron inclusion	CCCAAGGATTCTCCGTCTA	GCGAAGTAGTTTCCGATCTC
ttr-46	Intron inclusion	GCTACATCACTAGCGGAAAAG	CCATATAATCGGCATTGAAACC
phy-2	Intron inclusion	GGTTATCAAGGAGTTGGCTTC	CCAAATCTCCCTTCAACCAAG
rpl-12	IP efficacy	CCAAAGAAGATCGGAGAAGAC	GATTGGGCTGTTCCAAGG
rpl-10a	IP efficacy	GAACTACGACCCACAGAAGGACAA	GCGACGGAGAGGCAGAGAA
rsp-4	IP efficacy	TTTTATGAACGTCGTGATGCTG	TAGCGGGAGTTGGAACGG
B0495.8	Class I mRNA	GAGTATGTCCTCACGATATGG	GTTTACGATACGCTCGCAG
F45D11.1	Class I mRNA	GGGAAATCTCAAGGGTCAAC	GAAGAGAAGCCTTGTGAGC
linc-9	Class I mRNA	AAGACTACGCCCTGGAC	CTTCTCGAATTGTAGTGGTCG
pqn-70	Class I mRNA	GTGGAAAGGATACAATGGACC	GCTGTTCTTTCCGAATCCAG
rsp-2	Class I mRNA	CGACTAGATTCCGTTTGGTG	GCTTGGTTGACATTTCTCTTG
rsp-6	Class I mRNA	ATGGACGCCAAGGTGTACG	CGATCAAAAATCTCTTCGAGTTC
tdp-1	Class I mRNA	GGGTGCTACTGGTTTGAAG	TTATTTTCCCAGCCGTCAG
fbxa-33	Class I mRNA	TGACCTGCCGAAAAGTTTG	CAAGATCGGTGAAGTTAGCTC
nhr-109	Class I mRNA	GCAGGGAATGAAGGATTATACG	AATGGAATGTAGGAGTTGCG
rpl-7a	Class I mRNA	GTCGCTCATTTTACTTTTGGG	CTTCTGGACTTGAGTTTGGG
fib-1	Class II mRNA	CGTTGTCCCAATTGTCGAAG	AGGAAGTTTTGGGCATTGAG
pho-11	Class II mRNA	ACTATCCATACTCCCCACTC	GAGTCAACATCTGCCCTTG
F53B2.8	Class III mRNA	GGATTTGGGAGATGCTATGAG	CGTACCCTTTTCTGGTGTTTC
Y39B6A.21	Class III mRNA	GCCCTGGATAGTGTTAATTGTG	GCTGTCTTTGTCTCCGAAAAG
R08E5.3	Class IV mRNA	TCAGCGGAGACCAGATTT	CTCTACCATAGCGAACATCC
Y51A2D.13	Class IV mRNA	CGATAGTTATCAGGCAAACCG	CGCAGAAATCTCCAGTGTTAC
ZK970.7	Class IV mRNA	CATCTCCAGTTTCTCACCAG	CTTTAACATTAGCTGGGAGGG

References

1. Leeds, P., et al., *The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon.* Genes Dev, 1991. **5**(12A): p. 2303-14.

2. Leeds, P., et al., *Gene products that promote mRNA turnover in Saccharomyces cerevisiae.* Mol Cell Biol, 1992. **12**(5): p. 2165-77.

3. Hodgkin, J., et al., *A new kind of informational suppression in the nematode Caenorhabditis elegans.* Genetics, 1989. **123**(2): p. 301-13.

4. Pulak, R. and P. Anderson, *mRNA surveillance by the Caenorhabditis elegans smg genes.* Genes Dev, 1993. **7**(10): p. 1885-97.

5. Gatfield, D., et al., *Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways.* EMBO J, 2003. **22**(15): p. 3960-70.

6. Arciga-Reyes, L., et al., *UPF1 is required for nonsense-mediated mRNA decay (NMD) and RNAi in Arabidopsis.* Plant J, 2006. **47**(3): p. 480-9.

7. Wittkopp, N., et al., *Nonsense-mediated mRNA decay effectors are essential for zebrafish embryonic development and survival.* Mol Cell Biol, 2009. **29**(13): p. 3517-28.

8. Medghalchi, S.M., et al., *Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability.* Hum Mol Genet, 2001. **10**(2): p. 99-105.

9. Sun, X., et al., *A mutated human homologue to yeast Upf1 protein has a dominant-negative effect on the decay of nonsense-containing mRNAs in mammalian cells.* Proc Natl Acad Sci U S A, 1998. **95**(17): p. 10009-14.

10. Donahue, T.F., P.J. Farabaugh, and G.R. Fink, *Suppressible four-base glycine and proline codons in yeast.* Science, 1981. **212**(4493): p. 455-7.

11. Linde, L., et al., *Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin.* J Clin Invest, 2007. **117**(3): p. 683-92.

12. Wickens, M. and P. Stephenson, *Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation.* Science, 1984. **226**(4678): p. 1045-51.

13. Houseley, J. and D. Tollervey, *The many pathways of RNA degradation.* Cell, 2009. **136**(4): p. 763-76.

14. Moore, M.J. and N.J. Proudfoot, *Pre-mRNA processing reaches back to transcription and ahead to translation.* Cell, 2009. **136**(4): p. 688-700.

15. Chang, J.C., et al., *Suppression of the nonsense mutation in homozygous beta 0 thalassaemia.* Nature, 1979. **281**(5732): p. 602-3.

16. Losson, R. and F. Lacroute, *Interference of nonsense mutations with eukaryotic messenger RNA stability.* Proc Natl Acad Sci U S A, 1979. **76**(10): p. 5134-7.

17. Mitrovich, Q.M. and P. Anderson, *mRNA surveillance of expressed pseudogenes in C. elegans.* Curr Biol, 2005. **15**(10): p. 963-7.

18. Mendell, J.T., et al., *Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise.* Nat Genet, 2004. **36**(10): p. 1073-8.

19. Garcia, D., S. Garcia, and O. Voinnet, *Nonsense-mediated decay serves as a general viral restriction mechanism in plants.* Cell Host Microbe, 2014. **16**(3): p. 391-402.

20. Balistreri, G., et al., *The host nonsense-mediated mRNA decay pathway restricts Mammalian RNA virus replication.* Cell Host Microbe, 2014. **16**(3): p. 403-11.

21. Smith, J.E. and K.E. Baker, *Nonsense-mediated RNA decay--a switch and dial for regulating gene expression.* Bioessays, 2015. **37**(6): p. 612-23.

22. Peccarelli, M. and B.W. Kebaara, *Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway.* Eukaryot Cell, 2014. **13**(9): p. 1126-35.

23. He, F., et al., *Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5′ to 3′ mRNA decay pathways in yeast.* Mol Cell, 2003. **12**(6): p. 1439-52.

24. Metzstein, M.M. and M.A. Krasnow, *Functions of the nonsense-mediated mRNA decay pathway in Drosophila development.* PLoS Genet, 2006. **2**(12): p. e180.

25. Ramani, A.K., et al., *High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective Caenorhabditis elegans.* Genome Biol, 2009. **10**(9): p. R101.

26. Rayson, S., et al., *A role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in Arabidopsis thaliana NMD mutants.* PLoS One, 2012. **7**(2): p. e31917.

27. Rehwinkel, J., et al., *Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets.* RNA, 2005. **11**(10): p. 1530-44.

28. Yepiskoposyan, H., et al., *Autoregulation of the nonsense-mediated mRNA decay pathway in human cells.* RNA, 2011. **17**(12): p. 2108-18.

29. Wittmann, J., E.M. Hol, and H.M. Jack, *hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay.* Mol Cell Biol, 2006. **26**(4): p. 1272-87.

30. Kebaara, B.W. and A.L. Atkin, *Long 3'-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in Saccharomyces cerevisiae.* Nucleic Acids Res, 2009. **37**(9): p. 2771-8.

31. Schweingruber, C., et al., *Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells.* Biochim Biophys Acta, 2013. **1829**(6-7): p. 612-23.

32. Cali, B.M., et al., *smg-7 is required for mRNA surveillance in Caenorhabditis elegans.* Genetics, 1999. **151**(2): p. 605-16.

33. Casadio, A., et al., *Identification and characterization of novel factors that act in the nonsense-mediated mRNA decay pathway in nematodes, flies and mammals.* EMBO Rep, 2015. **16**(1): p. 71-8.

34. Longman, D., et al., *Mechanistic insights and identification of two novel factors in the C. elegans NMD pathway.* Genes Dev, 2007. **21**(9): p. 1075-85.

35. Yamashita, A., et al., *SMG-8 and SMG-9, two novel subunits of the SMG-1 complex, regulate remodeling of the mRNA surveillance complex during nonsense-mediated mRNA decay.* Genes Dev, 2009. **23**(9): p. 1091-105.

36. Rosains, J. and S.E. Mango, *Genetic characterization of smg-8 mutants reveals no role in C. elegans nonsense mediated decay.* PLoS One, 2012. **7**(11): p. e49490.

37. Johns, L., et al., *Caenorhabditis elegans SMG-2 selectively marks mRNAs containing premature translation termination codons.* Mol Cell Biol, 2007. **27**(16): p. 5630-8.

38. Page, M.F., et al., *SMG-2 is a phosphorylated protein required for mRNA surveillance in Caenorhabditis elegans and related to Upf1p of yeast.* Mol Cell Biol, 1999. **19**(9): p. 5943-51.

39. Ohnishi, T., et al., *Phosphorylation of hUPF1 induces formation of mRNA surveillance complexes containing hSMG-5 and hSMG-7.* Mol Cell, 2003. **12**(5): p. 1187-200.

40. Chiu, S.Y., et al., *Characterization of human Smg5/7a: a protein with similarities to Caenorhabditis elegans SMG5 and SMG7 that functions in the dephosphorylation of Upf1.* RNA, 2003. **9**(1): p. 77-87.

41. Yamashita, A., et al., *Human SMG-1, a novel phosphatidylinositol 3-kinase-related protein kinase, associates with components of the mRNA surveillance complex and is involved in the regulation of nonsense-mediated mRNA decay.* Genes Dev, 2001. **15**(17): p. 2215-28.

42. Grimson, A., et al., *SMG-1 is a phosphatidylinositol kinase-related protein kinase required for nonsense-mediated mRNA Decay in Caenorhabditis elegans.* Mol Cell Biol, 2004. **24**(17): p. 7483-90.

43. Anders, K.R., A. Grimson, and P. Anderson, *SMG-5, required for C.elegans nonsense-mediated mRNA decay, associates with SMG-2 and protein phosphatase 2A.* EMBO J, 2003. **22**(3): p. 641-50.

44. Glavan, F., et al., *Structures of the PIN domains of SMG6 and SMG5 reveal a nuclease within the mRNA surveillance complex.* EMBO J, 2006. **25**(21): p. 5117-25.

45. Huntzinger, E., et al., *SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan.* RNA, 2008. **14**(12): p. 2609-17.

46. Isken, O. and L.E. Maquat, *The multiple lives of NMD factors: balancing roles in gene and genome regulation.* Nat Rev Genet, 2008. **9**(9): p. 699-712.

47. Kim, Y.K., et al., *Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay.* Cell, 2005. **120**(2): p. 195-208.

48. Park, E., M.L. Gleghorn, and L.E. Maquat, *Staufen2 functions in Staufen1-mediated mRNA decay by binding to itself and its paralog and promoting UPF1 helicase but not ATPase activity.* Proc Natl Acad Sci U S A, 2013. **110**(2): p. 405-12.

49. Kaygun, H. and W.F. Marzluff, *Regulated degradation of replication-dependent histone mRNAs requires both ATR and Upf1.* Nat Struct Mol Biol, 2005. **12**(9): p. 794-800.

50. Carter, M.S., et al., *A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts in vivo is reversed by protein synthesis inhibitors in vitro.* J Biol Chem, 1995. **270**(48): p. 28995-9003.

51. Thermann, R., et al., *Binary specification of nonsense codons by splicing and cytoplasmic translation.* EMBO J, 1998. **17**(12): p. 3484-94.

52. Hoshino, S., et al., *The eukaryotic polypeptide chain releasing factor (eRF3/GSPT) carrying the translation termination signal to the 3'-Poly(A) tail of mRNA. Direct association of erf3/GSPT with polyadenylate-binding protein.* J Biol Chem, 1999. **274**(24): p. 16677-80.

53. Uchida, N., et al., *A novel role of the mammalian GSPT/eRF3 associating with poly(A)-binding protein in Cap/Poly(A)-dependent translation.* J Biol Chem, 2002. **277**(52): p. 50286-92.

54. Singh, G., I. Rebbapragada, and J. Lykke-Andersen, *A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay.* PLoS Biol, 2008. **6**(4): p. e111.

55. Stalder, L. and O. Muhlemann, *Transcriptional silencing of nonsense codon-containing immunoglobulin micro genes requires translation of its mRNA.* J Biol Chem, 2007. **282**(22): p. 16079-85.

56. Behm-Ansmant, I., et al., *A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay.* EMBO J, 2007. **26**(6): p. 1591-601.

57. Nagy, E. and L.E. Maquat, *A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.* Trends Biochem Sci, 1998. **23**(6): p. 198-9.

58. Kashima, I., et al., *Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay.* Genes Dev, 2006. **20**(3): p. 355-67.

59. Buhler, M., et al., *EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length.* Nat Struct Mol Biol, 2006. **13**(5): p. 462-4.

60. Wang, W., et al., *The role of Upf proteins in modulating the translation read-through of nonsense-containing transcripts.* EMBO J, 2001. **20**(4): p. 880-90.

61. Czaplinski, K., et al., *The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs.* Genes Dev, 1998. **12**(11): p. 1665-77.

62. Bhattacharya, A., et al., *Characterization of the biochemical properties of the human Upf1 gene product that is involved in nonsense-mediated mRNA decay.* RNA, 2000. **6**(9): p. 1226-35.

63. Czaplinski, K., et al., *Purification and characterization of the Upf1 protein: a factor involved in translation and mRNA degradation.* RNA, 1995. **1**(6): p. 610-23.

64. Le Hir, H., et al., *The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay.* EMBO J, 2001. **20**(17): p. 4987-97.

65. Eberle, A.B., et al., *Posttranscriptional gene regulation by spatial rearrangement of the 3' untranslated region.* PLoS Biol, 2008. **6**(4): p. e92.

66. Zhang, J., et al., *Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm.* RNA, 1998. **4**(7): p. 801-15.

67. Sauliere, J., et al., *CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex.* Nat Struct Mol Biol, 2012. **19**(11): p. 1124-31.

68. Singh, G., et al., *The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus.* Cell, 2012. **151**(4): p. 750-64.

69. Okada-Katsuhata, Y., et al., *N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD.* Nucleic Acids Res, 2012. **40**(3): p. 1251-66.

70. Deniaud, A., et al., *A network of SMG-8, SMG-9 and SMG-1 C-terminal insertion domain regulates UPF1 substrate recruitment and phosphorylation.* Nucleic Acids Res, 2015. **43**(15): p. 7600-11.

71. Chamieh, H., et al., *NMD factors UPF2 and UPF3 bridge UPF1 to the exon junction complex and stimulate its RNA helicase activity.* Nat Struct Mol Biol, 2008. **15**(1): p. 85-93.

72. Chakrabarti, S., et al., *Molecular mechanisms for the RNA-dependent ATPase activity of Upf1 and its regulation by Upf2.* Mol Cell, 2011. **41**(6): p. 693-703.

73. Eberle, A.B., et al., *SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells.* Nat Struct Mol Biol, 2009. **16**(1): p. 49-55.

74. Gatfield, D. and E. Izaurralde, *Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila.* Nature, 2004. **429**(6991): p. 575-8.

75. Kashima, I., et al., *SMG6 interacts with the exon junction complex via two conserved EJC-binding motifs (EBMs) required for nonsense-mediated mRNA decay.* Genes Dev, 2010. **24**(21): p. 2440-50.

76. Unterholzner, L. and E. Izaurralde, *SMG7 acts as a molecular link between mRNA surveillance and mRNA decay.* Mol Cell, 2004. **16**(4): p. 587-96.

77. Ezzeddine, N., et al., *Human TOB, an antiproliferative transcription factor, is a poly(A)-binding protein-dependent positive regulator of cytoplasmic mRNA deadenylation.* Mol Cell Biol, 2007. **27**(22): p. 7791-801.

78. Yamashita, A., et al., *Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover.* Nat Struct Mol Biol, 2005. **12**(12): p. 1054-63.

79. Loh, B., S. Jonas, and E. Izaurralde, *The SMG5-SMG7 heterodimer directly recruits the CCR4-NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2.* Genes Dev, 2013. **27**(19): p. 2125-38.

80. Osawa, M., et al., *Biological role of the two overlapping poly(A)-binding protein interacting motifs 2 (PAM2) of eukaryotic releasing factor eRF3 in mRNA decay.* RNA, 2012. **18**(11): p. 1957-67.

81. Cho, H., et al., *SMG5-PNRC2 is functionally dominant compared with SMG5-SMG7 in mammalian nonsense-mediated mRNA decay.* Nucleic Acids Res, 2013. **41**(2): p. 1319-28.

82. Cho, H., K.M. Kim, and Y.K. Kim, *Human proline-rich nuclear receptor coregulatory protein 2 mediates an interaction between mRNA surveillance machinery and decapping complex.* Mol Cell, 2009. **33**(1): p. 75-86.

83. Muhlrad, D. and R. Parker, *Premature translational termination triggers mRNA decapping.* Nature, 1994. **370**(6490): p. 578-81.

84. Kuroha, K., T. Tatematsu, and T. Inada, *Upf1 stimulates degradation of the product derived from aberrant messenger RNA containing a specific nonsense mutation by the proteasome.* EMBO Rep, 2009. **10**(11): p. 1265-71.

85. Takahashi, S., et al., *Upf1 potentially serves as a RING-related E3 ubiquitin ligase via its association with Upf3 in yeast.* RNA, 2008. **14**(9): p. 1950-8.

86. Amrani, N., et al., *A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay.* Nature, 2004. **432**(7013): p. 112-8.

87. Peixeiro, I., et al., *Interaction of PABPC1 with the translation initiation complex is critical to the NMD resistance of AUG-proximal nonsense mutations.* Nucleic Acids Res, 2012. **40**(3): p. 1160-73.

88. Ghosh, S., et al., *Translational competence of ribosomes released from a premature termination codon is modulated by NMD factors.* RNA, 2010. **16**(9): p. 1832-47.

89. Celik, A., S. Kervestin, and A. Jacobson, *NMD: At the crossroads between translation termination and ribosome recycling.* Biochimie, 2015. **114**: p. 2-9.

90. Guan, Q., et al., *Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast.* PLoS Genet, 2006. **2**(11): p. e203.

91. Rehwinkel, J., J. Raes, and E. Izaurralde, *Nonsense-mediated mRNA decay: Target genes and functional diversification of effectors.* Trends Biochem Sci, 2006. **31**(11): p. 639-46.

92. Tani, H., et al., *Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability.* RNA Biol, 2012. **9**(11): p. 1370-9.

93. Schmidt, S.A., et al., *Identification of SMG6 cleavage sites and a preferred RNA cleavage motif by global analysis of endogenous NMD targets in human cells.* Nucleic Acids Res, 2015. **43**(1): p. 309-23.

94. Johansson, M.J., et al., *Association of yeast Upf1p with direct substrates of the NMD pathway.* Proc Natl Acad Sci U S A, 2007. **104**(52): p. 20872-7.

95. Hurt, J.A., A.D. Robertson, and C.B. Burge, *Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay.* Genome Res, 2013. **23**(10): p. 1636-50.

96. Chapin, A., et al., *In vivo determination of direct targets of the nonsense-mediated decay pathway in Drosophila.* G3 (Bethesda), 2014. **4**(3): p. 485-96.

97. Matia-Gonzalez, A.M., et al., *Functional characterization of Upf1 targets in Schizosaccharomyces pombe.* RNA Biol, 2013. **10**(6): p. 1057-65.

98. Maderazo, A.B., et al., *Nonsense-containing mRNAs that accumulate in the absence of a functional nonsense-mediated mRNA decay pathway are destabilized rapidly upon its restitution.* Mol Cell Biol, 2003. **23**(3): p. 842-51.

99. Barberan-Soler, S., N.J. Lambert, and A.M. Zahler, *Global analysis of alternative splicing uncovers developmental regulation of nonsense-mediated decay in C. elegans.* RNA, 2009. **15**(9): p. 1652-60.

100. Weischenfeldt, J., et al., *Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns.* Genome Biol, 2012. **13**(5): p. R35.

101. Maquat, L.E. and G.G. Carmichael, *Quality control of mRNA function.* Cell, 2001. **104**(2): p. 173-6.

102. Cali, B.M. and P. Anderson, *mRNA surveillance mitigates genetic dominance in Caenorhabditis elegans.* Mol Gen Genet, 1998. **260**(2-3): p. 176-84.

103. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

104. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.

105. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

106. Hardcastle, T.J. and K.A. Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 2010. **11**: p. 422.

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

108. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

109. Seyednasrollah, F., A. Laiho, and L.L. Elo, *Comparison of software packages for detecting differential expression in RNA-seq studies.* Brief Bioinform, 2015. **16**(1): p. 59-70.

110. Lareau, L.F., et al., *The coupling of alternative splicing and nonsense-mediated mRNA decay.* Adv Exp Med Biol, 2007. **623**: p. 190-211.

111. Ni, J.Z., et al., *Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay.* Genes Dev, 2007. **21**(6): p. 708-18.

112. Filichkin, S.A., et al., *Environmental stresses modulate abundance and timing of alternatively spliced circadian transcripts in Arabidopsis.* Mol Plant, 2015. **8**(2): p. 207-27.

113. Gardner, L.B. and P.G. Corn, *Hypoxic regulation of mRNA expression.* Cell Cycle, 2008. **7**(13): p. 1916-24.

114. Garre, E., et al., *Nonsense-mediated mRNA decay controls the changes in yeast ribosomal protein pre-mRNAs levels upon osmotic stress.* PLoS One, 2013. **8**(4): p. e61240.

115. Kalyna, M., et al., *Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis.* Nucleic Acids Res, 2012. **40**(6): p. 2454-69.

116. Karam, R., et al., *The unfolded protein response is shaped by the NMD pathway.* EMBO Rep, 2015. **16**(5): p. 599-609.

117. Sakaki, K., et al., *RNA surveillance is required for endoplasmic reticulum homeostasis.* Proc Natl Acad Sci U S A, 2012. **109**(21): p. 8079-84.

118. Wang, D., et al., *Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis.* Mol Cell Biol, 2011. **31**(17): p. 3670-80.

119. Brumbaugh, K.M., et al., *The mRNA surveillance protein hSMG-1 functions in genotoxic stress response pathways in mammalian cells.* Mol Cell, 2004. **14**(5): p. 585-98.

120. Masse, I., et al., *A novel role for the SMG-1 kinase in lifespan and oxidative stress resistance in Caenorhabditis elegans.* PLoS One, 2008. **3**(10): p. e3354.

121. Lelivelt, M.J. and M.R. Culbertson, *Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome.* Mol Cell Biol, 1999. **19**(10): p. 6710-9.

122. Harrison, P.M., N. Echols, and M.B. Gerstein, *Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome.* Nucleic Acids Res, 2001. **29**(3): p. 818-30.

123. Mitrovich, Q.M. and P. Anderson, *Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans.* Genes Dev, 2000. **14**(17): p. 2173-84.

124. Morrison, M., K.S. Harris, and M.B. Roth, *smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans.* Proc Natl Acad Sci U S A, 1997. **94**(18): p. 9782-5.

125. Sureau, A., et al., *SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs.* EMBO J, 2001. **20**(7): p. 1785-96.

126. Wollerton, M.C., et al., *Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay.* Mol Cell, 2004. **13**(1): p. 91-100.

127. Lewis, B.P., R.E. Green, and S.E. Brenner, *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A, 2003. **100**(1): p. 189-92.

128. Green, R.E., et al., *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes.* Bioinformatics, 2003. **19 Suppl 1**: p. i118-21.

129. Pan, Q., et al., *Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression.* Genes Dev, 2006. **20**(2): p. 153-8.

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

131. Kipreos, E.T. and M. Pagano, *The F-box protein family.* Genome Biol, 2000. **1**(5): p. REVIEWS3002.

132. Taubert, S., J.D. Ward, and K.R. Yamamoto, *Nuclear hormone receptors in nematodes: evolution and function.* Mol Cell Endocrinol, 2011. **334**(1-2): p. 49-55.

133. Sluder, A.E., et al., *The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes.* Genome Res, 1999. **9**(2): p. 103-20.

134. Haerty, W., et al., *Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution.* BMC Genomics, 2008. **9**: p. 399.

135. de Lima Morais, D.A. and P.M. Harrison, *Large-scale evidence for conservation of NMD candidature across mammals.* PLoS One, 2010. **5**(7): p. e11695.

136. Decourty, L., et al., *Long open reading frame transcripts escape nonsense-mediated mRNA decay in yeast.* Cell Rep, 2014. **6**(4): p. 593-8.

137. Dunbar, T.L., et al., *C. elegans detects pathogen-induced translational inhibition to activate immune signaling.* Cell Host Microbe, 2012. **11**(4): p. 375-86.

138. Lee, M.H. and T. Schedl, *Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in C. elegans.* Genes Dev, 2004. **18**(9): p. 1047-59.

139. Kertesz, S., et al., *Both introns and long 3'-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants.* Nucleic Acids Res, 2006. **34**(21): p. 6147-57.

140. Zund, D., et al., *Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3′ UTRs.* Nat Struct Mol Biol, 2013. **20**(8): p. 936-43.

141. Cuccurese, M., et al., *Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression.* Nucleic Acids Res, 2005. **33**(18): p. 5965-77.

142. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.

143. Gu, Z., et al., *Extent of gene duplication in the genomes of Drosophila, nematode, and yeast.* Mol Biol Evol, 2002. **19**(3): p. 256-62.

144. Anders, S., et al., *Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.* Nat Protoc, 2013. **8**(9): p. 1765-86.

145. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

146. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. **10**: p. 48.

147. Mi, H., A. Muruganujan, and P.D. Thomas, *PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.* Nucleic Acids Res, 2013. **41**(Database issue): p. D377-86.

Appendix 1: High Confidence NMD Substrates

Gene name	*smg(-)* > N2 log2 F.C.	IP > Input log2 F.C.
aat-4	1.36	2.25
adr-1	1.07	3.79
agr-1	1.85	2.61
AH10.4	2.19	1.28
alkb-8	Not Sig.	2.28
aly-2	Not Sig.	2.51
anr-24	2.56	3.59
arf-1.1	3.19	3.83
arl-7	3.78	3.59
arrd-21	2.32	3.30
arrd-23	1.91	3.23
asna-2	3.72	3.93
asp-11	2.33	3.12
B0035.3	1.45	3.09
B0238.1	1.57	3.42
B0250.12	1.46	1.95
B0250.4	2.71	4.28
B0250.7	2.01	4.25
B0284.5	1.86	1.49
B0285.4	1.94	3.05
B0286.3	Not Sig.	2.15
B0391.10	Not Sig.	0.75
B0416.10	3.82	3.36
B0495.8	1.81	4.49
B0495.9	2.22	2.96
B0507.9	1.93	1.67
B0511.11	Not Sig.	Not Sig.
B0511.2	2.11	1.44
B0545.4	1.86	1.99
bath-14	2.28	3.10
bath-18	3.43	4.34
bath-25	2.50	3.52
bath-29	Not Sig.	1.26
bath-6	1.90	3.10
bbs-8	1.59	3.94
best-8	2.03	4.42
brc-1	Not Sig.	1.62
btb-1	3.32	2.81
btb-3	3.18	2.56
btb-4	2.10	1.05
C01B9.4	Not Sig.	Not Sig.
C01B9.5	3.40	1.59
C01G10.10	1.44	2.95
C01G12.12	2.98	2.17
C01H6.6	1.91	2.21
C02B8.1	1.66	1.81
C02G6.1	1.18	3.50
C03B1.7	1.49	2.98
C03H12.1	Not Sig.	Not Sig.
C05B5.5	2.24	2.23
C05B5.9	2.31	3.18
C05C8.1	1.25	1.78
C06C3.12	Not Sig.	3.84
C06C3.4	1.88	3.18
C06G3.8	Not Sig.	0.92
C08B6.10	1.25	3.04
C08F11.7	3.32	2.35
C09E7.6	2.61	4.70
C10B5.1	2.06	3.36
C12D8.1	1.88	3.82
C14C6.12	5.95	2.46
C16A11.5	1.29	4.04
C16C4.18	2.67	4.38
C16C8.20	2.45	3.59
C17C3.1	Not Sig.	1.80
C17C3.15	2.74	4.25
C17G1.5	1.45	3.23
C23G10.5	2.42	2.48
C24F3.2	Not Sig.	1.40
C25A1.4	1.33	3.09
C25A6.1	3.55	2.78
C25H3.12	3.87	3.87
C26B9.1	2.22	2.54
C27A7.2	2.25	2.46
C27A7.3	1.67	1.41
C28H8.2	2.56	2.50

C29E4.13	1.16	2.22	cbl-1	1.70	1.97
C29F9.3	1.96	2.01	cco-2	Not Sig.	-0.90
C29F9.4	1.57	2.15	CE7X_3.2	3.65	3.44
C30E1.9	2.95	3.20	ceh-19	1.65	1.52
C32E12.4	1.68	4.25	ceh-22	1.70	3.45
C33E10.8	3.75	3.04	ceh-88	1.09	3.36
C34C12.12	2.26	2.41	che-13	3.63	4.00
C34C6.2	1.80	3.26	chil-13	1.58	2.30
C34D4.10	3.41	4.73	chil-3	2.35	3.67
C35A5.11	2.51	2.77	chil-7	3.90	3.51
C35B1.2	2.33	4.21	clec-149	Not Sig.	0.75
C35D10.8	2.01	2.35	clec-150	Not Sig.	1.29
C35D6.5	3.56	3.04	clec-221	2.13	1.78
C36A4.12	1.91	1.47	clec-23	2.74	3.51
C37C3.2	2.00	1.85	clec-69	3.48	3.84
C37C3.7	2.08	2.98	clec-75	2.11	3.42
C40H1.2	3.44	1.62	clp-3	Not Sig.	0.94
C40H5.8	4.42	3.92	col-182	3.10	4.22
C42C1.13	1.12	2.47	col-53	2.12	4.31
C44C10.5	3.52	4.29	col-55	3.62	2.37
C44E4.5	Not Sig.	Not Sig.	coq-3	1.07	2.48
C44H9.6	1.54	3.17	cpr-6	Not Sig.	Not Sig.
C45E5.3	2.03	3.11	csp-2	1.61	2.58
C46G7.5	2.19	1.39	cul-6	1.34	3.26
C46H3.1	2.37	2.49	cutl-14	1.74	2.02
C48D1.5	1.90	3.03	cyl-1	1.25	4.64
C49C3.3	3.94	2.47	cyp-32B1	1.47	3.68
C50D2.10	2.53	2.39	cyp-33D1	2.74	3.06
C50D2.8	Not Sig.	1.24	D1022.4	1.15	2.04
C52D10.10	4.57	2.35	D1037.1	1.06	3.71
C53A5.13	1.81	3.19	D1054.6	2.26	3.24
C53C11.1	3.38	2.67	ddo-2	1.34	4.18
C53H9.2	3.27	4.74	deg-1	Not Sig.	Not Sig.
C54C8.4	2.83	2.10	dpy-7	Not Sig.	3.09
C54D10.12	2.95	2.28	dyla-1	2.17	2.84
C55A6.10	1.27	3.17	E01A2.2	1.48	4.56
C56A3.10	3.28	2.55	E01A2.5	1.32	3.45
C56E6.4	1.80	3.21	E02C12.8	3.62	3.87
C56G3.2	2.57	3.68	E02H9.2	1.56	2.66
cal-5	Not Sig.	Not Sig.	E03A3.10	1.65	1.88
cal-6	3.08	2.56	ears-2	1.18	2.34
cam-1	Not Sig.	-0.56	ech-2	1.78	3.91
cars-2	2.32	1.55	EEED8.13	3.96	3.27

EEED8.18	4.14	2.92	F23D12.9	4.92	3.53
EEED8.3	3.00	3.23	F23F12.3	Not Sig.	0.65
eef-1G	Not Sig.	2.20	F25G6.9	1.50	2.75
EGAP9.4	5.03	1.63	F26F2.15	4.12	3.18
egl-38	2.24	3.39	F26F4.8	1.48	2.62
eor-2	Not Sig.	1.85	F27B3.7	1.59	3.84
F01G12.1	3.93	1.85	F27D4.3	2.79	2.94
F02C12.3	4.36	3.65	F28F8.10	2.25	3.06
F02C12.7	4.73	3.62	F28F8.9	1.30	2.60
F02D10.3	2.08	3.49	F28G4.3	3.05	2.48
F07E5.9	2.87	3.54	F29D10.3	2.19	2.92
F07F6.2	1.76	2.62	F29G9.7	4.75	3.99
F08B4.3	3.30	3.86	F31E9.3	1.89	2.29
F08F8.7	Not Sig.	2.50	F32B4.4	1.75	4.31
F09D12.2	2.27	3.64	F32D8.5	Not Sig.	0.83
F09E5.8	Not Sig.	2.63	F33E2.4	1.72	3.38
F09F3.8	3.38	4.08	F33E2.8	2.86	4.04
F10G2.7	2.10	2.85	F35E12.4	2.81	1.99
F13G3.6	1.50	3.18	F35E2.7	3.02	2.98
F13H6.4	3.79	4.56	F35F10.4	1.75	3.29
F13H8.9	1.66	1.84	F36G9.4	3.39	3.64
F14D2.14	3.48	3.45	F37C4.6	Not Sig.	3.49
F14D2.5	1.75	1.82	F37H8.2	Not Sig.	1.90
F14D2.8	3.57	3.07	F39E9.1	2.37	2.18
F14E5.1	2.32	2.27	F40D4.17	3.27	2.87
F15H9.7	2.01	3.09	F40E3.7	3.54	3.82
F16A11.2	1.14	3.13	F40H6.6	2.62	2.26
F16B4.5	2.02	4.39	F41C3.6	2.48	2.50
F18A11.6	1.68	4.36	F42H10.6	Not Sig.	3.89
F18C5.9	5.42	2.47	F43D2.3	2.99	2.78
F19B2.12	3.16	4.19	F43G6.3	3.21	4.15
F19F10.12	1.07	2.95	F44F1.4	1.88	2.85
F19G12.8	1.94	1.65	F44F1.5	3.65	5.12
F20C5.5	2.39	4.08	F45B8.5	3.04	2.10
F20G2.7	1.47	1.94	F45D11.1	3.72	4.66
F21A9.1	1.84	3.73	F45D3.1	3.21	1.22
F21F8.5	3.28	3.87	F45H10.3	Not Sig.	-0.61
F21F8.6	4.12	3.78	F46B6.4	2.25	4.12
F22D6.8	2.27	2.32	F46F11.7	1.57	3.32
F22D6.9	1.53	1.23	F49E8.7	1.47	3.21
F22E12.3	1.91	2.84	F52C6.12	3.29	4.53
F22E5.13	3.43	2.47	F52C6.13	3.69	4.50
F22G12.8	2.30	2.24	F52C6.4	3.17	1.70

F52H2.7	1.70	2.06
F53B2.7	2.11	4.09
F53F4.8	2.88	4.36
F54A3.6	Not Sig.	2.74
F54C8.4	2.02	4.19
F54D10.4	1.75	2.67
F54G2.1	Not Sig.	Not Sig.
F54G2.4	2.97	4.14
F55A3.2	1.10	2.50
F55G11.1	2.59	2.00
F55G11.10	3.18	3.83
F55G11.3	5.53	2.54
F56A8.9	1.94	2.22
F56B3.2	1.93	4.23
F56C11.3	Not Sig.	2.26
F57C12.6	2.03	3.25
F57G9.6	3.10	1.42
F58G1.2	2.42	4.88
F58H1.3	2.03	3.07
F59A7.7	1.87	1.86
F59A7.8	2.31	3.92
F59B2.3	1.68	2.66
F59D6.7	2.09	3.37
F59G1.4	2.60	4.76
F59H6.3	2.57	2.21
fbxa-104	2.11	2.15
fbxa-106	1.82	1.96
fbxa-11	2.91	4.14
fbxa-121	2.20	2.35
fbxa-122	2.19	1.82
fbxa-123	2.04	1.53
fbxa-124	3.10	3.38
fbxa-129	2.83	3.85
fbxa-13	1.95	2.40
fbxa-140	3.25	3.63
fbxa-144	2.28	2.11
fbxa-146	3.25	3.74
fbxa-161	2.76	3.40
fbxa-168	2.32	4.08
fbxa-169	1.92	1.68
fbxa-175	2.42	3.84
fbxa-180	2.51	4.16
fbxa-187	2.63	3.84
fbxa-191	3.23	3.38
fbxa-192	Not Sig.	-0.66
fbxa-193	2.30	4.11
fbxa-198	1.45	1.19
fbxa-20	1.91	2.50
fbxa-209	2.70	2.08
fbxa-211	3.15	3.77
fbxa-212	3.39	3.35
fbxa-217	1.60	2.06
fbxa-33	2.54	2.66
fbxa-34	2.03	2.99
fbxa-4	1.71	2.37
fbxa-57	4.26	2.96
fbxa-59	Not Sig.	-0.74
fbxa-77	3.83	2.75
fbxa-8	6.22	2.86
fbxa-80	2.13	3.90
fbxa-93	2.76	2.77
fbxa-94	2.91	3.90
fbxa-97	2.03	3.42
fbxb-8	2.63	3.10
fbxc-46	3.73	4.13
fbxc-48	1.80	2.70
glb-24	2.21	2.05
glb-5	1.89	2.29
glr-6	2.63	3.58
gly-11	1.59	2.07
gpx-4	3.22	3.03
gyg-3	3.41	2.61
H10D12.2	1.54	4.28
H10D18.6	2.52	4.34
H28G03.2	2.80	3.87
H40L08.2	2.27	3.25
him-14	Not Sig.	0.73
his-63	1.88	2.09
his-73	2.82	2.54
hmit-1.1	Not Sig.	0.98
hpo-10	2.10	4.67
ife-5	2.06	3.64
islo-1	1.47	2.64
JC8.4	2.33	2.89
jmjd-4	1.58	3.25
K02E7.10	2.97	3.05

K03H6.2	Not Sig.	1.93
K05F6.4	2.70	2.48
K06B9.2	3.43	3.67
K06G5.3	2.92	4.84
K07B1.7	1.82	3.34
K07C5.4	Not Sig.	3.84
K07D4.1	2.88	2.56
K08C9.8	3.88	4.16
K08D12.3	1.14	3.14
K08D8.7	1.74	3.22
K08F11.6	3.36	3.54
K09F6.4	1.24	3.09
K09H11.10	2.93	1.99
K09H9.5	Not Sig.	1.31
K10C9.9	4.35	4.27
K10G4.13	3.47	3.64
K10G4.4	1.96	1.84
K11B4.2	Not Sig.	1.63
K11D12.11	2.91	1.77
K11D2.5	1.61	3.21
K11H12.3	3.34	2.48
K12B6.8	1.28	2.34
kel-1	1.70	2.50
klo-1	1.24	3.16
klp-15	1.31	2.50
klp-20	1.35	4.37
lec-7	2.16	2.77
lgc-44	1.70	3.63
lgc-6	2.61	1.73
lin-26	Not Sig.	0.65
lin-56	1.33	1.28
linc-125	2.21	2.67
linc-20	4.14	2.20
linc-29	1.85	1.50
linc-85	1.86	1.78
linc-9	3.71	2.15
lipl-4	1.96	2.62
lipl-5	Not Sig.	Not Sig.
lips-14	4.09	3.24
lips-16	3.71	2.48
lir-1	3.20	4.79
lir-2	3.22	3.78
M02B7.2	2.34	3.14
M03A1.8	2.09	2.43
M03F4.6	1.22	2.34
M04F3.2	2.52	2.23
M116.2	2.71	2.53
M18.6	1.16	2.80
mai-1	2.61	1.68
math-21	1.51	2.98
math-43	3.40	2.25
math-44	Not Sig.	0.93
micu-1	2.04	3.80
mir-238	3.28	1.87
miro-3	1.97	4.21
mltn-2	2.86	5.02
mltn-4	2.03	3.99
mog-1	Not Sig.	Not Sig.
nas-39	1.51	4.00
nhr-104	1.77	2.38
nhr-109	2.30	3.44
nhr-110	1.80	1.56
nhr-133	1.63	3.54
nhr-135	1.17	3.18
nhr-151	1.50	2.09
nhr-194	2.18	2.95
nhr-195	2.76	3.42
nhr-210	1.69	2.93
nhr-220	3.85	3.57
nhr-227	1.91	2.31
nhr-238	1.83	3.55
nhr-240	2.37	1.66
nhr-50	1.78	2.30
nhr-87	2.35	4.23
nmat-1	2.00	2.31
nono-1	1.08	4.15
nos-2	1.78	1.88
oac-42	Not Sig.	1.11
odc-1	4.09	2.34
orc-4	1.30	3.55
osm-1	2.57	3.72
par-1	Not Sig.	Not Sig.
pbo-6	2.61	2.60
phy-2	Not Sig.	Not Sig.
pik-1	1.68	1.88
pme-6	Not Sig.	1.94

polh-1	1.83	3.64
ppfr-1	Not Sig.	Not Sig.
pqbp-1.2	3.28	3.65
pqn-68	1.61	1.32
pqn-70	1.31	3.10
ptr-18	Not Sig.	Not Sig.
R01B10.6	1.13	2.98
R03D7.2	1.96	5.06
R05D3.3	2.02	4.57
R05D3.9	Not Sig.	Not Sig.
R06C1.4	Not Sig.	2.00
R06C7.2	2.51	2.25
R07B1.5	3.15	3.78
R07B1.6	Not Sig.	3.23
R07B1.7	3.17	3.33
R07H5.11	3.36	3.22
R09D1.9	2.94	2.41
R10D12.13	3.22	4.78
R10E8.1	1.92	4.00
R10E8.3	2.22	3.81
R10E8.4	Not Sig.	3.39
R10E8.7	2.59	3.77
R119.1	1.43	2.79
R11G10.3	Not Sig.	Not Sig.
R11H6.6	4.73	4.12
R12C12.8	3.21	3.25
R186.1	2.58	4.06
R74.10	Not Sig.	3.10
rab-39	Not Sig.	Not Sig.
rcor-1	Not Sig.	0.57
rnf-113	2.25	4.07
rnp-7	1.27	4.16
rpl-30	Not Sig.	2.67
rpl-36	Not Sig.	-1.90
rpl-5	Not Sig.	-0.57
rpl-6	Not Sig.	Not Sig.
rpl-7A	1.26	3.86
rpn-6.2	1.95	3.07
rps-22	Not Sig.	2.93
rps-24	Not Sig.	-1.07
rsp-1	1.12	3.92
rsp-2	1.23	3.96
rsp-4	Not Sig.	1.92
rsp-6	1.26	2.76
rsp-7	1.77	4.32
rsp-8	1.37	3.81
rsr-1	1.11	4.78
sams-3	1.79	4.66
scbp-2	3.25	2.93
seu-1	1.94	4.96
smd-1	2.54	2.69
smg-1	1.07	2.43
sna-1	1.09	1.83
snpc-3.3	2.06	3.05
spds-1	1.36	2.38
sra-30	Not Sig.	Not Sig.
srp-5	3.06	3.33
srr-3	2.70	2.87
srz-33	Not Sig.	Not Sig.
stip-1	1.80	3.77
str-216	2.79	2.55
sup-12	1.31	1.66
sym-4	1.55	2.67
T01B11.2	Not Sig.	3.13
T01C8.3	2.33	3.30
T01D1.8	1.84	2.86
T01H8.2	1.30	2.01
T02G5.11	2.30	1.93
T02G5.4	1.89	3.38
T05F1.2	1.38	3.52
T05H10.3	2.22	2.84
T07C12.12	1.64	3.14
T07C4.12	1.19	2.86
T07D10.3	1.73	1.69
T07D10.8	1.90	2.36
T07F10.5	1.79	1.25
T07F10.6	Not Sig.	Not Sig.
T08B2.4	2.35	3.12
T08B2.5	1.24	4.27
T08D2.2	2.18	3.49
T08E11.8	1.72	1.92
T09B4.3	2.18	2.66
T09F3.4	3.09	2.73
T09F5.12	2.26	3.80
T12B5.9	2.73	4.78
T13F2.4	1.72	1.33

T15B7.17	1.60	2.13
T16A9.3	1.39	2.19
T16G1.3	3.28	2.79
T16H12.1	3.76	4.58
T20D4.13	3.63	3.21
T20D4.20	2.96	1.86
T20D4.9	2.98	3.26
T20G5.15	3.22	2.68
T22C8.4	2.36	2.23
T23E1.2	1.83	2.34
T23F11.6	3.12	2.07
T23G7.2	1.28	2.76
T25B2.4	2.86	1.70
T25F10.6	Not Sig.	Not Sig.
T25G12.1	2.14	4.14
T25G12.2	2.71	3.44
T26C11.9	3.12	3.26
T26H5.4	2.38	2.81
T27E7.10	2.96	3.25
T28A8.8	2.53	1.23
tars-1	Not Sig.	-1.00
tdp-1	1.30	3.48
tpa-1	Not Sig.	Not Sig.
tpra-1	Not Sig.	-0.88
trpl-1	1.93	3.44
trpp-10	Not Sig.	2.31
try-10	3.51	2.31
try-5	Not Sig.	0.82
ttr-46	Not Sig.	-1.39
ugt-31	1.58	1.98
ugt-46	Not Sig.	1.90
ugt-60	1.50	2.76
ugt-7	2.18	2.98
use-1	Not Sig.	2.21
W02B3.7	Not Sig.	Not Sig.
W02F12.1	3.49	3.07
W03F9.2	1.16	1.64
W03G11.3	1.55	3.00
W03H9.2	1.66	3.71
W04A8.2	1.89	3.74
W05G11.2	1.89	2.08
W05H12.2	2.00	3.20
W06A11.4	2.31	1.60
W07E6.5	2.22	2.19
W09C5.7	1.24	1.64
W09C5.8	Not Sig.	-0.83
W09G3.1	1.84	1.71
W09G3.7	1.70	4.26
Y102A5C.35	1.80	2.34
Y102A5C.5	2.71	2.01
Y102A5C.6	2.63	2.01
Y105C5B.20	3.21	1.64
Y105E8A.28	3.06	4.10
Y105E8A.32	2.68	3.68
Y106G6D.5	2.09	2.51
Y106G6D.6	1.78	2.51
Y110A7A.7	2.04	1.97
Y111B2A.10	1.47	4.03
Y111B2A.21	1.79	3.43
Y111B2A.25	2.50	5.01
Y111B2A.27	3.04	4.23
Y113G7B.14	2.42	1.76
Y116A8A.10	2.40	2.34
Y11D7A.10	Not Sig.	Not Sig.
Y17G9A.3	2.01	2.58
Y17G9A.91	3.23	3.09
Y17G9B.4	2.61	3.11
Y17G9B.8	1.24	3.19
Y18D10A.22	2.60	2.61
Y20C6A.1	1.62	2.47
Y22D7AL.4	1.65	1.22
Y32F6A.5	Not Sig.	Not Sig.
Y32F6A.6	2.14	2.04
Y32G9A.10	3.73	2.25
Y32G9A.13	3.67	1.89
Y32H12A.6	2.01	4.25
Y34F4.7	2.27	2.40
Y37B11A.3	2.21	2.22
Y37E11B.1	1.33	2.78
Y37H2A.16	1.80	3.50
Y37H2A.18	2.22	2.90
Y37H2B.1	2.78	3.10
Y37H9A.4	3.42	1.92
Y38A10A.11	1.95	2.72
Y38F1A.4	2.15	4.12
Y38H6C.9	2.44	3.00

Y39A1A.3	Not Sig.	2.07
Y39B6A.34	1.49	1.73
Y39B6A.40	1.30	1.25
Y39B6A.43	1.37	2.92
Y39B6A.49	3.27	2.73
Y39C12A.1	Not Sig.	1.88
Y39C12A.2	2.15	2.52
Y39E4B.14	1.95	3.61
Y39G10AR.9	1.22	1.99
Y40A1A.3	3.04	3.72
Y40C5A.4	1.49	2.89
Y41E3.11	1.37	4.45
Y43F11A.6	2.15	4.83
Y43F8B.23	2.46	2.96
Y45F10D.10	2.03	2.09
Y46G5A.43	2.03	3.25
Y47G6A.21	Not Sig.	Not Sig.
Y47H10A.6	1.42	2.47
Y48B6A.16	2.04	1.26
Y48C3A.5	2.49	3.65
Y48G1BL.4	2.87	3.43
Y48G1BM.6	1.58	2.47
Y49A10A.1	2.91	4.04
Y51F10.2	1.49	3.35
Y53C10A.1	2.69	2.87
Y54G11A.7	1.96	3.32
Y54G2A.50	1.24	3.32
Y55F3AM.21	2.86	1.83
Y55F3BR.10	2.73	2.12
Y55F3C.13	3.05	4.07
Y57A10A.1	2.14	4.21
Y57A10A.6	2.00	4.04
Y57A10A.7	1.89	4.47
Y57G7A.8	1.96	2.87
Y5H2B.1	2.06	2.06
Y61A9LA.3	1.43	2.26
Y66D12A.10	Not Sig.	3.50
Y67D2.4	1.19	2.39
Y67H2A.9	2.27	3.69
Y69A2AR.12	2.15	3.06
Y6G8.2	1.75	2.32
Y71A12B.14	1.54	2.90
Y71A12B.19	2.01	3.01
Y71F9AL.4	1.22	2.36
Y71H2AM.14	1.37	1.71
Y71H2B.11	2.30	2.77
Y73B3A.18	5.14	4.08
Y73B3A.20	3.41	3.16
Y73B3A.21	1.64	3.62
Y73F8A.32	3.32	4.92
Y73F8A.33	Not Sig.	3.65
Y73F8A.37	3.02	4.25
Y75B12A.1	2.11	2.19
Y7A9C.1	2.03	3.84
Y80D3A.11	2.68	2.43
Y80D3A.9	1.35	2.64
Y80D4G.1	1.89	2.76
Y82E9BL.19	Not Sig.	Not Sig.
Y82E9BL.9	3.72	2.60
Y82E9BR.25	1.92	2.36
yars-1	2.36	4.31
zak-1	1.36	3.48
ZC196.1	2.22	2.87
ZC196.5	1.93	1.30
ZC239.6	1.85	1.70
ZC262.2	1.18	2.83
ZC262.4	1.63	2.76
ZC328.5	3.74	2.81
zfp-3	Not Sig.	2.12
zip-12	2.02	3.54
ZK1025.1	1.72	2.88
ZK1055.5	2.52	2.40
ZK1055.6	1.79	3.93
ZK1127.12	3.06	3.17
ZK1128.1	1.50	4.03
ZK1225.7	2.71	2.87
ZK1240.2	1.45	2.88
ZK1240.9	3.40	1.97
ZK180.3	1.16	2.11
ZK185.1	1.43	3.17
ZK185.5	1.39	3.53
ZK228.4	1.88	3.66
ZK6.11	Not Sig.	1.39
ZK637.6	1.78	2.79
ZK662.8	2.09	3.23
ZK816.4	3.02	3.20

ZK896.4	1.74	2.68

* “Not Sig.” indicates the gene was identified only in the intron-level analysis

Appendix 2: Class II RNAs from Gene-Level edgeR Analysis

2L52.1	asd-1	B0554.7	C04B4.1	C09E7.7
4R79.2	asd-2	bas-1	C04E12.10	C09G5.13
aars-1	ash-2	basl-2	C04E12.2	C09H10.10
abf-1	ast-1	bath-22	C04E12.6	C10A4.1
abt-6	atf-8	bath-27	C04F12.6	C10A4.2
abtm-1	attf-6	bath-36	C04F5.9	C10E2.2
abts-1	B0001.3	bath-46	C05C8.5	C11E4.6
acdh-4	B0001.7	bath-47	C05C8.8	C11E4.7
acdh-6	B0019.2	bbs-1	C05D11.13	C11H1.9
acin-1	B0024.10	BE0003N10.3	C05D12.1	C12D5.9
acl-6	B0035.12	ben-1	C05D12.4	C13A2.3
acl-7	B0041.1	best-16	C05D9.4	C14A11.5
acox-1	B0041.11	best-18	C05G5.1	C14B1.7
acp-1	B0205.1	best-19	C06A1.12	C14C11.4
acp-5	B0212.3	bir-2	C06A1.4	C15A11.2
acr-11	B0218.5	bpl-1	C06A8.2	C15A11.7
acr-20	B0222.11	brc-1	C06C3.10	C15B12.4
acs-12	B0238.13	brd-1	C06C3.12	C15B12.8
acs-21	B0250.3	btb-13	C06C3.5	C15C6.3
acs-5	B0273.3	btb-17	C06C3.7	C15C7.6
aex-5	B0280.11	btb-18	C06C3.8	C15F1.5
afmd-1	B0284.2	btb-9	C06C3.9	C15H7.4
age-1	B0285.11	bus-18	C06E1.11	C16B8.2
agl-1	B0285.6	bus-19	C06G3.9	C16C8.14
agt-1	B0286.1	C01A2.1	C07A12.7	C16C8.22
AH9.1	B0286.3	C01B10.4	C07A4.2	C17B7.5
ahr-1	B0294.3	C01B9.1	C07B5.4	C17C3.1
alh-10	B0303.11	C01F6.14	C07D8.5	C17E4.6
alh-6	B0334.10	C01G5.4	C07E3.8	C17H1.1
alkb-8	B0334.5	C01G5.9	C07F11.2	C18A3.2
aly-2	B0353.1	C01G6.9	C07G3.8	C18B12.6
aly-3	B0365.2	C01H6.2	C07H6.4	C18D11.3
aman-2	B0391.8	C01H6.4	C08A9.8	C18D4.8
amt-1	B0393.5	C01H6.8	C08B6.4	C18H2.2
anoh-2	B0412.3	C02F12.5	C08F11.14	C18H2.3
anr-35	B0454.9	C02F4.4	C08F8.3	C18H9.5
ape-1	B0457.2	C02F5.7	C08G5.1	C23G10.8
apt-9	B0457.3	C03A3.3	C08H9.3	C23H4.3
arl-13	B0462.1	C03A7.12	C08H9.9	C23H5.15
arrd-11	B0507.1	C03A7.13	C09B9.7	C24F3.2
arrd-12	B0507.8	C03C10.7	C09E7.10	C24H10.3
arrd-19	B0546.3	C03E10.3	C09E7.5	C24H10.4
arrd-20	B0554.4	C03H5.5	C09E7.6	C25A1.1

C25D7.14
C25D7.2
C25F9.2
C25G4.3
C25G6.3
C25H3.1
C26C6.8
C26D10.4
C27A12.9
C27B7.7
C27D6.11
C27F2.4
C27F2.9
C27H5.2
C27H6.3
C29A12.6
C29F5.1
C29F7.1
C30F12.2
C30F12.5
C30F2.2
C30F8.3
C31C9.6
C31G12.4
C31H1.1
C31H1.2
C32B5.13
C32H11.3
C33B4.2
C33D9.9
C33E10.3
C34B2.4
C34B4.2
C34C6.10
C34C6.7
C34F11.3
C34H4.5
C35A5.4
C35B8.3
C35D10.5
C36A4.14
C36B7.4
C36B7.5
C36C9.5
C36E6.1
C37A5.7
C37C3.1
C38C3.7
C39B10.3
C39B10.4
C39D10.7
C39D10.8
C39F7.1
C40H1.8
C41A3.2
C41H7.5
C42C1.2
C42D4.13
C44B12.3
C44B12.9
C44H4.4
C44H4.6
C45B2.6
C45G7.13
C45G9.15
C45H4.13
C46A5.1
C46C11.4
C46F11.4
C46H3.2
C47E8.3
C47F8.5
C47G2.8
C48D1.7
C49A9.1
C49C3.4
C49G7.1
C49G9.1
C49H3.4
C50D2.6
C50D2.8
C50D2.9
C50F4.10
C50H2.13
C50H2.7
C52A10.2
C52D10.1
C52D10.12
C52E2.4
C52E2.5
C53D5.5
C53D6.7
C54F6.16
C54G4.4
C54G4.9
C55A6.12
C55C3.3
C55C3.7
C56A3.8
C56C10.11
C56E10.3
C56G2.15
cacn-1
cash-1
cbs-1
ccb-2
ccpp-6
cct-5
CD4.5
cdc-6
cdf-1
cdh-8
cdt-2
CE7X_3.1
ceeh-2
ceh-10
ceh-18
ceh-2
ceh-24
ceh-27
ceh-34
ceh-6
cft-1
cgt-2
chd-1
che-11
chil-12
chil-24
chk-1
chl-1
cht-1
cin-4
cku-70
cky-1
clec-107
clec-113
clec-118
clec-12
clec-126
clec-139
clec-155
clec-164
clec-170
clec-180
clec-184
clec-185
clec-198
clec-201
clec-205
clec-217
clec-220
clec-233
clec-236
clec-243
clec-245
clec-246
clec-26
clec-264
clec-266
clec-30
clec-32
clec-36
clec-43
clec-44
clec-49
clec-62
clec-70
clec-77
clec-78
clec-8
clec-85
clec-91
clec-99
clh-4
clh-6
clk-2
clp-4
clpf-1
cng-1
coel-1
cogc-1
cogc-2
cogc-8
coh-4
col-118
col-128
col-150
col-151
col-172
col-47
col-58
col-61
col-69
col-75
col-78
col-83
col-87
comt-1
copb-1
coq-2
coq-8
cor-1
cosa-1
cox-11
cpf-2
cpg-3
cpsf-1
cpt-2
cra-1
crb-1
csr-1
cTel7X.1
ctg-1
cth-1
cutl-10
cutl-18
cutl-2
cutl-20
cutl-27
cutl-8
cux-7
cya-2
cyk-7
cyp-13A3
cyp-14A2
cyp-14A4
cyp-23A1
cyp-25A4
cyp-25A5
cyp-29A2

cyp-31A1	dsh-1	exos-9	F15E6.9	F27B3.7
cyp-33C12	dsl-1	exp-1	F15H9.1	F27C8.2
cyp-33C3	dsl-3	eya-1	F16B12.1	F28A10.9
cyp-34A3	duo-2	F01D4.3	F16B12.6	F28C1.1
cyp-35A2	duxl-1	F01D5.7	F16C3.2	F28C6.4
cyp-43A1	dyf-17	F01E11.17	F16C3.3	F28E10.16
cyp-44A1	dyf-3	F01G10.10	F16H6.5	F28F8.7
D2030.12	E_BE45912.2	F01G10.4	F16H6.6	F28H6.6
D2030.7	E01G4.5	F07C4.12	F16H6.7	F29C4.4
D2045.9	E01G6.3	F07E5.7	F17A9.2	F29G6.1
D2096.1	E02D9.1	F08D12.12	F17A9.7	F29G6.2
D2096.10	E03H4.8	F08F1.4	F17E9.14	F30B5.4
D2096.11	E04D5.4	F08F1.9	F18A1.1	F30B5.7
daao-1	E04F6.2	F08F8.7	F18E9.4	F31A3.5
dab-1	eat-4	F08G2.7	F19B10.10	F31C3.6
daf-10	ech-8	F08G2.8	F19B6.1	F31E8.6
dap-3	EEED8.14	F09E5.10	F19B6.3	F31F4.1
dars-2	EEED8.16	F09E5.8	F19F10.11	F31F7.2
daz-1	eef-1G	F09E5.9	F19F10.9	F32B4.1
DC2.5	efhd-1	F09F7.5	F19H6.6	F32D8.1
ddb-1	EGAP9.3	F09F7.7	F20H11.1	F32D8.12
del-1	egas-1	F10A3.4	F22B3.10	F32D8.8
del-5	egas-4	F10D7.5	F22B8.7	F32E10.9
denn-4	egl-15	F10E7.11	F22F1.2	F32H2.12
dgk-3	egl-17	F10E9.5	F22F4.1	F33A8.10
dhc-3	ekl-5	F11A5.16	F22G12.5	F33A8.6
dhod-1	elp-1	F11C1.4	F23A7.5	F35D2.2
dhp-1	elpc-1	F11C3.1	F23C8.1	F35E2.10
dhrs-4	elpc-3	F11D5.5	F23C8.7	F35E2.3
dhs-26	elt-7	F11D5.6	F23F1.4	F35F11.1
dhs-27	emb-30	F11E6.4	F23H11.2	F35H12.5
dhs-8	emb-4	F11F1.1	F25A2.1	F35H12.6
dic-1	eng-1	F12A10.8	F25B4.5	F36A2.3
div-1	enu-3	F12B6.2	F25B5.6	F36H12.2
dlc-5	eor-2	F13E9.11	F25E2.2	F36H5.10
dnj-10	ercc-1	F13E9.5	F25E5.4	F36H5.13
dnj-17	eri-3	F13E9.9	F25F8.1	F37C4.6
dnj-25	eri-6	F13G3.3	F25G6.1	F37H8.2
dod-6	erl-1	F13H6.3	F25H2.12	F37H8.6
dpf-4	ess-2	F13H8.6	F25H2.14	F38A1.6
dpt-1	etr-1	F14B4.1	F26A1.4	F38A6.4
dpy-19	ets-5	F14D12.1	F26A1.9	F38B2.3
dpy-20	evl-18	F14D2.2	F26A3.7	F38B6.4
dpy-27	exc-7	F14D7.1	F26B1.2	F39B2.7
dpy-7	exos-1	F14H8.2	F26D11.13	F39H12.2
dre-1	exos-2	F15E6.6	F26F4.9	F40E12.2

F40F11.4
F40F4.7
F40H6.5
F41C3.7
F41C6.4
F41D3.6
F41D9.2
F41E7.2
F41E7.9
F41G4.4
F41G4.8
F41H10.3
F42C5.3
F42G4.6
F42H10.6
F42H11.1
F43C11.5
F43C11.7
F43D2.2
F43E2.6
F44B9.9
F44E2.7
F44E7.7
F44G4.7
F45E10.2
F45E4.11
F46B3.22
F46B6.6
F46C5.2
F46F11.8
F46F2.1
F46F5.4
F47B10.9
F47B7.6
F47F6.4
F47G3.1
F48E3.8
F48E3.9
F48E8.2
F48F5.2
F48G7.4
F49B2.4
F49D11.6
F49E10.4
F49E2.2
F49E8.6
F52A8.3
F52B10.3
F52F10.2
F52G2.3
F53A10.2
F53A2.7
F53B1.6
F53B3.3
F53C3.8
F53E10.1
F53F1.3
F53F10.1
F53F8.7
F53G12.13
F53G12.4
F53G12.9
F53G2.2
F53G2.8
F53H1.2
F53H2.1
F54A3.6
F54B11.4
F54D10.3
F54D10.8
F54D11.3
F54D5.11
F54D5.2
F54D5.7
F54D8.6
F54E7.6
F54F11.1
F54F12.1
F54G2.2
F54H12.2
F55A11.4
F55A3.7
F55A4.3
F55A4.4
F55C12.6
F55C5.6
F55D12.2
F55D12.5
F55F3.2
F55G1.1
F55G7.1
F56B6.6
F56C11.3
F56C11.6
F56D1.2
F56D2.2
F56D2.5
F56D5.2
F56G4.6
F56G4.7
F57B9.1
F57B9.8
F57C9.4
F57C9.7
F58A4.12
F58B6.1
F58D5.6
F58D5.9
F58E10.3
F58E2.3
F58E2.4
F58F9.10
F58G11.2
F58H1.2
F59A1.11
F59A7.5
F59B2.9
F59E12.15
F59E12.6
F59F5.7
faah-4
farl-11
fars-2
fbf-2
fbxa-110
fbxa-117
fbxa-137
fbxa-141
fbxa-149
fbxa-154
fbxa-158
fbxa-189
fbxa-195
fbxa-196
fbxa-201
fbxa-203
fbxa-207
fbxa-21
fbxa-210
fbxa-218
fbxa-5
fbxa-50
fbxa-61
fbxa-64
fbxa-69
fbxa-85
fbxb-7
fbxb-80
fbxc-12
fbxc-19
fbxc-33
fbxc-41
fbxc-55
fbxc-57
fcd-2
fib-1
fkh-6
flad-1
flp-6
flr-4
fmo-3
fmo-4
fncm-1
fog-1
fos-1
fox-1
fre-1
frpr-10
frpr-13
frpr-19
fust-1
gasr-8
gba-2
gcp-2.1
gcp-2.3
gcst-1
gcy-11
gcy-18
gcy-21
gcy-23
gcy-31
gcy-33
gcy-34
gcy-35
gcy-36
gcy-6
gcy-9
gip-1
glb-14
glb-2
glb-4
gld-2
glr-2
glr-3
glr-8
gly-1
gly-12
gnrr-2
got-2.1
gpc-1
grd-2
grd-7
grl-14
grl-2
grld-1
gsnl-1
gst-19
gst-26
gst-43
gst-44
gsto-1
gtl-2
H01G02.4
H01M10.2
H03G16.3
H04D03.2
H04D03.3
H04J21.1
H06H21.8
H06I04.5
H06I04.9
H16D19.5
H21P03.2
H31G24.3
H39E20.1
H40L08.3
haao-1
haf-6
haly-1
hecw-1

hel-1
herc-1
hgo-1
hhat-1
him-17
him-18
him-19
hlh-1
hlh-11
hmgr-1
hmgs-1
hog-1
hot-3
hot-4
hot-7
hot-9
hpl-1
hpo-22
hpo-26
hpo-32
hpo-36
hpo-38
hpo-40
hpo-41
hpo-42
hrp-1
hrp-2
hrpf-1
hrpf-2
hsd-3
hsp-70
hus-1
icln-1
idhg-2
ifa-3
ifc-1
ift-81
ifta-2
iglr-3
ikb-1
ilys-3
ilys-4
imb-3
irld-57
irld-6
jbts-14
jmjd-3.3
K01A11.1
K01A6.5
K01B6.4
K01C8.2
K01D12.3
K01G5.3
K01G5.9
K02A11.2
K02A6.3
K02C4.5
K02D10.1
K02D10.2
K02E11.9
K03B4.9
K03H6.2
K04A8.10
K04A8.3
K04F1.1
K04F10.1
K04F10.2
K04G2.7
K05C4.9
K05F1.5
K05G3.1
K06A5.1
K06B9.4
K07A1.9
K07C5.10
K07C5.4
K07E8.6
K07F5.8
K07H8.5
K07H8.9
K08B12.3
K08C7.1
K08C9.7
K08D10.9
K08D9.2
K08E3.4
K08E5.1
K08E7.5
K08F8.5
K08F9.3
K09A9.8
K09B11.5
K09C4.4
K09C6.8
K09D9.12
K09E2.2
K09E2.3
K09E4.3
K09F6.10
K09F6.13
K09H9.7
K10B3.6
K10C3.4
K10D11.2
K10D3.10
K10G4.5
K10H10.10
K10H10.5
K11B4.2
K11D12.9
K11D2.1
K11E4.1
K11E4.2
K11G12.6
K11G9.2
K11G9.5
K11H12.8
K11H3.4
K12C11.7
K12D12.4
K12H6.12
kap-1
kat-1
kel-8
kgb-2
klo-2
klp-10
klp-13
klp-6
klp-8
kqt-2
kqt-3
kri-1
kup-1
kvs-4
lab-1
lact-3
lact-4
lact-5
lact-8
lact-9
lev-9
lgc-11
lgc-12
lgc-15
lgc-26
lgc-31
lgc-43
lgc-45
lge-1
lig-1
lim-4
lim-6
lin-11
linc-101
linc-91
lips-1
lnp-1
lpr-2
lpr-7
lron-7
lrx-1
lst-1
lst-3
lst-4
lst-5
M01A10.1
M01A12.4
M01E5.1
M01E5.2
M01F1.4
M01F1.8
M01H9.2
M02D8.5
M02F4.1
M03A8.3
M03C11.3
M03E7.2
M04B2.2
M04B2.7
M04G7.2
M05D6.6
M116.5
M117.6
M142.3
M153.5
M162.12
M162.5
M176.9
M195.2
M60.2
madf-11
madf-8
maea-1
mam-5
mam-7
mam-8
math-26
math-27
math-3
math-35
math-39
math-46
mboa-2
mbtr-1
mcp-1
mcrs-1
mcu-1
mdt-1.2
mdt-26
mdt-28
mec-1
mec-10
mec-17
mecr-1
mel-46
memb-2
mes-2
metr-1
mett-10
mgl-1
mib-1
mig-14
mig-17
mig-22
mig-38
miro-2
mks-1
mks-3
mksr-2

mlh-1
mlt-4
mlt-9
mltn-1
mltn-3
mltn-8
mmaa-1
moc-1
moc-2
moc-3
mocs-1
mog-3
mps-3
mrps-18C
msh-4
msh-5
msp-42
msp-74
mtd-1
mth-2
mthf-1
mtk-1
mtrr-1
mut-15
nac-3
nars-1
nars-2
nas-29
nas-38
nas-4
nas-6
nas-9
nasp-1
ncs-5
nep-12
nep-16
nep-18
nex-3
nft-1
nhr-103
nhr-112
nhr-114
nhr-127
nhr-128
nhr-130
nhr-142
nhr-146
nhr-147
nhr-15
nhr-156
nhr-158
nhr-179
nhr-18
nhr-184
nhr-190
nhr-193
nhr-206
nhr-212
nhr-219
nhr-228
nhr-231
nhr-232
nhr-234
nhr-237
nhr-239
nhr-243
nhr-256
nhr-284
nhr-5
nhr-57
nhr-58
nhr-59
nhr-62
nhr-64
nhr-65
nhr-73
nhr-78
nhr-79
nhr-8
nhr-80
nhr-83
nhr-84
nhr-90
nhr-92
nhr-93
nhr-98
nipi-4
nlr-1
nmr-1
nmr-2
noca-1
nop-1
nos-3
npp-16
npp-8
npr-30
npr-8
nrde-4
nth-1
nuaf-3
nxf-1
oac-12
oac-13
oac-18
oac-22
oac-23
oac-33
oac-42
oac-47
obr-1
ocr-3
odr-10
orai-1
osm-5
pah-1
pam-1
par-4
pars-2
pat-12
pax-2
pax-3
pcf-11
pcs-1
PDB1.1
pdcd-2
pde-12
pdr-1
pes-4
pfs-2
pgl-2
pgl-3
pgp-12
pgp-15
pha-1
phf-14
phf-15
pho-11
phy-3
phy-4
pif-1
piga-1
pin-2
pitr-1
pitr-3
plc-2
plc-4
pme-3
pme-6
pms-2
pnc-2
polg-1
ppw-1
pqn-11
pqn-18
pqn-42
pqn-53
prg-2
pri-1
prl-1
prmt-5
prmt-7
prom-1
prp-6
prpf-4
psr-1
ptd-2
ptr-12
ptr-15
ptr-4
pxn-1
R01H10.4
R01H10.5
R02E12.4
R02F11.2
R03D7.3
R03H10.4
R04E5.9
R05A10.5
R05A10.7
R05D3.12
R05D3.8
R05G6.5
R05G9R.1
R05H10.1
R06A10.1
R06B10.1
R06B10.2
R06C1.4
R07A4.4
R07B1.6
R07B1.8
R07B7.9
R07D5.2
R08E3.3
R08E5.1
R09A1.2
R09D1.12
R09E10.9
R09F10.5
R102.7
R10E8.4
R10E8.6
R119.5
R11D1.1
R11D1.10
R11F4.3
R11G1.2
R12B2.2
R12B2.3
R12E2.8
R13A1.10
R13D7.2
R151.4
R155.3
R155.4
R74.10
rab-11.2
rab-33
rad-51
ran-2
rde-10
rfc-1
rgs-10
rhgf-2
rig-6
rilp-1
rmd-4
rnh-1.0
rnp-1
rnp-6

rol-3
rop-1
rpa-1
rpl-1
rpl-12
rpl-3
rpl-30
rps-22
rrf-2
rsp-4
rsp-5
sacy-1
sams-2
sams-4
sams-5
sao-1
sas-4
sas-6
sav-1
scav-3
scl-16
scl-5
scm-1
sdha-2
sdz-3
sec-12
secs-1
selb-1
set-20
set-23
set-31
set-5
sfa-1
sft-1
sfxn-5
sgca-1
shc-1
slcf-2
slo-2
slx-1
sma-2
smf-2
smg-3
smg-8
smgl-2
smu-2
snf-4
snf-6
snf-7
snf-8
snt-7
soc-1
sodh-2
sox-2
spd-3
spe-39
spe-44
spe-5
spe-9
sphk-1
sprr-2
spsb-2
sptf-2
sptl-2
srd-7
sre-4
sre-43
srh-2
srh-48
sri-16
sri-30
sri-46
srj-14
srj-23
srp-1
srp-3
srp-9
srt-20
srt-39
srw-85
stdh-4
stn-1
stn-2
str-259
such-1
sulp-3
sulp-8
sup-10
sup-9
svh-1
swah-1
swp-1
sygl-1
syx-16
syx-4
T01B11.2
T01B7.5
T01G6.10
T02C12.2
T02G6.10
T02H6.1
T02H6.9
T03E6.8
T03G11.3
T04A6.5
T04A8.7
T04C9.11
T05A7.11
T05A7.3
T05A8.6
T05B11.7
T05B4.4
T05C3.6
T05D4.5
T05E11.7
T05E7.1
T05F1.11
T05F1.8
T05H4.10
T05H4.15
T05H4.3
T05H4.4
T05H4.7
T06A4.1
T06A4.3
T06G6.8
T07A5.1
T07D3.9
T07E3.2
T07G12.2
T07G12.3
T08A9.6
T08D2.5
T08G11.3
T08G3.4
T09B9.3
T09E11.11
T09F3.5
T10B11.2
T10B11.7
T10B5.2
T10E9.4
T10G3.2
T10H4.13
T11A5.6
T11G6.5
T12A2.1
T12F5.2
T13A10.1
T13C5.9
T13F3.7
T14D7.1
T14F9.2
T14G8.2
T15H9.4
T16G12.6
T19D12.3
T19D7.7
T20B12.1
T20D4.8
T20F10.5
T21C9.6
T21D12.11
T21D12.7
T21E8.7
T21H3.5
T22C1.5
T22F3.10
T22F7.5
T23B12.8
T23E1.1
T23F1.5
T23F4.5
T23G4.127
T23G4.2
T23G5.3
T23H2.3
T24C4.3
T25B2.1
T25B2.3
T25D10.4
T25G12.6
T26A8.2
T26E4.4
T26E4.5
T26F2.2
T26G10.7
T26H2.10
T27A1.2
T27A3.7
T27C5.12
T27E4.1
T28A8.5
T28B4.2
T28F3.5
T28F4.1
taf-11.3
taf-6.1
tag-244
tag-250
tag-253
tag-257
tag-30
tag-65
tba-7
tbc-16
tctn-1
teg-1
tep-1
tftc-1
tiar-1
tiar-3
toca-2
tos-1
trf-1
trp-2
try-3
try-6
tsp-8
ttr-12
ttr-3
ttr-49
ttx-1
twk-22
twk-25
twk-45
twk-48
twk-49
ubc-19
ubl-1

ugt-10	W04E12.7	Y11D7A.1	Y39A3A.2	Y47D3A.5
ugt-27	W04E12.9	Y13C8A.4	Y39A3A.3	Y47D9A.1
ugt-37	W04G3.17	Y17D7C.3	Y39A3CL.7	Y47D9A.3
ugt-46	W04G5.5	Y17G7B.19	Y39B6A.18	Y47G6A.25
ugt-50	W04G5.8	Y17G7B.22	Y39B6A.29	Y47G6A.3
ugt-52	W05F2.7	Y17G9B.5	Y39B6A.31	Y47G6A.30
ugt-56	W05G11.4	Y18H1A.8	Y39B6A.33	Y47H10A.4
ugt-65	W05H7.2	Y18H1A.9	Y39B6A.37	Y47H9C.14
unc-25	W06B4.2	Y22D7AL.11	Y39B6A.42	Y48A6B.8
unc-3	W06F12.2	Y22D7AL.6	Y39B6A.9	Y48A6C.1
unc-30	W06H8.2	Y22D7AR.12	Y39C12A.1	Y48B6A.1
unc-33	W09B6.4	Y23H5A.2	Y39E4B.2	Y48C3A.14
unc-42	W09D6.1	Y23H5A.8	Y39G10AR.16	Y48C3A.18
unc-62	W09G12.7	Y23H5B.2	Y39G10AR.21	Y48E1B.3
unc-86	W09G12.8	Y23H5B.3	Y39G8B.5	Y48E1C.1
unc-94	W09H1.3	Y24D9B.1	Y40B10A.4	Y48G1A.1
use-1	W09H1.4	Y25C1A.8	Y40D12A.1	Y48G1BL.7
uso-1	W10G11.17	Y27F2A.6	Y40H7A.7	Y48G8AL.10
usp-5	W10G11.19	Y32B12C.3	Y41D4A.1	Y49A10A.2
vab-23	wago-11	Y32B12C.5	Y41D4A.6	Y49A3A.4
vab-3	wago-2	Y32G9A.11	Y41D4B.15	Y49E10.4
vap-1	wago-5	Y32G9A.3	Y41D4B.17	Y49E10.7
vars-1	wht-7	Y32G9B.1	Y41D4B.18	Y49F6B.9
VB0395L.1	wip-1	Y34B4A.4	Y41E3.3	Y4C6B.4
ver-1	wrt-8	Y34B4A.8	Y41E3.5	Y50D4C.2
vha-1	xbx-1	Y34D9A.3	Y41G9A.10	Y50D7A.12
vha-11	Y102A5C.2	Y37B11A.2	Y43B11AR.3	Y50D7A.13
vms-1	Y104H12D.2	Y37D8A.2	Y43D4A.5	Y51A2D.18
vps-33.1	Y104H12D.4	Y37D8A.6	Y43E12A.3	Y51A2D.7
VY35H6BL.1	Y105C5A.15	Y37E11B.5	Y43F4B.9	Y51B9A.10
VY35H6BL.2	Y105C5B.25	Y37F4.6	Y43F8A.3	Y51H1A.2
W01A11.7	Y105C5B.7	Y37H2A.7	Y43F8B.15	Y51H4A.1
W01A8.7	Y105C5B.8	Y37H9A.1	Y43F8B.9	Y51H4A.24
W01B6.3	Y105E8A.1	Y37H9A.2	Y43F8C.17	Y51H7C.10
W01B6.4	Y105E8A.13	Y38A10A.2	Y45F10B.8	Y52D5A.1
W02A2.4	Y105E8A.20	Y38F1A.6	Y45F10B.9	Y52E8A.1
W02B12.10	Y106G6A.1	Y38F2AL.12	Y45G12B.3	Y53C10A.2
W02B12.11	Y106G6G.1	Y38F2AL.6	Y45G12C.1	Y53F4B.17
W02D7.6	Y108G3AL.3	Y38F2AR.12	Y45G5AM.2	Y53F4B.18
W02H3.2	Y110A2AL.2	Y38H6C.13	Y45G5AM.9	Y53F4B.21
W03C9.6	Y111B2A.24	Y38H6C.17	Y46E12BL.2	Y53F4B.27
W03D8.2	Y116A8C.10	Y38H6C.20	Y46G5A.18	Y53F4B.38
W03F9.3	Y116A8C.25	Y38H8A.5	Y46G5A.29	Y53F4B.42
W04C9.5	Y116A8C.40	Y38H8A.8	Y46G5A.6	Y53F4B.6
W04D2.4	Y116F11B.10	Y39A1A.18	Y47D3A.14	Y53G8AR.8
W04D2.6	Y116F11B.2	Y39A1A.3	Y47D3A.28	Y53G8B.3

Y54E5A.6
Y54E5A.8
Y54E5B.2
Y54F10BM.1
2
Y54G11A.14
Y54G11A.3
Y54G11A.4
Y54G2A.17
Y54G2A.24
Y54G2A.26
Y54G2A.32
Y54G2A.36
Y54G2A.37
Y54G2A.40
Y54G2A.49
Y54G2A.73
Y54H5A.2
Y55B1BL.1
Y55F3AM.13
Y55F3AM.3
Y55F3BR.2
Y55F3BR.6
Y56A3A.11
Y56A3A.2
Y57A10A.10
Y57A10A.2
Y57A10A.24
Y57A10A.4
Y57E12AL.3
Y57G11B.1
Y57G11C.22
Y57G11C.28
Y57G11C.36
Y57G11C.38
Y57G11C.5
Y57G11C.9
Y59A8B.8
Y59E9AL.5
Y61A9LA.4
Y61A9LA.5
Y62E10A.17
Y62E10A.19
Y62H9A.13
Y65B4A.10
Y65B4BL.7
Y66D12A.10
Y66D12A.15
Y66D12A.19
Y67D2.6
Y67D8B.4
Y67H2A.2
Y67H2A.7
Y69A2AR.14
Y69A2AR.16
Y69A2AR.32
Y69A2AR.8
Y69H2.10
Y69H2.7
Y71A12B.10
Y71A12B.13
Y71A12B.18
Y71D11A.3
Y71F9AL.8
Y71F9B.9
Y71G10AL.1
Y71G12A.2
Y71G12B.10
Y71G12B.38
Y71H2AM.10
Y71H2AM.15
Y71H2AM.24
Y71H2AM.25
Y71H2B.5
Y73B3A.1
Y73B3A.10
Y73B3A.13
Y73B3A.2
Y73B3A.3
Y73B3A.9
Y73B3A.t1
Y73B3B.5
Y73B6BL.29
Y73E7A.5
Y73F8A.26
Y73F8A.33
Y73F8A.35
Y75B8A.14
Y75B8A.31
Y75B8A.32
Y75B8A.33
Y77E11A.7
Y79H2A.2
Y79H2A.3
Y7A5A.11
Y81G3A.4
Y82E9BL.12
Y82E9BR.18
Y82E9BR.2
Y82E9BR.20
Y8A9A.3
Y92H12BL.1
Y94H6A.3
Y95B8A.6
Y95D11A.2
Y97E10AL.2
ZC123.4
ZC13.2
ZC239.1
ZC247.2
ZC416.6
ZC443.1
ZC449.1
ZC449.2
ZC449.5
ZC47.8
ZC477.5
ZC513.1
ZC513.2
ZC8.6
zfp-3
zhit-1
zhp-3
zim-1
zip-11
zip-3
zip-4
ZK1010.10
ZK1010.2
ZK1025.4
ZK1058.3
ZK1098.1
ZK1098.2
ZK1127.4
ZK1127.6
ZK1193.2
ZK1225.1
ZK1225.2
ZK1236.1
ZK1240.11
ZK1290.10
ZK1320.14
ZK177.4
ZK185.3
ZK185.4
ZK250.13
ZK287.7
ZK381.2
ZK418.13
ZK418.8
ZK418.9
ZK484.6
ZK54.3
ZK546.2
ZK546.5
ZK546.7
ZK550.7
ZK593.9
ZK616.1
ZK632.5
ZK637.15
ZK669.2
ZK686.6
ZK742.2
ZK783.3
ZK783.5
ZK783.6
ZK792.1
ZK822.1
ZK822.6
ZK856.11
ZK858.10
ZK899.3
ZK945.4
ZK993.2
ztf-15
ztf-4
zwl-1
zyg-1

Appendix 3: Genes Indirectly Affected by NMD

Class III mRNAs

linc-15	T24C4.5	Y105C5B.15	T06D10.3	ZK105.1
linc-64	ugt-64	T16G1.5	F40H3.3	F39F10.4
mir-4938	cpi-1	Y75B8A.28	nstp-9	col-36
Y66H1A.8	clec-66	fbxa-1	T08E11.1	F33H12.7
Y50D4A.5	C27C7.1	lys-3	oac-32	K08B4.7
T07G12.8	fat-5	C49H3.12	K08D8.3	C39B5.10
tsp-2	ttr-25	hpo-15	Y6G8.5	clec-4
tsp-1	lys-2	T19C9.8	B0024.4	M162.9
Y57A10A.3	math-14	F44G3.10	F53B2.8	bath-12
F52C6.3	H11L12.1	btb-16	M02E1.3	Y73F8A.1173
clc-1	M02H5.8	T24C4.4	Y39A3B.3	fbxa-70
smg-5	ZK678.3	F16G10.15	K07H8.11	C08F11.6
W01A8.8	K06G5.1	asp-17	F37C12.21	Y39B6A.21

Class IV mRNAs

D1086.9	F12A10.7	C50F7.5	ftn-1	cnc-4
F53A9.6	F48C1.9	thn-1	tre-5	Y43C5A.3
fip-1	cnc-11	F14F9.3	Y51A2D.13	T05E7.4
cnc-6	nhx-6	R08E5.3	F53C3.4	M163.1
F48C1.11	R11A5.3	comt-4	C13G3.1	ttr-17
Y39A3B.7	hsp-12.6	F53C3.6	F59C12.4	R06C1.6
C18D11.6	fipr-26	tyr-5	F11E6.11	Y57A10A.26
numr-1	F32G8.3	B0205.13	F19B10.13	F22B5.4
ZK970.7	F49H6.5	pud-4	nlp-29	F18G5.6
ZC21.10	W02D7.11	nlp-34	clec-146	C33A12.3
H39E23.3	F53A9.1	fbxa-163	W02D9.10	
far-4	F53B7.7	irg-2	fbxa-24	
F16B4.7	Y46H3A.5	K07B1.8	sqrd-1	
C27B7.9	cnc-2	cnc-7	T28A11.2	

Appendix 4: Panther Statistically Overrepresented GO Terms

High Confidence NMD Substrates

GO Term	# Pos in Background	#Pos in List	Total Background	Total High Conf	% of expressed	% of list	P-value	Change
mRNA splicing, via spliceosome	36	10	8365	414	0.43%	2.42%	0.0214	5.61
RNA splicing, via transesterification rxns w/ bulged adenosine as nucleophile	36	10	8365	414	0.43%	2.42%	2.14E-02	5.61
RNA splicing, via transesterification reactions	38	10	8365	414	0.45%	2.42%	3.36E-02	5.32
structural constituent of ribosome	149	19	8365	414	1.78%	4.59%	3.11E-02	2.58
RNA binding	353	37	8365	414	4.22%	8.94%	0.0141	2.12
translation	361	37	8365	414	4.32%	8.94%	0.00593	2.07
organonitrogen compound metabolic process	716	65	8365	414	8.56%	15.70%	2.14E-03	1.83
nucleobase-containing compound metabolic process	1200	95	8365	414	14.35%	22.95%	2.54E-03	1.60
heterocycle metabolic process	1245	98	8365	414	14.88%	23.67%	2.13E-03	1.59
gene expression	1018	80	8365	414	12.17%	19.32%	0.0265	1.59
cellular aromatic compound metabolic process	1239	97	8365	414	14.81%	23.43%	3.10E-03	1.58
nucleic acid metabolic process	997	78	8365	414	11.92%	18.84%	4.02E-02	1.58
nitrogen compound metabolic process	1628	127	8365	414	19.46%	30.68%	4.62E-05	1.58
organic cyclic compound metabolic process	1275	99	8365	414	15.24%	23.91%	3.37E-03	1.57
RNA metabolic process	1054	81	8365	414	12.60%	19.57%	0.00747	1.55
cellular nitrogen compound metabolic process	1491	114	8365	414	17.82%	27.54%	9.13E-04	1.54
nucleic acid binding	1331	101	8365	414	15.91%	24.40%	0.000811	1.53
hydrolase activity	1182	88	8365	414	14.13%	21.26%	3.86E-02	1.50
hydrolase activity	1120	82	8365	414	13.39%	19.81%	2.63E-02	1.48
primary metabolic process	2845	185	8365	414	34.01%	44.69%	5.62E-03	1.31
organic substance metabolic process	2916	187	8365	414	34.86%	45.17%	1.23E-02	1.30
primary metabolic process	3426	206	8365	414	40.96%	49.76%	3.45E-02	1.21

Class II RNAs

GO Term	# Pos in Background	#Pos in List	Total Background	Total Class II	% of expressed	% of list	P-value	Change
mRNA splicing, via spliceosome	120	39	8365	1406	1.43%	2.77%	0.0215	1.93
mRNA processing	208	58	8365	1406	2.49%	4.13%	0.0349	1.66
RNA binding	354	91	8365	1406	4.23%	6.47%	0.009	1.53
sequence-specific DNA binding transcription factor activity	586	134	8365	1406	7.01%	9.53%	0.0349	1.36
nucleic acid binding transcription factor activity	599	136	8365	1406	7.16%	9.67%	4.19E-02	1.35
RNA metabolic process	1054	238	8365	1406	12.60%	16.93%	0.000311	1.34
organic cyclic compound metabolic process	1275	286	8365	1406	15.24%	20.34%	2.42E-04	1.33
heterocycle metabolic process	1245	279	8365	1406	14.88%	19.84%	3.90E-04	1.33
nucleic acid binding	1277	284	8365	1406	15.27%	20.20%	3.15E-04	1.32
DNA binding	715	159	8365	1406	8.55%	11.31%	3.29E-02	1.32
cellular aromatic compound metabolic process	1239	272	8365	1406	14.81%	19.35%	2.98E-03	1.31
nucleic acid binding	1331	292	8365	1406	15.91%	20.77%	0.000135	1.31
nucleobase-containing compound metabolic process	1200	261	8365	1406	14.35%	18.56%	1.00E-02	1.29
small molecule binding	1267	272	8365	1406	15.15%	19.35%	8.93E-03	1.28
nucleobase-containing compound metabolic process	1607	344	8365	1406	19.21%	24.47%	1.35E-04	1.27
nitrogen compound metabolic process	1628	347	8365	1406	19.46%	24.68%	1.20E-03	1.27
nucleotide binding	1221	259	8365	1406	14.60%	18.42%	3.43E-02	1.26
nucleoside phosphate binding	1221	259	8365	1406	14.60%	18.42%	3.43E-02	1.26
cellular nitrogen compound metabolic process	1491	313	8365	1406	17.82%	22.26%	1.76E-02	1.25
organic cyclic compound binding	2352	484	8365	1406	28.12%	34.42%	1.05E-04	1.22
heterocyclic compound binding	2345	482	8365	1406	28.03%	34.28%	1.31E-04	1.22
primary metabolic process	3426	644	8365	1406	40.96%	45.80%	2.48E-02	1.12

Indirectly Affected RNAs

GO Term	# Pos in Background	#Pos in List	Total Background	Total Indirect	% of expressed	% of list	P-value	Change
defense response to fungus	18	5	8365	84	0.22%	5.95%	1.60E-03	27.66
response to fungus	19	5	8365	84	0.23%	5.95%	2.08E-03	26.21
defense response to other organism	106	9	8365	84	1.27%	10.71%	1.70E-03	8.46
response to other organism	107	9	8365	84	1.28%	10.71%	1.84E-03	8.38
response to biotic stimulus	107	9	8365	84	1.28%	10.71%	1.84E-03	8.38
response to external biotic stimulus	107	9	8365	84	1.28%	10.71%	1.84E-03	8.38
innate immune response	154	11	8365	84	1.84%	13.10%	0.000574	7.11
immune response	155	11	8365	84	1.85%	13.10%	0.000611	7.07
immune system process	156	11	8365	84	1.86%	13.10%	0.000651	7.02
defense response	246	12	8365	84	2.94%	14.29%	0.00853	4.86