Chapter 5

Nonsense-mediated mRNA decay is a regulator of developmental gene expression

5.1. Introduction

Having established robust protocols to use tiling array and sequence data to identify structural and expression changes for genes we next sought to apply these techniques to furthering our understanding of nonsense-mediated mRNA decay (NMD).

As previously stated, NMD is best understood as a surveillance mechanism which detects and degrades transcripts containing an in-frame premature termination codon (PTC) (reviewed in Behm-Ansmant *et al*., 2007b; Chang *et al*., 2007; Mango, 2001). Study of the NMD pathway by numerous groups, however, has indicated that NMD regulates wild-type transcripts as well as aberrant transcripts containing PTCs. Recent studies have indicated that alternative splicing and NMD appear to be highly linked. This is to say that classes of splicing activators (e.g. the SR genes) have specific PTC-introducing exons that lead to NMD-targeting on inclusion (Lareau *et al.*, 2007; Ni *et al.*, 2007; Saltzman *et al.*, 2008). It has therefore been suggested that NMD may play a role in maintaining homeostasis of splicing factors. It is currently unclear if there are any further biological roles of NMD. Expression analyses in *S. cerevisiae*, *Drosophila melanogaster*, and human cells have revealed non-orthologous sets of NMD targets, indicating no clear role for NMD in any other biological process (Guan *et al*., 2006; He *et al*., 2003; Lelivelt and Culbertson, 1999; Mendell *et al*., 2004; Rehwinkel *et al*., 2005). All of these studies consider changes in transcript levels between wild-type and NMD-perturbed conditions. They do not comprehensively consider the transcript structures of NMD targets or how these structures and targets change throughout a defined biological process such as development. In order to address these questions we have interrogated

the transcriptome of the nematode worm *C. elegans* at multiple developmental stages, comparing the wild-type reference strain (Bristol N2) to strains carrying a lesion in key NMD effectors, SMG-1(the central kinase) and SMG-5 (a key phosphatase). Specifically we have used Affymetrix GeneChip® *C. elegans* Tiling 1.0R Arrays to interrogate the transcriptome of *smg-1(r861)* mutant animals at L3, L4, young adult and gravid adult stages and the *smg-5(r860)* mutant animals at L4 stage. Furthermore we have used the Illumina ultra-high density sequencing platform to generate transcriptome sequence data at L4 and young adult stages in both NMD mutants and N2. As in chapter 4, the timecourse hybridizations on Affymetrix GeneChip® *C. elegans* Tiling 1.0R Arrays were performed in the laboratory of T.R. Gingeras, Affymetrix Inc., Santa Clara, CA., USA, owing to the fact that these arrays were not yet commercially available at the time the experiments were performed. All subsequent hybridizations were performed by the author. Much of the informatics analysis was performed in association with Arun Ramani, a postdoctoral researcher in the Fraser lab.

In this chapter I will describe how the methodologies detailed in chapter 4 have been applied to uncovering the transcripts regulated by NMD and how the structural features of these transcripts are different between NMD targeted and non-targeted forms. I will then detail the further analyses that have revealed the underlying causes of these transcripts being targeted and how each cause contributes to the global repertoire of NMD targets. I will then demonstrate how the way transcript levels change across development indicates roles for NMD in regulation of operonic gene expression and developmental gene expression.

5.2. The targets of NMD

I first sought to identify the transcripts that differ in abundance between wild-type and the NMD mutants. As discussed in chapter 4, for our tiling array data genes were considered expressed if ≥50% of unique exons had ≥50% of probes above background. The background threshold is set to include the top 5% of non-genic probes on the array. The gene intensity value relating to such genes is the median probe intensity of all probes above background in exons with ≥50% of probes above background. An average of 7028 genes were detected at any stage in any strain considered in this study. This covered a total of 50% (9515/19169) of all coding genes annotated in WS150. The fold-change in intensity between N2 and each NMD mutant for each gene was calculated to reveal NMD regulated genes. Where a gene is called as expressed in only one of the two conditions being compared then a gene is still called as NMD regulated if its intensity is greater than the fold-change being considered above background.

At any individual developmental stage, ~13% (1235/9515) of all genes detected produce transcripts which differ by at least 1.5-fold in intensity between wild-type and *smg-1(r861)* worms. In the vast majority of cases (75% overall), transcript levels are higher in *smg-1(r861)* suggesting that they are indeed true NMD targets. To confirm that these targets are not specific to *smg-1(r861)*, we also made comparisons with L4 *smg-5(r860)* mutant animals. We find that the majority (318/437, ~73%) of genes whose expression differs by ≥1.5-fold between wild-type and *smg-1(r861)* animals also differ between

wild-type and *smg-5(r860)* animals, confirming these differences are indeed the result of loss of NMD.

5.3. Structural features which define NMD targets

Both tiling arrays and ultra-high density sequence data give information on transcript structure. This is to say that when the resulting signal is aligned to the genome differences in intensity can be observed across a genic structure, indicating differential inclusion, truncation or elongation of exonic structures. At 1bp resolution ultra-high density sequence data is likely to give more accurate information than tiling arrays. It is important, however, that we understand the limitations of our platforms and interpret our data with this in mind.

A deficiency of ultra-high density transcriptome sequencing relative to capillary sequencing of RT-PCR products is read length. Our tiling and ultra-high density sequence data allow us to predict structural changes that lead to NMD targeting at up to bp resolution indicating exactly what is transcribed, but defining connectivity between distant reads and annotated structures is a more complex issue. If RT-PCR is done for a gene, the PCR products purified and individually sequenced then the connectivity over the read acquired is clear. This is only so for each 35bp read acquired using the Illumina platform and whilst connectivity can be inferred by the presence of overlapping reads, inevitably there will be cases where structures terminate in regions where there are overlapping reads. The analysis discussed in chapter 4 to reveal connectivity of exons in a gene is based on identifying reads that span annotated exon junctions. In the case of

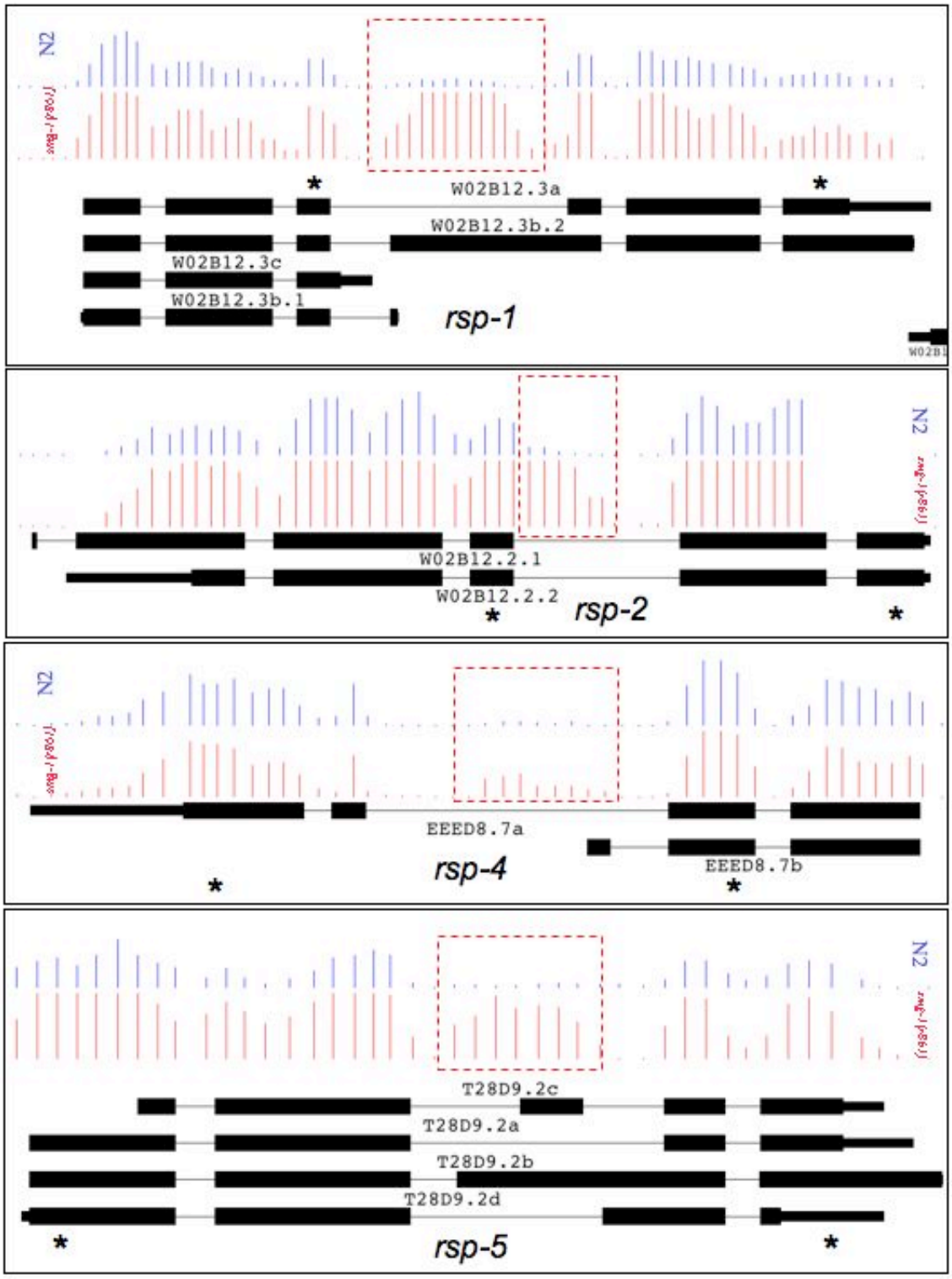
NMD targets produced by alternative splicing, the repertoire of splice sites used and junctions present is inevitably beyond what is annotated. Consequently it is far more complex and computationally intensive to look for unique reads that span two regions of a gene, undefined other than that they are both in expressed regions of genes that appear to be NMD targets. Furthermore the linking of any two reads does not mean that those two sequences are always linked in transcripts. Coupled with the fact that the data produced by either technology used here are not strand specific, it is inevitable that whilst the data that we have produced is extremely useful it cannot give us complete information on all isoforms of all genes.

In wild-type animals NMD targeted transcripts are produced and degraded whereas in the NMD mutants these transcripts are retained. In some cases multiple transcript isoforms of the same genes will be produced, not all of which are NMD targets. The simplest explanation for genes that appear to be NMD targets is that the structural change between the transcript present in the wild-type animal and NMD mutant, as observed in the tiling or ultra-high density sequence data is likely to be causative. In these cases the novel or extensions of known exons can be tested to see if they have stop codons in all frames or may lead to a frame shift. This does not identify the causative PTC. The compelling factor then is that splice forms appear to exist in NMD deficient animals that are undetectable in wild-type animals. Aware of the drawbacks of our datasets I sought to test how the structural changes we observe in our tiling and sequence data compare to those seen by RT-PCR.

The most well characterized targets of NMD are perhaps the SR genes. The SR genes are a family of splicing factors, which are conserved from yeast to human. In all organisms investigated there is evidence that members of this gene family are NMD regulated due to the production of splice forms containing PTCs (Lareau *et al*., 2007; Morrison *et al*., 1997). In *C. elegans* there are eight members of this gene family (*rsp-1* to *rsp-8*) of which *rsp-2* and *rsp-4* were previously known to be NMD regulated (Longman *et al*., 2000; Morrison *et al*., 1997). These genes are interesting in terms of this study for two reasons. Firstly they act as a set of positive controls, demonstrating that NMD truly is perturbed in the mutant strains. Secondly it provides a set of controls to test our ability to detect true NMD targets with our tiling data but the necessity for sequence data to pinpoint the likely cause of NMD targeting in some cases. Specifically, the primary in-frame stop codon, which initiates NMD targeting in one isoform of *rsp-5* appears to be produced by a four-nucleotide extension of exon 2. Whilst this was observed in our sequence data it could not have been determined from the tiling array data.

Of the eight *C. elegans* SR genes seven appear to have NMD-targeted splice forms as indicated by our tiling and sequence data (*rsp-3* does not). In order to determine how these genes with deleterious splice forms compared between RT-PCR and our chosen technologies I focused on the seven SR family genes that appear to be NMD regulated. RT-PCR was done across a region at least spanning the regions indicated to be differentially included by our tiling and ultra-high density sequence data. Figure 5.1 indicates the number of isoforms of each gene detected in both the wild-type animal and *smg-1(r861)* by RT-PCR (manifested as bands on a gel), along with the transcript

structures indicated by the tiling array data. In every case there is one clear isoform present in N2 and at least one larger isoform present in *smg-1(r861)*, in some cases two. Clearly the different isoforms cannot be completely identified within the tiling or sequence data, however, in most cases the predicted maximum size of the NMD-targeted transcript from the tiling data matches the size of a band on the gel. Where multiple larger isoforms exist in *smg-1(r861)* they are best explained by splice events occurring in the novel or extended exon regions observed in the tiling data. Such events cannot be determined by our tiling data or current sequence data analysis. Theoretically, however, all splice junctions should be represented in Illumina sequence data of sufficient depth and should be identifiable in due course.


N2
rsp-1 mut
W02B12.3a
W02B12.3b.2
W02B12.3c
W02B12.3b.1
rsp-1
N2
rsp-2 mut
W02B12.2.1
W02B12.2.2
rsp-2
N2
rsp-4 mut
EEED8.7a
EEED8.7b
rsp-4
N2
rsp-5 mut
T28D9.2c
T28D9.2a
T28D9.2b
T28D9.2d
rsp-5

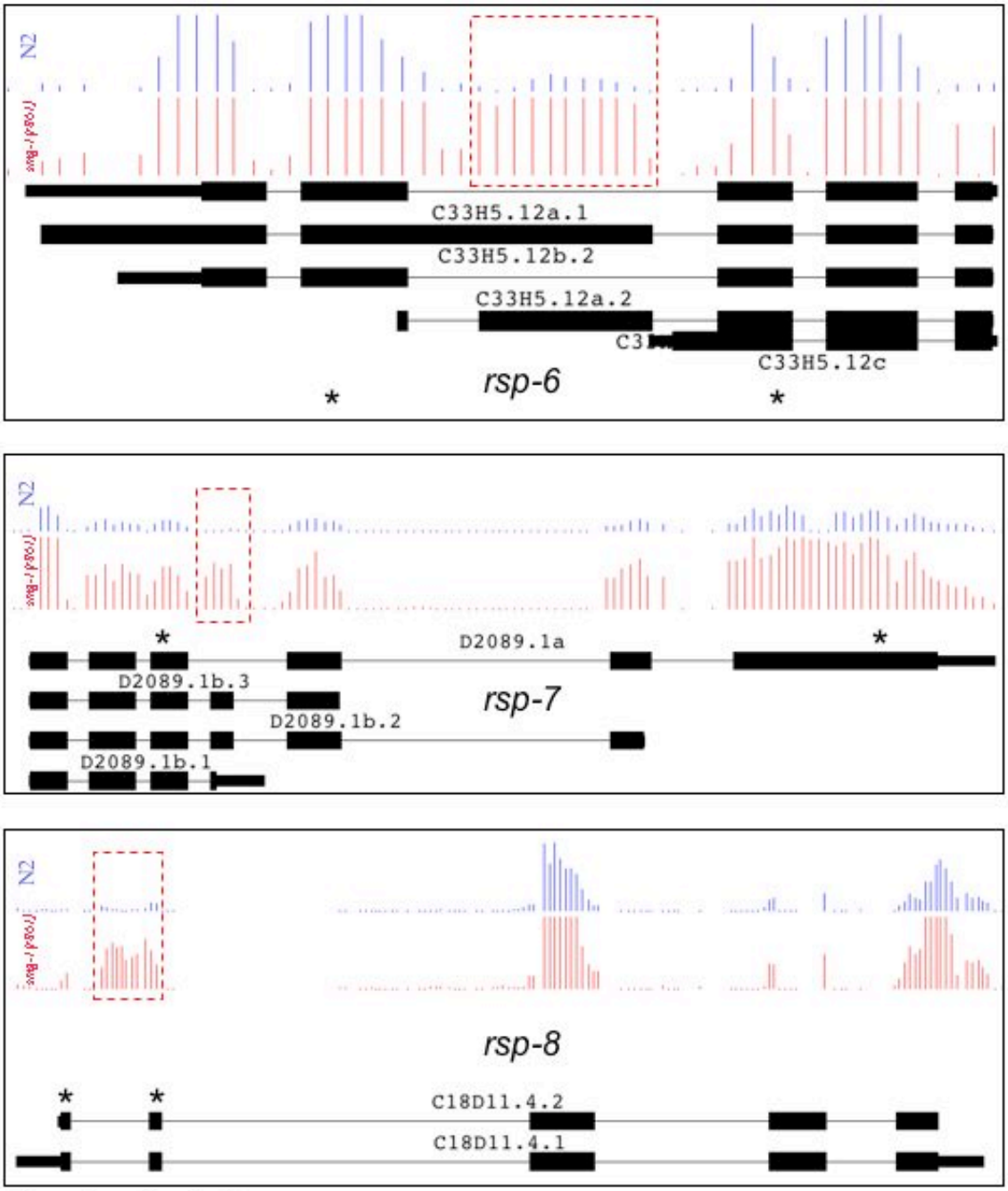


Figure legend overleaf.

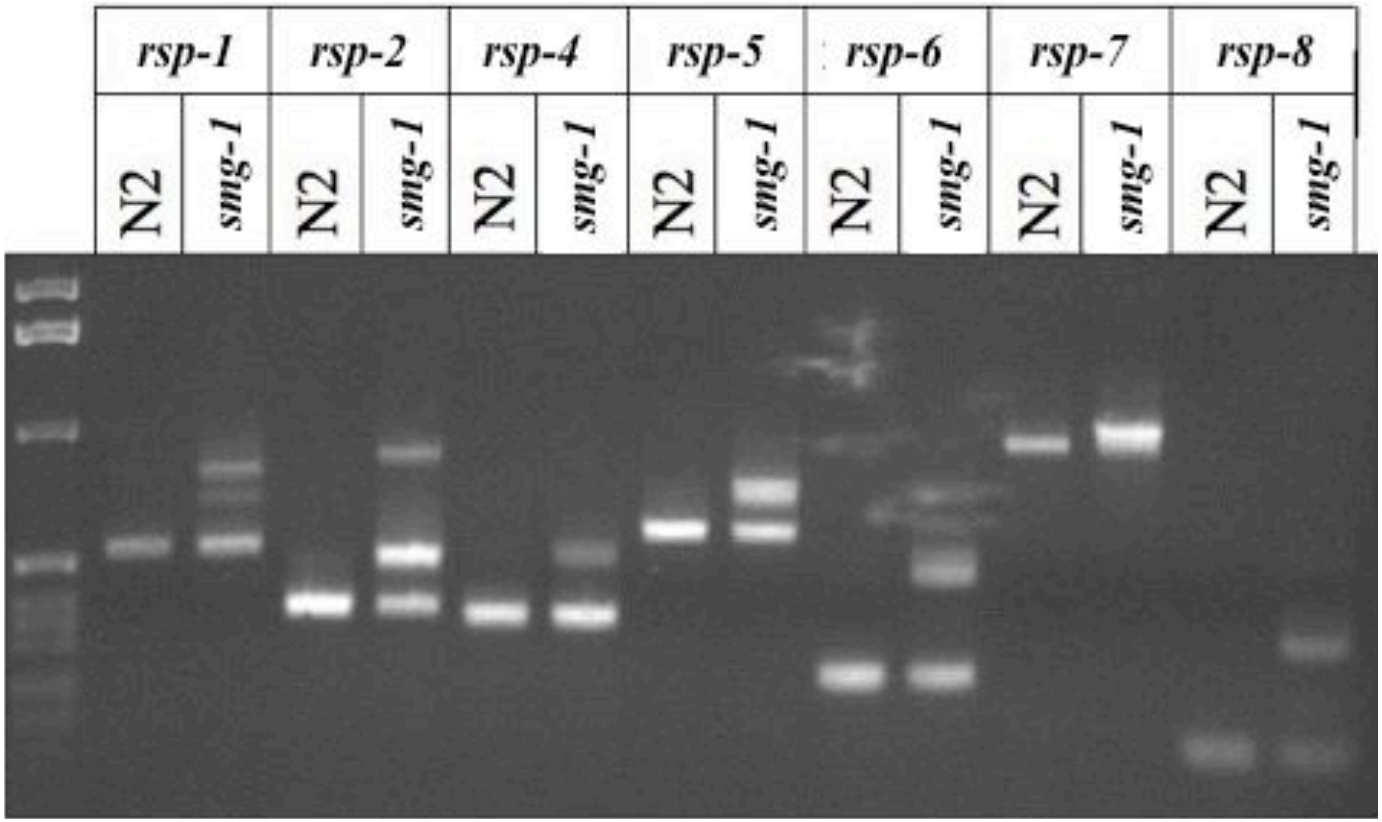


Figure 5.1. Structural changes in SR gene transcripts leading to NMD. Pages 1-2 of this figure show the normalized probe signal for each SR gene shown to be NMD regulated by our tiling array data (in order *rsp-1* to *rsp-8*). Gene annotations are in black. Normalized probe intensities derived from N2 L4 stage animals is in blue and *smg-1(r861)* L4 animals in red. The visually identified structural difference between the N2 and *smg-1(r861)* transcript(s) is indicated by the red box. RT-PCR to amplify across this region was performed between flanking exons and the PCR products run on a gel. The positions of the primers used for RT-PCR are indicated with asterisks. As can be seen in the above gel image, a single band was detected for each gene in N2 but at least one additional larger product was seen in *smg-1 (r861)*. This suggests that NMD-targeted isoforms of these genes are produced. In most cases the largest band correlates with the inclusion of the full novel structure but intermediate bands imply that multiple splice events occur within.

Clearly then, the absolute structural identity of transcripts that are NMD targets cannot be accurately determined from our datasets. Key structural differences between NMD targeted transcripts that are retained in NMD mutants and transcripts detected in wild-type animals can however be inferred from these data. This is what I am going to discuss in the following paragraphs.

Given the structural features of transcripts previously identified to lead to NMD targeting, I am going to discuss the presence and lengths of 5' and 3' UTRs of NMD regulated genes, the presence of upstream AUGs (uAUGs), and the prevalence of alternative spliceforms seen between wild-type and *smg-1(r861)*. The relative intensities of individual exons can be used as in chapter 4 to infer changes in major spliceform that lead to NMD targeting, or the use of an alternative promoter to include or exclude the 5' UTR or exon(s). Furthermore the length and sequences of annotated 3' and 5' UTRs for detected NMD targets allow us to assess how these features compare to the annotated transcriptome as a whole.

Firstly, considering UTR length - UTRs are defined regions of transcripts that are known to have regulatory roles. It therefore follows that they may have a role in determining whether a transcript is NMD regulated. Since we have no clear, easily testable notion of how this would occur at the level of sequence (other than by the presence of a uAUG), we tested whether UTR length appears to be a determinant of NMD targeting. We find that of the 13% of genes called as NMD regulated at >1.5-fold, 30% and 17% can be

classified as having a >1.5-fold longer than average 5' or 3' UTR respectively as compared to 11% and 10% for all genes.

Next we examined the likelihood of transcripts with a 5' UTR containing a uAUG leading to a uORF. This is likely to lead to NMD targeting due to the resulting frame-shift leading to an in-frame PTC. We find that 18% of genes (220/1235) called as NMD regulated at >1.5-fold contain at least one uAUG, versus ~10% of annotated genes. This represents a statistically significant enrichment (p-value < $1x10^{-4}$).

Regarding UTR length – the average length of both 5' and 3' UTRs of NMD targeted transcripts is longer than the average length of both structures across all genes with annotated UTRs. It follows then that the average total UTR length is also greater for NMD-targeted genes. Recent research in human suggests that recognition of a termination codon as premature is dependent on its distance from the ribonucleoprotein environment of the 3' end of the transcript (Eberle *et al*., 2008). Intriguingly, the distance between PTC and 3' end (>420nt) appears critical for NMD targeting. Clearly then the 3' UTR length and NMD are highly linked and our observation that NMD targets have longer 3' UTRs is logical. How the 3' UTR relates to NMD, however, is clearly a complex issue as is the regulatory role of 3' UTRs in general. Not all transcripts with long 3' UTRs appear to be NMD regulated and so it seems reasonable to hypothesize that there is a duality of function whereby 3' UTRs may predispose some transcripts to NMD as a function of their length and others protect the transcript from NMD as a function of their sequence. This could either be by formation of a secondary

structure, which brings the 3' end closer to the termination codon or by the recruitment of factors that inhibit NMD. Further research is required, however, to test whether this is so. That said, our observation that NMD targets are enriched for long 3' UTRs is not statistically significant (p-value = 0.512). The observation of longer than average 5' UTRs is, however (p-value < $1x10^{-4}$). Not only is it the case that NMD targets are more likely to have longer than average 5' UTRs, but also that the greater the fold change in regulation of a gene and the more developmental stages at which they are called as NMD-regulated, the longer the 5' UTR. Effectively then magnitude of NMD regulation correlates well with increased UTR length. This is most likely due to an increased likelihood that a transcript contains a uAUG, the longer its 5' UTR is. This is represented in figure 5.2.

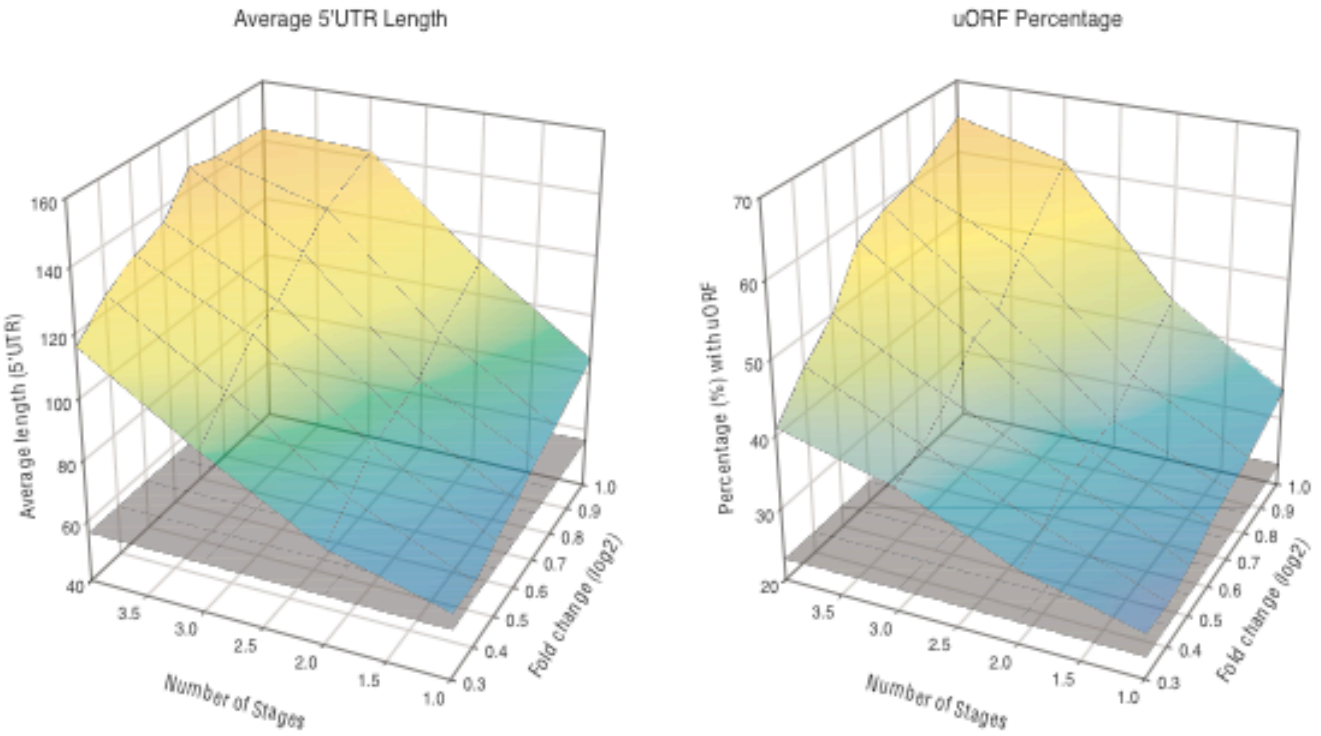


Figure 5.2. Increasing 5' UTR length correlates with increased magnitude of NMD. Each graph demonstrates how the characteristic labelled above increases with increasing fold-change of gene intensity between N2 and *smg-1(r861)* and increasing number of stages at which that fold-change occurs. The average length/occurrence of the characteristic considered is represented by the grey square – the average length of annotated 5' UTRs and the percentage of annotated 5' UTRs containing a uAUG respectively. The plots clearly show that the greater the extent of NMD-regulation of a gene, both in terms of fold-change and number of stages at which it is regulated the longer the 5' UTR.

Next we tested how prevalent differences in major spliceform between N2 and *smg-1(r861)* are at any stage for genes called as NMD regulated at >1.5-fold. Such a difference in spliceform may indicate that a transcript is retained in the NMD mutant, the splicing of which has led to a PTC. We find that ~33% of genes (406/1235) show a >1.5-fold change in relative exon intensity between N2 and *smg-1(r861)*. Genes presenting a change in major spliceform encompass transcripts exhibiting the differential inclusion of an annotated exon, a novel exon overlapping an annotated exon or the use of alternative splice sites within an annotated exon. Genes that are alternatively spliced to include a non-annotated "poison exon" which leads to a PTC will not be detected since this method only considers annotated exons. The number of transcripts alternatively spliced leading to NMD at this threshold cutoff may therefore be higher.

5.4. Translation initiation and NMD

It has long been recognized that there is a link between the nucleic acid environment of a translation initiation codon (AUG) and the efficiency with which translation is initiated at that point. A consensus sequence has been defined based on the relative enrichment of individual nucleotides in the region of translation initiation codons. This identifies the key nucleotides that ensure efficient recognition of the translation start site. This consensus is often called the Kozak consensus sequence after pioneer in the field Marilyn Kozak (Kozak, 1984; Kozak, 1986; Kozak, 1987). Variations on the common consensus occur between eukaryotes. In *C. elegans* the key nucleotide is recognized to be an A nucleotide at the -3 position, where the A of the AUG is +1 (figure 5.3). As previously stated the link between this consensus sequence and translation efficiency is well

established. Thus far, however, a direct link between such a consensus and NMD has not been reported. It has been recognized, however, that leaky scanning by the ribosome leading to translation initiation at an internal AUG leading to a frame shift and in-frame PTC leads to NMD targeting.

We wanted to test whether there is a strong link between the nucleotide environment of the annotated start codon and NMD targeting of transcripts by assessing the relative enrichment of nucleotides within the flanking regions of the AUG at different magnitudes of transcript fold change between N2 and *smg-1(r861)*. As illustrated in figure 5.4, the greater the fold increase in transcript levels in *smg-1(r861)* over N2, the less likely that transcript is to have an A nucleotide at the -3 position. This suggests that detected targets of NMD are more likely to be subject to leaky scanning and NMD. Whilst this is an interesting observation, it is not completely surprising. Intriguingly, however, the genes upregulated in N2 above *smg-1(r861)*, are more likely to have an A nucleotide at the -3 position. They are therefore stronger candidates for translation initiation at the correct site and therefore less susceptible to NMD due to "leaky" translation. This may suggest that the transcripts that appear to be upregulated in N2 are actually technical artefacts. More specifically, the nature of the normalization may mean that transcript levels that are actually equal in both N2 and *smg-1(r861)* appear higher in N2 because the vast majority of differentially expressed genes are higher in *smg-1(r861)*. The fact that the probe intensities are effectively scaled to the same mean therefore leads intensities to be artificially low for *smg-1(r861)*. This would not be a serious problem in terms of this study as at worst it would lead to a higher false negative rate in terms of NMD target

discovery but would not invalidate genes being called as NMD regulated. Importantly then, if transcripts which appear higher in N2 are in fact the genes which are not NMD regulated, which is supported by their stronger translation initiation consensus, then this suggests that the majority (if not all) transcripts are NMD regulated to some extent as a function of the translation initiation consensus. If this is so then the evolutionary value of this is clear. It is critical that the transcript level of individual genes is tightly regulated. This regulation is inevitably a combination of transcript production and degradation. Variation of the 5' UTR nucleotides within the Kozak consensus would therefore act via NMD to control transcript and protein levels within the cell. On this level alone NMD would therefore be a bona fide regulator of gene expression.

That there is a statistically significant association of diminished Kozak consensus with NMD suggests that the translation initiation sequences at a uAUG could also be critical in determining the extent to which a transcript is NMD regulated. Specifically, if a transcript has a strong translation initiation sequence at a uAUG it may be more likely to be strongly NMD regulated than if it has a weak translation initiation sequence and a strong translation initiation sequence at the true AUG. This is because it may increase the likelihood of translation of a uORF. This is a question that should be addressed in the future.

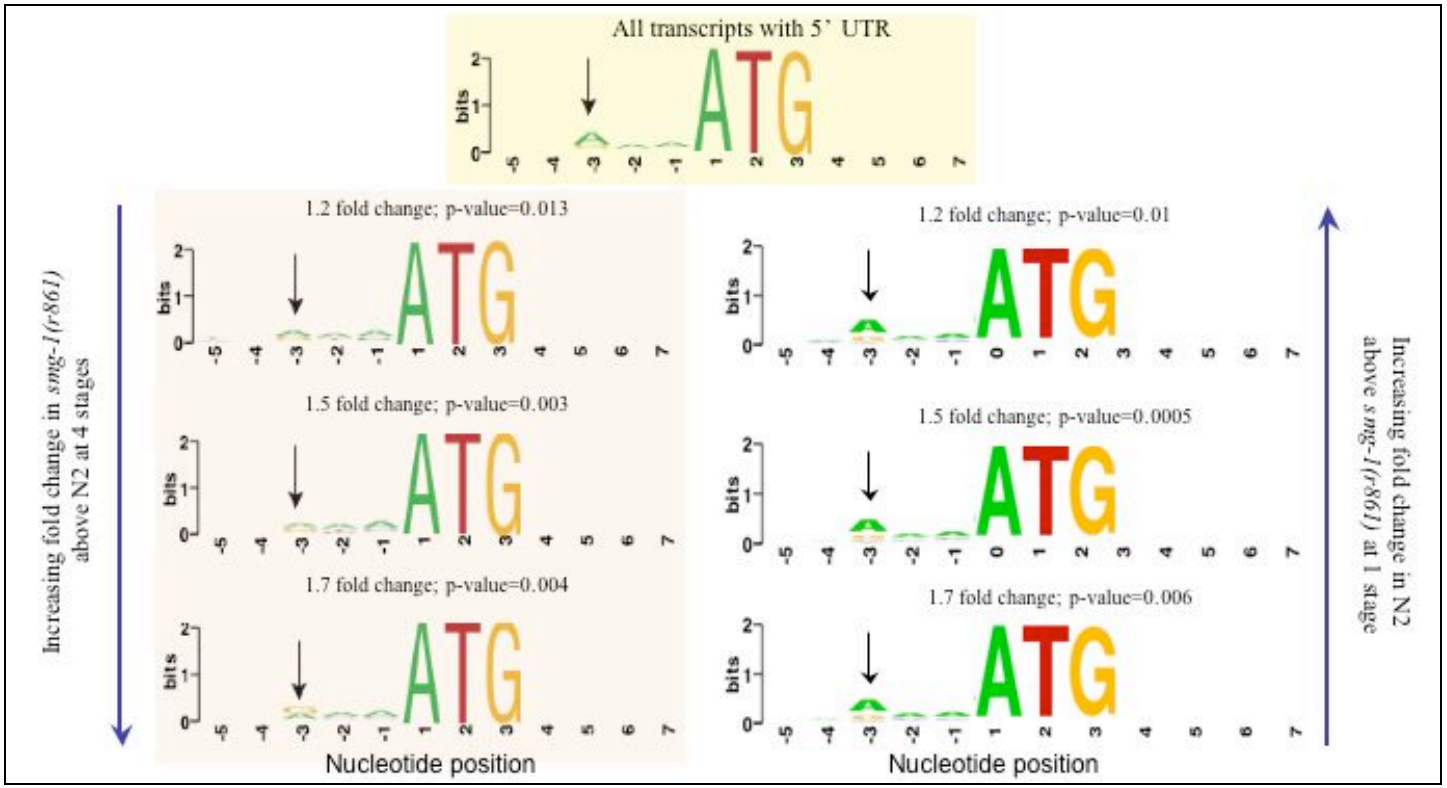


Figure 5.3. An A nucleotide -3 of the annotated start codon correlates with NMD regulation. Surveying the consensus sequence around all annotated start codons in transcripts reveals an enrichment for an A nucleotide -3 of the annotated start codon. This enrichment diminishes with increased NMD regulation in transcripts higher in *smg-1(r861)*. Shown is increasing mean fold change of transcript in *smg-1(r861)* above N2 across all four stages. The significance of change in enrichment of the A at -3 between NMD regulated and all genes was determined by chi-square test. Conversely, a significant enrichment of an A nucleotide at the -3 position in genes upregulated in N2 above *smg-1(r861)* is seen. Note that the analysis of genes upregulated in N2 above *smg-1(r861)* is limited to changes seen at any one stage due to too few genes being thusly regulated at all stages. The overall height of the stack at each position indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each nucleotide at that position (Schneider and Stephens, 1990). Nucleotide enrichment plots were generated using WebLogo (Crooks *et al*., 2004).

5.5. NMD regulates the expression of genes in operons

C. elegans and related species appear to be rare amongst animals in that they have operons. Operons consist of contiguous genes, which are transcribed as polycistronic pre-mRNAs, which are trans-spliced to form mature monocistronic mRNAs. Current evidence suggests that there are more than 1000 operons, each containing between 2 and 8 genes and encompassing ~15% of annotated genes (Blumenthal *et al*., 2002).

Operonic genes appear to fall into functionally related clusters of genes involved in transcription, splicing and translation as well as mitochondrial function. Regulation of operonic gene expression is clearly complex. That regulators of such key functions appear to be co-regulated themselves in operons is not surprising, beyond the fact that this does not seem to be the case in other animals. The nature of any such regulation, however, is not well understood. One of the critical open questions regarding operons is how the detected levels of co-transcribed genes are often different. Whilst it appears unlikely that one single known pathway or process governs the inequity of gene expression within all operons, one of our goals was to test whether NMD is involved in such regulation.

We examined whether the transcript levels of any two genes within an operon, which are unequal in the wild-type transcriptome become equalized in the NMD-deficient transcriptome (figure 5.4). Of the 651 operons for which there is a ≥1.5-fold change in expression between genes in N2 at any stage, ~8% (50) of these operons show equalization of gene expression (<1.1-fold difference) in *smg-1(r861)*. This demonstrates that whilst NMD is not the only mechanism by which operonic

transcripts are regulated, it is a bona fide mechanism by which correct transcript levels are maintained for operonic genes.

Clearly NMD represents only one method of regulation of transcript levels of operonic genes. Operons are not statistically significantly enriched for NMD regulated genes relative to the genome as a whole. The critical factor is that NMD is a hitherto unrecognized mechanism by which this specific set of genes is regulated.

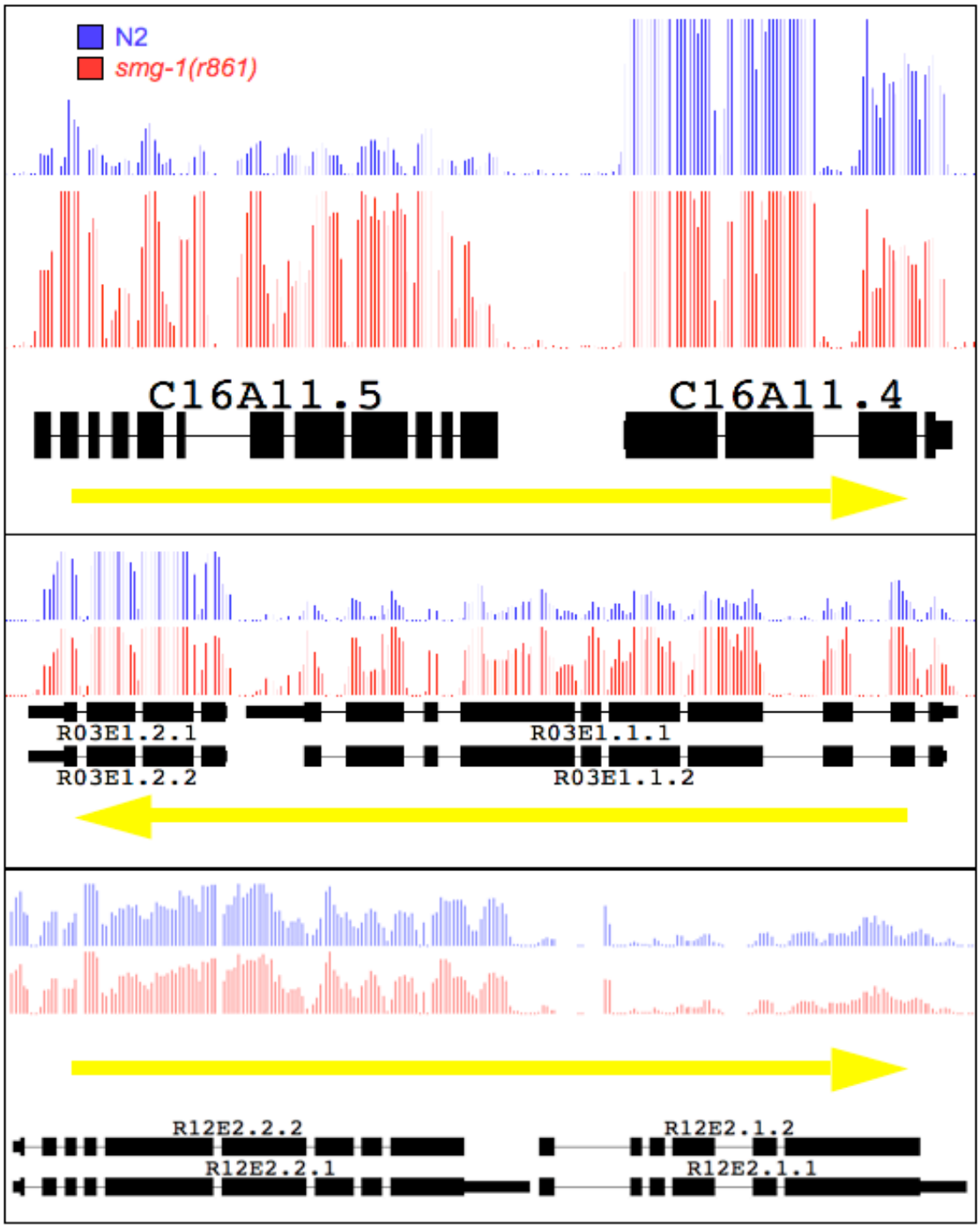


Figure 5.4. Examples of operonic gene regulation by NMD. Each segment shows tiling array data relating to an operon of two genes and the direction of transcription. The top and middle operons show clear equalisation of transcript levels within the operon on NMD perturbation. This is not the case in the bottom example, demonstrating that NMD is not the sole regulator of transcript levels of operonic genes.

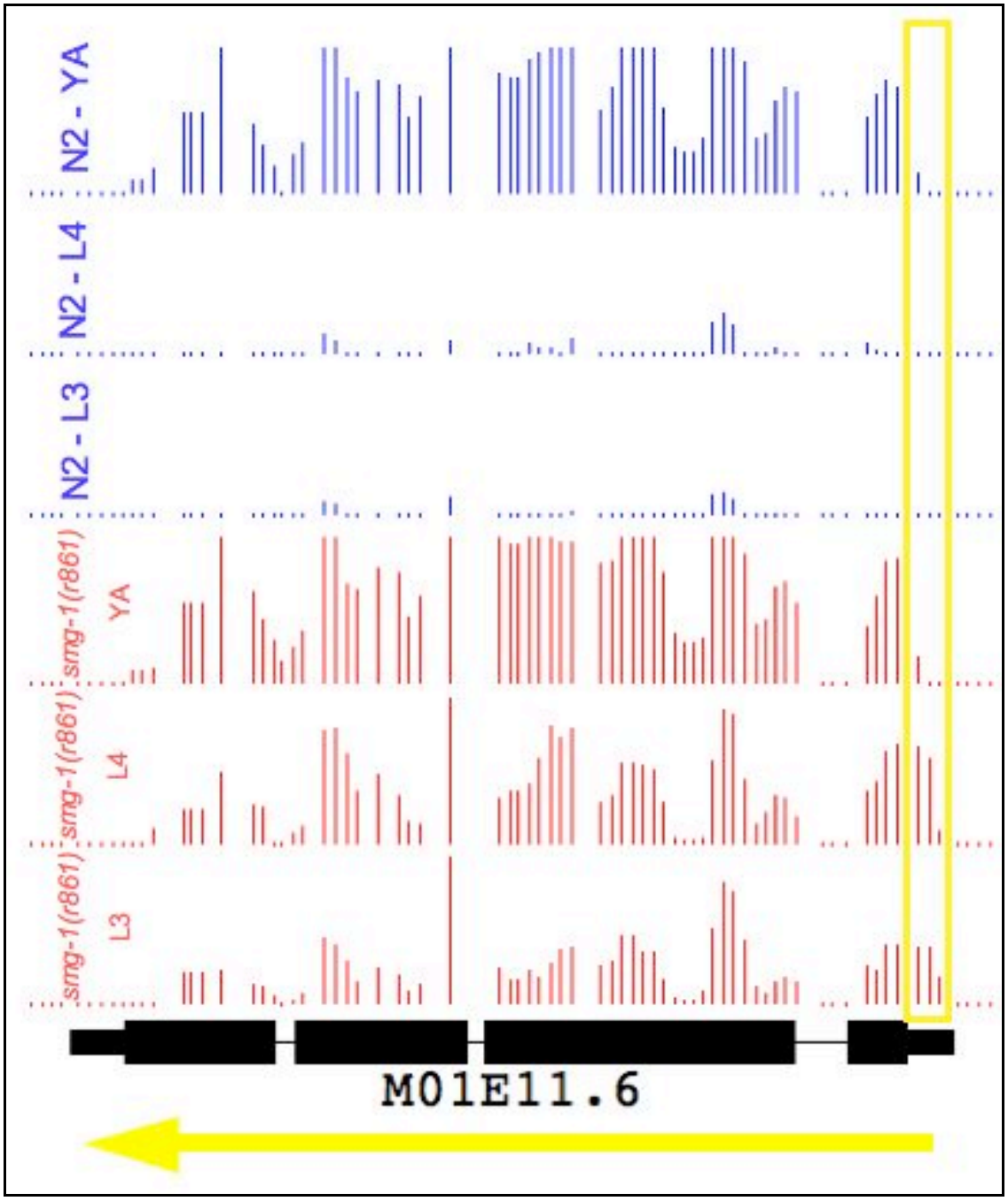


Figure 5.5. NMD regulation via a shift in promoter usage. *klp-15* (M01E11.6) is transcribed at all developmental stages considered, but degraded at L2-L4. The 5' UTR of *klp-15* contains an AUG with an A nucleotide at the -3 position. The annotated start codon does not have an A nucleotide at the -3 position. The change in probe signal across exon 1 implies that a switch in promoter site at the young adult stage to omit the uAUG leads to the transcript no longer being NMD targeted. Note – absent probes are the result of the inability to design unique probes in that region, not low signal.

5.6. NMD regulates developmental gene expression

Browsing of the tiling array data revealed a number of genes that, whilst expressed at similar levels across development in the NMD mutants, were absent or severely reduced at specific stages in N2 (example in figure 5.5). Assessment of the structural features of these transcripts revealed obvious changes that lead to NMD targeting, such as a shift in promoter site to include a uAUG, or the differential inclusion of a novel or alternative exon. We sought to systematically probe our dataset for genes that exhibit expression indicating that they are regulated by NMD in a developmentally controlled manner – in other words genes for which the correct timing of expression is detectably controlled by NMD.

We identified the sets of genes whose expression changed between any two consecutive developmental stages in wild-type animals and examined whether these expression changes require NMD, that is, if we see the same change in *smg-1(r861)* animals. In total 3222 genes (~34% of detected) change expression by >2-fold between any two consecutive developmental stages. We refer to the genes that require NMD for this change as NMD-regulated and those that do not as NMD-neutral. 318 (~10%) of these expression changes are strongly reduced (i.e. differ by <1.1-fold between stages) in the *smg-1(r861)* animals i.e. are NMD-dependent. We conclude that in these cases, the expression change is mediated by NMD. The simplest explanation for this is that there are two transcript forms synthesised from such genes — a 'normal' form, which is not an NMD target, and a form that is degraded via NMD. A change in expression in such cases is not due to a change in transcription rate, but instead from a change in transcript structure from viable to NMD-targeted form.

To ensure that these changes in transcript abundance are a direct result of NMD and not as a secondary effect of the regulation of other genes, we compared the frequency with which we observe structural changes in the 318 NMD-regulated genes with the 2,904 NMD-neutral genes. If the expression changes of the NMD-regulated genes are indeed driven by regulated structural changes, we would expect these genes to be enriched for such structural changes relative to the NMD-neutral genes. We refer to the time point where the expression is low in wild-type but not in *smg-1(r861)* worms as t_{diff} and the time where the expression is identical in both strains as t_{same}. We only compare transcript structures in the *smg-1(r861)* animals, since at t_{diff}, the transcript that is NMD targeted is degraded and thus not detectable in the wild-type animals.

Given our list of NMD-regulated genes, first we compared the splice index (SI) of exons of genes that are NMD regulated against exons of NMD-neutral genes. As in chapter 4, this is done in order to detect a change in major spliceform. SI = $(E_i/G_i)t_1/(E_i/G_i)t_2$ where E_i is the median probe intensity above background of the exon, G_i of the gene and t_1 and t_2 are the different timepoints. SI therefore is the fold-change of intensity of an exon relative to the whole gene between the conditions being compared. We find that 25% (p-value < 0.003) of these genes have at least one exon with SI >2-fold compared to 15% of NMD-neutral genes. Secondly, we compared the exons of regulated genes in t_{diff} versus t_{same} for probe distribution. We specifically compared the number of exons in each set with less than 50% of probes above threshold. While 25% (p-value < $1x10^{-4}$) of exons of genes at t_{diff} have less than 50% probes greater than threshold only 10% of exons of genes in t_{same} do so. These

bulk analyses immediately suggest that there are structural characteristics of the genes we discover which are significantly different from random.

Next we sought to determine the false positive rate of discovery, the percentage of genes for which we believe the expression change and the subset of those for which we can determine the likely structural change leading to NMD targeting. We deemed that the interpretation of changes in gene structure beyond the analyses previously performed, as well as determining false positive rate could best be achieved through manual annotation. To do this we focused on the genes that require NMD for the expression change between L4 and young adult stages. We define 100 genes thus by the previously mentioned criteria. We visualized the normalized probe data in Affymetrix Integrated Genome Browser (IGB) to assess the characteristics of these 100 genes. We consider that 13% of the genes discovered are probable false positives, as a result of a single (or small number of) probe(s) dropping below threshold leading to an exon being disregarded and consequently the gene not being called as expressed.

Determining potential NMD causative structural changes in the tiling data was problematic due to the limitations of the data itself and our ability to visualize it. An example of this is that the levels of each gene were not scaled relative to each other between developmental stages. It was therefore difficult to determine changes in the relative levels of each exon (or part thereof) between stages. In addition to observing the normalized data track in IGB therefore, we created other tracks to better represent changes in probe signal. Firstly, all probes corresponding to each gene were scaled to the highest probe in the gene. This was to bring the distribution of probe signal across

the gene into the same range at both stages. We then subtracted the scaled probe signal of young adult from L4 to visualize the structural changes. This is not a perfect method of determining structural changes as it requires that the highest probe is representative, but appeared to be the best available to aid in the manual annotation of gene changes. Examples of our manual annotations and the structural changes determined are shown in figure 5.6.

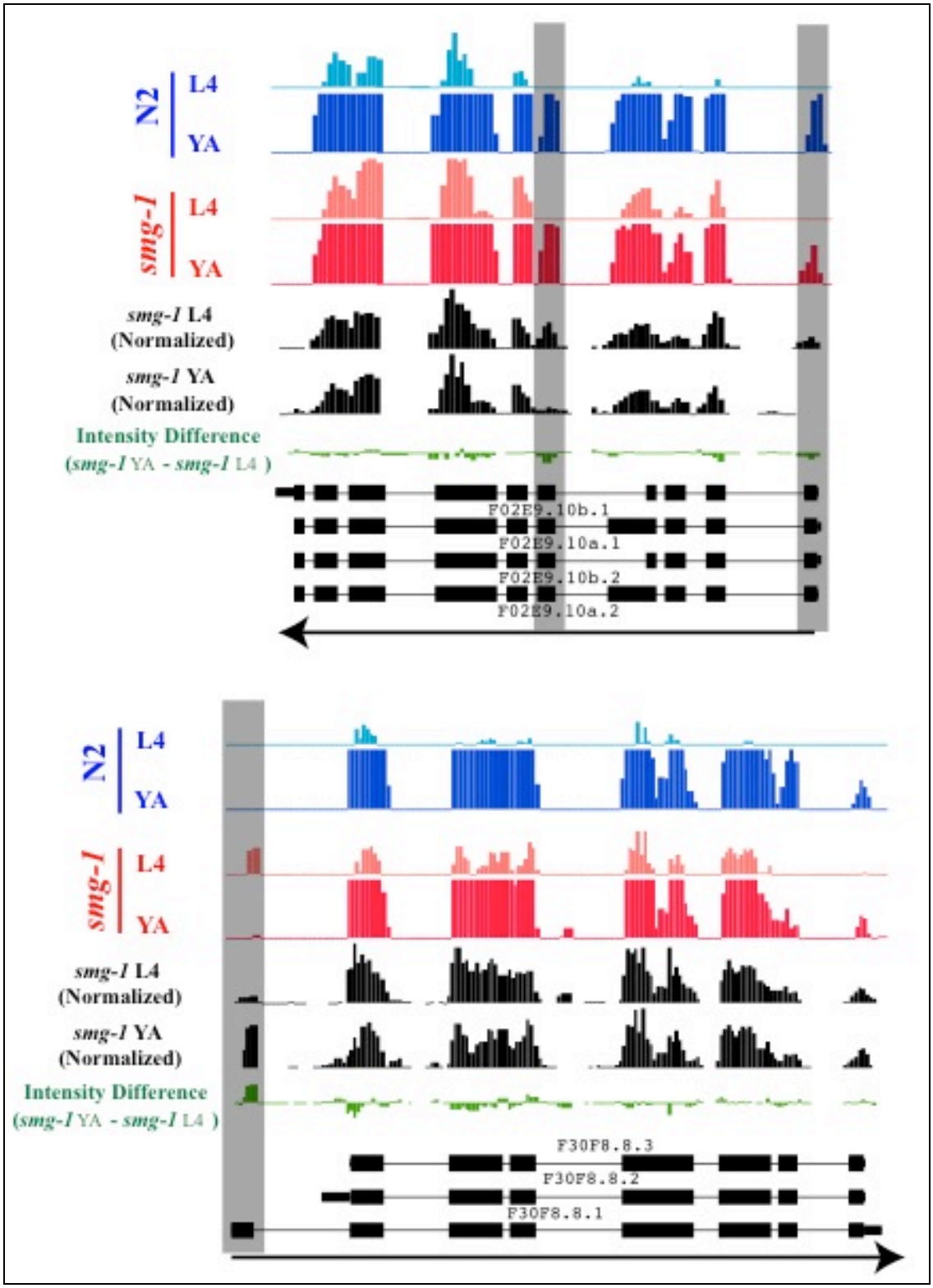


Figure legend overleaf.

Figure 5.6. Structural changes leading to NMD targeting. Manual annotation of transcripts indicates structural changes between two consecutive developmental stages leading to stage-specific gene regulation by NMD. The identity of each data track is colour coded and indicated on the left (N2 – blue and *smg-1* – red). To make valid comparisons between the *smg-1* developmental stages the probe intensities for each gene were normalized to the most intense probe in each gene (tracks shown in black). Arrows indicates direction of transcription and grey boxes indicate likely structural changes. Comparable data tracks are scaled equivalently**.** For F30F8.8.3 (*taf-5* – TBP associated transcription factor) the inclusion of an alternate 5' start appears to be the key structural change. NB – blue and red tracks are not scaled across the full range of probe intensities, rather they are scaled to visualize the structural difference.

Of the genes that are not called as false positives we find clear evidence for changes in transcript structures between t_{diff} and t_{same} for over 50% of the genes examined (44/87). It is important to point out that our ability to call structural changes is limited by the resolution of the array. We resolved that we would only consider structural changes of at least two probes. Since the resolution of the array is 35bp we therefore only consider clear changes of ≥70bp. This will inevitably lead to a false negative rate in our structural calls. Furthermore, a number of genes appeared to have unannotated 5' UTRs or 5' exons. Inclusion of such structures could potentially lead to translation of a uORF. We do not consider such features in our assessment of structural change, however, as their connectivity to the annotated gene is undetermined. We therefore believe that there are likely to be genes for which there are NMD causative structural changes that are not detected in our manual annotation.

In summary, we determine that NMD is required for ~10% of developmentally regulated expression changes. Approximately 50% of these genes show clear structural changes in the transcripts between the two developmental stages probed at the resolution of our tiling data manually and by computational criteria. At least this

number of expression changes are therefore likely to be a direct result of NMD rather than indirect regulation via loss/gain of a transcription factor or equivalent. We conclude that NMD is required for the correct developmental timing of expression of these genes and thence NMD is a bona fide regulator of developmental gene expression.

5.7. GLD-1 as a protector of transcripts from NMD

The RNA binding protein GLD-1 has previously been proposed as a protector of transcripts from NMD by preventing the translation of uORFs through binding to hexameric binding elements in the 5' UTR (Lee and Schedl, 2004; Ryder *et al*., 2004). As discussed in chapter 1, GLD-1 is a key regulator in germline development, acting as a translation inhibitor to control transition between mitosis and meiosis and is also involved in gametogenesis. Previously only one gene (*gna-2*) has been demonstrated to be protected from NMD by GLD-1 (Lee and Schedl, 2004). This is thought to be through binding of GLD-1 to the 5' UTR, thus preventing the translation of uORFs. I undertook to search for other NMD protected transcripts by microarray analysis of *gld-1(RNAi)* in both N2 and *smg-1(r861)*, at L4 stage in biological triplicate using the same tiling arrays as previous. The rationale behind the experiment is that transcripts predisposed to NMD but protected by GLD-1 would not be detected in our original timecourse but would be on *gld-1* knockdown. I identified 117 genes that were >2-fold upregulated in *smg-1(r861);gld-1(RNAi)* over *gld-1(RNAi)* in N2, indicating that they are targets of NMD. Of these 117 genes 44 were not previously identified as NMD targets in our original timecourse at >1.5-fold regulation (table 5.1). These genes therefore correspond exactly to potential candidates of GLD-1 protection from NMD. 16 of these 44 genes were not

sufficiently represented at any stage in the timecourse for them to be determined as NMD regulated or otherwise. This may be a result of the physiological change caused by *gld-1(RNAi)*, potentially resulting in an increased ability to detect transcripts enriched in the mitotic germline. Of the 44 novel NMD targets 10 have an annotated 5' UTR containing a uAUG. 4 of these UTRs are in genes detected but not NMD regulated in the timecourse. I searched for STAR-binding elements (SBEs), the hexameric sequences that GLD-1 is thought to bind in the 5' UTRs of these 10 genes.

The SBE is so called as GLD-1 is a member a of conserved family of RNA binding proteins containing the STAR/GSG domain. The hexameric motif was defined in two forms by Ryder *et al*., (2004) – the conservative UACU(C/A)A, most high affinity form and the relaxed (U>G>C/A)A(C>A)U(C/A>U)A form. Ryder *et al*. confirmed the *in vivo* activity of the range of binding motifs in the germline, verifying that transcripts containing these motifs in their 3' and 5' UTRs co-immunoprecipitate with GLD-1. The NMD protected GLD-1 target published by Lee and Schedl (2005) contains both a conservative and relaxed form of the motif in its 5' UTR (UACUCA and CACTAA). Of the 10 uAUG containing 5' UTRs revealed in our array data 4 contained at least one SBE. One gene, *pac-1* (C04D8.1) contained two SBEs, both of a higher-affinity relaxed form (GAATAA and GAATCA). Of the four uORF containing genes with 5' UTR SBEs only *pac-1* is represented in the Nematode Expression Pattern Database (NEXTDB). NEXTDB is a freely accessible database of RNA *in situ* hybridizations performed by the Kohara lab, National Institute of Genetics, Japan. The *in situ* data for *pac-1* clearly demonstrates that it is a germline enriched transcript and so is highly likely to be regulated by GLD-1 via its SBEs. The other three genes containing SBEs in their uAUG-containing 5' UTRs were *arf-1.1*

(AAATAA), C49A9.4 (GAATCA) and Y73B3A.20 (AAATCA). None of these three genes were sufficiently detected in the original timecourse to be called as NMD regulated or otherwise. Further to this, *pac-1* is the only of these four genes that has a strong translation initiation sequence at its uAUG and a weak translation initiation sequence at its true AUG. This suggests that *pac-1* is highly prone to NMD. *pac-1* is therefore by far our strongest hit. Further work would be required to confirm its association with GLD-1, however, such as comparison of RNA *in situ* hybridizations in N2, *smg-1(r861)* and both strains with RNAi against *gld-1*.

Common name	GeneID	Max. fold-change in timecourse	uORF	5' UTR	uAUG	AnnAUG at true AUG	AnnAUG at uAUG	3' UTR
arf-1.1	WBGene00000190	NA	Yes	Yes	Yes	Yes	No	No
B0513.2	WBGene00007195	1.42	No	No	NA	NA	NA	No
pac-1	WBGene00015418	1.13	Yes	Yes	Yes	No	Yes	No
C05D12.4	WBGene00007341	NA	Yes	Yes	Yes	No	Yes	No
C40H1.2	WBGene00008038	NA	No	No	NA	NA	NA	Yes
C45B2.8	WBGene00016662	NA	Yes	Yes	Yes	No	No	No
C49A9.4	WBGene00016758	NA	Yes	Yes	Yes	No	No	Yes
D1037.1	WBGene00017025	1.48	No	No	NA	NA	NA	No
F01F1.2	WBGene00017159	1.23	No	Yes	No	NA	NA	Yes
F10C2.4	WBGene00008645	1.35	No	Yes	No	NA	NA	Yes
F38H4.1	WBGene00009545	NA	No	No	NA	NA	NA	Yes
F39E9.1	WBGene00018194	NA	No	No	NA	NA	NA	Yes
gst-43	WBGene00001791	NA	No	No	NA	NA	NA	No
lpd-8	WBGene00003064	1.09	Yes	Yes	Yes	Yes	Yes	No
math-14	WBGene00015828	NA	Yes	Yes	Yes	Yes	Yes	No
math-20	WBGene00016555	NA	No	Yes	No	NA	NA	No
math-41	WBGene00020360	1.02	No	Yes	No	NA	NA	No
pgp-12	WBGene00004006	NA	No	No	NA	NA	NA	No
pqn-68	WBGene00004151	NA	No	No	NA	NA	NA	Yes
rgs-4	WBGene00004347	1.32	No	No	NA	NA	NA	No
suf-1	WBGene00006307	1.41	No	Yes	No	NA	NA	Yes
T04F3.1	WBGene00011436	1.16	No	No	NA	NA	NA	Yes
T06A10.4	WBGene00020287	1.05	No	Yes	No	NA	NA	Yes
T08B2.4	WBGene00020345	NA	No	No	NA	NA	NA	Yes
T14G11.3	WBGene00020511	1.13	No	Yes	No	NA	NA	Yes
T15H9.1	WBGene00011787	1.09	No	No	NA	NA	NA	Yes
T16G12.3	WBGene00011804	NA	No	No	NA	NA	NA	Yes
T20B12.1	WBGene00020600	1.34	No	No	NA	NA	NA	Yes
T20D4.11	WBGene00020617	1.27	No	No	NA	NA	NA	Yes
T27F6.4	WBGene00012104	1.30	No	Yes	No	NA	NA	Yes
tag-202	WBGene00009002	1.34	Yes	Yes	Yes	Yes	Yes	Yes
tag-317	WBGene00007107	1.13	No	Yes	No	NA	NA	Yes
Y10G11A.1	WBGene00012423	1.03	No	Yes	No	NA	NA	Yes
Y17G7B.20	WBGene00012471	1.23	Yes	Yes	In frame uAUG	Yes	Yes	Yes
Y32G9A.13	WBGene00044517	NA	No	No	NA	NA	NA	Yes
Y41D4B.18	WBGene00021520	1.11	No	No	NA	NA	NA	No
Y48A6B.2	WBGene00012963	NA	No	No	NA	No	No	No
Y48G1C.12	WBGene00044345	1.23	No	No	NA	NA	NA	Yes
Y51A2D.4	WBGene00013073	1.06	No	Yes	No	NA	NA	No
Y73B3A.20	WBGene00022221	NA	Yes	Yes	Yes	Yes	Yes	Yes
Y73B3A.5	WBGene00022207	1.03	No	No	NA	NA	NA	No
Y76A2B.5	WBGene00013577	1.08	No	Yes	No	NA	NA	Yes
Y95B8A.8	WBGene00022388	1.28	No	No	NA	NA	NA	No
ZK180.4	WBGene00022678	1.15	No	Yes	No	NA	NA	Yes

Table 5.1. Novel NMD regulated genes detected on *gld-1(RNAi)*. 44 genes were detected as NMD regulated >2-fold on *gld-1(RNAi)* and <1.5-fold without *gld-1* knockdown. Presence of annotated UTRs (yes/no), uORFs and translation start site sequence are indicated.

5.8. Discussion

NMD has long been considered to be a process by which aberrant PTC containing transcripts, arising though mutation or incorrect post-transcriptional processing are detected and degraded. Causative events of NMD targeting such as splicing errors and "leaky" translation have previously been reported. We have used our tiling data to test the individual contributions of these events to the global repetoire of NMD targets in *C. elegans*. The most interesting findings when considering structural features of all NMD targeted genes regulated at any stage relate to uAUGs and the translation initiation consensus. The sequence local to the translation start site appears to be indicative of whether a gene will be NMD regulated by determining whether translation will proceed from that codon or an internal site. Evolutionary modification of the key nucleotides may therefore occur to regulate transcript and protein levels in the cell. The presence of a uAUG may also lead to NMD by leading to the translation of a uORF. This too appears to be modulated by the nucleic acid environment of the two start sites. The likelihood of a transcript being NMD regulated as a consequence of the position of translation initiation is therefore determined by the presence of a uAUG and the consensus surrounding both the uAUG and the annotated AUG.

If the presence of a uAUG and the sequence surrounding that and the annotated start site are used by the cell to determine the extent of NMD regulation of a gene then it follows that the length of the 3' UTR might also be used to predispose a gene to some measure of NMD regulation. More specifically, if the effect of varying the sequence surrounding the true AUG and/or uAUG is used by evolution to determine the extent to which a transcript is NMD regulated, then perhaps evolution has also acted to vary

the length of 3' UTRs to the same ends. This is based on the assumption that distance of the termination codon to the poly(A) tail is critical to NMD targeting. It would therefore be reasonable to expect that NMD regulated genes would be significantly enriched for long 3' UTRs. That we do not find this may be indicative of the complexity of the regulatory properties of 3' UTRs, or our continued lack of understanding of what defines a PTC, at least in *C. elegans*.

The prevalence of spliceforms that appear to lead to NMD targeting is also intriguing. Though it is not possible to determine this from our data, it would be interesting to know if this is in part indicative that splicing is a generally low-fidelity process. It may be that splicing of transcripts to a deleterious form at a given time in development has evolved as a form of gene regulation. Alternatively it may be that splicing factors, which are themselves NMD regulated (e.g. the SR genes) direct the splicing of many transcripts to a deleterious form when they are dysregulated as in an NMD deficient background.

Whatever the underlying causes or evolutionary pressures leading to a gene being NMD regulated, it now appears that NMD is much more than a mechanism by which aberrant or incorrectly processed transcripts are degraded. That a set of developmentally regulated gene expression changes appear to be NMD-dependent is intriguing. It clearly demonstrates that NMD is a bona fide mechanism of gene regulation implying that evolutionary pressures have led to NMD being a specific regulator of gene expression as well as a transcript quality control mechanism. Though it seems reasonable to assume that this property of NMD is likely to be conserved it would be of great value were a similar study to be undertaken in another

organism. Expression analyses of *Drosophila* embryogenesis or meiosis in *Schizosaccharomyces pombe* appear to be obvious choices for such a study.

A simple model for the regulation of gene expression could be the following: Transcripts are either predisposed to NMD or not at the level of transcription, depending on uAUGs and the strength of the translation initiation site. Predisposition to NMD could later be introduced at the level of splicing. Temporal regulation of gene expression by NMD is controlled both at the level of transcription and splicing, allowing the cell to switch between viable and deleterious transcript forms. This model is represented in figure 5.7.

Regarding the protection of transcripts from NMD by RNA binding proteins – the extent of this regulation is still unknown and worthy of future investigation. The array experiment detailed in 5.7 yielded few potential candidates of such regulation. This is likely to be indicative of many things, including the limits of detection of this array platform, but also potentially the limited extent of this regulation by GLD-1. It is likely that more transcripts could be detected at other stages and by using a purpose designed expression microarray rather than tiling arrays. The use of the same tiled microarray for this experiment as used previously was to acquire comparable data.

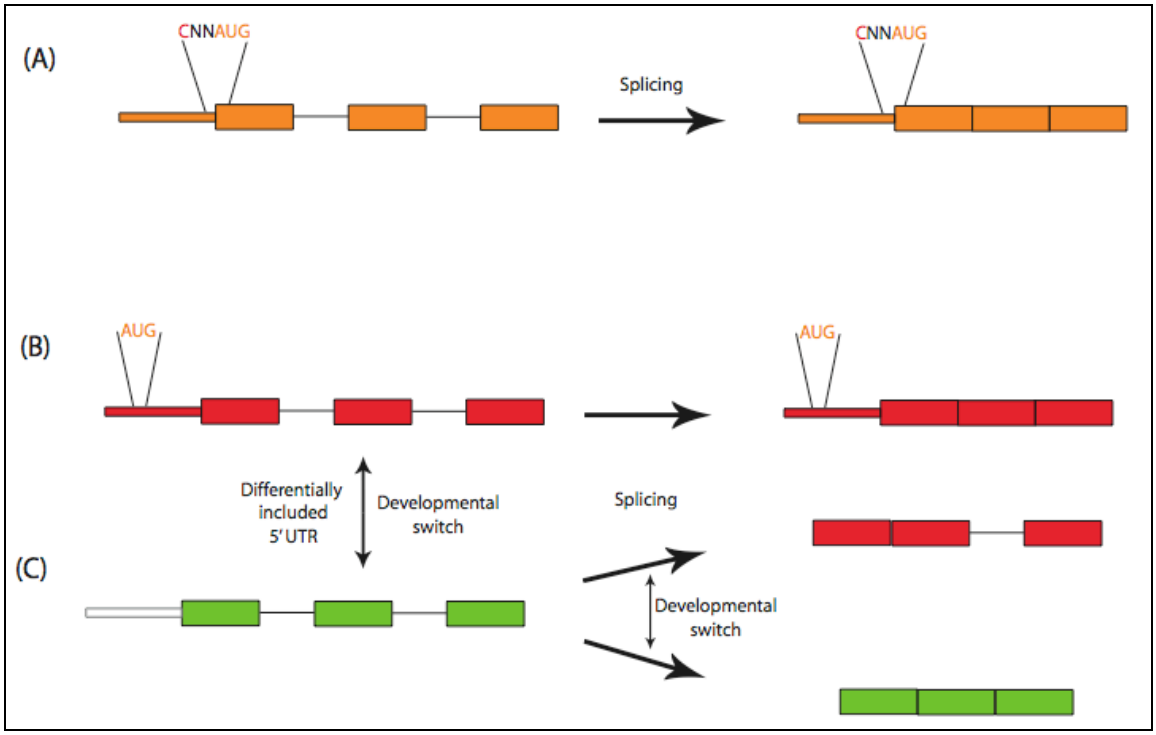


Figure 5.7. Model of gene regulation by NMD. Transcripts may have structural characteristics that predispose them to NMD, introduced either at the level of transcription or splicing. The presence of a weak translation initiation motif leads to the translation machinery occasionally skipping the correct start, leading to NMD (A). The presence of an upstream AUG (uAUG) in the 5' UTR of a transcript leads to translation of a uORF and NMD (B). The level of regulation of such transcripts may be determined in part by the translation recognition sequence at both the uAUG and correct AUG. Transcripts which are otherwise not predisposed to NMD may be spliced to normal or deleterious forms (C). NMD-dependent developmental regulation of gene expression is controlled by the regulated inclusion or exclusion of a uAUG containing 5' UTR or stage-specific splicing of transcripts to a deleterious form.

Given that we are considering genes candidates of GLD-1 regulation if they are NMD regulated in a GLD-1 dependent way, contain STAR-binding sites and are expressed in the germline perhaps an alternative approach would be more fruitful. If it is necessary to follow up any candidate genes from the array experiment with *in situ* hybridizations of the germline to see if the expression of the genes really is affected by NMD and loss of GLD-1 this is even more likely to be so. The approach to which I am insinuating would be to take all germline-expressed genes with 5' UTRs and search for uAUGs and STAR binding sites in those UTRs. Depending on the number

of candidate genes this yields one could proceed straight to *in situ* hybridizations of the germline for these RNAs in NMD and GLD-1 deficient animals without the necessity of a microarray experiment. An additional form of validation of GLD-1 targets would be the identification of all transcripts which co-immunoprecipitate with GLD-1. Both Ryder *et al*. and Lee and Schedl perform immunoprecipitation if GLD-1 followed by RT-PCR to confirm the binding of candidate transcripts. The detection of transcripts is limited by primers used for the RT-PCR and it seems logical that producing cDNAs from the recovered RNAs followed by microarray analysis or Illumina sequencing would reveal the transcripts present in a quantifiable way. The immunoprecipitation of ribonucleoproteins followed by the microarray analysis of bound mRNAs is known as RIP-chip and appears to be a very real option (Keene *et al*., 2006).

In summary then, our model is that the cell uses NMD to regulate gene expression via aspects of transcript sequence and programmed variation of transcript structure. This adds an extra level to steady-state regulation of gene expression, but also permits temporal regulation of gene expression by alteration of transcript structure. An extra dimension of this regulation is likely to be added by the protection of transcripts from NMD by RNAi binding proteins. This regulation may happen in a spatially and temporally controlled way. The extent of this regulation, however, is yet to be determined.