Analysis of the transcriptomes of wild-type and mutant *C. elegans*

Andrew Christopher Nelson

This dissertation is submitted for the
degree of Doctor of Philosophy
September 2008

Corpus Christi College
University of Cambridge

The Wellcome Trust Sanger Institute
Hinxton
Cambridge, UK

Declaration

I hereby declare that my dissertation contains material that has not been submitted for a degree or diploma or any other qualification at any other university. This thesis describes my own work and does not include work that has been done in collaboration, except when specifically indicated in the text.

Andrew C. Nelson

29/09/2008

Abstract

A key question in biology is how genotype can inform us of phenotype. For model organisms, most phenotypes reported have been at the level of the morphology and behaviour of the whole organism. However, recent advances in technology allow gene expression to be assessed on a genome-wide scale and pioneering work in yeast has shown that such expression profiles can be used as high density, quantitative phenotypes. I wanted to test whether expression profiles can also serve as useful phenotypes of whole animals rather than single cells. More specifically I sought to test whether the expression profiles resulting from perturbations of genes in one pathway looked more like those of other perturbations of the same pathway than another pathway. To do this I used two-colour DNA expression microarrays to survey gene expression in the nematode *Caenorhabditis elegans*. Expression profiles were produced for a number of different worm strains with mono-genic perturbations in different pathways involved in germline development. Clustering of the resulting expression profiles rediscovered the known pathways. This then allowed me to query perturbations of candidate modulators of EGF signalling against the compendium of expression profiles. I conclude that, as in yeast, expression profiles serve as reliable high-density phenotypes that allow meaningful biological comparisons to be drawn.

The quality of an expression microarray can only be as high as the gene annotations on which it is based. I therefore sought to evaluate how well characterised the transcribed genome of *C. elegans* is. To do this I used a combination of whole genome tiled microarrays and ultra-high density sequencing to assess the transcript complement of whole animals throughout development. We found that the vast majority (~95%) of expression is genic but the combinations and numbers of splice sites used are greater than previously predicted, suggesting that current annotations are largely complete, but that our knowledge of splice variation across development is still far from finished.

Whilst surveying transcripts in wild-type animals yields valuable data, it is known that there are many transcripts that are produced and subsequently degraded by the nonsense-mediated mRNA decay pathway (NMD). To identify these transcripts we compared the transcripts of wild-type animals to those of mutants of the NMD pathway. We find that ~13% of endogenous genes are NMD targets. The majority of these transcripts have upstream start codons in the 5' UTR or are alternatively spliced leading to a premature in-frame stop codon. Finally, we find that ~10% of all gene expression changes throughout development require NMD and thus that NMD is a bona fide regulator of gene expression.

List of Figures

List of Tables

Chapter 1
Introduction

1.1. Outline

On commencing my PhD studies using the nematode *Caenorhabditis elegans* there were two key questions that I wanted to address – 'How can evaluation of the transcriptome of an animal inform us of its physiological state?' and 'How well characterised is the transcriptome of *C. elegans*?' Recent advances in technologies to assess transcript levels, such as microarrays and ultra-high density sequencing make such goals more achievable and the outcome more comprehensive than was previously possible. In wild-type animals, however, the measured transcriptome is not completely representative of all transcripts produced. Rather post-transcriptional regulation leads to the degradation of certain transcripts. One such regulatory mechanism is nonsense-mediated mRNA decay (NMD), a pathway that detects and degrades transcripts with an in-frame premature termination codon. I therefore expanded my study to the NMD-deficient transcriptome in order to identify the targets of this pathway and to establish whether the structures of these targets and how these structures change throughout development indicate a role for NMD beyond that of surveillance mechanism.

Since the two questions I sought to address are distinctly different, although related, the following thesis is ordered accordingly, addressing the study of the former question to its conclusion in chapter 3 followed by the latter question and study of nonsense-mediated mRNA decay from chapters 4-5. Chapter 3 details the generation of expression phenotypes of genic perturbations using microarrays. The study itself focuses on signalling pathways that are involved in germline development and the comparison of candidate modulators of one of these signalling pathways to that of previously identified

components of the pathway. The methods of genic perturbation are mutation and RNAi. This introduction chapter therefore begins by introducing *C. elegans* as an appropriate and powerful model system for my studies. I will go on to discuss aspects of *C. elegans* physiology, focusing on the germline and vulva as systems for the study of inter- and intracellular signalling and their utility in my study. I will also discuss RNAi as a method of perturbing gene function in *C. elegans*.

Chapters 4-5 utilize methods of surveying the transcriptome using whole genome tiled microarrays and ultra-high density sequencing. I will therefore discuss the various methods of surveying gene expression, both at the level of annotated genes and for the genome as a whole.

1.2. *Caenorhabditis elegans* as a model system

The nematode *Caenorhabditis elegans*, a roundworm, was first established as a powerful model organism for genetic study in the laboratory of Sydney Brenner in the 1970s (Brenner, 1974). It has since become the tool of choice to a global community of research laboratories. Among the favourable attributes of *C. elegans* is a short life-cycle giving a rapid generation time of three days at room temperature. The animals develop though four larval stages (L1-L4) before reaching adulthood and becoming fertile. Adult worms are ~1mm in length and give rise to ~300 progeny. Worms can be maintained at minimal cost in the laboratory on agar plates or in liquid culture. Typically the worms are fed *Escherichia coli* but on starvation the animals enter a developmental programme that leads to a 'dauer stage' during which the worms can survive for months in the

absence of food. For long-term storage worms can also be frozen. *C. elegans* is therefore an extremely robust and practicable organism.

Under laboratory conditions *C. elegans* are maintained as hermaphrodites and reproduce by self-fertilisation, leading to a clonal population. It also contributes to the ease with which the animals can be propagated, as one hermaphrodite with unlimited food will lead to a population reproducing indefinitely. Furthermore it ensures that the measured differences between any treated population are as a result of the treatment alone. Classically gene function was established by performing genetic screens for mutants exhibiting a certain phenotype. Hermaphrodites are ideal for this as they automatically self-fertilize, negating the need for outcrossing in order to obtain homozygotes. This has lead to a vast collection of loss-of-function mutants, which are available to the global *C. elegans* community.

As a multi-cellular animal *C. elegans* is highly differentiated but its development is extremely well characterised. The essentially invariant somatic lineage of *C. elegans* gives rise to 959 cells in the adult, which encompasses the digestive and nervous systems, muscle, epidermis and other tissue types common to metazoans (Sulston and Horvitz, 1977; Sulston *et al*., 1983). The germline of the worm is a syncytium containing ~2000 nuclei in the adult (Kimble and White, 1981). The germline is a relatively well-studied tissue in terms of its development. Critically, the germline accounts for a large proportion of the transcripts in the adult animal and the expression of ~25% of genes are enriched in the germline. It is therefore highly amenable for study at the level of the

whole animal. Numerous expression analyses of this tissue have therefore already been performed, demonstrating the validity of such an approach. This is therefore the best tissue in which to study expression changes caused by the perturbation of different signalling pathways, as will be discussed in chapter 3.

1.2.1. The germline

Germ cells are specified during early embryogenesis, proliferating during larval development to form a multi-nucleate syncytium consisting of ~2000 nuclei in the adult germline. Although the germline is a syncytium the individual nuclei and the cytoplasm that surrounds them are often referred to as "germ cells" in the interests of conciseness. The developed germline consists of two gonad arms in the hermaphrodite, each comprised of multiple spatially distinct regions. The distal end of the germline contains a mitotic stem cell niche. As nuclei are produced they move through the germline reaching a transition zone where nuclei are stimulated to enter meiosis. Beyond the transition zone all nuclei are in transit through the meiotic cell-cycle prior to gametogenesis (Crittenden *et al*., 1994) (figure 1.1).

Broadly, between hatching and being a fully reproductive adult the germline of the worm goes through two phases – proliferation and maintenance. During L2 stage the number of mitotic nuclei increases and the germline elongates. During L3 stage germ cells continue to proliferate distally and undergo meiosis proximally starting at late-L3 stage. During L4 stage the germ cells continue to proliferate distally whilst spermatogenesis occurs proximally. Once the young adult stage is reached the germline ceases to proliferate but

mitotic nuclei still self renew in order to maintain the developed germline. Oogenesis proceeds proximally and the developed oocytes can be fertilized by the sperm produced during L4, leading to embryogenesis.

Three of the key pathways or machineries involved in germline development are the Notch pathway, the Ras/MAPK signalling arc and the RNA binding proteins that control the transition from mitosis to meiosis. These are the focus of the study detailed in chapter 3. The following sections of this chapter discuss germline development as a whole focusing on the role of these pathways.

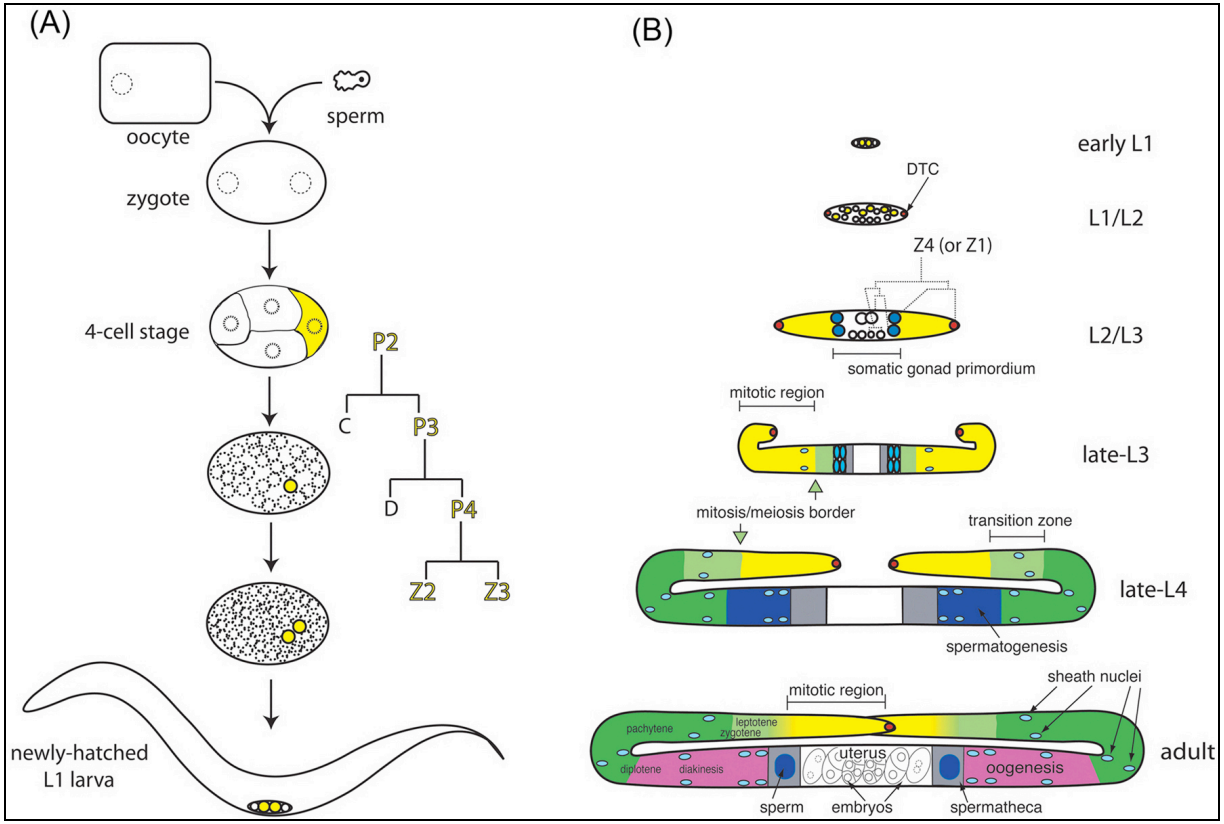


Figure 1.1. Cartoon representation of gonadogenesis. (A) Fertilization and the embryonic germ line: Fertilization of oocyte by sperm leads to embryonic development. Germline lineages are in yellow. (B) Post-embryonic hermaphrodite gonad development: Germline colour scheme: yellow = mitotic region, light green = transition (early prophase of meiosis I), dark green = pachytene, dark blue = spermatogenesis, and pink = oogenesis. In the adult, the mitosis/meiosis border is not sharp (mitotic and meiotic nuclei are interspersed at the border) as indicated here by a yellow/green color gradient. Somatic gonad color scheme: red = DTC, blue = sheath/spermatheca precursor cells, light blue = sheath nuclei, grey = spermatheca, and white = uterus. NB: Comparative size of gonads at different stages is not to scale. (Taken from Hubbard and Greenstein, 2005)

1.2.1.1. Germline specification, early development and Notch signalling

The germline is specified early during *C. elegans* embryogenesis, at the 4-cell stage. The cell designated P4 is the germline founder cell from which all germ cells are derived and which makes no contribution to the somatic lineage (Sulston *et al*., 1983). P4 undergoes only one cell division before the developed embryo hatches. This cell division gives rise to cells designated Z2 and Z3. These cells are flanked by Z1 and Z4, which give rise to the somatic gonad from which the distal tip cells (DTCs) are derived (Sulston *et al*., 1983). Post-embryonic germ cell divisions only begin when the nutritional environment is favourable (Kimble and Hirsh, 1979). Experimentally this is hugely advantageous as it means that vast quantities of worms can be hatched in the absence of food and will arrest. They then develop synchronously once food is supplied. It is by this method that all synchronous populations for expression study in this thesis were produced.

The gonad remains four cells until mid-L1, when Z1 and Z4 proliferate to form 12 somatic cells before L2 stage, including the DTCs. The fully developed somatic gonad consists of 143 cells forming structures such as the spermatheca and uterus (Kimble and Hirsh, 1979). During L2 and L3 stages the germline increases to ~100 nuclei. The majority of germline expansion and development occurs during L4 and young adult stages, giving a total germline complement of ~2000 nuclei (Kimble and White, 1981).

At the distal end of each gonad arm Notch pathway signalling from the DTC suppresses meiosis in the surrounding germline nuclei, thus establishing a distal mitotic zone in the germline. The DTC is known to be necessary and sufficient for the maintenance of this

mitotic stem cell niche as laser ablation of the DTC causes all mitotic nuclei to enter meiosis and duplication or transplantation of the DTC establishes new mitotic niches (Kimble and White, 1981).

There are two homologous Notch receptors in *C. elegans*, LIN-12 and GLP-1 known to share some redundant functions (Austin and Kimble, 1989; Lambie and Kimble, 1991; Yochem and Greenwald, 1989; Yochem *et al*., 1988). The receptors are activated by an overlapping set of ligands and activate transcription via association with nuclear proteins (Chen and Greenwald, 2004; Christensen *et al*., 1996; Petcherski and Kimble, 2000). These ligands are known as the Delta/Serrate/Lag2 (DSL) ligands.

The accepted model of Notch signalling is that the binding of the ligand by the receptor leads to the proteolytic cleavage of the intracellular domain of the receptor. The released domain then associates with transcriptional activators to drive the expression of their target genes (Schroeter *et al*., 1998). The Notch pathway as it is known to act in the germline consists of the Notch ligand LAG-2, Notch receptor GLP-1, and the pathway-specific transcriptional activators LAG-1 and SEL-8 (LAG-3) (figure 1.3). LAG-2 is expressed by the somatic DTC whereas GLP-1 is expressed in the germline. The location of these two key proteins is tightly regulated in two mechanistically distinct ways. LAG-2 is tethered to the surface of the DTC via a transmembrane domain. Expression of LAG-2 without the transmembrane domain leads to the establishment of ectopic mitotic regions within the germline (Fitzgerald and Greenwald, 1995; Henderson *et al*., 1997). *glp-1* mRNA exists throughout the germline. Its translation is repressed everywhere

other than at the distal end of the germline which I shall discuss later. Loss-of-function of any of the core Notch signalling components leads to the nuclei at the distal end of the germline entering meiosis. As a consequence the nuclei complement of the germline is not replenished and the worm is sterile (Austin and Kimble, 1987; Doyle *et al*., 2000; Lambie and Kimble, 1991; Petcherski and Kimble, 2000). Conversely, unregulated GLP-1 and LAG-2 are known to lead to unregulated germline mitoses and consequent germline tumours (Berry *et al*., 1997; Fitzgerald and Greenwald, 1995; Henderson *et al*., 1997; Pepper *et al*., 2003). The complete complement of Notch targets that lead to the suppression of meiosis is unknown. Genetic screens have revealed enhancers of *glp-1*, however, the mechanism of these interactions is yet to be fully explored (Qiao *et al*., 1995; Sundaram and Greenwald, 1993). Furthermore a protein that physically interacts with the intracellular domain of both LIN-12 and GLP-1 has been identified. Called EMB-5, it is thought to act downstream of GLP-1 and is required for correct germline development (Hubbard *et al*., 1996).

GLP-1 activation leads to the transcription of *fbf-2* via the four LAG-1 binding-sites in its 5' flanking region. *fbf-1*, however, does not appear to be transcribed in response to Notch signalling and the mechanism by which this occurs is unknown. FBF-1 and FBF-2 regulate each other to dictate the size of the mitotic region of the germline (Lamont *et al*., 2004). FBF-1 and FBF-2, known collectively as FBF are almost identical and largely functionally redundant. Loss of either protein leads to a fully functional germline, albeit with differing sizes of mitotic region. The double mutant, however, reveals that FBF is essential for the maintenance and not proliferation of the germline as the germline

develops normally until spermatogenesis when the mitotic nuclei enter meiosis rather than continuing to self-renew (Crittenden *et al*., 2002).

Notch signalling preserves the mitotic character of the distal end of the proliferating germline and the maintenance of this mitotic stem cell niche in the developed germline. The nuclei in this niche meet the criteria to be considered stem cells as they are self-renewing and produce differentiated progeny (Watt and Hogan, 2000). Notch signalling is conserved in metazoans and appears to be conserved in the role of promoting stem cell proliferation (Calvi *et al*., 2003; Gaiano and Fishell, 2002). Understanding how Notch signalling regulates stem cell proliferation and maintenance in *C. elegans* may therefore be very relevant to human biology.

1.2.1.2. Regulation of the mitosis/meiosis switch in germline development

The switch from mitosis to meiosis in the germline is regulated by a complex network of Notch effectors and suppressors. RNA binding proteins which regulate the mitosis/meiosis switch are another focus of chapter 3. A simplified network diagram of the regulation of the mitosis/meiosis switch is shown in figure 1.2.

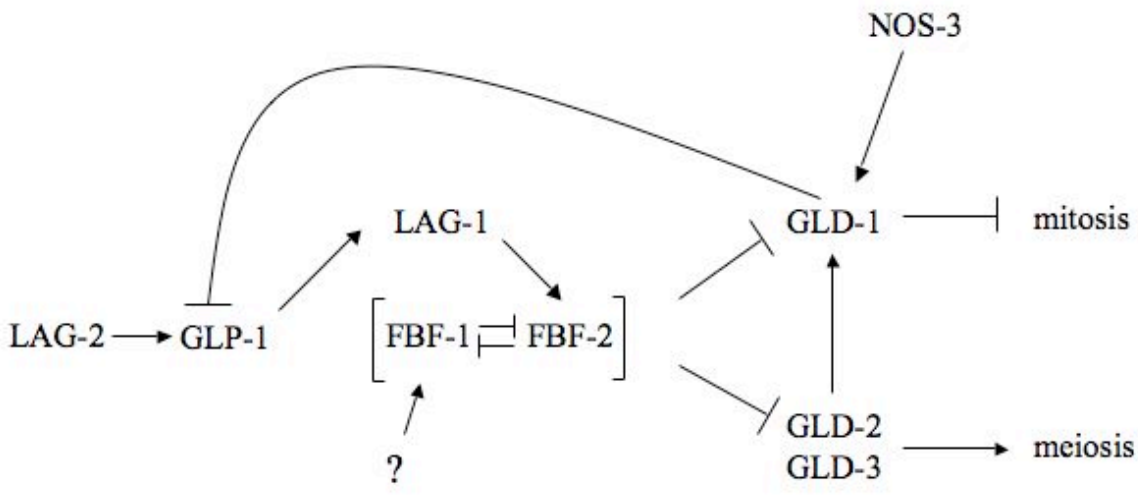


Figure 1.2. Regulation of the mitosis/meiosis decision by the interplay of pro- and anti-meiotic factors. Notch signalling activates anti-meiotic factors but is in turn suppressed by GLD-1 permitting entry into meiosis, stimulated by GLD-2 activation of pro-meiotic targets. This circuitry provides only a partial explanation of the mitosis/meiosis switch. (Modified from Hubbard and Greenstein, 2005)

As nuclei from the distal mitotic niche move more proximal GLD-1, GLD-2, GLD-3 and NOS-3 regulate entry into meiosis in a post-transcriptional way (Eckmann *et al*., 2004; Hansen *et al*., 2004a; Hansen *et al*., 2004b; Kadyk and Kimble, 1998). This sets up a transition zone in the germline consisting of nuclei undergoing mitosis and nuclei undergoing meiosis. *glp-1* mRNA is present throughout the germline but the protein is only found in the distal mitotic zone. Promotion of meiosis in the transition zone occurs (at least in part) due to the translational repression of *glp-1* mRNA by GLD-1, which binds its 3' UTR. This relieves the Notch controlled suppression of meiosis (Marin and Evans, 2003; Ryder *et al*., 2004). GLD-1, however, is suppressed by FBF, as is GLD-3, which acts as an activator of pro-meiotic targets (Crittenden *et al*., 2002; Eckmann *et al*., 2004). FBF-2 is spatially localized to the most distal end of the germline, thus

determining the position of the transition zone (Lamont *et al*., 2004). Activation of pro-meiotic targets by GLD-3 is thought to be via the poly(A) polymerase activity of GLD-2 on its target mRNAs, allowing them to be translated. This is supported by evidence that the two proteins physically interact *in vivo* and GLD-3 promotes GLD-2 activity *in vitro* (Eckmann *et al*., 2004; Eckmann *et al*., 2002; Wang *et al*., 2002). FBF may act in opposition to this to prevent meiosis by preventing GLD-3 expression and consequent binding to its targets, of which one is *gld-1* (Eckmann *et al*., 2004). FBF-1 and FBF-2 are members of the PUF family of RNA binding proteins. It is known in yeast and *Drosophila* that PUF proteins mark their targets for deadenylation and it is possible that the same occurs in *C. elegans* (Olivas and Parker, 2000; Wreden *et al*., 1997). The mechanism of the mitosis/meiosis switch therefore is one of FBF repression of pro-meiotic targets switching to GLD-2 activation of targets. This is not to say that there is significant overlap between GLD-2 and FBF targets. The targets of both FBF and GLD-2 are largely unknown but importantly it is known that they both regulate *gld-1*. The precise mechanism by which this switch occurs is unknown although a number of speculative models have been proposed. These models, however, are oversimplifications. It is known that loss of GLD-2 does not prevent entry into meiosis, nor does the loss of any of the individual components previously mentioned. The true mechanism by which the mitosis/meiosis switch occurs is therefore clearly extremely complicated and only partially understood.

Many of the proteins cited as being involved in the mitosis-meiosis switch also appear to be implicated in the sperm/oocyte fate decision and so germline development and sex-

determination appear to be highly linked processes. Gametogenesis begins at the L4 stage with spermatogenesis and switches to oogenesis from young adulthood. GLD-1 promotes spermatogenesis, as does GLD-3 (Eckmann *et al*., 2002; Francis *et al*., 1995). Additionally, GLD-2 loss-of-function is known to lead to cell cycle arrest in meiotic prophase and so may be required for spermatogenesis along with GLD-1 and GLD-3 (Kadyk and Kimble, 1998). FBF is involved in the switch from spermatogenesis to oogenesis and NOS-3 also promotes oogenesis (Kraemer *et al*., 1999; Zhang *et al*., 1997). This is in contrast to the mitosis/meiosis switch where NOS-3 acts in concert with GLD-1 to relieve Notch induced suppression of meiosis (Hansen *et al*., 2004b).

1.2.1.3. Progression beyond the pachytene stage of meiosis

Progression beyond the pachytene stage of the meiotic prophase requires mitogen-activated protein kinase (MAPK) signalling and is another focus of chapter 3. Loss-of-function of numerous components of the classical EGF/ras/MAPK signalling pathway result in sterile worms for this reason, as revealed by staining and detailed microscopy (Chang *et al*., 2000; Church *et al*., 1995; Hsu *et al*., 2002; Ohmachi *et al*., 2002). Figure 1.3 illustrates the canonical EGF/ras/MAPK signalling pathway, highlighting the components known to be required for pachytene release. The downstream targets of MPK-1 involved in meiosis are unknown. Likewise, neither are the upstream activators of SOS-1 known. Consequently from here onwards this signalling in the germline will be referred to as Ras/MAPK signalling as no upstream ligand or receptor tyrosine kinase has been defined. After exit from pachytene the meiotic nuclei become completely compartmentalised as cells and terminally differentiated as either sperm or oocytes.

Since many of the factors that are involved in pachytene release are yet to be determined this is clearly a research area with much remaining potential.

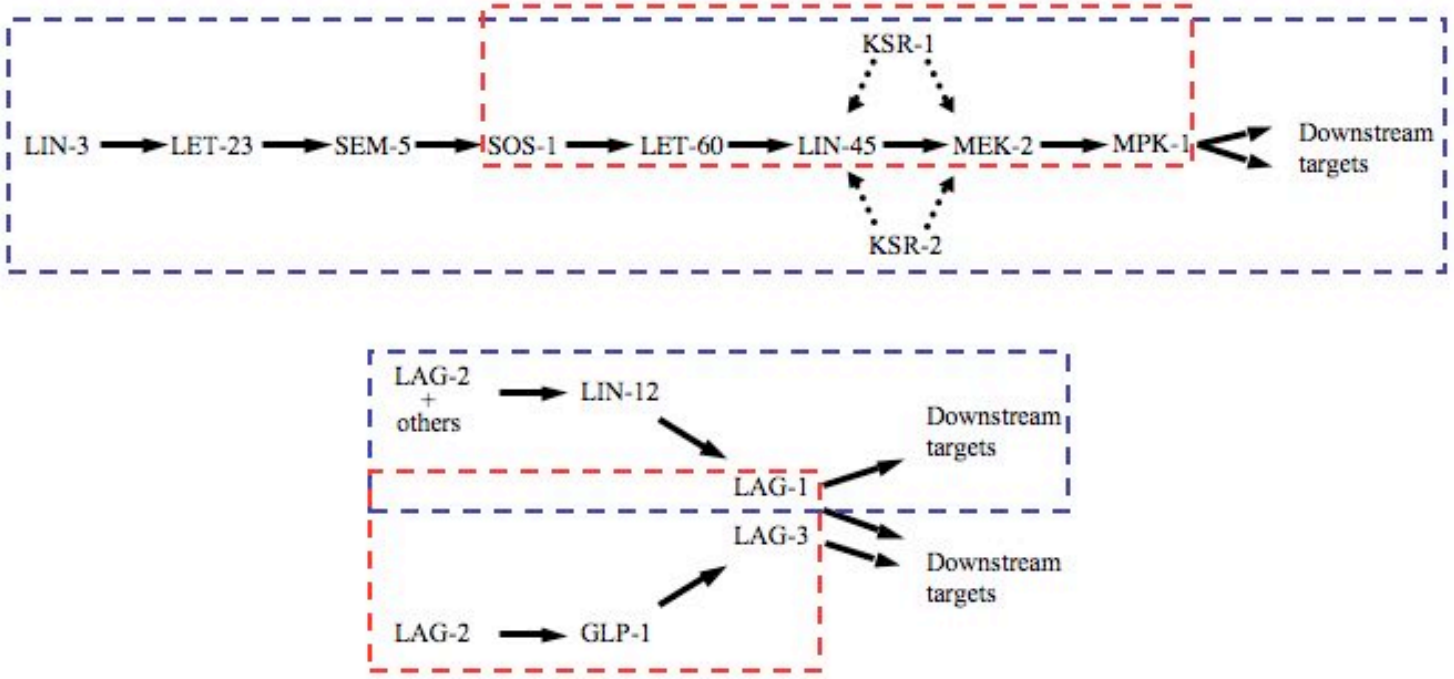


Figure 1.3. The canonical EGF/ras/MAPK and Notch signalling pathways as they are known to act in the vulva and germline. The EGF/ras/MAPK signalling pathway is shown at the top and Notch at the bottom. Components known to act in the vulva are outlined in blue and the germline in red. The classic model of the EGF/ras/MAPK pathway involves the activation of an RTK by ligand binding (components 2 and 1 in the flow-through), followed by a cascade of protein activations as indicated by the arrows. KSR-1 and KSR-2 act as scaffold proteins, which assist in the activation of LIN-45 and/or MEK-2, as indicated by the dotted arrows. Notch signalling acts by proteolytic cleavage of the intracellular domain of the receptor on ligand binding. The now free intracellular domain translocates to the nucleus and activates down-stream targets in consort with various transcriptional activators. Whereas many of the downstream targets of both signalling pathways in the vulva are known, downstream targets of these pathways in the germline are yet to be determined.

Here I have discussed a number of the key pathways and processes involved in germline development, and how their perturbation leads to germline defects and sterility. Some key common features of all three pathways and machineries discussed are that they are conserved between *C. elegans* and mammals, and their downstream targets and effectors in the *C. elegans* germline are either partially or completely unknown. This is therefore a potentially fertile area of research. Methods are clearly required to identify potential targets of these pathways and to confirm this role. That is the ultimate goal of chapter 3. The method of identifying potential candidates of involvement in these pathways is by screening for genes that modulate the phenotype of mutants in these pathways using RNAi. RNAi in the worm is a simple means of generating loss-of-function phenotypes as will be discussed later in this chapter. Another key feature of the Notch and EGF/ras/MAPK pathways is that they are known to act in other tissues in the worm. One of the best-studied tissues for which this is the case is the vulva. I will therefore go on to discuss the roles of Notch and EGF/ras/MAPK signalling in the vulva and how screens in this tissue can identify candidate modulators of these pathways. The chosen method of confirming the roles of candidate genes in the germline is by comparison of molecular phenotypes generated using expression microarrays with genic perturbations in these pathways. I will therefore go on to discuss the principles of DNA microarrays and their use as phenotyping tools.

1.2.2. The vulva

The *C. elegans* vulva is an extremely well studied tissue that shares signalling pathways that are involved in germline development. It provides a simple model of organogenesis involving the interaction of well-studied signalling pathways. The early identification of EGF/ras/MAPK signalling as being involved in vulval development was considered most interesting given that the EGF/ras/MAPK signalling pathway has long since been known to be dysregulated in many human cancers. This perhaps served as the catalyst for widespread study of the *C. elegans* vulva.

Vulval development begins with the specification of 6 multipotent vulval precursor cells (VPCs), designated P3.p – P8.p, along the ventral axis of the worm during L1 and L2. Whilst P5.p, P6.p and P7.p develop into the 22-cell vulva, the remaining three cells divide to produce cells that fuse with the syncytial epidermis. These cells were identified in ablation studies as having the potential to develop into vulval tissue in response to intercellular signalling events (Kimble, 1981; Sternberg and Horvitz, 1986; Sulston and White, 1980).

Key to the ability of VPCs to develop into the vulva is the expression of the Wnt- and EGF-responsive Hox gene *lin-39*. Expression of this gene in P3.p – P8.p is required to prevent fusion of these cells with the epidermis and in cooperation with *eff-1*, to permit correct cell division. Wnt signalling via *bar-1* has been shown to be required for *lin-39* expression. It has since been demonstrated that the expression of *lin-39* is co-ordinately regulated by Wnt and EGF signalling (Eisenmann *et al*., 1998). The EGF/ras/MAPK

signalling pathway therefore has a role in ensuring the competence of VPCs to generate vulval tissue.

Signals received from the anchor cell (AC) located above P6.p in the somatic gonad (see figure 1.4) leads to the specification of these cells as either 1^o, 2^o or 3^o in the order 3^o, 3^o, 2^o, 1^o, 2^o, 3^o. There are differing models for how the EGF signalling from the AC leads to the establishment of the different VPC fates, a graded signalling and sequential signalling model. The graded signalling model suggests that the different VPC cell fates are determined by the dose of the EGF signal (LIN-3) as a consequence of the distance of each cell from the AC (Katz *et al*., 1995; Sternberg and Horvitz, 1986). This model cannot be completely correct, however, as it has been demonstrated that only P6.p, which adopts the 1^o cell fate, need express the EGF receptor tyrosine kinase (RTK), LET-23, for correct vulval development to occur (Simske and Kim, 1995). This led to the theory of a sequential signalling model. This model postulates that specification of the 1^o cell leads to a consequent signal specifying the 2^o cell fate. This signal has been identified. Termed the "lateral signal", it has been demonstrated that LIN-12/Notch signalling from the 1^o cell leads to the adoption of 2^o fates in its flanking cells (Chen and Greenwald, 2004; Greenwald *et al*., 1983; Sternberg, 1988). Specifically, the Notch ligands LAG-2, APX-1 and DSL-1 signal from the 1^o cell to promote 2^o cell fates in the adjacent cells. This effect is dependent on the LIN-3 signal (Chen and Greenwald, 2004). Further evidence suggests that the downregulation of the Notch receptor, LIN-12 in the 1^o cell is required for the transmission of the lateral signal to the adjacent cells. This acts through the endocytosis of LIN-12 as a result of signalling via LET-23 inducing changes in

transcription (Shaye and Greenwald, 2002). This downregulation of LIN-12 in P6.p is important for the 2° cell specification of P5.p and P7.p. This may suggest that the sensitivity of P6.p to Notch signalling modulates the outcome of EGF signalling in this cell. Signalling via LIN-12 therefore appears to oppose the outcome of EGF signalling via LET-23. It seems reasonable to postulate therefore that the graded LIN-3 signal received by the cells destined for 2° cell fates is counteracted by Notch signalling from the 1° cell. It has since been demonstrated that a number of the targets of LIN-12 signalling in P5.p and P7.p are negative regulators of LET-23 signalling (Yoo *et al*., 2004). A model for the specification of 1° and 2° cell fates is one of the LIN-3 signal being received by P6.p leading to an upregulation of transmission of the Notch signal and a downregulation of reception of the Notch signal. P6.p is now specified as the 1° cell. Reception of the Notch signal by P5.p and P7.p leads to a counteraction of the LIN-3 signal received from the AC. This blocks the specification of the 1° fate while activation of LIN-12 targets leads to the specification of the 2° cells. This is graphically represented in figure 1.4. The 1° and each 2° VPC then go through a series of divisions resulting in 8 cells from the 1° VPC and 7 from each of the 2°, totalling 22 cells in the fully developed vulva. The 3° cells divide to produce cells, which then fuse with the syncytial epidermis.

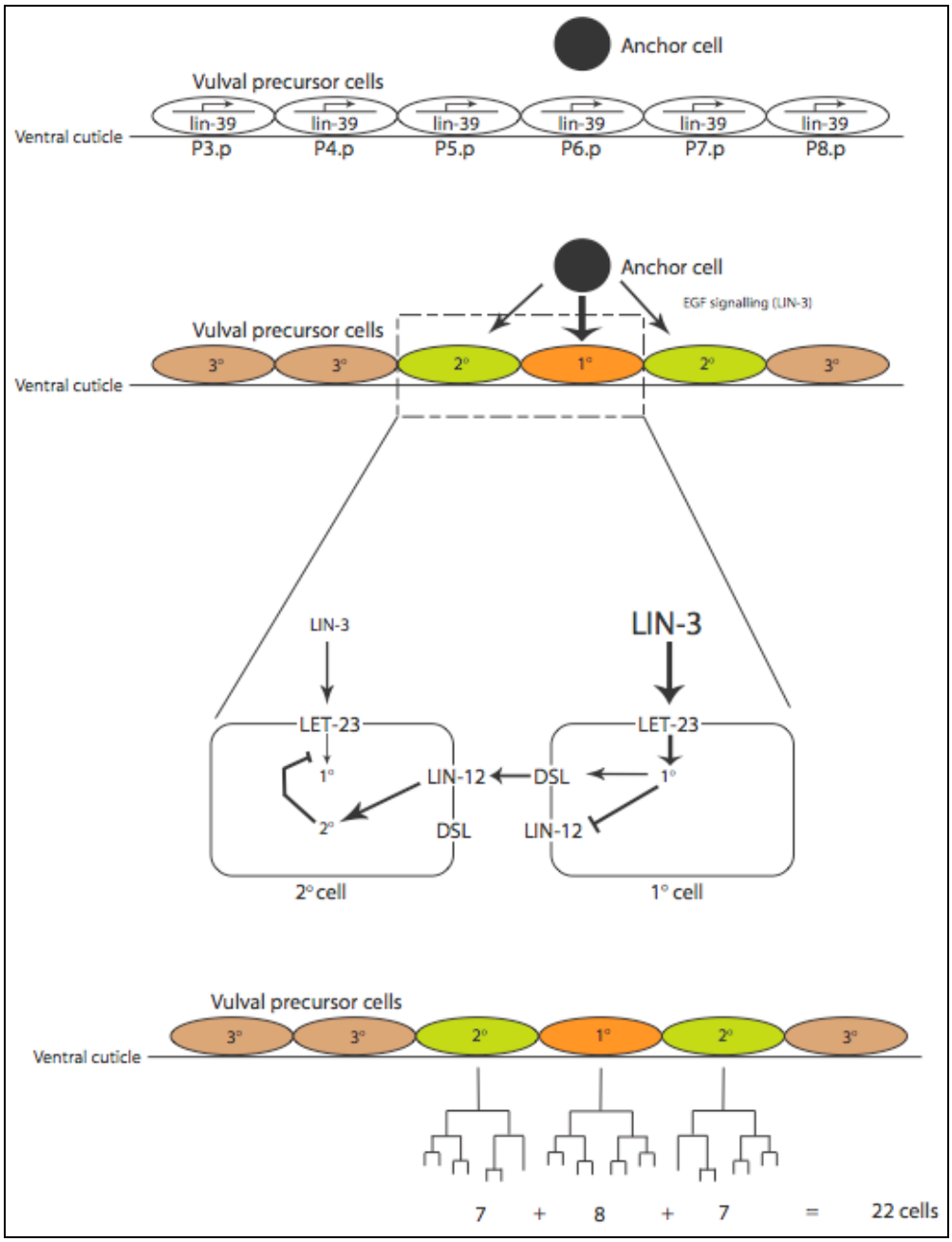


Figure 1.4. Vulval specification and lineage. Expression of *lin-39* imparts the potential on six cells (P3.p-P8.p) along the ventral axis of the worm to adopt vulval fates. EGF signalling from the anchor cell, part of the somatic gonad, leads to the specification of these cells as either 1^{o}, 2^{o} or 3^{o} as shown. EGF signalling leads to the specification of the 1^{o} cell fate in its closest VPC cell. This in turn leads to an increase in LIN-12/Notch signalling (DSL-type ligands) from the 1^{o} cell and a reduced sensitivity to LIN-12/Notch signalling. This LIN-12/Notch lateral signal received by the cells adjacent to the 1^{o} cell promotes 2^{o} cell specification whilst suppressing 1^{o} cell specification. This results in an invariant arrangement of cell fates. The 1^{o} and 2^{o} cells then go through a series of divisions to give a 22-cell vulva while 3^{o} cells produce cells which then fuse with the syncytial epidermis.

1.2.2.1. Identification of modulators of EGF and Notch signalling in the vulva

As discussed, both EGF and Notch signals are required for vulval development. Our interest in the vulva in the context of this thesis is as a tissue in which to identify candidate modulators of these pathways. Loss of regulation of either of these pathways leads to the acquisition of 1^o and 2^o fates by other cells and the development of pseudovulval protrusions consisting of 2^o cell descended tissue (Notch gain-of-function) or 1^o and 2^o cell descended tissue (EGF/ras/MAPK gain-of-function). The phenotype of animals exhibiting multiple vulvae is termed Muv. Mutants exhibiting these phenotypes have been identified in genetic screens for animals exhibiting vulval lineage defects (e.g. Ferguson and Horvitz, 1985; Han *et al*., 1990; Horvitz and Sulston, 1980). Identification of genic perturbations that modulate the Muv phenotype identifies candidate modulators of the dysregulated pathway leading to the phenotype. This has already been done to great effect (e.g. Bender *et al*., 2007; Han *et al*., 1990; Poulin *et al*., 2005; Wu and Han, 1994).

The mutations that lead to the Muv phenotype can be split into three categories; gain-of-function mutations of components of the pathways, loss-of-function of targets negatively regulated by the pathways, and loss-of-function mutations of suppressors of the pathways. Examples of the first type are clear, such as *let-60* and *lin-12* gain-of-function mutants. Loss-of-function *lin-1* and *lin-31* are examples of the second type. Both are transcriptional activators and direct targets of MPK-1 phosphorylation, leading to their inactivation. Loss-of-function *lin-1* and *lin-31* therefore mimic constitutively active EGF/ras/MAPK signalling in the vulva (Tan *et al*., 1998). The quintessential example of

the latter category is the synMuv genes. The synMuv genes (named for "Synthetic Multivulva") were originally identified as two redundant sets of genes which promote the specification of VPC fates when perturbed in combination (Ferguson and Horvitz, 1989). Evidence suggests that the synMuv genes act by opposing LIN-3 signalling from the hypodermis by repressing *lin-3* transcription, or transcription of genes upstream of *lin-3* (Cui *et al.*, 2006). Genetic mutants carrying lesions in both synMuv class A and class B genes therefore exhibit the Muv phenotype due to an increase in EGF signalling.

To reiterate, screening for genes that modulate the Muv phenotype in any of these classes of Muv mutants is to identify candidate modulators of the pathways involved in VPC specification and vulval development. The most straightforward method of performing such screens in *C. elegans* is by RNA-mediated interference (RNAi). This is a key tool in the context of this thesis. Our chosen method of providing further evidence of the involvement of these candidate genes in pathways is by comparison of perturbations of these genes to those confirmed to be involved in the pathways by microarray phenotype. The majority of these perturbations will be performed by RNAi owing to its ease of execution and the scarcity of appropriate genetic mutants. The precise rationale and methodology of the approach will be detailed in chapter 3. RNAi in *C. elegans* is discussed next.

1.3. RNA interference in *Caenorhabditis elegans*

RNA interference (RNAi) is a phenomenon by which introduction of double-stranded RNA (dsRNA) into a biological system gives a sequence-specific knock-down of the complementary mRNA. This phenomenon was first discovered in *C. elegans* when it was seen that injecting dsRNA into the germline or the extracellular cavity of the worm resulted in an interference effect throughout the animal, demonstrating the ability of the dsRNA to cross cell boundaries (Fire *et al*., 1998). It was then shown that feeding of worms with bacteria expressing dsRNA also gives the same systemic RNAi effect (Timmons and Fire, 1998). Finally it was discovered that soaking worms in a buffer containing dsRNA had the same effect, also having an effect in the progeny (Tabara *et al., 1998)*.

1.3.1. The mechanism of dsRNA-induced gene silencing in *C.elegans*

The mechanism giving the observed RNAi effect in *C. elegans* can be split into two different catagories – the spreading of dsRNA throughout the animal and the silencing effect of the dsRNA in the cell. Screens for mutants deficient in RNAi have uncovered genes in both categories. The first class consists of genes that were identified in mutants whose sensitivity to RNAi is dependent on the delivery method or location of dsRNA (Winston *et al*., 2002; Winston *et al*., 2007). The second class consists of genes that are absolutely essential for RNAi. Much of our current knowledge and understanding of the mechanism of gene silencing comes from genetic studies in *C. elegans* and plants, as well as biochemical studies on *Drosophila* embryonic and S2 cell extracts (reviewed in Boisvert and Simard, 2008; Filipowicz, 2005; Hannon, 2002; Joshua-Tor, 2006; Matzke

and Birchler, 2005; Zamore and Haley, 2005). Dicer, an evolutionarily conserved member of the RNase III ribonuclease family cleaves dsRNA into ~22nt fragments with a 2nt 3' overhang and a 5' phosphate group. The resulting small interfering RNAs (siRNAs) are then incorporated into the RNA-induced silencing complex (RISC), a ribonuclease-containing protein complex, which targets RNAs complementary to the siRNAs for degradation. Recognition of RISC targets is by base pairing between the siRNA and its target. Endonucleolytic degradation of the target is performed by Slicer, which is the catalytic core of RISC, in an ATP-dependent manner. Slicer is a member of the Argonaute family and contains two RNA binding domains, the Piwi and PAZ domains.

Although RNAi is a conserved phenomenon in metazoans and the core gene silencing machinery is conserved it is striking that the systemic nature of RNAi is not present in *Drosophila* or mammals. Further to this, it has been shown that the effects of RNAi in *C. elegans* can persist in subsequent generations by passage through the germline. This does not appear to be the case in *Drosophila* or mammals. It appears, therefore, that whilst the core machinery is conserved, there are key differences in the global mechanisms of RNAi between organisms that must reflect their different biological requirements.

RNAi in *C. elegans* does not share many of the key experimental problems observed in more complex organisms. In mammalian systems for example siRNAs must be added to cells in culture as introduction of longer dsRNA elicits the so called "interferon

response". Consequently larger dsRNA (typically ~1kb) is used for RNAi in *C. elegans*, which is cleaved by Dicer to produce many siRNAs targeting the same gene.

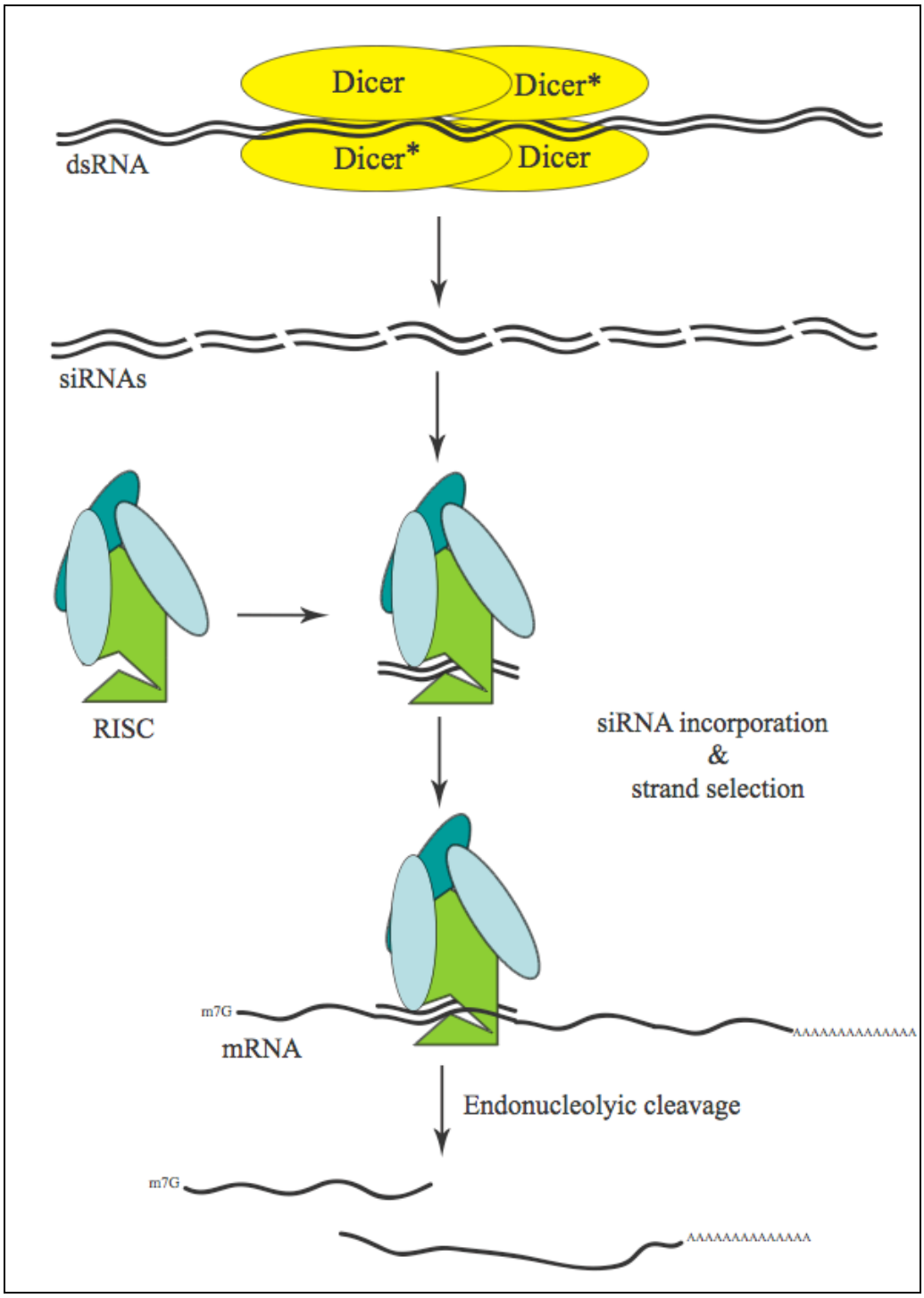


Figure 1.5. Mechanism of RNAi gene silencing. Two Dicer homo-dimers associate in anti-parallel orientation to cleave dsRNA. Only one catalytic centre in each Dicer homo-dimer is active (*). Active catalytic domains are spaced by ~22nt giving (siRNAs) of that length. siRNAs are incorporated into the RISC which is activated through unwinding of siRNAs. Watson-Crick base-pairing with siRNAs identifies homologous target mRNAs. The Piwi domain of the ribonuclease Slicer mediates cleavage of target mRNA.

1.3.2. RNAi by feeding

As previously stated, RNAi in the worm can be initiated by injection of or immersion in dsRNA, or by feeding worms with bacteria expressing dsRNA. The penetrances of RNAi phenotypes achieved by immersion or feeding are not as strong as by injection, but RNAi by feeding does have some key advantages (Timmons *et al*., 2001). Firstly, it does not require the costly *in vitro* synthesis of dsRNA. Secondly, the bacterial strains produced are a renewable resource that can be used indefinitely. There is no meaningful limit on the number of worms that can be fed a given bacterial strain, whereas only a relatively small number of animals can be injected in a given time period. In order to capitalize on these advantages a library of RNAi feeding strains each targeting one of 16,757 genes (~86% of annotated genes) was produced in the laboratory of Julie Ahringer (Fraser *et al*., 2000; Kamath *et al*., 2003; Kamath *et al*., 2001). This library has since been made available to the global community and was at my disposal for the duration of my PhD studies. The library consists of RNase III-deficient *Escherichia coli* strain HT115(DE3), transformed with a bacterial plasmid vector containing a 1-1.5kb PCR product corresponding to the gene of interest flanked by bacteriophage T7 promoters. HT115(DE3) is engineered to express T7 RNA polyerase under an isopropyl-β-D-thiogalactopyranoside- (IPTG-) inducible promoter (Timmons *et al*., 2001). Worms are then fed on agar plates containing IPTG and seeded with these bacteria and the loss-of-function phenotype assessed. Whilst the RNAi library was originally designed to provide one clone per gene, changes in gene predictions have since indicated that for some genes there are multiple clones.

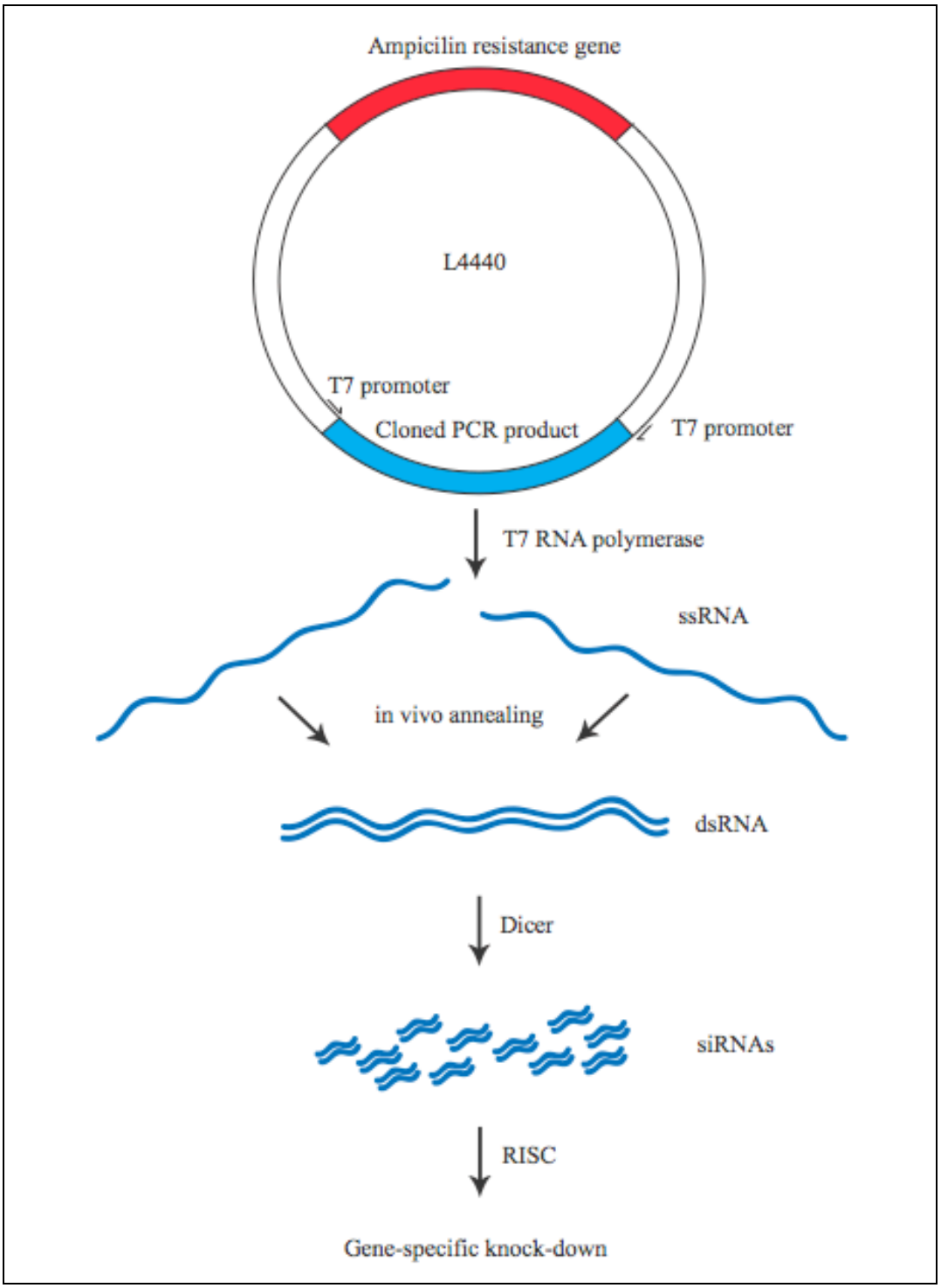


Figure 1.6. L4440 RNA interference feeding vector. A PCR product homologous to a target gene of interest is cloned between inverted T7 promoter sites. The vector is then transformed into an *Eschericia coli* strain expressing T7 RNA polymerase (HT115(DE3)), resulting in transcription of anti-parallel single-stranded RNAs. These RNAs anneal and form double-stranded RNAs (dsRNAs), which trigger RNA interference.

In the study detailed in chapter 3 the key method of phenotyping each individual RNAi perturbation is by microarray expression profiling. The next section discusses the essential qualities of DNA microarrays and their applications.

1.4. Microarray technologies

The sequencing of the genomes of many organisms demanded the creation of new technologies to capitalize on this advance. One such technology is the microarray that comes in numerous different formats and types and has many different applications. Consequently microarrays are the main platform used throughout the work contained in this thesis.

Broadly a DNA microarray is a large collection of DNA molecules arrayed on a solid support. Genomic microarrays are comprised of DNA molecules that tile a given region of the genome. Expression microarrays on the other hand contain DNA molecules, which are complementary to annotated genes. These DNA molecules, which are also referred to as "probes", may be PCR products derived from genomic DNA, synthetic oligonucleotides, or in the case of expression microarrays they may be derived from cloned cDNAs. In such cases the probes are then arrayed by a robot on glass slides treated in such a way that the probes adhere strongly (e.g. poly-L-lysine, epoxy or animo-reactive silane). These microarrays are generally used in two-colour applications, where a mixture of two samples each labelled with a different fluorophore (typically Cy3 and Cy5) are competitively hybridized against each other on the microarray and the ratios of the different fluorophores assessed (Duggan *et al*., 1999; Schena *et al*., 1995).

Other microarray types involve the *in situ* synthesis of oligonucleotides by photolithography, programmable optical mirrors or an ink-jet device (Hughes *et al*., 2001; Lipshutz *et al*., 1999). This has the potential of producing higher-density microarrays with a more consistent concentration of probe per spot. Such microarrays are often used for one-colour experiments where only one sample is hybridized per array and differences inferred between microarrays.

Both one- and two-colour microarrays are suitable for most applications. Choosing a microarray for a given application generally involves striking a balance between availability, cost and reliability. The two most common applications of microarrays are the assessment of the RNA complement of a sample and the assessment of the DNA complement of a sample. Assessing the RNA complement of a sample (which can be referred to as expression profiling) is effectively taking a measurement of the relative levels of all transcribed regions of the genome that can be detected using your microarray of choice. This represents the earliest use of DNA microarrays as reported in *Arabidopsis thaliana* (Schena *et al*., 1995). Expression studies using DNA microarrays have since been used to study many different aspects of biology such as tissue development (e.g. Reinke *et al*., 2000; Reinke and White, 2002), sex-specific aspects of development (e.g. Reinke *et al*., 2004), disease (e.g. Petricoin *et al*., 2002), elucidation of gene function (e.g. Hughes *et al*., 2000) and many others. The ability to draw direct comparisons between transcript complements either over time or between comparable conditions are key to all of these studies. An expression microarray where the probes are

designed against constitutive exons of annotated genes offers the simplest option in such studies, both in terms of experimental complexity and analysis. This assumes that gene predictions are correct and gives no information about the structure of the RNAs in the sample, nor does it provide information on novel RNAs. If any of these factors are relevant to the study then it is valid to use genomic microarrays of adequate resolution to provide a read-out of the RNA complement of a sample. Historically, however, RNA hybridizations of genomic microarrays have been used only to identify transcribed regions and not to compare gene intensities. This is due to the complexities of calculating a representative intensity from probes spanning all annotated exons, rather than focusing on 3' constitutive exons, which are more likely to be consistently represented in reverse transcribed cDNA.

Genomic microarrays are generally used to assess the DNA complement of a sample. This may be in order to assess the relative copy-numbers of different regions of the genome (e.g. Fiegler *et al*., 2003; Redon *et al*., 2006). It is also common for such arrays to be used to assess the enrichment of DNA molecules in a sample by the immunoprecipitation of chromatin components to which they are bound (e.g. Ercan *et al*., 2007; Horak and Snyder, 2002; Koch *et al*., 2007).

1.5. Microarrays as a phenotyping tool

The use of microarrays as a phenotyping tool is becoming progressively more prevalent (e.g. Booth *et al*., 2005; Hughes *et al*., 2000; Ishida *et al*., 2003; Wultsch *et al*., 2007; Zien *et al*., 2007). The application of DNA microarrays to measure expression and hence

provide a "molecular phenotype" for different cells and tissues has been useful in defining the molecular basis or response to a given condition by considering the gene expression that changes between any two conditions. Furthermore the relation of function between genes has been inferred through comparison of perturbation of individual genes. An approach that involves molecular phenotyping followed by hierarchical clustering both on conditions and on genes can therefore provide interesting information in two dimensions – both revealing relationships between conditions for which the phenotypes are acquired, and the molecular basis of the relationship revealed by the genes for which expression is similar, the former being driven by the latter.

In a classic of the genre Hughes *et al*., (2000) used the budding yeast *Saccharomyces cerevisiae* to generate molecular phenotypes for a large number of perturbations of genes with known function. Hierarchical clustering of these molecular phenotypes (or expression profiles) rediscovered the known cellular machineries to which these genes belong, manifested as discreet clusters within the complete cluster of profiles. The resulting compendium of expression profiles formed the basis for functional discovery of novel genes by comparison of their perturbed molecular phenotypes. Once the function of novel genes had been inferred by this approach it was then experimentally confirmed. These genes had been revealed to be involved in processes such as sterol metabolism, mitochondrial respiration and protein synthesis.

1.6. Aims of chapter 3

Whilst the Hughes *et al*. study was extremely valuable, both in proving the utility of its approach and in identifying gene function, it was limited to the biological repertoire of a single-celled organism. Should we wish to use such an approach to discover novel components of signalling pathways that are known to be dysregulated in cancer, for example, a metazoan system would be required. Such an animal would have to have numerous experimental advantages, such as a broad range of readily available loss-of-function mutants or the utility of rapidly generating them, and ease of producing appropriate samples. *C. elegans* is the only established model organism which is obviously a potential subject for such a study; a large repository of loss-of-function mutants already exists as well as an RNAi library which can deliver a systemic loss-of-function for almost any gene. The animal has a number of well-studied signalling pathways known to be conserved throughout metazoa and implicated in human disease. Whilst the isolation of individual tissues for study in *C. elegans* is problematic, there is strong precedent for expression analysis of a single tissue (the germline) at the level of whole animal. The involvement of the same signalling pathways in both germline and vulval development and established methods of screening for genes modulating the development of these tissues in conjunction with known pathways provides an independent means of inferring relatedness of gene function. Large-scale screening for genes which modulate both Muv and sterile phenotypes has revealed candidate modulators of signalling pathways involved in germline development. I therefore judged it feasible that adapting the approach taken by Hughes *et al*., to *C. elegans* and querying the resulting compendium with said candidate modulators may reveal novel genes

functioning in know signalling pathways in the germline. Chapter 3 details the establishment of this approach, its success and future potential in fulfilling this goal.

1.7. Transcriptome interrogation

A key aspect of modern biology is the mapping of transcriptomes and the application of this knowledge to different biological contexts in order to correlate gene expression with phenotype. As already intimated, the evaluation of transcript complement can give valuable information on either the biology underlying a phenotype, or serve as a tractable phenotype itself. The majority of expression studies performed to date refer to the current set of gene annotations and are limited by the accuracy of those annotations. Recent studies of the human, mouse, *Arabidopsis* and *Drosophila* transcriptomes have indicated substantially more widespread transcription than could be accounted for by the then current annotations (Bertone *et al*., 2004; Hanada *et al*., 2007; Manak *et al*., 2006; The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), 2006). Most of these studies were performed using tiled genomic microarrays, which can be used to assess the level of transcript corresponding to any region of the genome across which their probes are tiled, without the requirement for prior knowledge of the existing gene structures. Tiled genomic microarrays consist of probes arrayed at roughly equal distance across the region of the genome that they represent. They can therefore be used to detect RNA or DNA in a sample corresponding to those genomic coordinates regardless of gene annotations. I wanted to evaluate similarly the current gene annotations in *C. elegans*. The measured transcript complement of a cell or animal to the depth that is typically feasible, however, considers only the transcripts that are retained by the cell rather than all transcripts that are produced. This led me to interrogate the nonsense-mediated

mRNA decay deficient transcriptome and provided a valuable dataset for the study of this pathway by comparison with the wild-type transcriptome.

1.8. Nonsense-mediated mRNA decay

The process of gene expression is extremely complicated, with the potential for error at every stage. Eukaryotic cells have numerous surveillance mechanisms that ensure the fidelity of gene expression. Nonsense-mediated mRNA decay (NMD) is one such mechanism, which is conserved from yeast to human and acts at the level of translation (reviewed in Behm-Ansmant *et al*., 2007b; Chang *et al*., 2007; Mango, 2001). The NMD pathway targets and degrades mRNAs for which the position of translation initiation yields an in-frame premature termination codon (PTC). PTCs may arise from mutations in the coding gene, infidelity of transcription, export of improperly spliced transcripts from the nucleus, "leaky" translation (i.e. translation from a downstream start codon), or translation from a upstream start codon (uAUG) in the 5' UTR leading to an in-frame PTC (figure 1.7). Degradation of such transcripts ensures that truncated protein products that may have gain-of-function or dominant-negative characteristics do not accumulate in the cell. This explains the most well understood role of NMD - as a mechanism that ensures the fidelity of gene expression. It is unknown whether NMD has any consistent role in any other defined biological processes.

It is known that alternative splicing and NMD are highly coupled in humans. More than 75% of human pre-mRNAs are alternatively spliced (Harrow *et al*., 2006), of which perhaps a third give rise to at least one splice-form containing a PTC (Lewis *et al*., 2003).

NMD is also strongly implicated in human disease. Many known disease-associated mutations and variants result in mRNAs harbouring PTCs. The clinical outcome of harbouring such alleles is NMD dependent (Khajavi *et al*., 2006). Understanding the biological role of NMD and its underlying mechanism is therefore of immediate import.

Organism	**Yeast (*Saccharomyces cerevisiae*)**	**Nematode (*Caenorhabtitis elegans*)**	**Fruit fly (*Drosophila melanogaster*)**	**Mammal (*Homo sapiens*)**	**Plant (*Arabidopsis thaliana*)**
Effector	Upf1	SMG-2	UPF1	UPF1(RENT1)	UPF1(IBA1)
	Upf2	SMG-3	UPF2	UPF2	UPF2
	Upf3	SMG-4	UPF3	UPF3a/b	UPF3
		SMG-1	SMG1	SMG1	
		SMG-5	SMG5	SMG5	
		SMG-6	SMG6	SMG6	
		SMG-7		SMG7	
		SMGL-1		SMGL1(hNAG)	
		SMGL-2		SMGL2(hDHX34)	

Table 1.1. Components of the NMD machinery known to exist in model organisms. The core machinery of NMD is conserved from yeast to humans and expanded in mammals. Components in mammals are recognized to have divergent function.

The phenomenon of NMD was discovered almost simultaneously in human and *S. cerevisiae* in 1979 when it was first observed that nonsense mutations in a gene lead to a reduction in the corresponding mRNA rather than an accumulation of the truncated protein product (Chang and Kan, 1979; Losson and Lacroute, 1979). Further work then led to the discovery of the core NMD machinery of *Upf1*, *Upf2* and *Upf3* in *Saccharomyces cerevisiae* (Cui *et al*., 1995; Lee and Culbertson, 1995; Leeds *et al*., 1991), and the expanded metazoan machinery, all of which was first identified in *C. elegans* (Anders *et al*., 2003; Cali *et al*., 1999; Grimson *et al*., 2004; Hodgkin *et al*., 1989; Longman *et al*., 2007; Page *et al*., 1999). There are minor variations around the

core machinery in the different metazoans studied, as detailed in table 1.1. Whilst many of the components required for NMD are known, however, the mechanism by which they target transcripts for degradation is poorly understood. It is known that detection of NMD targets occurs in the first round of translation, leading to the phosphorylation of SMG-2 by SMG-1 and repeated rounds of phosphorylation by SMG-1 and dephosphorylation facilitated by SMG-5/6/7 (Anders *et al*., 2003; Chiu *et al*., 2003; Gatfield *et al*., 2003; Ohnishi *et al*., 2003; Yamashita *et al*., 2005). This is followed by degradation of the transcripts by seemingly evolutionarily diverged mechanisms (Gatfield and Izaurralde, 2004; Lejeune *et al*., 2003; Mitchell and Tollervey, 2003).

Key to understanding the mechanism of NMD is precise knowledge of what constitutes a PTC and how it is determined. Until recently it was held that in mammals PTCs are defined by their distance from the last exon junction complex (EJC), but in *Drosophila* and *C. elegans* NMD occurs in the absence or depletion of the EJC, suggesting that the EJC is not involved (Fribourg *et al*., 2003; Gatfield *et al*., 2003; Gehring *et al*., 2003; Longman *et al*., 2007; Lykke-Andersen *et al*., 2001). Recent research, however, has indicated that NMD still occurs in mammals in the absence of the EJC, rather distance between the PTC and the poly(A) tail may be a defining factor as in lower eukatyotes (Amrani *et al*., 2004; Behm-Ansmant *et al*., 2007a; Buhler *et al*., 2006; Longman *et al*., 2007). Questions remain regarding the structural features of transcripts that define termination codons as premature and that lead to the targeting of transcripts for degradation. It has been demonstrated in *S. cerevisiae*, *Drosophila* and human that tethering of poly(A) binding protein (PABP) downstream of a PTC prevents degradation

of the transcript by NMD (Amrani *et al*., 2004; Behm-Ansmant *et al*., 2007a; Singh *et al*., 2008). Using a system of folding back the poly(A) tract to different distances from a PTC Eberle *et al*., (2008) have provided evidence that strength of NMD targeting of a transcript is related to the distance of the PTC to the ribonucleoprotein (RNP) environment located at the 3' end of the transcript. Simultaneously, work by Singh *et al*., (2008) was published presenting evidence that 3' UTR associated factors are involved in either promoting or inhibiting the binding of UPF1 (SMG-2) to the terminating ribosome. Taken together this suggests that an in-frame termination codon at too great a distance from the relevant 3' end associated proteins would precipitate the degradation of such transcripts by NMD in humans as well as lower eukaryotes. This would then suggest that 3' UTR length is a key determinant of targeting for NMD. Studies inserting a false 3' UTR between a termination codon and poly(A) tract of transcripts have indicated that a distance of >420 nt between termination codon and poly(A) tract leads to NMD targeting in humans (Singh *et al*., 2008). There are, however, many natural human mRNAs with longer 3' UTRs. The simplest explanation of why such transcripts are not NMD substrates is that sequence motifs in the 3' UTR either lead to a secondary structure which brings 3' end associated proteins closer to the termination codon, or that they recruit other RNA binding proteins which antagonize the binding of UPF1 to the terminating ribosome. While both possibilities may be true, there is a lack of evidence to support either hypothesis. There is, however, evidence from studies in *S. cerevisiae* and *C. elegans* that generally support the hypothesis that RNA binding proteins protect PTC containing transcripts from NMD. The RNA binding proteins Pub1 in *S. cerevisiae* and GLD-1 in *C. elegans* have been shown to bind the 5' UTRs of transcripts containing

upstream open reading frames (uORFs) in a sequence specific manner (Lee and Schedl, 2004; Ruiz-Echevarria and Peltz, 2000; Ryder *et al*., 2004). mRNAs containing uORFs or upsteam start codons leading to a frame shift are natural substrates for NMD as they lead to translation termination at a PTC. Pub1 and GLD-1 have been shown to block access of the translational machinery to the upstream start of uORFs, thus protecting those transcripts from degradation. There has yet to be a comprehensive study of the targets of these RNA binding proteins. Furthermore there are likely to be many more RNA binding proteins that protect transcripts from NMD, either through masking incorrect translation start sites or preventing the binding of the NMD machinery to the terminating translation machinery.

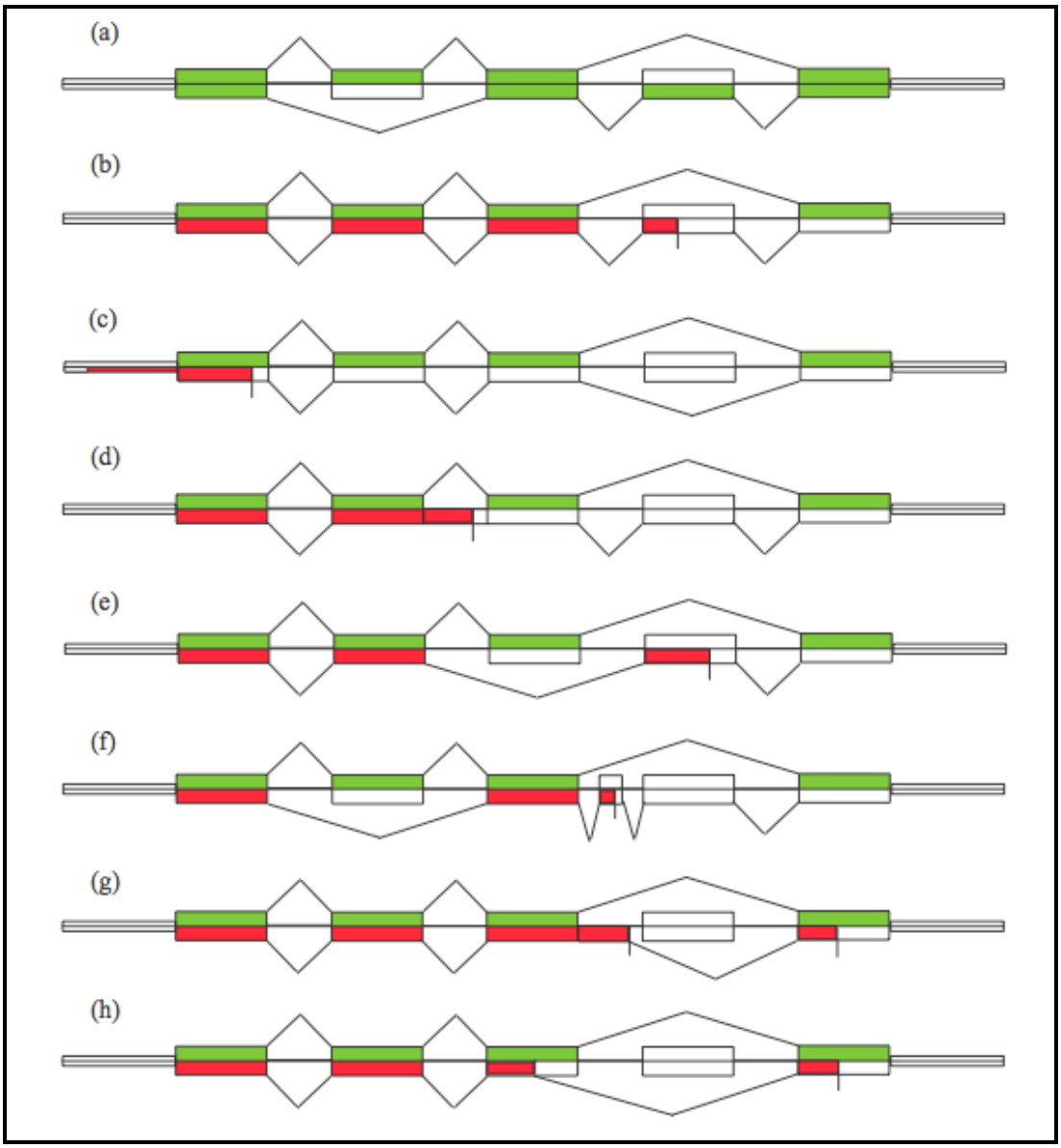


Figure 1.7. The recognized post-transcriptional causes of NMD targeting. Numerous translational and splicing events can lead to NMD targeting. The true coding ORF of the transcript not ending in a PTC is shown in green. The ORF ending in a detected PTC leading to NMD targeting is shown in red. (a) A pre-mRNA with two alternative viable spliceforms; (b) A non-viable spliceform utilizing only annotated exons leading to a PTC; (c) Translation from a uAUG leading to a frame-shift and consequently a PTC; (d) Intron retention leading to a PTC – this could either be in-frame in the intron or lead to a frame-shift and PTC; (e) Exon skipping leading to a frame-shift and PTC; (f) Splicing in of a poison exon containing a PTC or always resulting in a frame-shift and PTC; (g) Exon extension by use of an alternative splice-site leading to an in-frame PTC; (h) Exon truncation by use of an alternative splice-site leading to an in-frame PTC.

Many studies have indicated that there are endogenous transcripts, which are natural substrates for NMD. Whilst these targets appear to be involved in a particular biological process in each study, comparison of NMD regulated transcripts between organisms indicate non-orthologous, seemingly unrelated sets of genes are NMD regulated in each organism. Amongst the suggested roles of NMD as a result of these studies are the regulation of oxidative stress response and nutrient homeostasis (Gardner, 2008; Guan *et al*., 2006; He *et al*., 2003; Mendell *et al*., 2004; Rodriguez-Gabriel *et al*., 2006). Further work is required, however, to confirm these roles. Confirmed or otherwise, the potential for NMD to regulate other processes must still exist. It is becoming increasingly apparent, however, that NMD and splicing regulation are linked, with many splicing activators being NMD-regulated via inclusion of PTC-causing cassette exons (Lareau *et al.*, 2007; Ni *et al.*, 2007; Saltzman *et al.*, 2008). The question of whether NMD has a clear role in any other biological process and whether that role is conserved is still open.

Alternative roles of the components of the NMD machinery are also becoming clearer. For example, recent evidence suggests that SMG-1 plays roles in oxidative stress response in *C. elegans* as well as mammals (Masse *et al.*, 2008; Gehen *et al.*, 2008). SMG-1 has also been implicated in tumour necrosis factor alpha-induced apoptosis (Oliveira *et al.*, 2008) and telomere maintenance (Azzalin *et al.*, 2007). Components of the NMD machinery are also involved in Staufen mediated and histone RNA degradation pathways (Kim *et al.*, 2005; Kaygun *et al.*, 2005).

1.9. Methods of surveying the transcriptome

The two most obvious ways of surveying the transcriptome are by microarray analysis using tiled genomic microarrays and by sequencing of cDNAs. Recent advances in these two technological areas have allowed rapid sampling of transcriptomes at high resolution. The two platforms that have been utilized in the work presented in this thesis are Affymetrix GeneChip® *C. elegans* Tiling 1.0R Arrays and Illumina ultra-high density sequencing technology. These two technologies have produced highly complementary data sets, which will be discussed in depth in chapter 4 and beyond. Figure 1.8 illustrates a basic flow-through of the two technological applications. Briefly, (ds)cDNA is produced from RNA, fragmented and analyzed using the two different platforms.

The tiling arrays have 25mer probes arrayed at an average distance of 10bp giving complete coverage of the *C. elegans* genome at 35bp resolution. All of the probes are unique and so any regions of the genome for which it was not possible to design unique probes are not represented. Only a tiny fraction of the genome is not represented, however. The output of the array is ~3 million probe intensities which can be aligned along genomic coordinates and analyzed in order to define discreet regions of expression, referred to as transcription fragments or "transfrags". Because the arrays allow us to assess what is present in the sample relative to genomic coordinates a transfrag is most likely to correspond to an individual exon, rather than a whole gene. This therefore allows the user to both assess which regions of the genome are transcribed in any given condition and how these regions differ between conditions without reference to a set of genome annotations. Additionally, with knowledge of gene annotations one can assign

probe intensities to a gene and calculate a representative gene intensity, allowing the tiled genomic microarray to be used as an expression microarray. Comparison of probe signal to gene annotations allows the user to identify differing exon intensities within a gene and also to look for consistent differences in signal relative to annotations, which may indicate annotation errors. There are drawbacks to using tiling arrays relative to other approaches however. Firstly, the resolution of arrays is limited and therefore cannot be used to define precise structures such as exon-exon boundaries. They also cannot be used to call the presence/absence of structures (e.g.introns or exons) that are smaller than the resolution of the array or in regions to which no unique probes could be assigned. They do have the advantage of being cheaper than ultra-high density sequencing applications and also requiring substantially less starting material per experiment.

Illumina ultra-high density sequencing technology allows the generation of 1bp resolution data. The output of this technology is ~3 million 35bp reads per lane of a flow cell with a confidence score assigned to each base of a read. Unique reads can then be mapped to the genome or transcriptome with a confidence score and intensities calculated for each base relative to how frequently it is represented in aligned reads, thus equating to an expression score. Not only do these sequence data have ultimate resolution but also give information on connectivity, identification of reads overlapping exon boundaries inferring splice-junctions. Furthermore an aligned read is much more easy to interpret than the intensity of a probe on a microarray. Intensities derived from numbers of uniquely alignable reads require no background correction as is involved in microarray data analysis and so the full potential of the signal in the data is more likely to be tapped.

Additionally any noise that exists within the data is automatically discarded as non-alignable reads. There are drawbacks to this platform however. Because the great majority of total RNA extracted from a cell is ribosomal RNA, polyadenylated RNA must be purified from total RNA before it can be evaluated. Consequently non-polyadenylated, non-ribosomal RNAs that may be of interest are under-represented in sequenced samples. Furthermore a single lane in a flow cell does not provide the depth of coverage of the transcriptome that is provided by microarrays in terms of gene intensities. More specifically, a handful of reads mapping to a gene provide evidence of its presence in a sample but not an adequately gene intensity to allow accurate comparisons between samples. The two technologies utilized in this study therefore provide complementary datasets – one providing sufficient depth from which to infer gene expression changes and the other of sufficient resolution to accurately identify structural properties of the genes sufficiently represented.

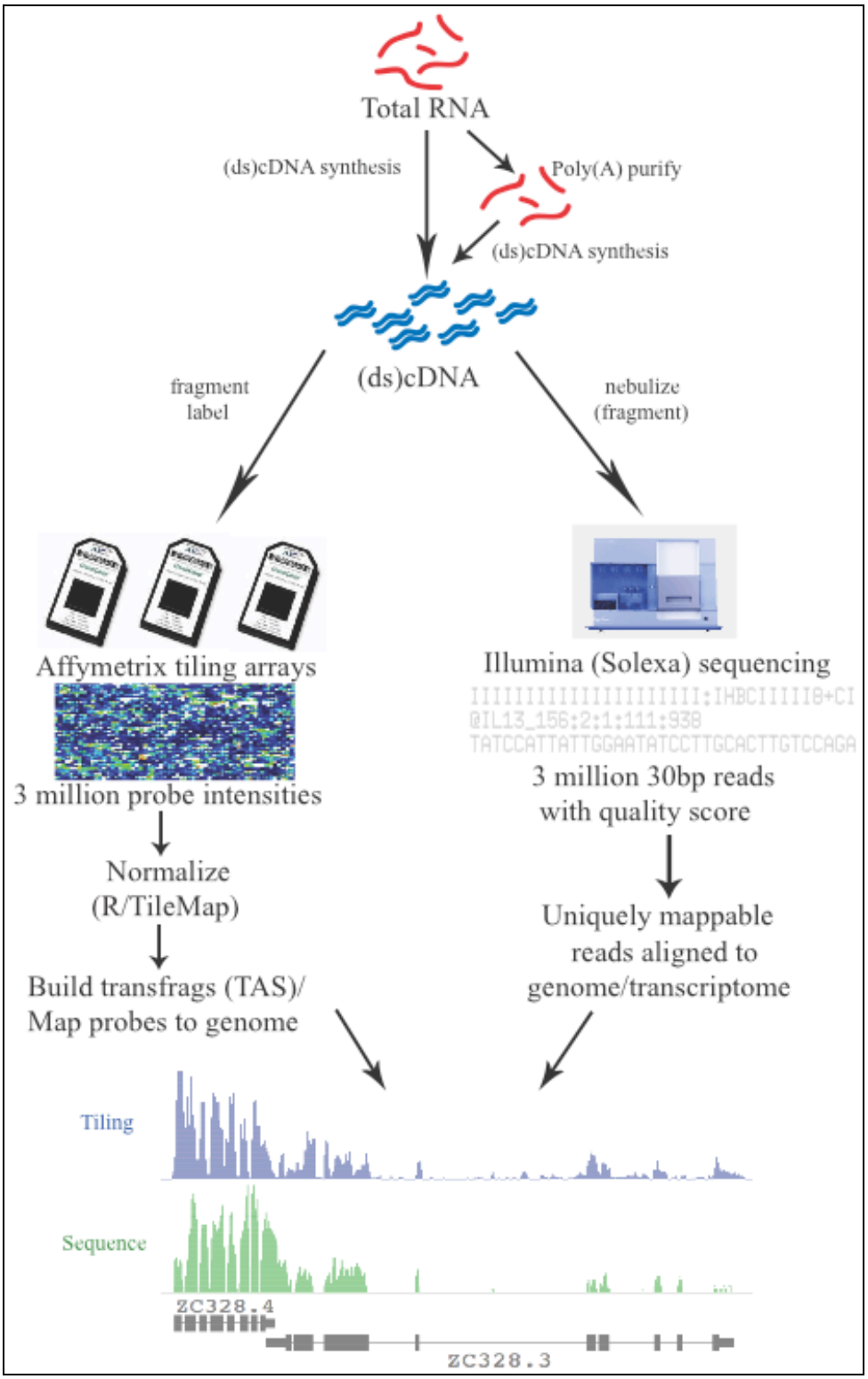


Figure 1.8. Technical flow-through of Affymetrix tiling array and Illumina sequencing technologies. The two independent technologies provide analogous and complementary datasets, tiling arrays of 35bp resolution and high depth, the sequencing of 1bp resolution but lower depth.

Chapter 2
Materials and Methods

2.1. Reagents

2.1.1. *C. elegans*

2.1.1.1. *C. elegans* strains

The following mutant strains were acquired from the *Caenorhabditis* Genetics Centre (CGC), University of Minnesota, USA (http://www.cbs.umn.edu/CGC/): Bristol N2, *glp-1(or178)*, *lag-2(q420)*, *emb-5(hc61)*, *smg-1(r861)* and *smg-5(r860)*.

2.1.1.2. Nematode Growth Medium (NGM) (Stiernagle, 2006)

NaCl	3g
Peptone	2.5g
Optional[a]: Agar	19g
dd H_2O	to 1L

The solution was autoclaved and cooled to 55°C before addition of:

Cholesterol solution (5 mg/ml in ethanol)	1ml
1M $CaCl_2$	1ml
1M $MgSO_4$	1ml
1M KH_2PO_4, pH6.0	25ml
Fungizone	800µl

in the order as written, with mixing thoroughly after addition of each component.

Solutions were sterile-filtrated through a membrane filter with a pore size of 0.2 µm.

[a] For preparation of agar plates, solution was poured into sterile Petri dishes.

2.1.1.3. M9 Buffer (Stiernagle, 2006)

1M KH_2PO_4	3g
1M Na_2HPO_4	6g
1M NaCl	5g
ddH_2O	to 1L

1ml 1M $MgSO_4$ was added after solution had been autoclaved to sterilize.

2.1.1.4. Freezing buffer (Stiernagle, 2006)

KH_2PO_4	3g
0.05M K_2HPO_4	129ml
0.05M KH_2PO_4	871ml
NaCl	5.85g
Glycerin	30% (v/v)

2.1.1.5. Bleach solution

1M NaOH	250µl
Sodium hypochlorite, available chlorine 10-13%	100µl
Autoclaved H_2O to	1000µl

2.1.2. Bacteria

2.1.2.1. RNAi feeding strains

Bacterial clones used for RNA interference (RNAi) experiments were selected from the Ahringer RNAi feeding library (Kamath *et al*., 2003) and *C. elegans* ORFeome collection (Rual *et al*., 2004).

2.1.2.2. **Luria-Bertani (LB) medium** (Bertani, 1951)

Bacto-tryptone	10g
Bacto-yeast extract	5g
NaCl	10g
Optional[a]: Bacto-Agar	15g
ddH2O	to 1L

pH was adjusted to 7.2 and solution was autoclaved to sterilize.

[a] For preparation of agar plates, solution was poured into sterile Petri dishes.

2.1.2.3. 2 x Tryptone / yeast extract (TY)

Bacto-tryptone	16g
Bacto-yeast extract	10g
NaCl	5g
dd H_2O	to 1L

pH was adjusted to 7.2 and solution was autoclaved to sterilize.

2.1.3. Buffers used for Affymetrix tiling microarray hybridization and processing

(From Affymetrix GeneChip® Whole Transcript (WT) Double-Stranded Target Assay Manual)

2.1.3.1. 12x MES Buffer

MES hydrate	64.61g
MES Sodium Salt	193.3g
Molecular Biology Grade Water	800ml

Mix, adjust volume to 1L and 0.22µm filter. Stored at 4°C in the dark.

2.1.3.2. 2x hybridization buffer

12X MES Stock Buffer	8.3ml
5M NaCl	17.7ml
0.5M EDTA	4.0ml
10% Tween-20	0.1ml
Molecular Biology Grade Water	19.9ml

Stored at 4°C in the dark.

2.1.3.3. Wash Buffer A

20X SSPE (Ambion)	300ml
10% Tween-20	1.0ml
Molecular Biology Grade Water	699ml

0.22µm filtered

2.1.3.4. Wash Buffer B

12X MES Stock Buffer	83.3ml
5M NaCl	5.2ml
10% Tween-20	1.0ml
Molecular Biology Grade Water	910.5ml

0.22µm filtered and stored at 4^{o}C in the dark.

2.1.3.5. 2x Stain Buffer

12X MES Stock Buffer	41.7ml
5M NaCl	92.5ml
10% Tween-20	2.5ml
Molecular Biology Grade Water	113.3ml

0.22µm filtered and stored at 4^{o}C in the dark.

2.1.3.6. Array Holding Buffer

12X MES Stock Buffer	8.3ml
5M NaCl	18.5ml
10% Tween-20	0.1ml
Molecular Biology Grade Water	73.1ml

Stored at 4^{o}C in the dark.

2.1.3.7. Streptavidin Phycoerythrin Stain Cocktail

2x Stain Buffer	300µl
50mg/ml BSA (Invitrogen)	24µl
1mg/ml Streptavidin Phycoerythrin (Molecular Probes)	6µl
Molecular Biology Grade Water	270µl

2.1.3.8. Antibody Stain Cocktail

2x Stain Buffer	300µl
50mg/ml BSA (Invitrogen)	24µl
10mg/ml grade Reagent Goat IgG (Sigma-Aldrich), made up in 150mM NaCl, stored at $4^{o}C$	6µl
0.5mg/ml goat anti-streptavidin biotinylated antibody (Vector Laboratories)	3.6µl
Molecular Biology Grade Water	266.4µl

2.1.4. 10x PCR reaction buffer

100 mM Tris-HCl

500 mM KCl

15 mM $MgCl_2$

pH 8.3 at 25°C

2.2. Protocols

2.2.1. Maintenance of *C. elegans* stocks

C. elegans was maintained on NGM agar plates seeded with OP50 *E. coli* according to standard protocols (Brenner, 1974). For maintenance of large worm populations in liquid culture HB101 *E. coli* grown in 2 x TY was resuspended in NGM. Freshly bleached embryos were then added to HB101 in NGM in conical flasks and shaken at 150r.p.m. at 15^{o}C.

2.2.2. Bleach sterilization of *C. elegans* strains and synchronization

Worms were washed off plates with M9 buffer and pelleted for 1 minute at 1000r.p.m.. Alternatively, worms in liquid culture were pelleted in the same way. The resulting pellet was then resuspended in freshly prepared bleach solution and incubated shaking at room temperature until the worms had broken apart, all of the carcass dissolved and only embryos remained. Embryos were then pelleted for 1 minute at 1000r.p.m. and washed twice with M9 buffer. Embryos were then left shaking for 26h at room temperature in order to obtain a synchronous population growth-arrested at mid-L1 stage (Stiernagle, 2006). Alternatively, if a synchronous population was not required embryos pelleted in M9 buffer were spotted on NGM plates or added to HB101 in NGM for liquid culture.

2.2.3. Freezing and recovery of *C. elegans* stocks

A population of worms containing L1 and L2 stage animals that were approaching starvation were washed off plates in M9 buffer, pelleted by centrifugation at 1000r.p.m.

for 1 minute, and resuspended in an equal volume of M9 buffer and freezing buffer. 1ml of suspension was aliquoted per 1.8ml cryovial. Cryovials were placed into freezing boxes filled with isopropanol to allow a gradual 1°C decrease in temperature per minute when placed at -70°C. Cryovials were stored at -70°C. For thawing, cryovials were placed at room temperature and worms were spotted onto NGM plates seeded with OP50 *E. coli* as soon as all ice had turned to liquid (Stiernagle, 2006).

2.2.4. RNAi by feeding on plates, RNA extraction and visual phenotyping

dsRNA expressing bacteria from glycerol stocks were streaked out onto LB agar plates containing 50mg/ml Amp and incubated overnight at 37^{o}C. The next day the resulting colonies were cultured overnight in 2xTY + 100mg/ml Amp and spotted onto NGM Single Peptone plates containing 50mg/ml Amp and 1mM IPTG, and left overnight to dry. Synchronised L1 stage worms in M9 buffer were then spotted onto plates (or NGM plates sans Amp and IPTG and seeded with OP50 as appropriate) and incubated at the appropriate temperature (20 ^{o}C or 25^{o}C) (Kamath *et al.*, 2003). Worms were washed off plates in M9 buffer at the appropriate timepoint and spun down and washed once in M9 buffer. Worms were then pelleted at 1000r.p.m. for one minute and the pellet resuspended in 4ml Trizol® (Invitrogen) per ml pellet. RNA was then prepared from the Trizol solution according to the manufacturer's protocol, the final pellet resuspended in nuclease-free water. The quantity of RNA was measured with a NanoDrop ND-1000. 100% ethanol was added to give a 70% ethanol solution and stored at -70 ^{o}C. For visual phenotyping, individual animals were transferred to wells of 12 well plates baring the same constituents as above and incubated at 25^{o}C for 24hrs beyond young adult stage.

Adult worms were then removed and the plates returned to the incubator for 24hrs, after which progeny were counted.

2.2.5. DAPI staining

For staining of nuclei with DAPI, whole intact worms were washed off plates in M9 buffer and fixed in cold (-20^{o}C) methanol for 5 min. Fixed worms were washed twice in M9 buffer, incubated 30 min in 100 ng/ml DAPI in M9 and washed two to three times in M9. Worms were then mounted on glass slides and imaged with a Leica SP5 confocal microscope.

2.2.6. Generation of mixed-stage RNA reference sample

Synchronous L1 stage animals were added to *E. coli* strain HB101 in NGM medium in conical flasks shaking at 15^{o}C. HB101 was added into the culture as appropriate such that the animals neither starved nor became anoxic. At the appropriate developmental stage the cultures were placed at 4^{o}C to allow the animals to settle. The animals were isolated and washed twice with M9. RNA was then extracted as in 2.2.4. This gave RNA from synchronous L2, L3, L4, young adult and gravid adult stage populations. The resulting RNA was then mixed and supplemented with RNA extracted from growth-arrested L1 stage animals and asynchronous embryos yielding sufficient RNA for ~1000 microarray hybridizations as detailed in 2.2.7.

2.2.7. RNA labelling and two-colour microarray hybridization

A direct labelling method was used to produce fluorescently labelled cDNA. In all cases the experimental sample (Cy3) was hybridized against a universal reference sample

(Cy5). 20µg of RNA precipitated from 70% ethanol stock by addition of 1/40th volume 3M Sodium Acetate and storage at -70^{o}C for at least 30 minutes. The samples were then spun down at 20800g in an Eppendorf 5415R cooled centrifuge at 4^{o}C. The pellets were then washed once in 70% ethanol and air-dried briefly. The pellets were then resuspended in 14.4µl and 1µl 0.5µg/µl oligo(dT)$_{12-18}$, heated to 70^{o}C for 10 minutes and then placed on ice. The following reagents were then added in order:

- 6.0µl 5 x first strand buffer (Invitrogen)
- 3.0µl 0.1M DDT (Invitrogen)
- 0.6µl dNTP mix (25mM dATP, 25mM dGTP, 25mM dTTP, 10mM dCTP)
- 3.0µl dCTP-Cy3 or dCTP-Cy5 (25mM GE Healthcare)
- 2.0µl Superscript II (Invitrogen)

The mixture was then incubated at 42^{o}C for 2 hours. 1.5µl 1M NaOH was added and incubated at 70^{o}C for 20 minutes to hydrolyse the RNA. 1.5µl HCl was then added to neutralize the solution. The cDNA was then purified from the mixture using QIAGEN PCR Purification columns according to the manufacturers instructions with an additional wash with buffer PE. The eluted cDNA was precipitated in 70% ethanol, 75mM Sodium Acetate at -20^{o}C with 8µg human Cot-1 DNA (Invitrogen), 2µg polyA DNA (Sigma) and 250µg sheared salmon sperm DNA (Ambion) for 30 minutes. The precipitated DNA was then spun down at 20800g for 5 minutes, the pellet washed in 70% ethanol and dried at 70^{o}C for 2 minutes. 10µl nuclease-free water was added to the pellet and heated to 70^{o}C for 5 minutes. 50µl hybridization buffer (50% Dionised Formamide, 5xSSC, 0.1% SDS and 0.1mg/ml BSA) was then added and incubated at 70^{o}C for a further 5 minutes. The

hybridization mix was then allowed to cool to room temperature for 10 minutes in the dark and then centrifuged for 5 minutes at 20800g at room temperature for 5 minutes. 55µl of hybridization mix was spotted on a covered slip of equal width of the microarray slide and sufficient width to cover the printed area. The printed side of the slide was then applied to the cover slip and the slide placed in a saddle in an Advalytix SlideBooster SB800. The SlideBooster had previously been pre-warmed to 42^{o}C with 500µl humidifying buffer (20% Formamide, 2xSSC) added to each reagent reservoir and 30µl coupling buffer (Advalytix) to each saddle. The microarrays were then incubated for 16-24hrs with sonication.

The cover slips were allowed to slide from the arrays in 0.1xSSC, 0.1% SDS. The arrays were washed twice for 15 minutes in 0.1xSSC, 0.1% SDS and then three times for 5 minutes in 0.1xSSC in a slide rack in a pyrex trough in the dark. The arrays were then centrifuged for 1 minute at 1000r.p.m. in slide racks to dry. The microarrays were scanned using a GenePix 4000B scanner at 5µm resolution. All wash buffers were made up with sterile HPLC water.

2.2.8. Affymetrix tiling microarray hybridization

Total RNA was cleaned using Rneasy columns (QIAGEN) according to manufacturers protocol and then Dnase I (Roche) treated with 10U for 30 minutes in 100µl 1x One-Phor-All buffer (Amersham). The RNA was then re-purified using Rneasy columns (QIAGEN).

1µl random hexamers (3µg/µl) were added to 15µg total RNA in 7µl nuclease-free water, and placed in a thermal cycler using the heated lid for the following protocol:

- 70°C for 5 minutes
- 25°C for 5 minutes
- 4°C for 2-10 minutes

The following reagents were then added to the mixture:

- 4µl 5X first strand buffer (Invitrogen)
- 2µl 100 mM DTT (Invitrogen)
- 1µl 10 mM dNTPs (Invitrogen)
- 1µl RNase Inhibitor (Ambion)
- 4µl Superscript II (Invitrogen)

The mixture was then placed in a thermal cycler using the heated lid for the following protocol:

- 25°C 10 minutes
- 42°C 90 minutes
- 70°C 10 minutes
- 4°C 2-10 minutes

The following reagents were then added to the mixture on ice:

- 7.3µl nuclease-free water
- 8µl 5X second strand buffer (Invitrogen)
- 2µl 10 mM dNTPs (Invitrogen)
- 1µl 10 U/ml E. coli DNA Ligase (Invitrogen)
- 1.2µl 10 U/ml E. coli DNA polymerase I (Invitrogen)

- 0.5µl 2 U/ml E. coli RNase H (Invitrogen)

The mixture was then placed in a thermal cycler for the following protocol:

- 16°C 2 hours (without heated lid)
- 75°C 15 minutes (with heated lid)
- 4°C at least 2 minutes

followed by the addition of:

- 10 U/ml RNase H (Epicentre)
- 5 and 20 U/ml RNase A/T1 cocktail (Ambion)

and incubation at 37°C for 20 minutes.

The (ds)cDNA was then purified using QIAGEN PCR Purification columns according to the manufacturers instructions, eluting in nuclease-free water. The eluted RNA was ethanol precipitated by dilution in 100% ethanol to 70% concentration, addition of 1/40th volume 3M Sodium Acetate and incubation at -70°C for at least 30 minutes. The precipitated (ds)cDNA was spun-down at 20800g for 15 minutes, washed once in 70% ethanol, air-dried and resuspended in nuclease-free water. 17µg of (ds)cDNA in 22µl water was digested by the addition of:

- 3µl 10x One-Phor-All buffer (GE Healthcare)
- 5µl Dnase I (Invitrogen) diluted to 0.17U/µl in 1x One-Phor-All buffer

and incubation in a thermal cycler with heater lid using the following protocol:

- 37°C 8 minutes
- 99°C 10 minutes

- 4°C at least 2 minutes

2µg (1.76µl) of (ds)cDNA was assessed on a 1% agarose gel in order to check that the majority of (ds)cDNA was in the desired 50-100bp size-range.

The (ds)cDNA was then labelled by the addition of:

- 17.96µl nuclease-free water
- 14µl 5X TdT buffer (Roche)
- 7µl 25 mM CoCl2 (Roche)
- 2.3µl Affymetrix DNA Labeling Reagent
- 0.5µl Terminal deoxytransferase (8000U; Roche)

followed by incubation at 37°C for 2 hours.

The following was then added to the above mixture:

- 4.17µl Affymetrix Control Oligonucleotide B2
- 2µl 10mg/ml Herring Sperm DNA (Promega)
- 2.5µl 50mg/ml Acetylated BSA (Invitrogen)
- 125µl 2x Hybridization Buffer
- 17.5µl DMSO
- 28.83µl Nuclease-free water

This hybridization cocktail was then heated to 99°C for 5 minutes, cooled to 45°C for 5 minutes, centrifuged at 20800g for one minute and then injected into an Affymetrix

GeneChip® *C. elegans* Tiling 1.0R Array. The array was hybridized for 16 hours in a 45^{o}C Affymetrix hybridization oven at 60 r.p.m..

Hybridized microarrays were washed and scanned according to chapter 5 of the "GeneChip® Whole Transcript (WT) Double-Stranded Target Assay Manual" (https://www.affymetrix.com/support/downloads/manuals/wt_dble_strand_target_assay_manual.pdf).

2.2.9. (ds)cDNA production for Illumina sequencing

Total RNA was cleaned using Rneasy columns (QIAGEN) and then Dnase I (Roche) treated with 10U for 30 minutes in 100ml 1x One-Phor-All buffer (Amersham). RNA was then re-purified using Rneasy columns (QIAGEN). mRNA was purified from total RNA using Oligotex midi kits (QIAGEN) according to the manufacturers protocol. (ds)cDNA was then produced using SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) and purified using a QIAGEN PCR Purification kit. 1µg of (ds)cDNA was then submitted for sequencing.

2.2.10. Reverse transcription and PCR

Total RNA was cleaned using Rneasy columns (QIAGEN) and then Dnase I (Roche) treated with 10U for 30 minutes in 100µl 1x One-Phor-All buffer (Amersham). RNA was then re-purified using Rneasy columns (QIAGEN). 5µg total RNA was then used to produce first-strand cDNA using SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) and purified using a QIAGEN PCR Purification kit. 5ng cDNA was used as template for amplification with gene-specific primers in the following PCR mix:

- 3µl 10x PCR reaction buffer
- 3µl 10mM dNTPs
- 2.7µl 1mg/ml BSA
- 0.4µl 5% (v/v) β–Mercaptoethanol
- 0.9µl 10mM primer mix
- 0.6µl Taq polymerase
- 1.5µl template
- 17.9µl nuclease-free water

using the following amplification conditions in a thermal cycler:

- 94°C for 5 minutes
- 30 cycles of:
    - 94°C for 30 seconds
    - 58°C for 30 seconds
    - 72°C for 2 minutes
- 72°C for 5 minutes
- Hold at 16°C

PCR products were then analysed on a 1% agarose ethidium bromide gel.

2.2.11. Two-colour expression microarray data analysis

GenePix Pro 5.0 was used to identify and isolate signal from spots above background and export data. The methodology from this point is described in detail in chapter 3. Briefly, the data were then normalized using a publicly available Perl script available here: http://www.sanger.ac.uk/PostGenomics/S_pombe/software/. Differentially expressed genes between each condition and wild-type were determined by Student's t-test.

Comparative ratios of means of biological replicates were calculated between the relevant conditions via the universal reference sample. Hierarchical clustering was then performed based on a correlation matrix of the differentially expressed genes within all conditions being compared. This was done using GeneSpring

2.2.12. Identifying transcribed regions and visualization of tiling microarray data

Raw spot intensity files (.CEL files) were quantile normalized and scaled in R. The normalized data were processed and then exported as .BAR files using Affymetrix Tiling Analysis Software (TAS) version 1.1 for visualization in Affymetrix Integrated Genome Browser (IGB). A background cut-off was calculated to include the top 5% of all non-genic probes for each condition and interval analysis then performed in TAS to identify transcribed regions above this cut-off. The maxgap and minrun parameters that define the transcribed regions are discussed in chapter 4.

2.2.13. Affymetrix tiling microarray expression data analysis

The raw data for all arrays to be compared were quantile normalized in R. All further data manipulations were performed in Perl. Probe signal was mapped to all genes and exons of the relevant genome release. A background threshold was then calculated for the mean signal of biological replicates in order to include the top 5% of extra-genic probes. Genes were considered expressed if ≥50% of probes were above background in ≥50% of unique exons. Gene intensities of median exonic probes above background within filtered exons were then calculated. Exon intensities used for the splicing analysis were the median probe intensity of probes above background in the exons for which ≥50% of probes were above background.

2.2.14. Illumina sequence data analysis

Sequence reads were then aligned to the genome using Maq (http://maq.sourceforge.net/) to both identify where reads align and the number of reads that overlap a given base pair. The output was then visualized relative to the genome using IGB. Gene intensities based on sequence reads were calculated as the median number of reads spanning a given base of a gene amongst bases for which there is at least one spanning read. Sequence reads spanning exon-exon boundaries were identified as detailed in chapter 4. Briefly, such reads were identified as reads not alignable to the genome using Maq but giving complete alignment across exon-exon boundaries using Maq.

Chapter 3

Microarray analysis of germline perturbations

3.1. Introduction

A classical approach to understanding gene function is to generate loss-of-function phenotypes. Such phenotypes, however, require correct characterisation. Most phenotypes in model organisms have previously been reported at the level of morphology, often requiring many different techniques to measure each parameter. My intention was to develop use of expression microarrays as a single phenotyping methodology to compare genic perturbations resulting in brood-size defects in *C. elegans*. In this chapter I present a detailed rationale behind this project and the utility of the approach as a pathway-specific phenotyping tool with potential future application.

RNA-mediated interference (RNAi) has proved a powerful tool for the generation of loss-of-function phenotypes. In particular the capability of perturbing gene function in *C. elegans* simply by the feeding of bacteria expressing dsRNA has led to the generation of an RNAi library consisting of clones targeting ~86% of annotated coding genes (Fraser *et al*., 2000; Kamath *et al*., 2003; Kamath *et al*., 2001). Whole-genome screens using the RNAi library have revealed loss-of-function phenotypes for many genes under laboratory conditions (Kamath *et al*., 2003). For example, hundreds of genes give brood-size defects by RNAi, indicating a deleterious effect on either germline development or gametogenesis. The observation of a sterile animal at low resolution in an RNAi screen, however, tells us almost nothing about gene function as there are many independent pathways and processes which when perturbed lead to germline defects. It is clear, therefore that a high-resolution phenotyping methodology is required. One possibility is through careful microscopic analysis of the worms themselves along with *in situ* and

immuno-stainings to assess the level, location and combinations of expression of certain key genes, which define the biological state of a tissue. This, however, requires prior knowledge of a number of molecular markers and antibodies against them. It would also require the careful dissection of the germline from many animals, drastically limiting throughput.

An alternative approach is to use microarray expression data to define phenotypes. This has been previously demonstrated in *Saccharomyces cerevisiae* to great effect. The expression profile of mutant strains can be considered as 'molecular phenotypes' — they are read-outs of the expression changes that result from a given mutation. These signatures are high density, since they cover all predicted genes, and quantitative, allowing more criteria to be tested than through staining. In *S. cerevisiae* this allowed genes to be clustered into related functional groupings according to similarities in the expression profiles, even for perturbations that were otherwise sub-phenotypic (Hughes *et al*., 2000). For example, mutations in genes involved in mating yield similar signatures, whereas mutations in genes involved in mitochondrial respiration clustered in a separate cluster. By building a compendium of expression signatures of mutations in genes of known pathways it was then possible to place novel genes into pathways by comparing their signatures with the compendium – for example if a novel gene has a signature that resembles that of the sterol biosynthesis pathway, it suggests that it plays a role in this pathway. This was groundbreaking work by Hughes *et al*. and provided the inspiration for our own study. Whilst yeast and human share many key aspects of eukaryotic life, however, as a single-celled organism yeast is of little use in the study of

cellular signalling and development. An approach such as this would therefore be more relevant to human biology if it were performed in a metazoan. Consequently I set out to validate a similar approach in the nematode *C. elegans*.

As previously discussed, the *C. elegans* germline is a well-studied, largely syncytial tissue with a number of genes and pathways known to control certain processes, for example, the Notch pathway is known to regulate the maintenance of the mitotic stem cell niche, the *gld* genes are known to be involved in the mitosis-meiosis switch and gametogenesis, and Ras/MAPK signalling controls exit from the pachytene stage of meiosis. Broadly the germline goes through two distinct phases – firstly it develops into the complete tissue capable of generating differentiated gametes; secondly it is then continually maintained such that the loss of nuclei to gametogenesis is balanced by proliferation of mitotic nuclei.

Historically, due to the complexities of isolating individual tissues or their RNAs the majority of microarray studies in *C. elegans* have been at the level of the whole animal. Gene expression in any individual tissue has therefore proven difficult to establish. Comparisons of different well-characterised loss-of-function mutants, however, have allowed tissue-specific gene expression to be assessed in the germline. This was aided by the facts that the germline accounts for around half the mass of the adult worm, the great majority of transcripts in the adult, and the expression of ~25% of genes is enriched in this tissue. Consequently changes in gene expression in the germline can be assessed at the level of the whole animal (Jiang *et al*., 2001; Reinke *et al*., 2004; Reinke *et al*., 2000).

For this reason the worm germline is an attractive tissue as the focus of our study. The published expression studies also provide us with an ideal dataset against which to compare our data.

3.2. Outline of Approach

As well as there being many well-studied mutant strains exhibiting brood-size defects, the existence of the *C. elegans* RNAi library permits the generation of loss-of-function animals for almost any gene in the genome. For genes of known function and loss-of-function phenotype, whilst the loss-of-function phenotypes generated by RNAi when visually observed at low resolution do not appear to be as strong as null mutant phenotypes, they nevertheless demonstrate some measure of brood-size defect, as would be expected based on prior knowledge. We therefore have the ability to generate loss-of-function phenotypes for most genes with established roles in germline development.

The stage in germline development at which a defect occurs dictates the extent of development and the mitotic/meiotic character of the germline. I decided to consider four different categories of perturbation in our initial compendium before making comparisons with novel genes. This includes expression profiles of perturbations of genes known to control the three aspects of germline development previously mentioned – maintenance of the mitotic stem cell niche, regulation of the mitosis-meiosis switch, and release from the pachytene stage of meiosis. Thus far, however, all of the genes considered are involved in signalling, transcription and regulation of individual transcripts. Furthermore they appear to have discreet roles in the biology of the animal. In order to provide a contrast to this I chose to perturb components of the basal cellular machinery to see if

they appear distinctly different by array profile. The majority of ribosomal components give completely sterile phenotypes by RNAi. RNAi knockdown of these genes may be expected to give comparable functional defects, reflected in the corresponding microarray expression profiles. Ribosomal knockdowns were therefore added to the study in order to determine whether specific clustering can be achieved and whether the clustering is pathway or strength specific.

To be more clear, the expectations of this study are that the phenotypes of animals deficient for a single component of a signalling pathway will be more similar to that of animals deficient in the same pathway than in another. By using microarrays to generate high-density loss-of-function phenotypes for components of numerous pathways involved in germline development followed by hierarchical clustering, we would expect to rediscover the known pathways as independent branches of the clustering. Novel genes of interest could then be tested against the resulting compendium to provide evidence of their role in a given pathway.

RNA extracted from young adults was used for all experiments in this study. The germline is fully developed by this stage and all of the genes mutated or knocked down in these experiments act before and during the young adult stage.

The two established methods of gene perturbation that could be used in this study are mutation and RNAi. As a long established organism for forward genetics many mutagenesis screens have been performed using ethyl methane sulphonate- (EMS-) or N-

ethyl-N-nitrosourea-(ENU-) induced mutagenesis followed by genetic screening. This has led to a large collection of genetic mutants, which are available to the global *C. elegans* community from the *C. elegans* Genetics Center, USA (http://www.cbs.umn.edu/CGC/). One potential drawback of using such mutants is the possibility of there being some background mutations caused by the mutagenesis, which may not have been removed by out-crossing. Although there are many genetic mutants available there are still many genes pertinent this study for which no genetic mutant is available. RNAi offers an alternative method of genic perturbation, and the RNAi library contains clones allowing the knowdown of the majority of individual coding genes. This therefore necessitates the use of RNAi in this study. RNAi, however, is likely to give less complete perturbation of gene function. I therefore decided to compare RNAi with genetic mutants where possible. The differing level of RNAi knockdown per gene results in a range of brood-size defects. It is also known that there can be a high level of animal-to-animal phenotypic variability on RNAi. The questions that need to be addressed in order to establish the utility of this approach are therefore:

1. Can we rediscover known pathways based on expression profiles (i.e. do different perturbations of the EGF pathway cluster together; do different perturbations of the Notch pathway cluster together and independently of the EGF pathway)?
2. Does RNAi phenocopy mutation (both physiologically and molecularly)?
3. How dependent is molecular phenotype on the strength of the visual phenotype (does strength of phenotype or the pathway that the gene acts in drive clustering)?
4. How dependent is molecular phenotype on the penetrance of a perturbation?

In order to answer these questions, for each gene perturbed I used microarrays to expression profile a population of ~10,000 animals in biological triplicate, DAPI stained whole adult animals to broadly assess the quantity of germline present and assessed the fecundity of 12 individual animals by visual phenotyping. Where multiple RNAi clones existed against a gene of interest in the RNAi library they were each used individually in order to compare different strengths of RNAi against the same gene. Each clone may give different levels of observed sterility owing to the fact that they give rise to a different set of siRNAs, giving different efficiencies and levels of transcript knockdown. The genic perturbations (genetic mutants and RNAi) used for this set of experiments are shown in table 3.1. Note that whilst *sem-5* is not confirmed to be required for progression beyond pachytene, it is upstream of *sos-1* in the canonical EGF/ras/MAPK signalling cascade and gives a brood-size defect by RNAi. Consequently it was included in the first round of experiments.

NOTCH PATHWAY	RIBOSOME	RAS/MAPK SIGNALLING	MITOSIS/MEIOSIS SWITCH AND GAMETOGENESIS
glp-1 (or178)	*rps-1 (RNAi)*	*sos-1 (cs41)*	*gld-1 (RNAi)*
lag-2 (q420)	*rps-14 (RNAi)*	*sos-1 (RNAi)* x3	*gld-2 (RNAi)* x3
emb-5 (hc61)	*rpl-20 (RNAi)*	*sem-5 (RNAi)*	
glp-1 (RNAi)	*rpl-21 (RNAi)*	*let-60 (RNAi)*	
lag-2 (RNAi) x2		*mpk-1 (RNAi)*	
emb-5 (RNAi)		*mek-2 (RNAi)*	
lin-12 (RNAi)		*lin-45 (RNAi)*	
lag-1 (RNAi)			

Table 3.1. Genes involved in germline development perturbed in this study. The nature of the perturbation is indicated in parentheses. The column headings indicate pathway or machinery categories into which the below genes fall.

The microarrays chosen for this study were two-colour synthetic oligonucleotide arrays acquired from Washington University in St. Louis, MO, USA. The microarray contains 22,490 70mer genic probes. Detailed specifications can be found here: http://genome.wustl.edu/genome/celegans/microarray/array_spec.cgi. All experimental samples (Cy3) were hybridized against the same mixed-stage reference sample (Cy5). Each perturbation was compared indirectly to wild-type via a mixed-stage reference sample. The wild-type array profile was derived from animals fed on a bacterial strain expressing a non-targeting dsRNA.

It is typical in expression studies using two-colour microarrays that two samples are compared directly by competitive hybridization to the same microarray. Dye swaps are performed in order to correct for the differing efficiencies of incorporation of labelled nucleotides into cDNA by the reverse transcriptase and the different quantum-yields of the two dyes. "Dye swaps" refers to performing a repeat hybridization of the same RNA samples with the fluorescent labels switched. This approach doubles the number of hybridizations that need to be performed which can be financially prohibitive. Comparison of experimental samples via a universal reference sample negates the need for dye swap hybridizations as the experimental sample is always labelled with the same dye. The key requirement of the mixed stage reference sample is that it provides signal above background for the vast majority of spots on the array such that the corresponding genes are included in the analysis. Comparison between any two conditions on different arrays can then easily be inferred via the reference sample as:

(condition A signal/reference signal) ÷ (condition B signal/reference signal) = condition A signal/ condition B signal.

3.3. Initial microarray data processing, normalisation and assessment of data quality

Since the key manner in which two samples are compared on a two-colour microarray is by the measured ratio of signal present per spot, it is necessary that the signal for both samples is sufficiently higher than the measured background such that the ratios can be considered reliable. For this reason low quality spots are filtered out prior to normalization. Further to this, complex experimental platforms such as microarrays are highly prone to experimental and systematic variation, which must be corrected for before accurate measures of expression changes can be drawn between arrays. An example of this is an imbalance of the two dyes on the array, which may result from the laser settings when scanning the array (experimental) but also the position of the spot on the array (systematic). The term "normalization" therefore refers to the correction for experimental and not biological variation between experiments.

There is no general consensus in the scientific community regarding the best method of data normalization. Multiple methods were therefore tested, each a variation on the well-established loess normalization (Yang *et al*., 2002). This can be done in a global way - normalizing all spots together, or in a block-wise way by dividing each microarray into "sub-arrays" and normalizing within the sub-arrays. Global and block-wise loess normalization, both with and without background subtraction was performed using DNMAD (Tarraga *et al*., 2008). Pearson correlation of normalized biological triplicates

was performed. This was to determine the degree of biological and technical reproducibility of experiments. The correlation between each of three independent replicates may allow the identification of an outlying sample, which should be removed. The difference in Pearson correlation between the same samples for different normalization techniques may also indicate which method best corrects for technical variation. Pearson correlation was improved by filtering out spots giving median intensities <150 in either detection channel. The rationale behind this is that lower intensity spots have a higher percentage error in detection, leading to more variability between replicates. This will, however, lead to the loss of good spots and the spots discarded will be different depending on the quality of array and the gain of the lasers on scanning.

Multiple technical replicates were performed of the wild-type sample against the reference sample and the robustness of the system was assessed by the Pearson correlation. This was found to be consistently 0.93-0.96. Assessment of correlations allowed us to compare the performance of normalization methods. All four of the above methods of normalization performed comparably for good arrays. Global loess performed less well for arrays that exhibited marked positional effects, such as the loss of dye intensity near the periphery of arrays.

An alternative normalization method based on a sliding square window surrounding each spot was also tested (Lyne *et al*., 2003). This method outperformed the others, as it uses smaller windows for normalization around the periphery of the array, allowing it to better

account for positional effects. This method also offers an alternative method of filtering out lower quality spots. Spots with < 50% of pixels > 2 SD above median local background signal in one or both channels are flagged absent, unless one channel showed > 95% of pixels > 2 SD above local background. Removal of spots is therefore more consistent and in-line with the quality of the individual arrays. It also retains spots that are highly expressed in one channel and therefore less susceptible to skewing. The script uses only the lower 55% of pixel intensities as this reduces the likelihood of skewing by bright pixels. This script is also more versatile, allowing the default settings to be altered in a graphical user interface. Alternatively large quantities of arrays can be processed at default settings using the command line. This script therefore not only reduces loss of good spots, but is also favourable should we set up a database for automated microarray analysis.

Table 3.2 shows the Pearson correlations between replicates for all arrays for which data is presented in this chapter. It demonstrates that removal of low intensity spots followed by normalization with DNMAD performs well for good quality arrays. The Lyne *et al*. method broadly performs less well for the same good quality arrays but better for the arrays that gave poor correlations using the previous method. The average correlation across all arrays with both methods is identical. The data for each replicate it therefore more likely to be consistent using the Lyne *et al*. normalization script. Critically, the Lyne *et al*. method of filtering poor quality spots permits on average 50% more genes to be considered. The Lyne *et al*. normalization method was therefore chosen for future use.

Arrays compared		Lyne *et al.*	Flagging spots <150 and DNMAD	Arrays compared		Lyne *et al.*	Flagging spots <150 and DNMAD
N2 control 1	N2 control 2	0.94	0.96	*lin-12* 1	*lin-12* 2	0.90	0.95
N2 control 1	N2 control 3	0.91	0.97	*lin-12* 1	*lin-12* 3	0.92	0.95
N2 control 2	N2 control 3	0.94	0.96	*lin-12* 2	*lin-12* 3	0.93	0.95
emb-5 1	*emb-5* 2	0.90	0.92	*lin-3* 1	*lin-3* 2	0.92	0.93
emb-5 1	*emb-5 3*	0.92	0.92	*lin-45* 1	*lin-45* 2	0.86	0.83
emb-5 2	*emb-5* 3	0.89	0.90	*lin-45* 1	*lin-45* 3	0.82	0.83
*emb-5** 1	*emb-5** 2	0.92	0.95	*lin-45* 2	*lin-45* 3	0.88	0.92
*emb-5** 1	*emb-5** 3	0.87	0.93	*mek-1* 1	*mek-1* 2	0.96	0.96
*emb-5** 2	*emb-5** 3	0.92	0.94	*mek-1* 1	*mek-1* 3	0.87	0.79
gld-1 1	*gld-1* 2	0.90	0.91	*mek-1* 2	*mek-1* 3	0.87	0.78
gld-1 1	*gld-1* 3	0.92	0.95	*mpk-1* 1	*mpk-1* 2	0.88	0.89
gld-1 2	*gld-1* 3	0.86	0.90	*mpk-1* 1	*mpk-1* 3	0.84	0.84
gld-2 a 1	*gld-2* a 2	0.90	0.92	*mpk-1* 2	*mpk-1* 3	0.82	0.80
gld-2 a 1	*gld-2* a 3	0.86	0.90	*rpl-20* 1	*rpl-20* 2	0.83	0.88
gld-2 a 2	*gld-2* a 3	0.92	0.90	*rpl-20* 1	*rpl-20* 3	0.84	0.86
gld-2 b 1	*gld-2* b 2	0.91	0.91	*rpl-20* 2	*rpl-20* 3	0.93	0.90
gld-2 b 1	*gld-2* b 3	0.86	0.92	*rpl-21* 1	*rpl-21* 2	0.88	0.88
gld-2 b 2	*gld-2* b 3	0.89	0.86	*rpl-21* 1	*rpl-21* 3	0.84	0.90
gld-2 c 1	*gld-2* c 2	0.92	0.92	*rpl-21* 2	*rpl-21* 3	0.90	0.90
gld-2 c 1	*gld-2* c 3	0.89	0.87	*rps-1* 1	*rps-1* 2	0.86	0.71
gld-2 c 2	*gld-2* c 3	0.83	0.81	*rps-1* 1	*rps-1* 3	0.79	0.68
glp-1 1	*glp-1* 2	0.83	0.79	*rps-1* 2	*rps-1* 3	0.89	0.93
glp-1 1	*glp-1* 3	0.83	0.86	*rps-14* 1	*rps-14* 2	0.91	0.89
glp-1 2	*glp-1* 3	0.93	0.83	*rps-14* 1	*rps-14* 3	0.94	0.92
*glp-1** 1	*glp-1** 2	0.96	0.95	*rps-14* 2	*rps-14* 3	0.92	0.96
*glp-1** 1	*glp-1** 3	0.91	0.92	*sem-5* 1	*sem-5* 2	0.82	0.90
*glp-1** 2	*glp-1** 3	0.92	0.94	*sem-5* 1	*sem-5* 3	0.92	0.93
lag-1 1	*lag-1* 2	0.87	0.86	*sem-5* 2	*sem-5* 3	0.84	0.89
lag-1 1	*lag-1* 3	0.90	0.86	*sos-1* a	*sos-1* a 2	0.77	0.93
lag-1 2	*lag-1* 3	0.94	0.92	*sos-1* a	*sos-1* a 3	0.92	0.96
la g-2 a 1	*la g-2* a 2	0.89	0.93	*sos-1* a	*sos-1* a 3	0.90	0.95
la g-2 a 1	*la g-2* a 3	0.87	0.86	*sos-1* b	*sos-1* b 2	0.84	0.85
la g-2 a 2	*la g-2* a 3	0.91	0.86	*sos-1* c	*sos-1* c 2	0.92	0.86
la g-2 b 1	*la g-2* b 2	0.88	0.89	*sos-1* c	*sos-1* c 3	0.90	0.96
la g-2 b 1	*la g-2* b 3	0.88	0.89	*sos-1* c	*sos-1* c 3	0.87	0.87
la g-2 b 2	*la g-2* b 3	0.89	0.87		Average	0.89	0.89
*lag-2** 1	*lag-2** 2	0.83	0.79	No. spots considered post-filtering		14404.67	9582.69
let-60 1	*let-60* 2	0.83	0.85				
let-60 1	*let-60* 3	0.88	0.89				
let-60 2	*let-60* 3	0.87	0.94				

Table 3.2. Relative Pearson correlations using different normalization methods. Correlations markedly improved by the Lyne *et al.*, method highlighted in yellow. Low quality arrays that were removed from the analysis are indicated in red. Replicate number is indicated after gene name. Letters between gene name and replicate number indicate use of different RNAi clones. An * indicates a genetic mutant rather than RNAi.

The mixed-stage reference sample against which all experimental samples were hybridized was derived from vast quantities of synchronous animals grown in liquid culture. The RNA extracted from the individual cultures before mixing providing us with known quantities of RNA derived from each developmental stage. A key property of the reference sample is that it must represent the vast majority of annotated genes such that the minimum number of spots will be filtered out prior to normalization. For any given microarray > 85% of spots that are filtered as low quality are filtered due to low signal for both dyes. Across all experiments, of the 22,490 genic spots on the array > 20,300 are represented post-normalization by the filtering criteria used. We therefore consider the mixed-stage reference sample to be of suitable quality for the study.

We have idealized our methodology for producing expression data for any given biological condition. We have determined that the materials that we are producing for microarray analysis are adequately consistent and our initial data processing is robust and practical. We will next determine the differential regulation of genes between the conditions for which data have been generated. This gives us a basis for comparison of the different genic perturbations.

3.4. Proof-of-principle experiments

As is clear from table 3.1, I examined the effect of RNAi knockdown for multiple components of different pathways. Where appropriate mutants were available I sought to compare the effects of perturbation by RNAi and mutation. I also used multiple RNAi clones to target certain genes in order to compare the effects of different strengths of

RNAi against the same gene. Further to this, in order to validate our microarray data I sought to compare it with relevant data produced by other labs.

For each biological condition expression-profiled, differentially expressed genes were identified using Student's t-test. All comparisons were to the reference strain N2 fed bacteria expressing non-targeting dsRNA. This provided us with filtered data for each condition, a means of testing how well RNAi phenocopies mutation and a means of benchmarking our data against published data. The number of genes differentially expressed between the wild-type control and each perturbation is shown in table 3.3.

To check that our methods give similar data to other groups I used the comparison of *glp-1(or178)* with reference strain Bristol N2 (wild-type control), which is analogous to the comparison of *glp-4(bn2)* to N2 by Reinke *et al*. (2004). Both mutants lack a germline, however, the molecular identity of *glp-4* is unknown. Genes more highly expressed in N2 relative to either *glp-4(bn2)* or *glp-1(or178)* can be considered to be germline enriched/intrinsic. Reinke *et al*. define 3143 genes thus using Student's t-test (p-value $\leq$ 0.01). We discover 4831 genes by the same method, encompassing 65% of the Reinke set. We consider this to be a very good overlap, given that this is a cross-platform comparison of a 20K PCR product array (Reinke *et al*.) versus our 22.5K synthetic oligo array. Furthermore the inevitable difference in precise timing at which RNA was harvested between the two labs and the fact that the Reinke *et al.* data is derived from worms fed on *Escherichia coli* strain OP50 and ours from animals fed on HT115(DE3) may further explain the discrepancies.

Gene Perturbed	Genes higher than in N2	Genes lower than in N2
emb-5	1258	1232
*emb-5**	3659	6322
glp-1	1846	2005
*glp-1**	3989	6898
lag-1	654	1757
lag-2 a	2123	2607
lag-2 b	2209	1953
*lag-2**	2614	4076
lin-12	1274	1606
gld-1	2212	1243
gld-2 a	1651	2557
gld-2 b	1951	993
gld-2 c	1629	1713
let-60	1571	1465
lin-45	1496	1155
mek-2	818	857
mpk-1	527	771
sem-5	1404	675
sos-1 a	3262	1811
sos-1 b	1798	516
sos-1 c	1723	2327
pkc-1 a	1031	960
pkc-1 b	1888	2602
rpl-20	2649	3111
rpl-21	1671	2159
rps-1	2408	2808
rps-14	1788	1107

Table 3.3. Genes upregulated and downregulated relative to N2 for each condition. The table shows the number of Genes upregulated and downregulated relative to N2 for each genic perturbation, as determined by Students t-test (p-value <0.05). An asterisk indicates a genetic mutant rather than RNAi. A letter after the gene name indicates use of different individual clones used for RNAi knockdown.

Each condition was compared by hierarchical clustering of calculated ratios of perturbation/wild-type control for each gene differentially expressed between the two conditions (p-value ≤0.05), as can be seen in figure 3.1. It is immediately apparent from the clustering achieved that we recapitulate the known biology, with the components of

the Notch, Ras/MAPK and ribosome gene categories each populating their own separate branch of the condition tree. The components of the mitosis-meiosis switch machinery do not form such a clear niche in the clustering however. This is not completely surprising as the complexities of the dual functions of this machinery means different strengths of perturbation are less likely to consistently generate physiologically analogous animals. Furthermore the consideration of only two genes (albeit one of them appearing three times) may not be adequate to resolve the pathway.

The hierarchical clustering of array profiles is based on a correlation matrix of the differentially expressed genes within all conditions being compared. The standard correlation between all conditions is calculated and each condition arranged in a clustering based on the relative relationship of each condition. This is also performed for each individual expressed gene, leading to a 2-dimensional clustering. For the majority of this chapter I will only discuss one dimension – the clustering achieved between conditions in order to determine the relatedness of perturbations.

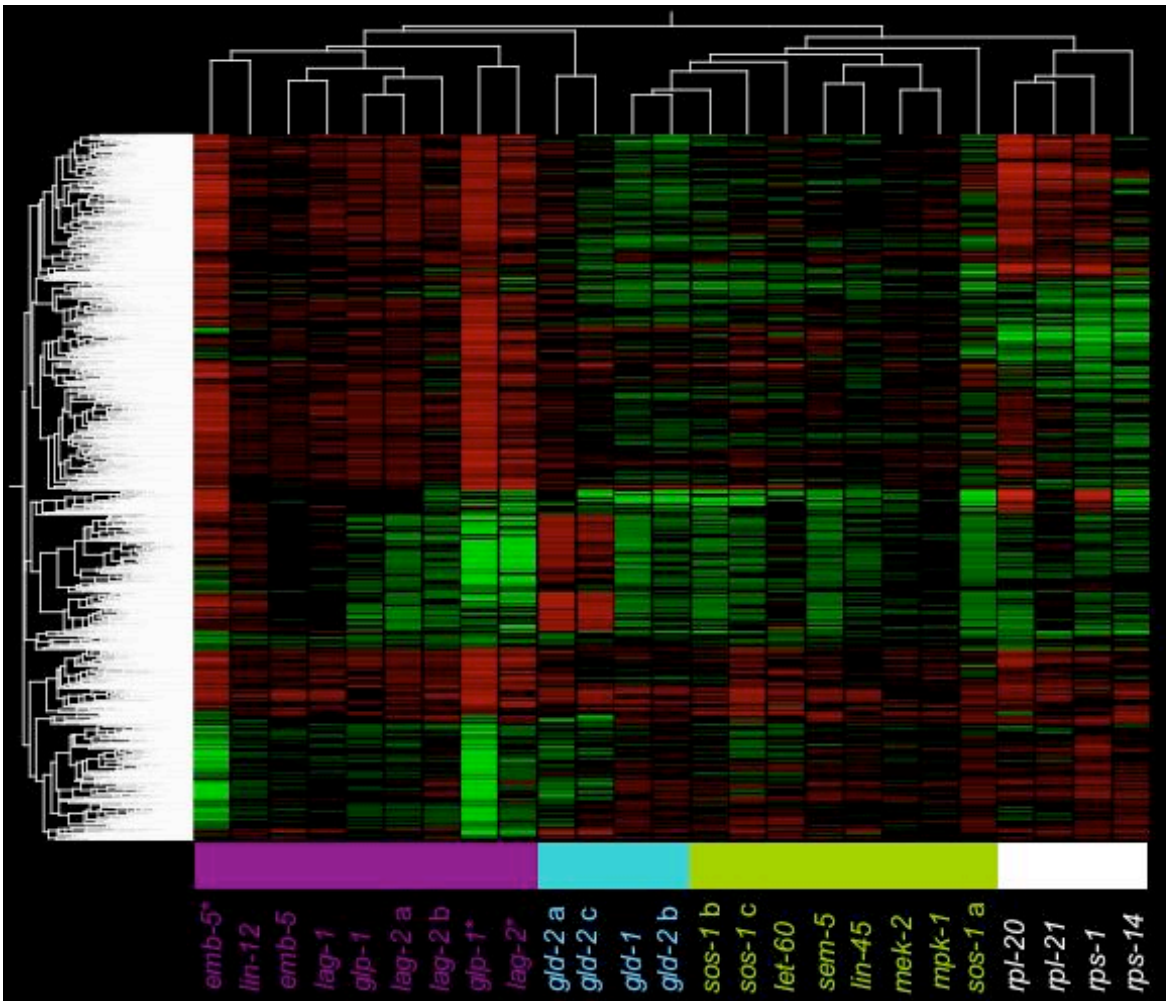


Figure 3.1. Clustering of differentially expressed genes between N2 and each genic perturbation. Calculated ratios of gene signal (perturbation/wild-type) for differentially expressed genes (Student's t-test p-value ≤0.05) were hierarchically clustered. The different pathways and machineries are colour-coded: purple – Notch; blue – mitosis-meiosis switch; green – EGF/ras/MAPK signalling; white – ribosome. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown. Genes upregulated in each condition relative to wild-type are represented in green and downregulated in red. The intensity of colour is analogous to the magnitude of regulation.

3.5. Low-resolution phenotypic analysis of pathway perturbations

It is necessary to establish that the clustering achieved is not simply indicative of strength of perturbation. Of all the genes discussed as having roles in germline development in chapter 1, the genes in table 3.1 are known to give brood-size defects by RNAi. In parallel with the production of each RNA sample the fecundity of 12 animals was measured relative to wild-type for each RNAi perturbation and mutant (figure 3.2). All of the mutants used in this study are temperature sensitive, having a relatively normal brood size at the permissive temperature and being 100% sterile and lacking a germline at the restrictive temperature. The variability in severity and penetrence of phenotype within pathways for the perturbations shown in figure 3.2 suggest that if pathways can be accurately rediscovered using these array profiles, then it is possible to cluster genes giving mild and variable perturbations into pathways. Figure 3.1 demonstrates that the strength of sterility is not driving the clustering as pathways are reliably rediscovered despite Notch and EGF perturbations giving overlapping ranges of sterility.

Animals representing all perturbations shown in table 3.1 have been DAPI stained and the germline imaged (figure 3.3). We find that whilst *glp-1(or178)* and *glp-1(RNAi)* cluster very closely and appear entirely distinct from *mpk-1(RNAi)* by array profile (as one would predict), by this method of staining they appear distinctly different. At up to 400x magnification all Notch mutants clearly have no germline. Notch perturbations by RNAi, however, are indistinguishable from the other perturbations studied at this magnification, even though their sterility ranges up to ~95%. This is understandable as the Notch mutants studied are temperature sensitive and having been grown from L1 at

the restrictive temperature are expected to almost completely negate gene function whereas RNAi has a cumulative effect over time and is unlikely to give 100% knock-down. This suggests that we may not have been able to recapitulate pathways by comparison of mutants and RNAi by staining alone. We have, however, already demonstrated that RNAi can reliably phenocopy mutation on a molecular level.

In conclusion, the clustering achieved appears to be pathway specific even though the extent and variability of brood-size defects overlaps between pathways for different genic perturbations. Whilst the quantity of germline present in genetic mutants and the equivalent RNAi animals can appear markedly different, on a molecular level the animals appear comparable. We therefore consider the methodology to be validated and ready for comparison with selected candidate genes.

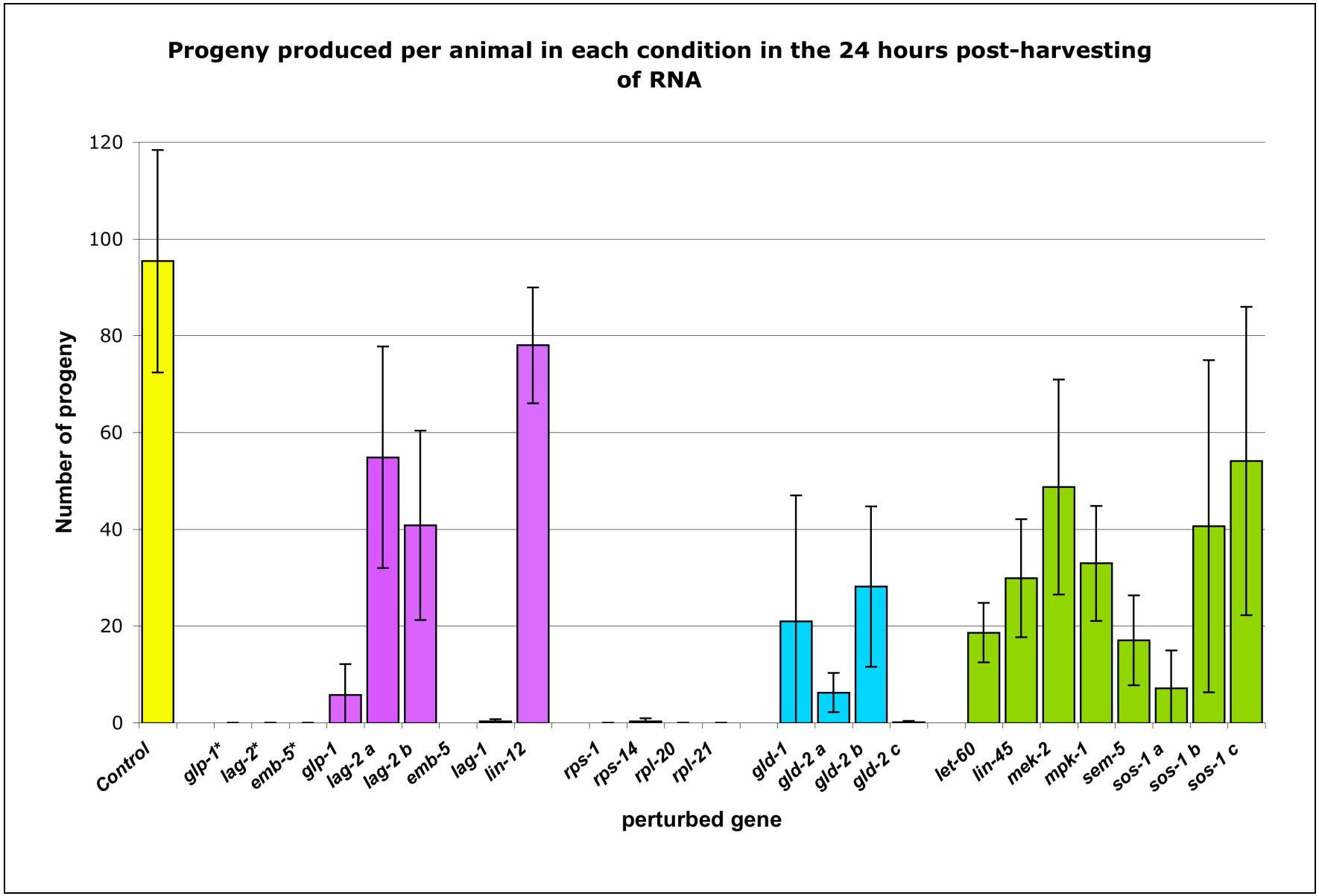

Progeny produced per animal in each condition in the 24 hours post-harvesting of RNA
Number of progeny
120
100
80
60
40
20
0
Control
glp-1*
lag-2*
emb-5*
glp-1
lag-2 a
lag-2 b
emb-5
lag-1
lin-12
rps-1
rps-14
rpl-20
rpl-21
gld-1
gld-2 a
gld-2 b
gld-2 c
let-60
lin-45
mek-2
mpk-1
sem-5
sos-1 a
sos-1 b
sos-1 c
perturbed gene

Figure 3.2. Relative fecundity of germline perturbations. The brood size in the 24 hours after RNA harvesting was assessed for 12 individual animals (3 from each replicate). The graph indicates the number of progeny for each RNAi perturbation, mutant and the wild-type control. Genes are separated and colour-coded according to pathway. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown.

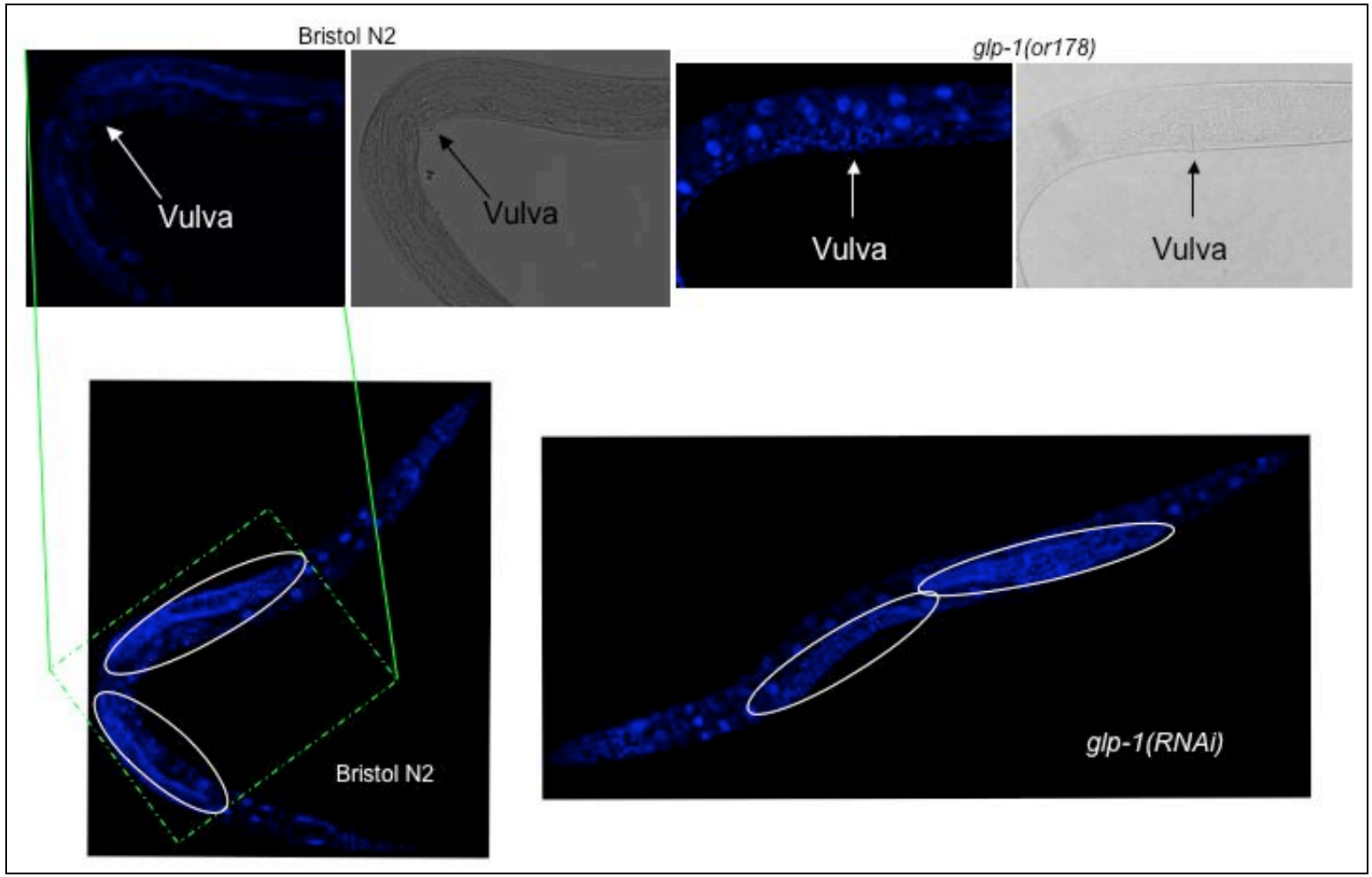


Figure 3.3. DAPI staining of whole animals to assess quantity of germline. This figure shows N2, *glp-1(or178)* and *glp-1(RNAi)* animals as labelled. It is clear that N2 and *glp-1(RNAi)* animals have two clear gonad arms (circled) stretching roughly equidistantly in both directions from the vulva. Higher magnification of this central portion of *glp-1(or178)* reveals no germline.

3.6. Identification of novel modulators of Ras/MAPK signalling in the germline

Once the compendium of well-characterized genes was established it was necessary to decide how to proceed. There were two clear options – (a) to add to the compendium perturbations of genes giving sterile animals by RNAi or mutation, but with no known link to any of the signalling pathways considered; (b) to query the compendium with candidate modulators of signalling pathways already represented in the compendium. These candidate modulators may either have been discovered in genetic interaction screens for genes that modulate the sterile phenotype of Notch and EGF/ras/MAPK signalling mutants or genes that modulate the multi-vulval (Muv) phenotype in mutants with activated EGF/ras/MAPK signalling. Both of these options appeared viable. The next step chosen was therefore to test candidate modulators revealed in vulval screens against the compendium for reasons discussed below.

As discussed in chapter 1, the *C. elegans* vulva is an extremely well studied tissue, serving as an exemplary model for how different signalling pathways combine to regulate the correct development of an individual tissue. Briefly, a set of vulval precursor cells (VPCs) exists along the ventral axis of the animal. EGF/ras/MAPK signalling to the correct cell leads to a cascade of events and the development of a single 22-cell vulva in the centre of the ventral axis, providing a breach between the uterus and the outside world (figure 1.4). Other cells with the potential to develop into the vulva exist along the ventral axis but do not receive adequate stimulus in wild-type animals, ensuring that only one vulval protrusion forms. Mutations leading to an increase in EGF/ras/MAPK

signalling, however, lead to the development of pseudo-vulvae along the ventral axis of the worm.

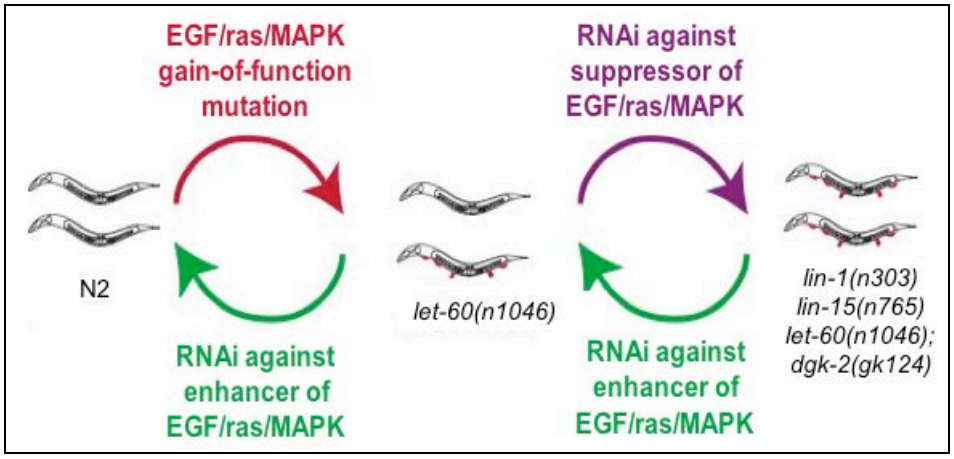


Figure 3.4. Screening for modulators of EGF/ras/MAPK signalling in the vulva. Wild-type animals have a single 22-cell vulva in the centre of their ventral axis. Gain-of-function ras (*let-60*) mutations lead to the formation of pseudo-vulval protrusions (red). RNAi against genes that enhance signalling via ras lead to a decrease in the number of Muv animals i.e. such genes are enhancers of ras signalling. Conversely, RNAi against genes that suppress the consequences of signalling through ras lead to an increase in the number of Muv animals.

In order to identify novel genes that may be involved in EGF/ras/MAPK signalling in *C. elegans*, RNAi screens in mutant animals exhibiting the multi-vulval (Muv) phenotype were performed by Catriona Crombie in the Fraser lab. Specifically, all genes annotated as being signalling (1121), transcription factor (500) or chromatin remodelling (216) genes (Kamath *et al*., 2003) were screened in multiple Muv mutants. Genes that gave a shift in the number of Muv worms by RNAi could be considered candidate modulators of signalling pathways involved in vulval patterning. Genes that when perturbed enhance the Muv phenotype are potential suppressors of EGF/ras/MAPK signalling. Conversely, genes that when perturbed suppress the Muv phenotype are potential enhancers of EGF/ras/MAPK signalling (figure 3.4).

I was specifically interested in genes that are potential enhancers of EGF/ras/MAPK signalling. I therefore selected candidate modulators identified in three different Muv mutants - *lin-1(n303)*, *lin-15(n765)* and *let-60(n1046);dgk-2(gk124)*. As a gain-of-function allele, *let-60(n1046)* gives a Muv phenotype due to increased EGF/ras/MAPK signalling causing more cells along the ventral axis of the worm to adopt 1^{o} VPC fates (see 1.2.2). ~60% of animals carrying this allele exhibit a Muv phenotype 20^{o}C. Genes that enhance or suppress the Muv phenotype can therefore be screened for in this background. A complexity of screening for modulators of the Muv phenotype in the *let-60(n1046)* gain-of-function mutant is that the penetrance of the Muv phenotype is variable, leading to noise in the screens. An unpublished observation made by Andrew Fraser was that crossing of the *let-60(n1046)* gain-of-function allele into a *dgk-2(gk124)* loss-of-function background led to a 100% Muv strain. This suggests that *dgk-2* is a suppressor of EGF/ras/MAPK signalling in the vulva. RNAi screens for suppressors of the Muv phenotype were therefore also performed in *let-60(n1046);dgk-2(gk124)* animals. *lin-1(n303)* and *lin-15(n765)* are both loss-of-function alleles. LIN-1 is a transcription factor and downstream target of EGF/ras/MAPK signalling. Phosphorylation by MPK-1 results in inactivation of LIN-1. *lin-1(n303)* is therefore akin to a EGF/ras/MAPK gain-of-function mutation. 100% of *lin-1(n303)* animals exhibit the Muv phenotype. The *lin-15(n765)* mutation also appears to lead to increased EGF/ras/MAPK signalling, again leading to a 100% Muv population. The *lin-15(n765)* mutation corresponds to loss-of-function of synMuv genes *lin-15A* and *lin-15B*. This may lead to a increase in *lin-3* signalling to the VPCs from neighbouring hypodermal

cells. As a consequence the VPCs that adopt a 3^o fate in wild-type animal adopt 1^o fates leading to pseudo-vulval protrusions.

GENE NAME	% MUV ANIMALS	GENETIC BACKGROUND	GENE FUNCTION
M01B12.5	20	*let-60(n1046);dgk-2(gk124)*	putative RIO kinase
R10D12.10	20	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
pkc-1 a	28	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
pkc-1 b	31	*let-60(n1046);dgk-2(gk124)*	Serine/threonine kinase
D2096.12	41	*let-60(n1046);dgk-2(gk124)*	Protein kinase
D2096.8	72	*let-60(n1046);dgk-2(gk124)*	Nucleosome assembly protein
K08F11.5	79	*let-60(n1046);dgk-2(gk124)*	Predicted Ras related/Rac-GTP binding protein
F27E5.2	53, 17, 15	*lin-1(n303), lin-15(n765), let-60(n1046);dgk-2(gk124)*	PAX transcription factor

Table 3.4. Selected genes suppressing the Muv phenotype in RNAi screens in 100% Muv mutants. Indicated are the genes against which RNAi was performed, the average % Muv animals across the three screens, and the mutant backgrounds in which the hits were observed. A letter following the gene name indicates multiple individual clones used to independently target the same gene.

A total of 24 novel genes were identified as consistently suppressing the Muv phenotype in three independent screens. All of these genes could potentially be tested against the compendium of expression profiles. A set of 7 genes (table 3.4) were initially selected for testing. RNAi against all of these genes except one gave severe morphological defects in the animals. This was problematic for two reasons – firstly it made the animals extremely difficult to stage accurately; secondly, it made it likely that there would be considerable changes in expression as a result of somatic defects. Consequently these

genes were discarded. The one selected candidate modulator of EGF/ras/MAPK signalling that yielded a seemingly wild-type phenotype with slight brood-size defects on RNAi in N2 was *pkc-1*. Two RNAi clones targeting *pkc-1* exist in the library, both of which reduce the severity of the Muv phonotype in the *let-60(n1046);dgk-2(gk124)* mutant. This implies that *pkc-1* may be an enhancer of EGF/ras/MAPK signalling. When RNAi against *pkc-1* using the two different clones was tested against our compendium of array profiles *pkc-1* clustered with the known EGF/ras/MAPK pathway in both cases (figure 3.5). The Muv screens and expression profiling of *pkc-1(RNAi)* therefore provide two independent forms of evidence that *pkc-1* is involved in EGF/ras/MAPK signalling in *C. elegans*.

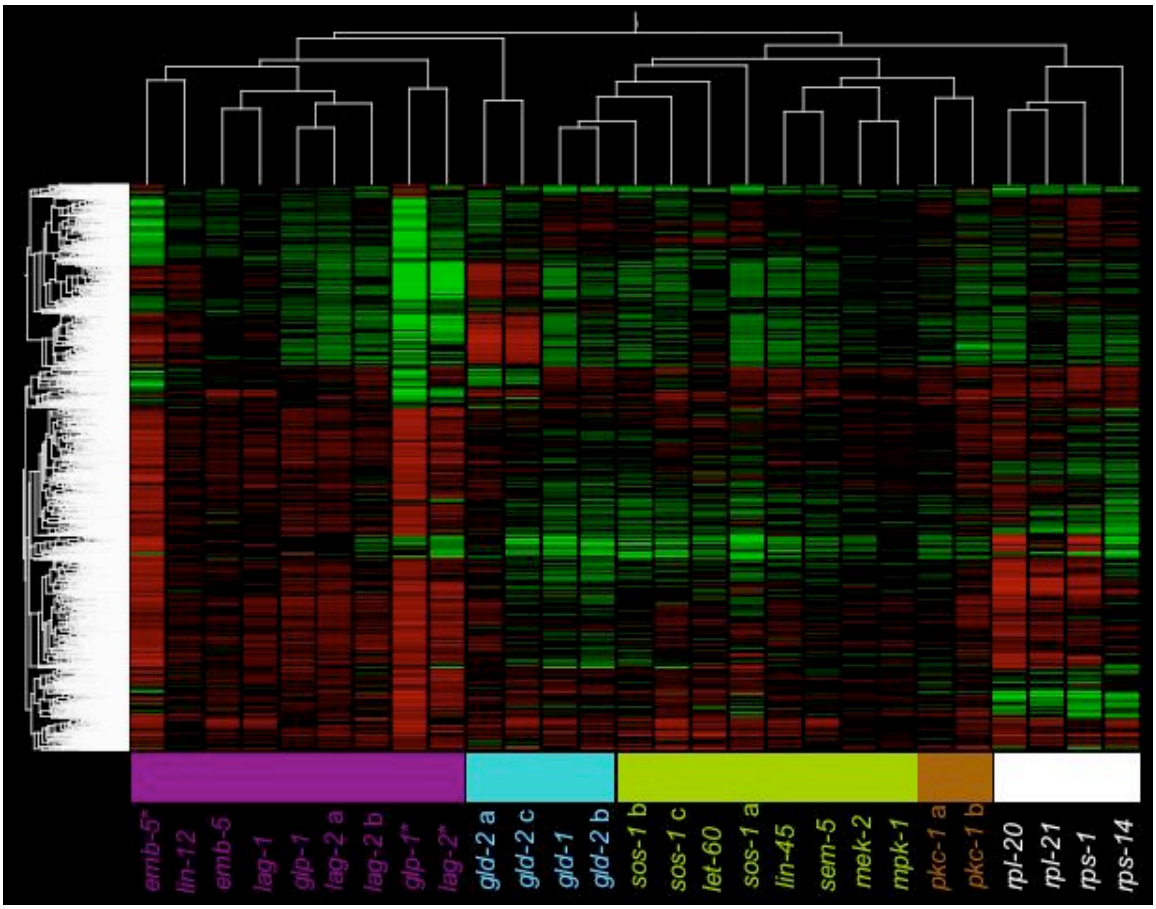


Figure 3.5. *pkc-1* clusters with the EGF/ras/MAPK signalling pathway. Genes and colour scheme are as in figure 3.1. The two RNAi experiments followed by expression profiling of *pkc-1* are labelled in brown. Lowercase letters following gene name indicates the use of different individual RNAi clones targeting the same gene. An asterisk indicates a genetic mutant rather than RNAi knockdown.

There is a well-established and conserved functional relationship between *pkc-1* and *dgk-2* (reviewed in Mellor and Parker, 1998; Merida *et al*., 2008; Nishizuka, 1984). *pkc-1* is an orthologue of mammalian protein kinase C, which is a diacylglycerol (DAG) dependent protein kinase. *dgk-2* is an orthologue of mammalian DAG-kinase, which phosphorylates DAG, converting it to phosphatidic acid. In this way it removes an essential factor for *pkc-1* activity (figure 3.6). Loss-of-function *dgk-2* therefore leads to increased *pkc-1* activity. That *dgk-2* loss-of-function increases the Muv phenotype in *let-60(n1046)* animals and *pkc-1(RNAi)* decreases it in *let-60(n1046);dgk-2(gk124)* is further

evidence that DAG signalling and EGF/ras/MAPK signalling are functionally related. Functional links between PKC and EGF/ras/MAPK signalling have previously been identified in mammalian and avian species (e.g. Banan *et al*., 2001; Crotty *et al*., 2006; Heo and Han, 2006; Lee *et al*., 2006a; Lee *et al*., 2006b; Sriraman *et al*., 2008).

The clustering of *pkc-1(RNAi)* as predicted amongst the other conditions in the compendium demonstrates our ability to provide further evidence of the signalling modulation indicated by the RNAi screens of the Muv phenotype. Our identification of *pkc-1* in this way represents a firm hit and will likely lead to further comparisons of screening-detected signalling modulators against our compendium.

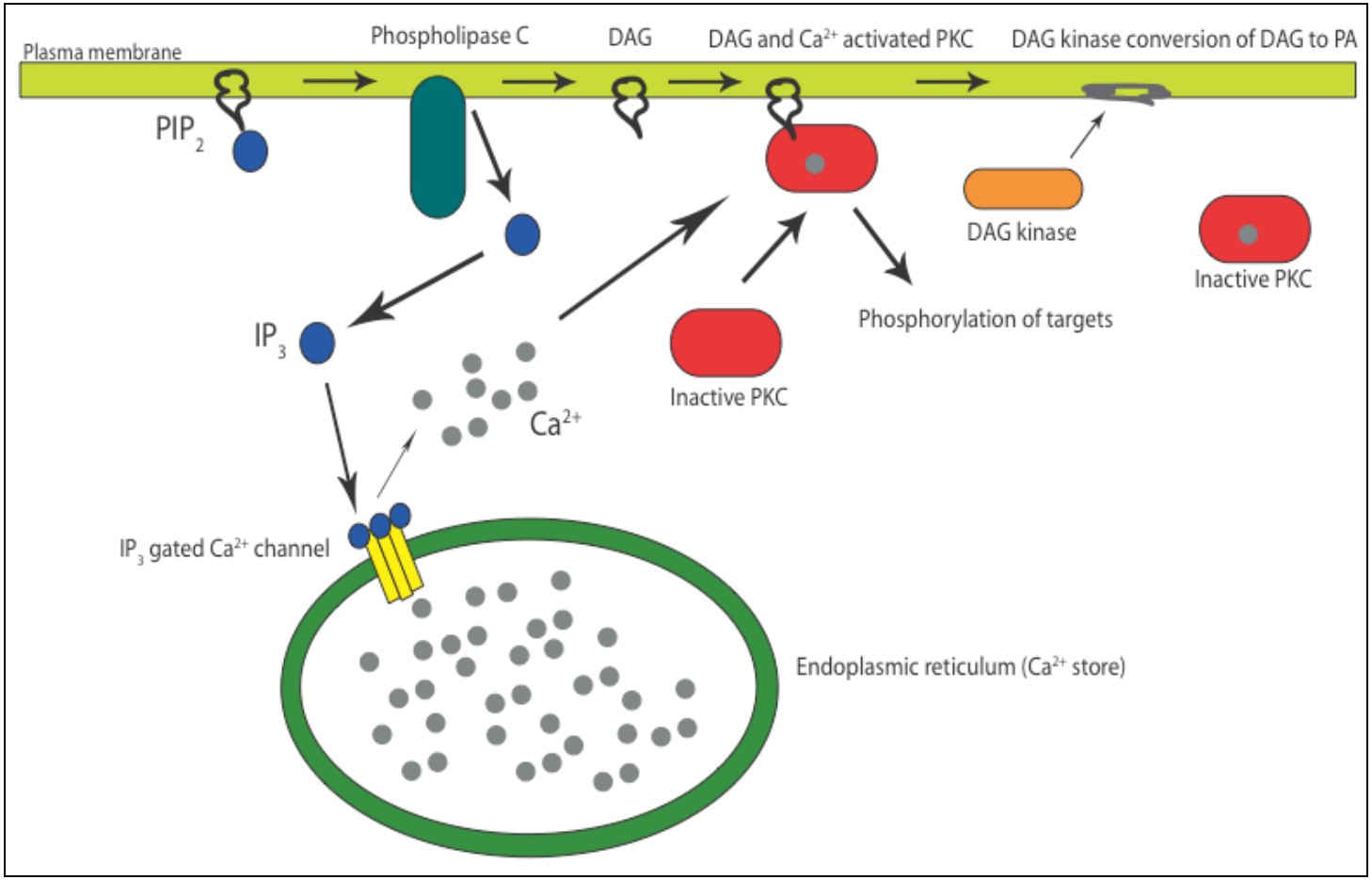


Figure 3.6. The activity of PKC is modulated by the activities of PLC and DGK. Phospholipase C (PLC) converts phosphatidylinositol bisphosphate (PIP_2) to inositol trisphosphate (IP_3) and diacylglycerol (DAG). Increased cellular IP_3 leads to the opening of IP_3 gated Ca^{2+} channels in the endoplasmic reticulum. Protein kinase C (PKC) is then activated by Ca^{2+} binding and tethering to the plasma membrane by DAG. DAG is converted to phosphatidic acid (PA) by DAG kinase (DGK). This results in PKC being released from the plasma membrane and inactivation.

3.7. The differentially expressed genes

It would be a missed opportunity to consider this data set only in terms of our ability to distinguish functional relationships between perturbed conditions. Rather, the genes that change in expression are likely to be of some interest in themselves. A number of papers from the Reinke and Kim labs over the years have used comparative expression profiling of mutant animals to identify genes enriched in the germline, gametes and both male and hermaphrodite soma (Jiang *et al*., 2001; Reinke, 2002; Reinke *et al*., 2004; Reinke *et al*., 2000). Our knowledge of the physiological changes caused as a result of perturbing these genes means that we know which parts of the germline should be enriched for each set of perturbations. We also have a number of perturbations in each category meaning that the number of times we see the same gene change in each can be a measure of our confidence that the expression of these genes is enriched in those regions. Specifically, genes upregulated in animals with Ras/MAPK signalling perturbations may be highly expressed in meiotic prophase. Conversely, the genes downregulated are likely to act after meiotic prophase, such as in gametogenesis. Genes downregulated on Notch perturbation are likely to be generally germline enriched genes. Upregulated genes may be enriched in the soma. Such lists of genes can be limited to genes specifically regulated only in certain conditions. For example, genes downregulated for every Notch perturbation but not downregulated for any other perturbation are highly likely to be mitotic-enriched genes. Genes upregulated for every Ras/MAPK perturbation and no other condition are more likely to be meiosis-enriched genes without contamination of soma-enriched genes. Genes up- or downregulated on either Notch or Ras/MAPK

perturbation and not the other or ribosomal perturbation are listed in appendix 1 (data CD), along with the number of perturbations of that class for which that regulation is seen.

As to the different general properties of the genes that fall into these classes, interpretation has proven difficult. Firstly, as is apparent from the clustering, the number of genes changing for any perturbation ranges from many hundreds to many thousands. Too much is changing for individual processes to be singled out. There is little functional information assigned to many genes and that which is, is often derived from their differential expression patterns observed in microarray experiments (e.g. sperm enriched genes). The identification of such genes being under-represented in a compendium of germline perturbations is not novel and of little biological value. An obvious analysis would be to see if any of our resulting gene lists are significantly enriched for any Gene Ontology (GO) terms – a set of definitions relating to gene properties or function. In *C. elegans* this is a fruitless endeavour as there are insufficient GO terms assigned to genes such that any statistical inference can be made. This is not to say that there is no value in this differential expression information beyond its ability to drive clustering of conditions. Numerous recent studies have applied the knowledge of common expression patterns amongst comparable conditions as the source data for biological network construction (Beer and Tavazoie, 2004; Freeman *et al*., 2007). This dataset may be ideally suited to such analysis, a possibility that is worth pursuing in future.

3.8. Discussion

Considering the progress made to this point it seems sensible to compare the approach relative to a more conventional staining approach. Array profiling is a powerful methodology and offers potential advantages over a staining approach for a number of reasons. Firstly, previous work as well as the data presented here has shown that the animal-to-animal variability of RNAi means that methodologies considering populations rather than individuals are more clean and powerful. Each RNA sample used in this study is derived from ~10,000 worms, many more than could be analysed post-staining for mitotic/meiotic markers. Secondly, microarrays offer an established technological platform that can test vastly more parameters than maximally 4-colour histological staining. It also lends itself to straightforward statistical analysis, which is preferable to counting large numbers of nuclei and attempting to categorise perturbations based on morphology and staining. The wealth of signalling components that lead to sterility may indicate hitherto unrecognized pathways and machineries involved in germline development. Our ultimate goal was to categorize such genes, which could potentially be beyond the capacity of current histological staining methods. A potential defect of this methodology, however, is that it is likely to be insensitive to physiological changes affecting only a few cells. Such changes are more likely to be identified by a detailed staining approach.

The rediscovery of the known biological machineries by clustering of the array profiles is firm evidence of our ability to place genes in pathways based on biological function. Since the clustering is inevitably very plastic and subject to change depending upon the

array profiles added to it, a method of testing the robustness of the clusters should perhaps be applied. An example of this would be a bootstrap approach. This would involve multiple rounds of removing random sets of genes and reclustering. The ability of the pathways to remain together in isolation within the clustering under these circumstances may act as an indicator of how strong the associations are within the clustering. It may also identify the key genes, which drive the clustering.

The obvious next step is the querying of more genes against the compendium. As previously stated, the list of candidates is vast including all signalling and transcription factor genes giving sterile phenotypes for as yet undetermined reasons. This list could be limited to genes that give sterile genetic interactions with components of the Notch or Ras/MAPK pathways i.e. genes that increase the brood-size defect of Notch or Ras/MAPK mutants by RNAi. A complexity of this is that genes identified in genetic interactions screens often interact with components of both pathways and others (Lehner *et al*., 2006). This hints at the complexities of interpreting genetic interactions but perhaps this expression approach represents an opportune system to study this.

It is clear that any such inference of gene function via a compendium such as this requires additional forms of evidence before inference can be considered confirmed. An obvious way in which this could be done is detailed dissection and staining of germlines. A number of markers have been suggested for immuno-staining of germlines (Crittenden and Kimble, 2008). These markers can be used to determine the relative quantities of each region of the germline. For example, GLP-1, FBF-1, FBF-2 or CEP-1 could be used

to mark the mitotic region. HIM-3 could be used as a marker of meiotic prophase, whilst RME-2 and SP56 mark the oocytes and sperm respectively. Staining of *pkc-1(RNAi)* germlines represents an obvious candidate for such staining. In this case we would expect to see an increase in the HIM-3 stained regions and decrease in RME-2 and SP56 stained regions relative to wild-type. That said, it is the complexities and limited resolution of this that was the motivation for this project in the first place. The limited brood-size defect for some of the conditions that appear in the compendium may indicate that germline staining may be inconclusive. The reality, however, appears to be that in order to assign genes to pathways at least a subset of novel genes added to the compendium would have to be evaluated in this way. For a subset of genes to exist many more genes would have to be tested against the compendium. Whilst obvious candidate genes exist, it was necessary to weigh the value of pursuing this project further against the potential of other projects to bear fruit. The project detailed in the following chapters was running concurrently with this in order to provide a fall-back position should this project have proven unworkable. Although this project appears far from unworkable it was not pursued further as it was deemed of lower potential than that which follows.

Chapter 4
Analysis of the wild-type *C. elegans* transcriptome

4.1. Introduction

The *C. elegans* genome was the first of any metazoan to be completely sequenced, this feat having been achieved in 1998 (C. elegans Sequencing Consortium, 1998). Furthermore it was only the second eukaryotic genome to be completed, after *S. cerevisiae*. Annotation of the ~100Mb genome of *C. elegans* is excellent and arguably more advanced than that of other animals. Regardless of this a completely stable set of gene annotations has not yet been achieved, with new releases (albeit with only minor changes) every month or so. My intention was to determine how well gene annotations corresponded to the transcribed regions of the *C. elegans* genome using whole-genome Affymetrix GeneChip® *C. elegans* Tiling 1.0R Arrays. Similar studies done in *Arabidopsis thaliana*, *Drosophila melanogaster* and humans had revealed that vastly more of each genome is transcribed than could be accounted for by then current annotations (Bertone *et al*., 2004; Hanada *et al*., 2007; Manak *et al*., 2006; The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group), 2006). The genome of *C. elegans* is already considered to be transcriptionally dense, with ~62% of the genome thought to be genic and ~33% exonic (WS150 release of Wormbase). The Affymetrix tiling arrays used can survey the transcriptome to a resolution of 35bp. When this project was conceived these microarrays were not yet commercially available. This project was therefore a collaboration with the laboratory of T.R. Gingeras, Affymetrix Inc., Santa Clara, CA., USA where the microarray hybridizations were performed. The informatics was performed in association with Arun Ramani, a postdoctoral researcher in the Fraser lab.

In order to achieve adequate cover of the transcriptome for this study total RNA from six different developmental stages in the *C. elegans* life-cycle, specifically embryos, L2, L3, L4, young adults and gravid adults was hybridized in at least duplicate. This RNA was derived from the wild-type reference strain Bristol N2. Not only was this done to give us maximum coverage of the transcriptome, but also to give an adequate data set for comparison with the NMD-deficient transcriptome, as will be seen in chapter 5. The output of this platform is a set of probe intensities for the ~3 million probes arrayed on each chip, analysis of which reveals the regions of the genome for which transcript is present in the sample.

The use of tiled microarrays allows us to survey all transcribed regions of the genome and therefore examine how transcript structures change as well as transcript levels. Historically, however, single colour tiled microarrays have not been used to generate gene intensities and determine differential expression between conditions. With no established methodology and pipeline by which to do this it was required that we develop our own analysis strategy. Also, as with any other technology platform, validation of the output was required before the data could be considered reliable. One possible method of validation would be exhaustive RT-PCR and sequencing to confirm the existence and identity of novel transcribed regions and structural changes indicated by the tiling data. A superior alternative now available to us, however, is ultra-high density sequencing of cDNAs. This automatically allows validation of novel features and gives information on connectivity of structures by identification of reads that span exon-exon boundaries.

Furthermore the number of reads that map to a given structure act as an expression value with which we can compare gene intensity values derived from tiling data. This therefore allows validation of transcript prediction and intensity values simultaneously. Consequently, we produced ultra-high density sequence data using the Illumina platform for two developmental stages individually (L4 and young adult), as well as a mixed stage sample containing RNA derived from all developmental stages in the worm lifecycle in order to give us maximum coverage of the transcriptome at the depth available. The Illumina sequence data have the advantage of being of greater resolution than the tiling array data but could not adequately replace the tiling array data, being of insufficient depth (i.e. insufficient number of unique reads) and providing stage-specific information at fewer stages. The purification of RNA for sequencing and tiling array analysis excludes RNAs <200nts. Consequently such RNAs are not represented in the data.

In this chapter I will demonstrate the quality of the tiling array data by comparison with the sequence data. I will then present the protocols established using the two forms of data produced and how they inform us on the current state of gene annotations. I will discuss how our data relate to the density and accuracy of gene predictions as well as how they can be used to predict changes in splice forms and connectivity between annotated and predicted structures.

4.2. Tiling array data normalization

All Affymetrix GeneChip® *C. elegans* Tiling 1.0R Array data presented in this thesis was quantile normalized prior to use. Quantile normalization is a standard approach

applied to one-colour microarray data (Bolstad *et al*., 2003). As discussed in chapter 3, there are two forms of variation that occur between individual microarray experiments – biological variation and technical variation. The goal of normalization is to reduce technical variation. Differences in labeling efficiency of samples, quantity of material hybridized and the gain of lasers used to scan the arrays are all examples of what introduces technical variation. The key assumption made by quantile normalization is that the true biology-driven distribution of probe intensities on a one-colour microarray is the same between all arrays. Quantile normalization takes all probes on an array and sorts them in order of intensity. This is done for all arrays that are to be compared. The mean of the probes for each array at each sorted position then becomes the normalized probe intensity at that position (e.g. the tenth highest probe intensity on all arrays is now the same – the mean of the non-normalized intensities). Each array now has the exact same probe intensities but the intensities are not assigned to the same probe, rather the ranking of intensities for each probe within an array is the same as before but the distribution of probe intensities is now the same for all arrays. Consequently the mean probe signal for all arrays is also the same. All microarrays are now comparable.

4.3. Defining regions of tiling array signal along genomic coordinates

In order to call regions of the genome as expressed using tiling array data it is first necessary to define the methodology and criteria by which this is to be done. There are two distinct ways in which this has been done in previous studies, each with its advantages and disadvantages. A method originally implemented by Wolfgang Huber at the EBI, involves aligning the signal acquired along genomic coordinates and then

dividing the signal into runs of probes showing similar intensities, thus defining transcript and intron-exon boundaries (David *et al*., 2006). This methodology has the advantage of not using gene annotations as a reference and is therefore completely unbiased. A disadvantage is that it requires the user to pre-define the number of partitions that should be drawn in the signal, which is distinctly problematic without reference to a defined set of controls, such as annotated gene structures. Knowledge of the annotated gene structures would permit optimization in order to ensure that expressed exons are not partitioned or fused during the analysis, ensuring that an accurate number of partitions are drawn in the data. Ultimately, however, the number of transcribed regions called by this method is defined by the user rather than the data, which may not be the best method for the purposes of transcript discovery where the user cannot know in advance how many regions of expression to expect.

An alternative way of defining regions of signal is by identifying runs of probes above a calculated background. Again, theoretically this requires no prior knowledge of or reference to annotated gene structures but the complexities of the methodology eventually demand optimization of the technique relative to a set of controls, of which annotated genes are likely to be best. An assumption when optimizing this technique therefore, is that the gene annotations used for comparison are close to correct. This is appropriate for the purposes of transcript discovery, as it makes no assumptions as to the number of genomic regions that correspond to a retained RNA but does ensure that the number of regions discovered is represented as accurately as possible relative to the known characteristics of the transcriptome. The output of such an analysis is discreet

regions of the genome for which transcript exists at a detectable level. Such detected regions are referred to as transcribed fragments or "transfrags" (figure 4.1). Satisfied that this was the most appropriate methodology to identify transcribed regions of the genome, this was the approach we used.

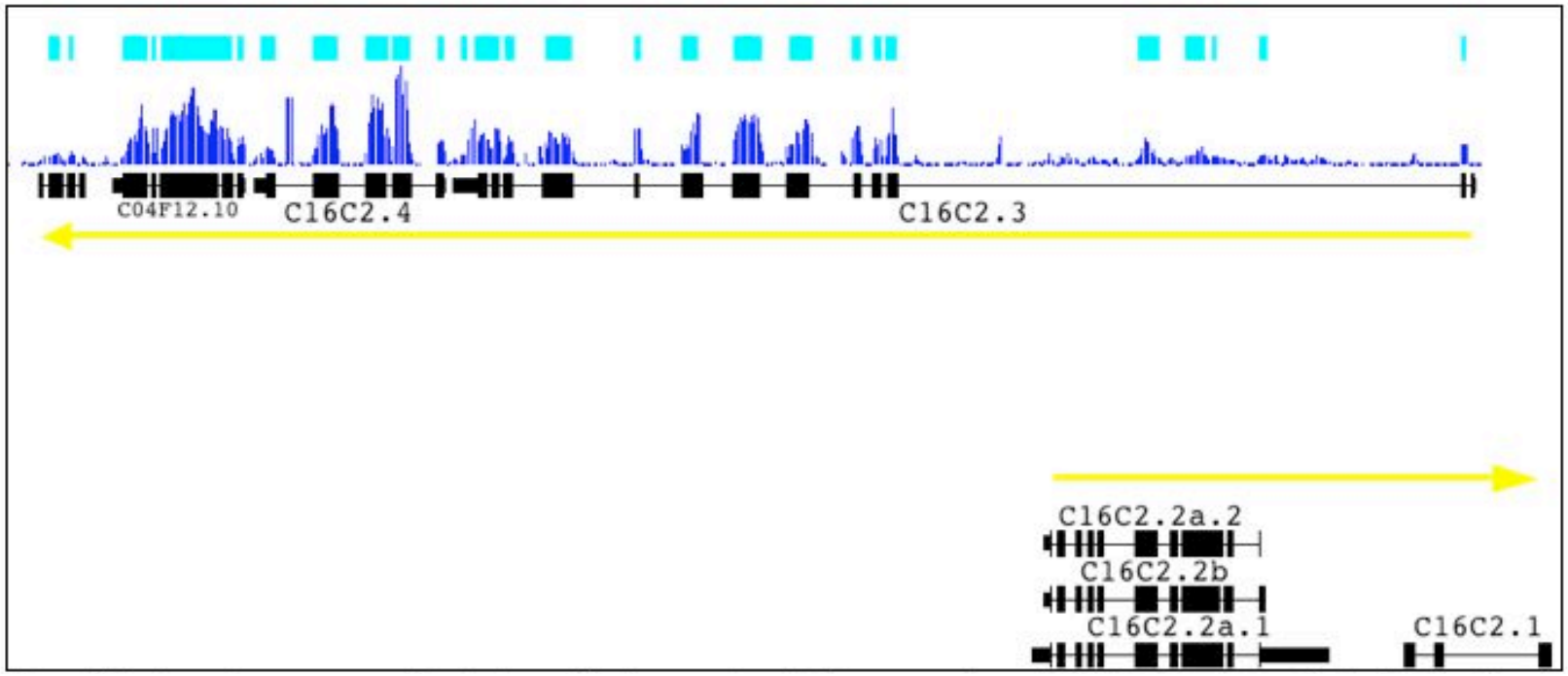


Figure 4.1. Transfrags corresponding to transcribed genes. Annotated genes are shown in black and are transcribed in the direction of their neighbouring yellow arrow. Normalized probe signal is shown in dark blue and the transfrags generated from that signal in light blue. As can be seen, transfrags broadly represent individual exons. There is not necessarily a transfrag for every exon for lowly expressed genes and transfrags may not represent full-length exons. Where short introns exist such that few or no probes map to that structure then exons may be merged into a single transfrag. Broadly, however, one transfrag = one exon; one exon = one transfrag.

4.4. Idealizing parameters for building transfrags

The interval analysis that defines transfrags for any given data set was performed using Affymetrix Tiling Analysis Software (TAS) version 1.1. Prior to interval analysis the data from each replicate are quantile normalized together in R (http://www.r-project.org). The three key parameters that then need to be defined for the interval analysis are the background, the maximum gap (maxgap) and the minimum run (minrun). The background is the threshold above which a probe intensity is considered. The minrun represents the number of consecutive probes that must be above background before a transfrag can be identified spanning that region, in terms of the number of bases of genome represented by those probes. The maxgap is the maximum amount of genome for which there is no signal above background that can be tolerated before a transfrag is terminated. In optimizing the interval analysis relative to gene annotations there are three assumptions that are made. The first is that for each expressed exon (i.e. exon to which a transfrag maps) there should be only one transfrag. If exons are being artificially split into numerous corresponding transfrags this is an indication that the maxgap is too low. Alternatively it could be that the minrun is too low and therefore low-level random noise is being called as transfrags. The second key assumption is that for each transfrag that maps to a gene, it should only span one exon. If a transfrag spans multiple exons then maxgap is likely to be too high, leading to the artificial fusion of transfrags. All of this assumes that the background threshold has been set such that noise is maximally reduced without loss of real signal. Background threshold was calculated to include the top 5% of non-genic probes on the array. This is summarized in figure 4.2. In order to satisfy the "one exon, one transfrag; one transfrag, one exon" optimization strategy a range of

maxgap and minrun combinations were tested and the combination most closely matching the criteria was selected. This was maxgap = 35bp, minrun = 70bp. As the tiling array is made up of 25mer probes tiled at an average genomic distance of 10bp, this is effectively a minrun of two probes and a maxgap of one probe.

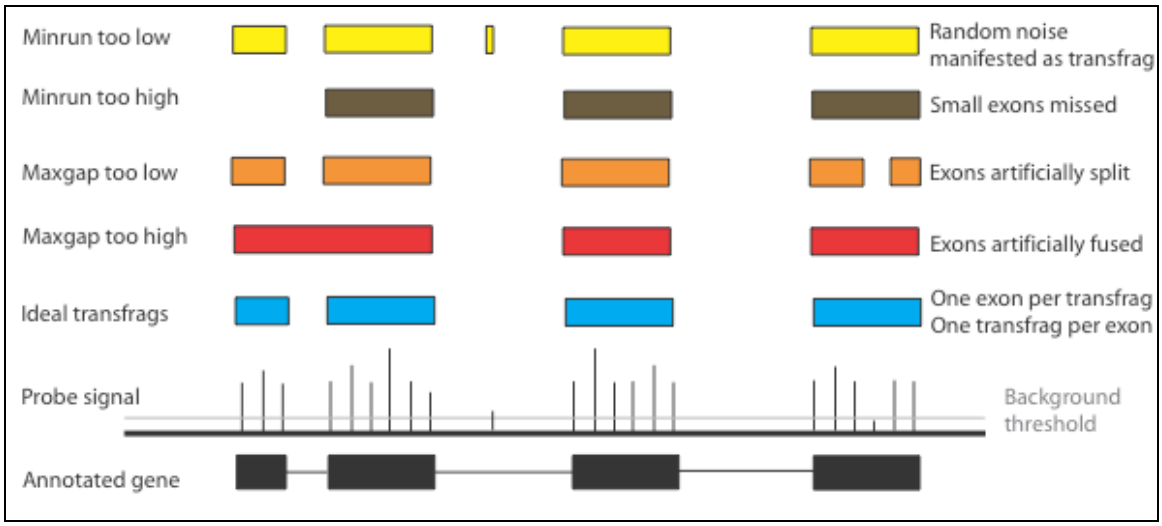


Figure 4.2. Selection of transfrag building parameters schematic. The parameters for building transfrags to represent transcribed regions of the genome were optimized such that one transfrag corresponded to one exon and one exon corresponded to one transfrag. This required that exons were not artificially fused or split by the use of inappropriate maxgap and minrun values.

4.5. Comparison of transfrags with the genome

Each transfrag was classified as either overlapping an annotated gene (genic) or not (extra-genic). The genic transfrags were then further classified as exonic if overlapping an exon. The number and percentage of transfrags within each category detected at each stage is shown in table 4.1.

Stage	Total transfrags	Genic	Percent	Exonic	Percent	Extra-genic	Percent
Embryo	36205	34886	96.36	33610	92.83	1319	3.64
L2	57564	53778	93.42	49499	85.99	3786	6.58
L3	49968	47717	95.50	45219	90.50	2251	4.50
L4	45770	43804	95.70	42050	91.87	1966	4.30
Young adult	46126	44139	95.69	42644	92.45	1987	4.31
Gravid adult	43507	41439	95.25	40045	92.04	2068	4.75

Table 4.1. Transfrag distribution at each developmental stage.

As is clear from table 1, the vast majority of transfrags detected are genic suggesting that the *C. elegans* genome is well annotated and there is not much novel transcription. This will be discussed further at the end of the chapter.

4.6. Measuring gene expression using tiling arrays

By the specification of the microarray design there is a probe every ~35bp, thus there are many probes per gene. Owing to the constraints of the array design, however, probes are not idealised and all behave differently. Furthermore for any given condition the probes that cover a gene which are above background may be different as a consequence of both biological and technical variability. Probes on a microarray are considered to behave differently as a consequence of their different binding capabilities owing to their different nucleotide constituents. The problem of how to derive a gene intensity from a set of probe intensities is therefore not as simple as taking the mean or median intensity across all probes above background as different probes will be used for each calculation. There are two possible methods of reducing technical variability introduced by using different probes for such a calculation. One approach is to correct for probe behaviour and the

other is to consider only exons and genes for which there is a sufficiently high number of probes for which there is signal above background. In the latter case the variability in individual probe intensity should be neutralized by the use of many probes.

The method of correcting for probe behaviour that has previously been used is to correct probe intensities from cDNA hybridizations relative to probe intensities derived from hybridization of genomic DNA (David *et al*., 2006). Hybridized genomic DNA is theoretically present at a ratio of 1:1 between probes and so the consequent probe intensities are representative of the binding characteristics of each probe. By this method all probe intensities should become more consistent relative to each other within a transcribed structure. Fewer probes should therefore be required to give a representative gene or exon intensity. Ultimately, however, this approach requires the optimization and performance of genomic hybridizations for potentially minimal gain as before an exon or gene could confidently called as expressed it is desirable that the majority of probes within any structure to be considered are above background. Furthermore for structures with relatively few probes above background it is possible that they are expressed at a low level but low intensity probes are more susceptible to errors in detection regardless of correction for probe behaviour. Calculating a gene intensity based on a small number of such probes is therefore inadvisable. Consequently we opted to stringently filter structures for which the majority of probes were above background and calculate intensities accordingly. Our criteria for doing this were to consider only exons for which more than 50% of probes were above background and only genes for which more than 50% of unique exons matched this criterion. We consider this to be reasonable as genes

not matching these criteria are generally too lowly expressed and probe intensities too close to the background cut-off to be considered accurate. The gene intensity is then taken as the median intensity of the probes filtered by the above criteria. Median rather than mean intensity was used, as this method is less susceptible to skewing by outlying probe intensities. The background threshold is calculated to include the top 5% of non-genic probes. A schematic of how gene intensities is calculated from both tiling array and Illumina sequence data is shown in figure 4.3.

4.7. Measuring expression using ultra-high density sequence data

The output of Illumina sequencing technology is ~3 million 35bp reads per sample sequenced. Alignment of reads uniquely mappable to the genome or annotated transcriptome leads to a certain number of reads overlapping each nucleotide. Each nucleotide can therefore be given an intensity score, which is the number of times it occurs in mapped reads. Gene intensities from sequence data are therefore calculated as the median number of reads that map to a nucleotide for which there is at least one read.

Table 4.2 shows the number of genes at each stage for which an intensity score can be derived by the criteria discussed for each of the technologies. The generation of gene intensities allows comparisons to be drawn between conditions to infer changes in overall gene expression or change in major splice form of genes. It is necessary, however, to demonstrate that these gene intensities are truly representative before such analyses can be undertaken. To this end gene intensities derived from tiling data were compared to gene intensities derived from sequence data. If these intensities look similar this is solid

evidence that the derived gene intensities are reliable and representative of true transcript abundance.

Stage	Tiling	Sequence	Overlap
Embryo	4471	NA	NA
L2	7323	NA	NA
L3	7208	NA	NA
L4	6355	7043	5164
Young adult	7220	6716	5681
Gravid adult	6577	NA	NA

Table 4.2. Number of genes called as expressed by each technology and the overlap between these lists.

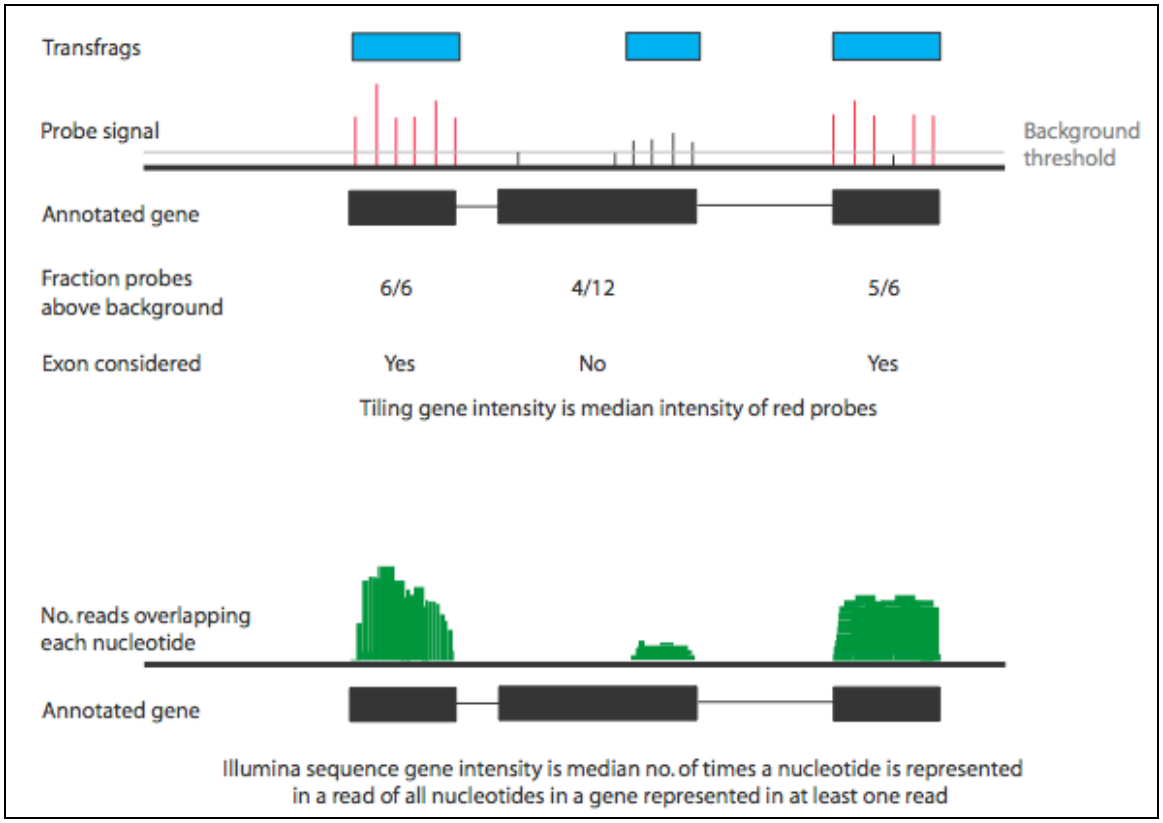


Figure 4.3. Calculating gene intensity values from tiling array and Illumina sequence data. For the tiling array data the gene intensity is the median probe intensity of all probes above background in exons for which ≥50% are above background (red probes). The background threshold is calculated to include the top 5% of non-genic probes on the array. The gene intensity derived sequence data is based on the number of times a base within a gene is represented within reads uniquely alignable to the genome. The gene intensity is the median number of times a single base is represented of all bases represented at least once within the gene.

Figure 4.4 shows the plot of gene intensities derived from the two different technologies. Gene intensities from the tiling data were binned at 0.1 increments of gene intensity (log_2 scale) and the mean gene intensity calculated. This was then plotted against the mean of gene intensities for the same genes in the sequence data. The plot indicates that there is good agreement (R = 0.82). Consequently we consider the intensities derived from our tiling data to be representative and usable. This correlation is greater than that previously reported for analogous comparisons made in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* (Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008) (R = 0.68 and R = 0.48). The manner in which tiling and sequence expression scores were calculated between these studies and that presented here are different. Critically both of these studies compensate for the inevitable 5' drop-off observed in the sequence data caused by oligo(dT) priming, by calculating the sequence expression scores based on n (30 and 300) 3' coding nucleotides. This is feasible given sequence data of sufficient depth such that the 3' end for genes for which there are reads are always represented. Our data are not of this depth and consequently expression scores are the median count of detected nucleotides.

Despite the clear correlation between gene intensities generated by the two technologies as exhibited in figure 4.4, there are clear discrepancies, especially in the top bins. There are a number of potential causes of this. Firstly, only polyadenylated transcripts are considered by the sequencing technologies whereas total RNA is hybridized to the microarrays. Secondly, the technical difference between the two technologies, such as

amplification of the (ds)cDNA for sequencing are likely to lead to discrepancies. The former difference may be the more likely cause as the discrepancies are most marked for the most abundant transcripts. The correlation observed, however, is most striking leading us to believe that the derived gene intensities are representative and usable.

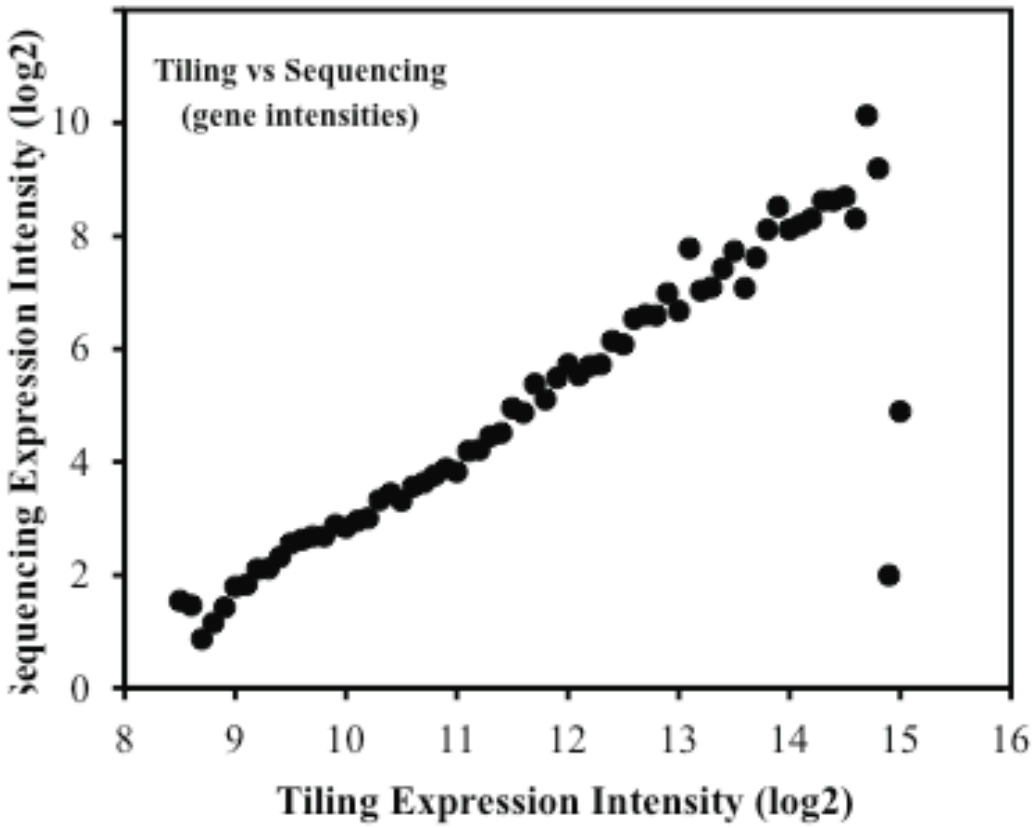


Figure 4.4. Correlation of gene intensities derived from tiling array and sequence data. Gene intensities from the tiling data were binned at 0.1 increments of gene intensity ($\log_2$ scale) and the mean gene intensity calculated. This was then plotted against the mean of gene intensities for the same genes in the sequence data. R = 0.82. This demonstrates good agreement between gene intensities derived from both technologies, thus validating our approach.

4.8. Validation of tiling data by sequence data

One method of validating the novel transfrags identified from the tiling array data would be exhaustive RT-PCR. This, however, would be time consuming and complicated by the fact that validation of small structures requires prior knowledge of their connectivity to surrounding structures. A more favourable alternative therefore is comparison of tiling data with ultra-high density sequence data. Not only does this allow the validation of novel transfrags, but should also allow them to be connected to other structures by identifying sequence reads which overlap transfrags. The number of transfrags identified by tiling arrays and validated by sequencing for stages at which we have stage-specific sequence data is shown in table 4.3. The ability of the sequence data to validate the tiling data is inevitably dependent on the depth of sequencing. It is clear then that the greater the intensity of the transfrag the more likely it is to be validated by the sequence data. The stringent background threshold set prior to the identification of transfrags, however, leads us to believe that were the sequence data of greater depth the rate of transfrag validation would have been consistently high across a greater range of transfrag intensities. We therefore consider our tiling data to be adequately validated and of a very high quality. That said, the marked difference in the fraction of genic and non-genic transfrags validated suggests that there may be a high rate of false discovery of novel transfrags.

The precise overlap between genes detected by the two technologies at all stages is illustrated in figure 4.5. Importantly, this is for genes called as expressed by the 50% criteria, rather than genes that have overlapping transfrags. It is these genes that will be

considered from this point on. The discrepancies between the two technologies are inevitably due to the differences in depth as well as stringency of the two technologies. The tiling data represents signal for more individual transcripts and is therefore of a greater depth than the sequence data. The presence of only one uniquely mappable read corresponding to a gene in the sequence data, however, is enough for that gene to be considered expressed whereas a transcript detected at a low level on the tiling array is more likely to be discarded as noise. Further to this, total RNA was hybridized to the tiling arrays whereas polyA+ RNA was used for sequencing in order to eliminate reads derived from rRNA. It is therefore inevitable that there will be differences in coverage by the two technologies.

Stage	Total transfrags	Genic	Exonic	Extra-genic	Total validated by seq	Percent
L4	45770	43804	42050	1966	42502	92.86
Young adult	46126	44139	42644	1987	42074	91.22
Stage	**Genic validated by seq**	**Percent**	**Exonic validated by seq**	**Percent**	**Extra-genic validated by seq**	**Percent**
L4	41521	94.79	40529	96.38	981	49.90
Young adult	40974	92.83	40152	94.16	1100	55.36

Table 4.3. Tiling array transfrags confirmed by sequencing. This table represents the proportions of genic, exonic and extra-genic transfrags validated by sequencing for the stages at which we have stage-specific sequence information. We note that more genic than extra-genic (novel) transfrags are validated by the sequence data. This may be due both to noise in our data and novel transcripts being expressed beneath the level of detection by the sequence data. ~91-93% of all transfrags are validated at stages for which we have stage-specific sequence data. We therefore consider our tiling data to be of high quality.

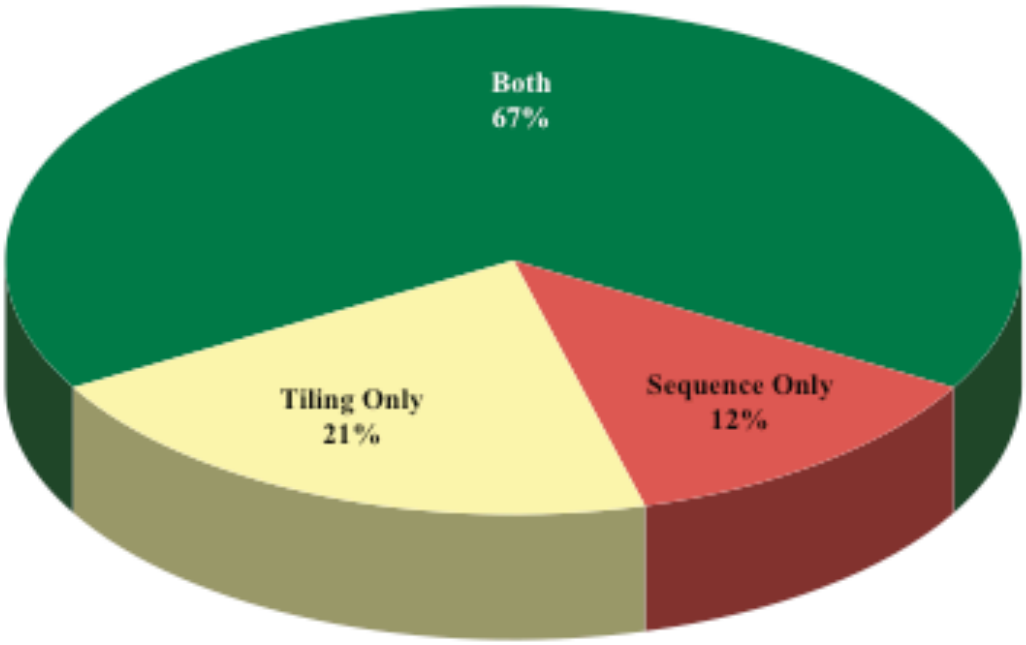


Figure 4.5. Overlap between genes detected by tiling arrays and by sequencing. Genes are defined as expressed in the tiling array data if ≥50% of probes per exon are above background and ≥50% of unique exons match that criterion. For a gene to be detected in the sequence data at least one uniquely mappable read must map to the gene.

4.9. Addressing alternative splicing using tiling data

High density tiling data theoretically allows the comparison of each exon of a gene in terms of expression level and in so doing, the identification of changes in major spliceform between conditions. Tiling arrays, however, can only provide data that allow the user to comment on differential inclusion of a given exon within the repertoire of splice forms of a gene. It gives no information in terms of connectivity of an exon to the other exons within a gene. Here we use our tiling data to generate a "splice index" (SI) for the change in expression of an exon relative to the expressed gene between conditions. More specifically, 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. The SI is used to infer a major change in splice form. It is essentially a

measure of how the intensity of a given exon changes relative to the whole gene between developmental stages. This then allows us to compare the genes for which a change in spliceform is detected to those for which different splice forms are known.

It is essential that intensity values can be assigned to exons with high confidence. For this reason exons with fewer than three probes were omitted from the analysis. At least one exon changes at least 2-fold in 5% of detected genes, which is to say that 5% of detected genes clearly exhibit a change in major isoform across development. While 18% of annotated genes have at least two annotated isoforms, this is the first systematic analysis of how these isoforms change across development. Of the 870 genes that show a change in spliceform between any two stages (>2-fold change for any given exon), 459 have multiple annotated isoforms in WS150. The remaining 47% of the genes we detect by this method are therefore not predicted to be alternatively spliced in WS150. These 411 genes, however, correspond to only 2% of annotated coding genes. This therefore does not conclusively demonstrate that alternative splicing is grossly underrepresented in current gene annotations. Since we detect less than 5% of genes to be alternatively spliced at this fold-change in SI, however, it may be that a relaxation of this threshold would show that the trend continues as the gene list expands. This would inevitably lead to an increase in false discovery, however, and it seems that our sequence data may offer a better alternative in addressing alternative splicing.

4.10. Addressing alternative splicing using sequence data

We have demonstrated that tiling array data can be used to indicate changes in major splice form for expressed genes between conditions. Connectivity of exons, however, cannot be inferred from tiling data. High-density sequence data can be used to this end by looking for reads that span exon-exon boundaries. This is the single biggest advantage of sequence data over tiling data – the information it provides on connectivity within expressed structures. The proportion of reads that span any set of exon boundaries relative to another may give an indication of the relative combinations of exons used in a given condition. This would be extremely useful in that it gives information on exon connectivity within transcripts. The methodology involves identifying sequence reads that do not map to the genome. These reads are then aligned with all combinations of adjacent and non-adjacent exons for all annotated isoforms of all genes using Maq. The output of this is reads that map to annotated exon-exon junctions and reads which span previously unidentified exon-exon junctions for annotated exons. The technique is therefore limited by the accuracy and completeness of exon boundary annotations. A schematic of the approach is shown in figure 4.6 and an example of the output in figure 4.7. A summary of the number of reads mapping to the genome and spanning annotated and non-annotated exon-exon boundaries across all samples is shown in table 4.4.

Ultimately sequence data at the required depth may render tiling data completely redundant in addressing alternative splicing. A constant issue which we face in our tiling analysis is what fold-changes are reasonable cut-offs, allowing us to call events. This is not an issue with sequence data for this analysis. If we identify a uniquely mappable read

spanning an exon-exon junction it is reasonable to assume that those exons are connected, allowing us to determine changes in spliceform. This also allows us to identify exons that are connected where no such connectivity exists in current gene annotations. Though the depth of our sequence data is inadequate to comprehensively map all splice events and spliceform changes at this stage, our current data show unannotated splice events for ~1% of detected genes. Critically, ~80% of genes identified to have alternative splicing in this manner also were found to have at least one exon with a splice index ≥1.5 by tiling analysis, confirming that the changes in transcript structure that we monitor by tiling analysis are likely to be real. Thus the high resolution mapping of the transcriptome using tiling arrays and the gene expression levels and transcript structural features that we derive from these data appear to be accurate.

	NO. OF NUCLEOTIDES	PERCENTAGE OF TOTAL
Nucleotides aligned to genome at ≥Q30	621610325	73.02
Nucleotides aligned to annotated transcriptome at ≥Q30	36085206	4.24
Nucleotides aligned to non-adjacent exons at ≥Q30	47205	0.01
Non-aligned nucleotides	193584559	22.74
Total Nucleotides	851327295	100.00

Table 4.4. Reads mapping to the genome and spanning exon-exon boundaries. Shown are the number of nucleotides across all samples mapped to the genome, and transcriptome as described above using Maq version 0.6.6 at a mapping quality of ≥Q30.

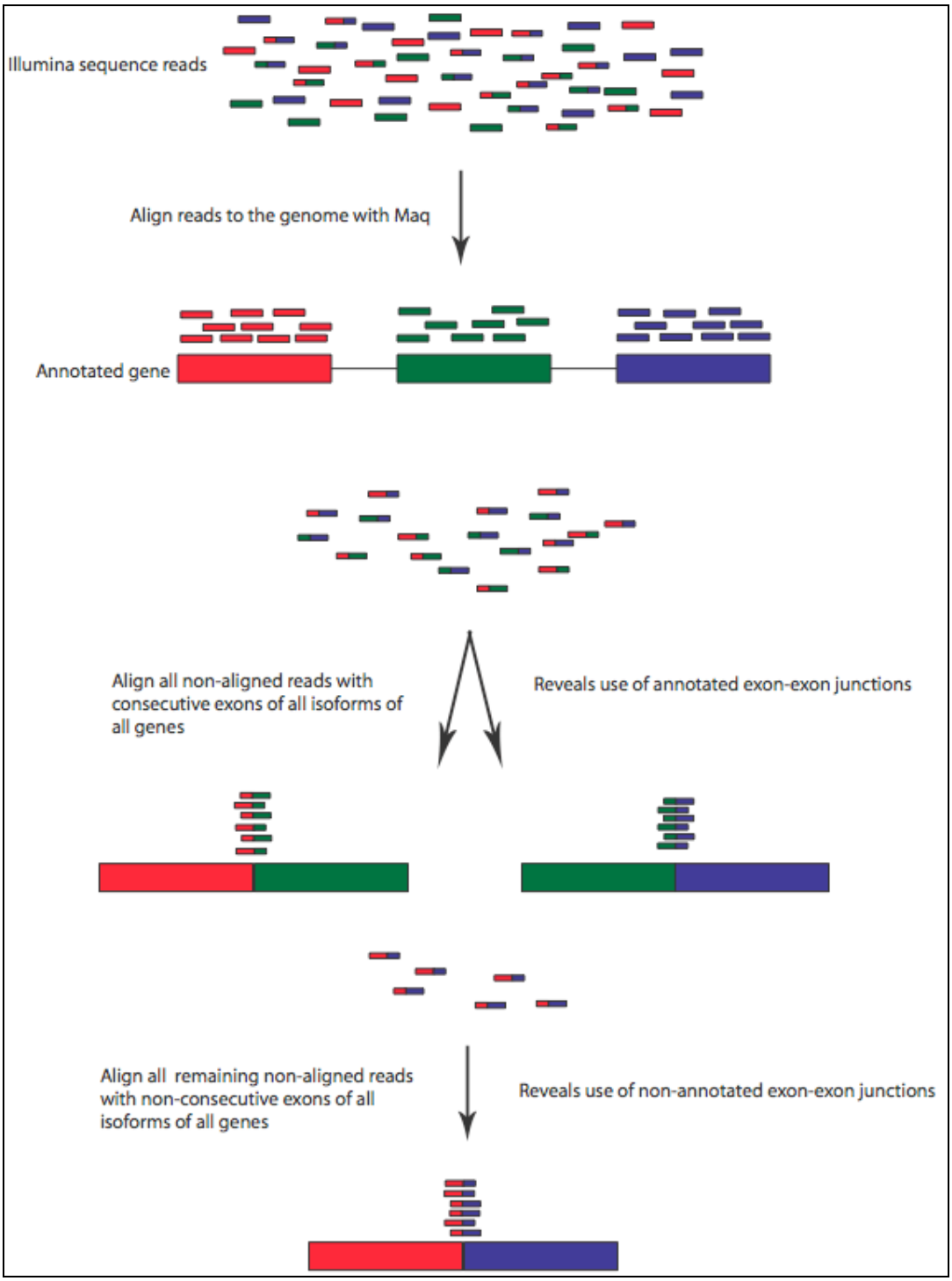


Figure 4.6. Use of Illumina sequence reads to identify utilized exon-exon junctions. Alignment of uniquely mappable reads to the genome using Maq removes reads not spanning exon boundaries and can be used to generate gene intensities. The remaining reads are aligned with consecutive exons for all isoforms of all genes. This reveals annotated exon-exon junctions. The remaining reads are then aligned to all combinations of non-consecutive exons for all isoforms of all genes. This reveals non-annotated exon-exon junctions.

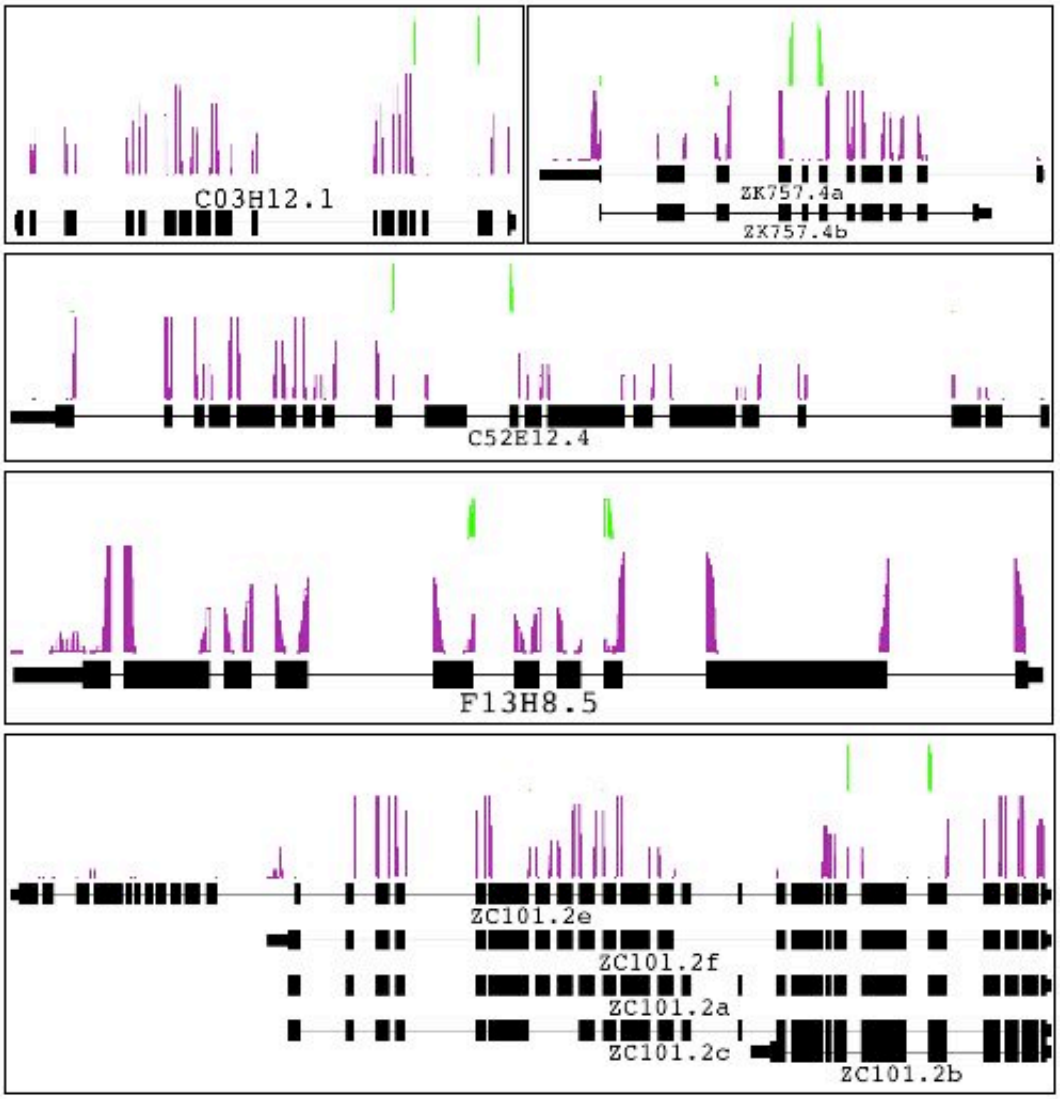


Figure 4.7. Ultra-high density sequence reads reveal novel splice junctions. Illumina sequence reads which cannot be aligned to the genome are aligned to adjacent annotated exons and all combinations of non-adjacent exons for all isoforms of all genes with Maq. Reads spanning annotated exon boundaries are shown in purple. Novel exon boundaries are shown in green. Relative numbers of reads spanning each exon-exon junction may reveal relative usage. At the current depth of sequencing 1% of genes appear to undergo at least one novel splice event.

4.11. Discussion

The work presented here clearly demonstrates the utility and complementarity of these technologies in forwarding our knowledge and understanding of gene annotations. It represents the first splicing analysis of its kind in *C. elegans* and the potential to become the most comprehensive analysis of its kind in any organism. The utility of the approaches developed in this work have been clearly demonstrated, as has the redundancy of tiling array data given the resolution and connectivity information of ultra-high density sequence data. Tiling array data, however, represents a more cost-effective approach in addressing the same questions. Sequence data to a greater depth will be required in order to more completely identify the complete repertoire of exon-exon junctions. At the time of printing this thesis sequence data had been produced at 15x the depth of the data utilized here. The tools are now in place to utilize these data to great effect. The splicing analysis performed using the tiling data does imply that alternative splicing is far more prevalent than can be accounted for by current annotations. It would be most interesting to see if this is further borne out by the newly acquired sequence data.

The dearth of novel transfrags detected in the tiling data and the low frequency of their validation by the sequence data suggest that the genome of *C. elegans* is well annotated. We do, however, provide clear evidence for novel transcription. Furthermore we do not discount the possibility that many of the novel transfrags which were not validated are expressed at a low level are beneath the level of detection of our sequence data, the scarcity of these transcripts being in part causative of their prior anonymity. Also, certain developmental stages are inevitably under-represented in the sequence data, leading to a

reduced possibility of validating novel transfrags detected at these stages. It will be interesting to see how many more transfrags are validated by the newly acquired sequence data.

The identity of these novel transfrags as additional exons of annotated genes, entire novel coding genes, or non-coding transcripts is yet to be tackled. This can be addressed using the sequence data but represents a more complex problem than the study of connectivity between annotated exons. Our splicing analysis thus far had involved looking for reads that span annotated exon boundaries. No such boundaries are defined by the transfrags or novel sequence reads mapped to the genome. A shotgun approach to assembling sequence reads into transcripts may represent the best possibility of connecting transcribed units. Whatever the approach taken and whomever implements it, it is likely to be extremely complex and computationally intensive.

Regarding the scarcity of novel transfrags, validated or otherwise, relative to analogous studies in other organisms – perhaps this is not surprising. The density of gene annotation in *C. elegans* surpasses that of human and *Drosophila*. Furthermore our study considered whole animals, i.e. all cells and tissues at once. If there are low levels of tissue-specific expression of novel genes we were unlikely to detect them in this study. Regardless, the output of this study has proven it a worthwhile undertaking and an ideal dataset for comparison with the nonsense-mediated mRNA decay transcriptome as we shall see. In terms of data quality, it is noted that markedly fewer genes are detected for any condition than were seen using two-colour microarrays in chapter 3. Importantly, the

expression microarrays used in chapter 3 have only one probe per gene, and are 70mers rather than the 25mer probes on Affymetrix arrays. The increased specificity per probed, coupled with the greater probe number per gene give us greater confidence in the output of the Affymetrix microarrays, even if the depth of detection is lesser. Our confidence in these data is further strengthened by the correlation of gene intensities between our tiling array and Illumina sequence data. We therefore consider the datasets presented in this chapter and the methodologies applied to them to be an ideal framework for comparison with the nonsense-mediated mRNA decay deficient transcriptome.

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

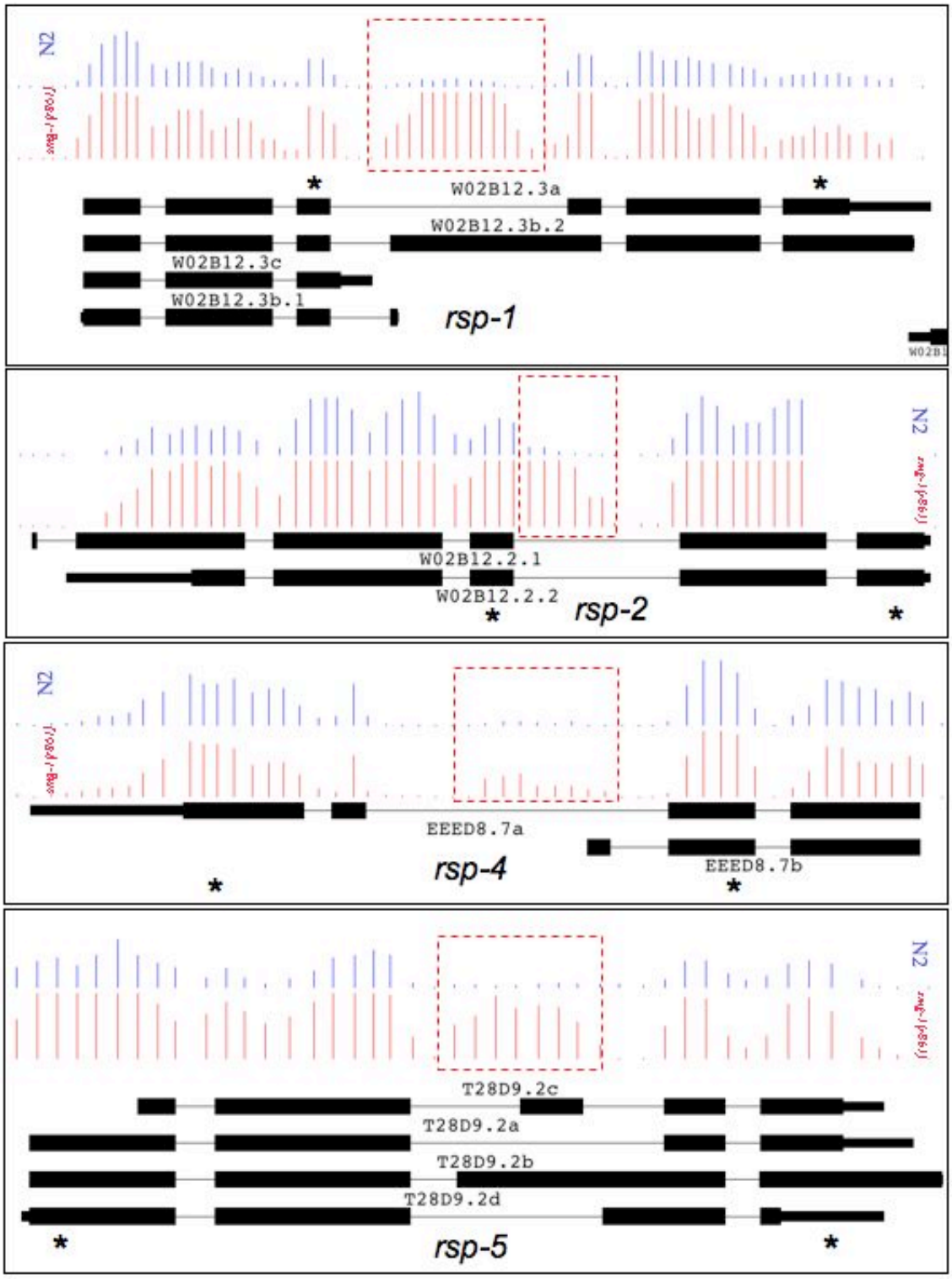
NMD targets produced by alternative splicing, the repertoire of splice sites used and junctions present is inevitably beyond what is annotated. Consequently it is far more complex and computationally intensive to look for unique reads that span two regions of a gene, undefined other than that they are 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

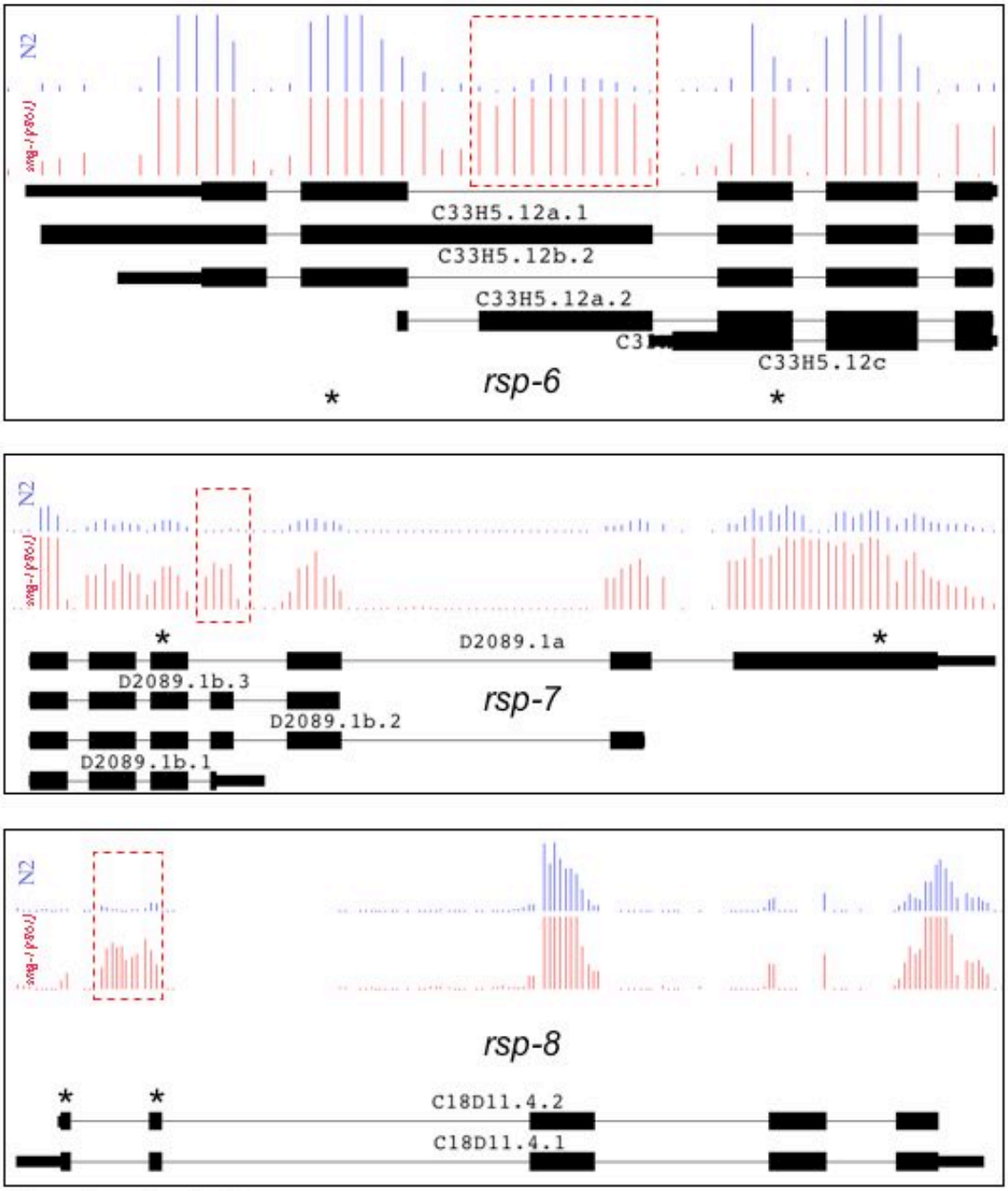


Figure legend overleaf.

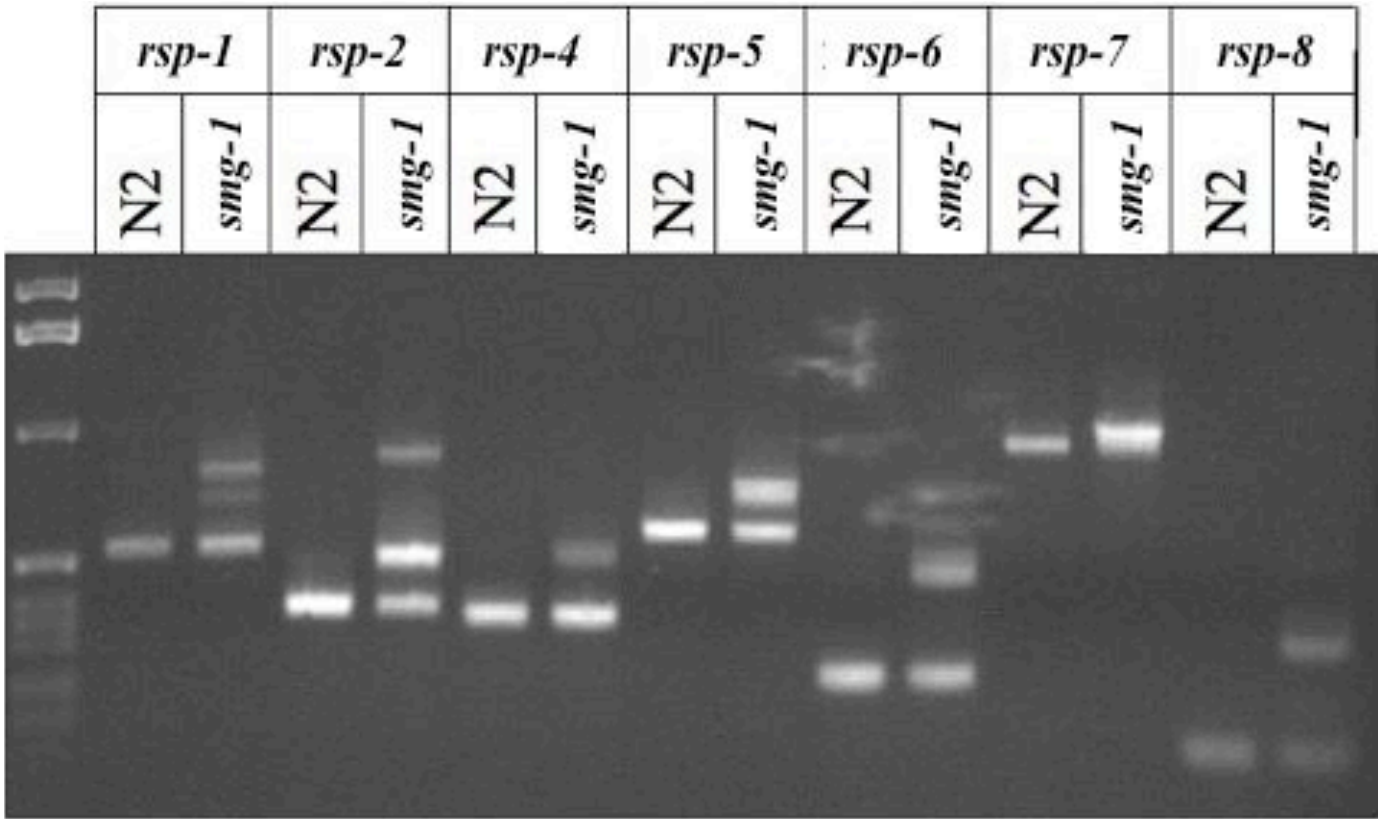


Figure 5.1. Structural changes in SR gene transcripts leading to NMD. Pages 1-2 of this figure show the normalized probe signal for each SR gene shown to be NMD regulated by our tiling array data (in order *rsp-1* to *rsp-8*). Gene annotations are in black. Normalized probe intensities derived from N2 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.

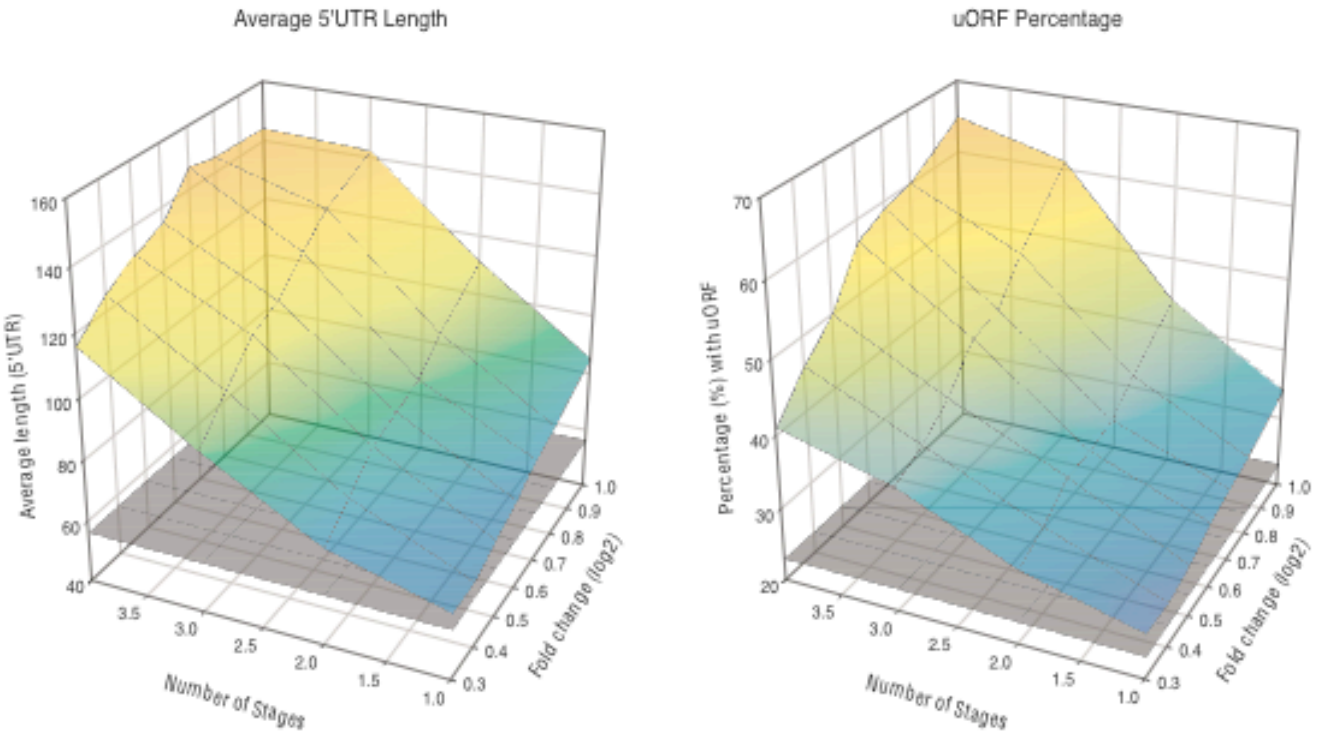


Figure 5.2. Increasing 5' UTR length correlates with increased magnitude of NMD. Each graph demonstrates how the characteristic labelled above increases with increasing fold-change of gene intensity between N2 and *smg-1(r861)* and increasing number of stages at which that fold-change occurs. The average 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.

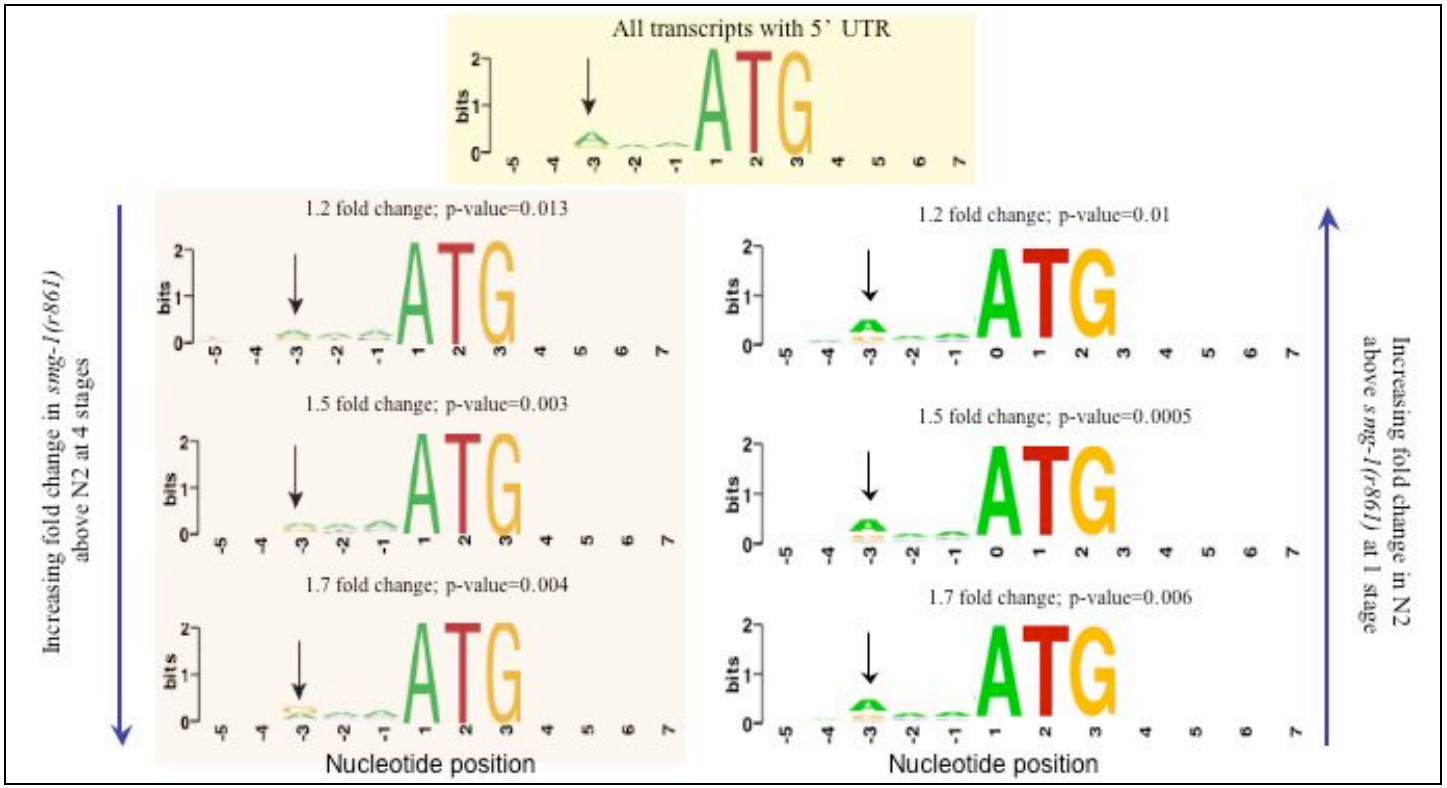


Figure 5.3. An A nucleotide -3 of the annotated start codon correlates with NMD regulation. Surveying the consensus sequence around all annotated start codons in transcripts reveals an enrichment for an A nucleotide -3 of the annotated start codon. This enrichment diminishes with increased NMD regulation 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.

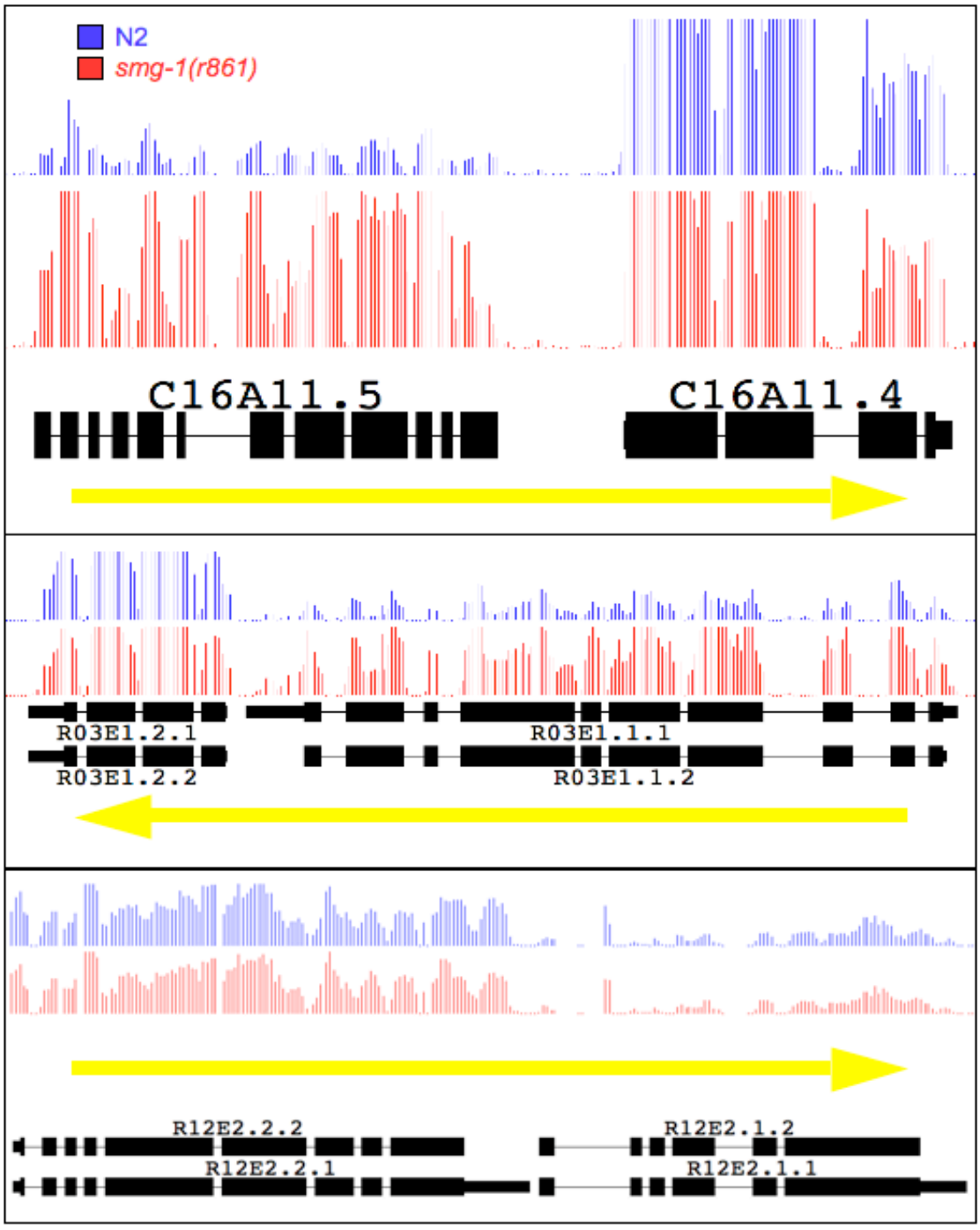


Figure 5.4. Examples of operonic gene regulation by NMD. Each segment shows tiling array data relating to an operon of two genes and the direction of transcription. The top and middle operons show clear equalisation of transcript levels within the operon on NMD perturbation. This is not the case in the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

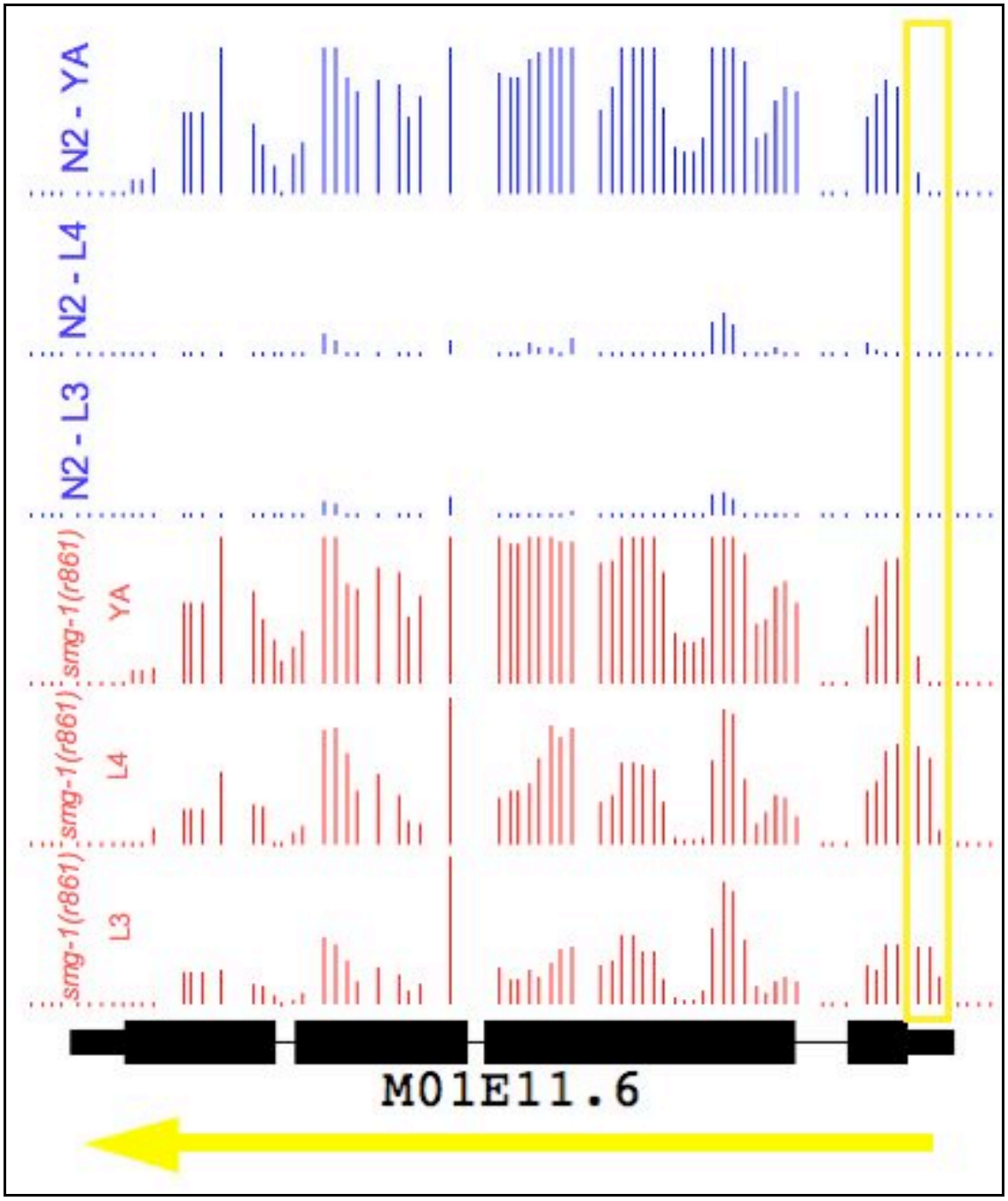


Figure 5.5. NMD regulation via a shift in promoter usage. *klp-15* (M01E11.6) is transcribed at all developmental stages considered, but degraded at L2-L4. The 5' UTR of *klp-15* contains an AUG with an A nucleotide at the -3 position. The annotated start codon does not have an A nucleotide at the -3 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.

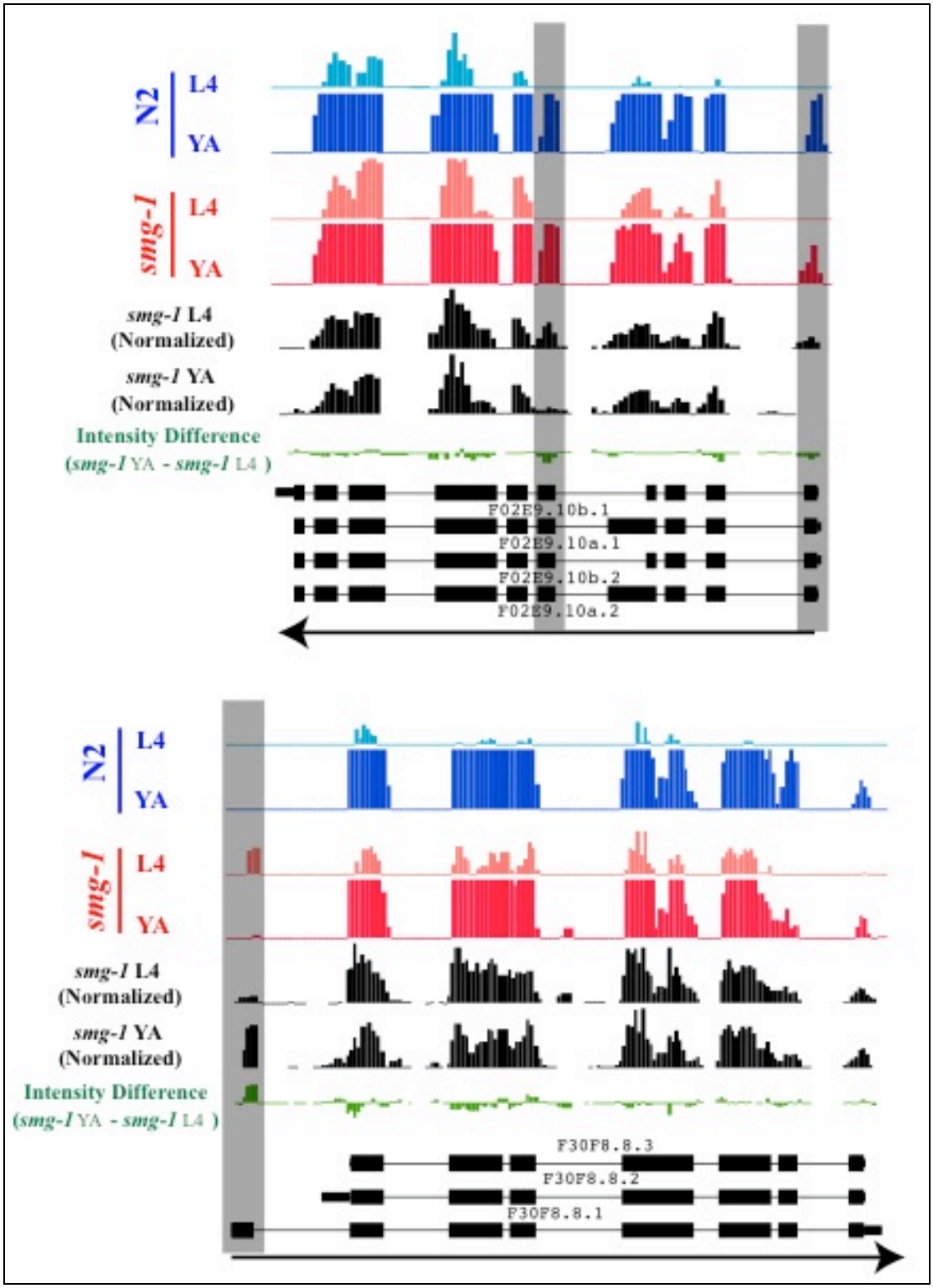


Figure legend overleaf.

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.

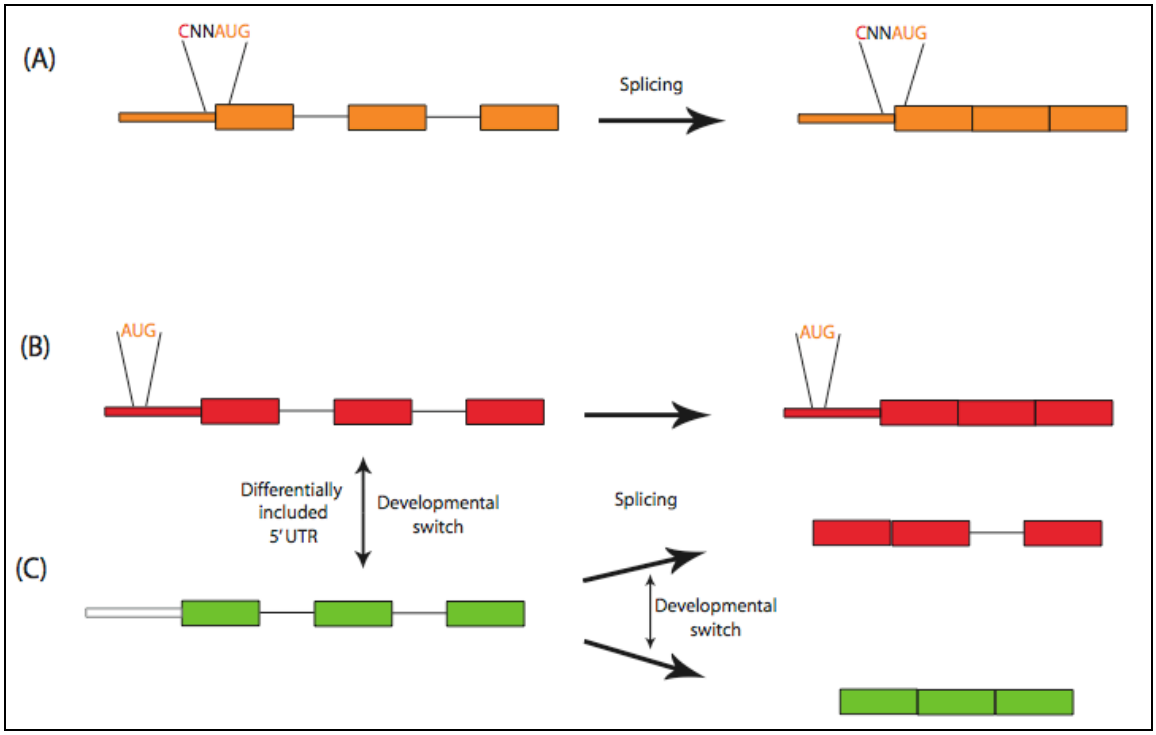


Figure 5.7. Model of gene regulation by NMD. Transcripts may have structural characteristics that predispose them to NMD, introduced either at the level of transcription or splicing. The presence of a weak translation initiation motif leads to the translation machinery occasionally skipping the correct 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.

Chapter 6
General Discussion and Future Work

The work detailed in this thesis represents clear progress in the fields to which it belongs. The approaches applied are either novel or are significant improvements on previous studies and have already yielded valuable results. The data and approaches taken also strongly indicate that future pursuit of the ultimate aims of these projects is worthwhile and are very likely to prove fruitful. I will now discuss the projects detailed in the previous chapters individually, focusing on the outcomes thus far and the future potential of the projects.

Expression profiles have been used with tremendous success in yeast as a means of describing the phenotype of different mutant strains. Comparing the expression profiles of two different mutants provides a high-resolution way to ask whether the two genes are likely to act in the same pathway – genes that act in the same pathway or complex have very similar profiles. This is a powerful approach to identify how novel genes act – if they share profiles with well-characterised genes, one can infer that they act in similar processes. The key question that I sought to address is whether this kind of approach can be used in a far more complex system than yeast, in a whole animal. To investigate this, I used standard two-colour array technology to generate expression profiles for a number of worm populations, either carrying loss-of-function mutations in genes known to play key roles in different signalling pathways affecting germline development, or having had these genes targeted through RNAi. I then used standard clustering algorithms to compare these expression profiles and used this to ask several questions: do two genes in the same pathway tend to cluster together? Do genes in different pathways cluster in different branches? Does the expression profile of one perturbation of a gene cluster very near another perturbed expression profile of the same gene? Finally, since all the profiles were of perturbations of genes affecting

brood size, I asked whether the clustering was directly correlated with strength of phenotype i.e. do genes of similar brood size cluster together?

I found first that genes in similar pathways do tend to cluster together far more strongly than associations between genes of different pathways. The implication of this is clear: if a novel gene is known to affect germline function then we can discover how it acts by comparing its profile with that of all those examined; if it clusters with several genes of a known pathway, it is highly likely to act in that pathway. Second, I found that in general the RNAi phenotype of a gene, as monitored by its perturbed expression profile, looks very similar to its genetic loss-of-function mutant – this is reassuring. Finally, I found that genes cluster independent of RNAi phenotype strength – clustering is thus driven by the underlying pathway affected and not by the extent to which it is affected. I thus concluded that expression profiles are indeed effective tools for identifying the mechanism of action of a novel gene.

To test whether we can really use the expression profile compendium to confirm how a novel gene acts, I turned to a dataset generated by Catriona Crombie, a postdoc in the Fraser lab. She had isolated a number of mutants that are candidate modulators of the EGF/ras/MAPK signalling pathway in the *C. elegans* vulva. Since EGF/ras/MAPK signalling is also required for germline development, I reasoned that I could confirm the role of these novel modulators by testing whether their expression profiles clustered with those of known EGF/ras/MAPK pathway genes. I selected one of these genes, *pkc-1*, to test this and find that it does indeed cluster tightly with other genes in this pathway, thus confirming both the approach and the pathway in which *pkc-1* appears to act. This result evidently requires further follow-up, for example by

detailed staining and microscopy, but it is very encouraging for such an approach. There is a wealth of genes identified as potential modulators of EGF/ras/MAPK and Notch pathway signalling revealed by screens in our lab and others. These are now candidates for testing against our compendium of expression profiles to provide further evidence of their roles in these pathways.

This approach appears to have much potential in adding evidence for the roles of genes in germline development. Importantly, in justifying the nature of the approach, it appears to be sensitive to even small weak gene perturbations resulting in only slight brood-size defects and considers populations rather than individuals. This may suggest that other approaches may be less sensitive to such changes in phenotype. But is there much value in increasing our knowledge of *C. elegans* germline development? Obviously the primary motivation behind any such biological study should be the downstream development of our understanding of human biology. That the signalling pathways being considered in the study are conserved from worms to humans and that the identified involvement of PKC signalling with EGF/ras/MAPK signalling had already been demonstrated in mammals suggests that there is relevant potential in this methodology. Identification of modulators of these pathways in *C. elegans* may be to identify candidate genes in human disease where these pathways are dysregulated, such as in cancer. The next step is clearly to increase the size of the compendium with other known regulators of germline development and gametogenesis as well as novel genes giving brood-size defects and novel genes, which are candidate regulators of the pathways of interest.

Our interrogation of the *C. elegans* transcriptome appears to have been similarly fruitful. We used tiled microarrays and ultra-high density sequencing to assess genome-wide transcription at multiple stages during worm development. We initially set out to address two key issues – whether there is a substantial amount of transcription beyond current annotations, and how complete current splicing annotations are. Widespread novel transcription has recently been shown to exist in a number of other organisms. Using whole genome tiled microarrays we have demonstrated that throughout development only ~5% of expressed regions of the genome lie outside annotated structures. This is reassuring in that it increases confidence in current gene annotations. It does, however, demonstrate that there are regions of novel transcription, which require further characterization. Ultra-high density sequencing technologies appear to be an ideal tool to do this, offering greater resolution than tiling array data and providing connectivity data in the form of reads that span exon-exon boundaries. We have used these data to identify reads spanning exon-exon boundaries of exons annotated as connected as well as novel exon-exon boundaries. Thus far our data indicate that novel splice events occur for ~1% of annotated genes. Critically, however, this approach is limited by the depth of coverage of the transcriptome provided by our sequence data. We have recently acquired sequence data to a greater depth, which will allow more thorough identification of novel splice events. Whilst ultra-high density sequence data offers a better option in studying splicing than tiled microarray data, our microarray data have nevertheless given us an interesting insight into the extent of unannotated splicing. Using our tiling array data to look at changes in relative exon intensities throughout development we have identified genes that exhibit major changes in exon use, indicating alternative spliceforms. Of the genes exhibiting novel splicing events in

our sequence data ~80% were also identified in our tiling analysis leading us to believe that the genes we discover using the tiling data are alternatively spliced. ~50% of the genes identified as alternatively spliced at high confidence using our tiling data have only one annotated isoform, suggesting that annotation of spliceforms is far less complete than that of transcription as a whole. It will be very interesting to see if this trend continues when our sequence data analysis is extended to our newly acquired data set. The approach discussed to study alternative splicing using sequence data could also be expanded to catalogue trans-splicing events by identifying reads that span independently transcribed structures. A further application of our sequence data may be to uncover the identity of novel transcribed regions in terms of their connectivity to already annotated genes, or each other as previously unannotated spliced or unspliced transcripts. This is a far more complex problem than studying connectivity of annotated exons. It will require an approach that does not rely on gene annotations. A shotgun approach to assemble sequence reads may offer a possible method of connecting novel structures and may also allow better annotation of exon boundaries in already annotated genes and novel splice sites within annotated introns and exons.

The quality of our data and the approaches applied represent a major step forward in transcriptome analysis towards the ultimate set of gene annotations. The value of this is difficult to overestimate. Identification of all genes may lead to the discovery of transcripts and proteins of novel function. Knowledge of all isoforms of all genes will allow a more complete study of protein structure and consequent biological properties and how these change between different conditions. It will also lead to improvement in approaches to quantify transcript levels by allowing more comprehensive

transcriptome coverage by expression microarrays. Any benefit to microarray design, however, may be short-lived. It seems that the key advantage of microarrays over ultra-high density sequencing is the cost-differential for the same depth of coverage of the transcriptome. Were funds unlimited it is difficult to identify many applications for which microarrays would be the preferred platform. Should ultra-high density sequencing become more affordable and of higher throughput then, the use of microarrays may become a thing of the past.

Our interrogation of the wild-type *C. elegans* transcriptome and the approaches applied to it provided the ideal framework for comparison with the NMD-deficient transcriptome. Our motivation in studying the NMD was to determine whether the identity of NMD targets, their structures and how those structures change could provide an insight into the role and mechanism of NMD. *C. elegans* appeared to be an ideal system in which to do this as it allowed us to study NMD in a dynamic biological environment i.e. throughout development. Whole genome tiling array data was produced for comparison with our wild-type dataset. Comparison of the resulting gene intensities between wild-type and NMD-deficient animals revealed genes regulated by NMD. Analysis of the properties of these transcripts confirmed features that have previously been reported as being NMD causative, such as identification of alternative spliceforms and transcripts containing uORFs. Interestingly the strength of the annotated translation initiation sequence appears to be critical to the predisposition of transcripts to NMD. NMD may therefore act to regulate steady-state transcription in accordance with the strength of the translation initiation sequence. Whilst translation initiation events occurring after the annotated translation initiation site leading to an in-frame premature termination codon have been recognized to lead

to NMD, this direct relationship of the strength of the translation initiation sequence at the annotated start site and NMD was previously unrecognized and is likely to be of great importance. It is known that many disease-associated mutations and variants result in mRNAs harbouring PTCs. The clinical outcome of harbouring such alleles is NMD dependent (Khajavi *et al*., 2006). Sequence variation at the translation initiation site may occur leading to effective under- or over-expression of a transcript due to a shift in its susceptibility to NMD. It may therefore have a significant link to human disease.

Amongst the repertoire of NMD targets are operonic genes. This is most interesting as it is known that whilst genes in operons are transcribed at equal levels, the measured abundance of transcripts for genes in the same operon are often different. Whilst not a complete explanation for this inequity of effective expression, NMD does appear to one mechanism by which this occurs.

Perhaps the most interesting finding of the work detailed in this thesis is the requirement of NMD for ~10% of developmentally regulated gene expression changes via regulated changes in transcript structure. Such structural changes may be a switch in spliceform or a shift in the position of transcription initiation to include or omit a uAUG. This demonstrates that the timing of gene expression is not dictated by the rate of transcription alone, rather in some cases the position of transcription initiation and also splicing events may act via NMD to dictate the effective level of gene expression in a temporally controlled manner. This represents a hitherto unrecognized mechanism of gene expression regulation and will inevitably alter perception of how such regulation is achieved. Whilst it seems likely that this method

of gene expression regulation occurs in wild-type animals, we cannot discount the possibility that the changes in transcript structure that lead to NMD targeting are through the action of splicing and transcription factors which are also NMD regulated. If this is so then much of the signal may be artifactual. This does appear extremely unlikely but we cannot discount the possibility. Testing this through identifying targets of NMD regulated splicing and transcription factors would be a huge undertaking and assumes that we have already comprehensively identified all such factors, which is unlikely. As previously stated then, possibly the best method of validating this finding and adding value to it would be repeating the study in another biological system where NMD is not essential such as yeast or fly. The same possibility of NMD regulation of transcripts due to regulation of upstream factors would still stand however.

Taken together our findings regarding NMD will have a significant impact on current perception of NMD and have real potential to influence perception of human disease biology and gene regulation. Though NMD is not essential in yeast, fly and worm it appears to required for mouse embryonic development (Medghalchi *et al*., 2001). NMD being required for correct developmental gene expression rather than just acting as a surveillance mechanism may serve as a partial or complete explanation of this. The possibility of variation at the translation initiation site leading to variation in effective gene expression via NMD and consequent modulation of a disease phenotype seems very real and worthy of investigation.

The output of all of these individual studies is indicative of the continuing value of *C. elegans* as a tool for large-scale biological studies. Whilst the requirement of

performing expression analyses at the level of the whole animal may often be cited as a disadvantage, it was the ease of studying expression throughout development and in the germline at the level of a whole animal that led to the manner of all of these studies. The utility and ease of RNAi in *C. elegans* and the wealth of genetic mutants were also key advantages that were essential to these studies. That the use of *C. elegans* continues to be just as valid as technology and biological research moves on demonstrated that *C. elegans* remains at the cusp of cutting-edge research and is likely to for the foreseeable future.

References

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D.A., Ghosh, S. and Jacobson, A. (2004) A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature*, **432**, 112-118.

Anders, K.R., Grimson, A. and Anderson, P. (2003) SMG-5, required for C.elegans nonsense-mediated mRNA decay, associates with SMG-2 and protein phosphatase 2A. *Embo J*, **22**, 641-650.

Austin, J. and Kimble, J. (1987) glp-1 is required in the germ line for regulation of the decision between mitosis and meiosis in C. elegans. *Cell*, **51**, 589-599.

Austin, J. and Kimble, J. (1989) Transcript analysis of glp-1 and lin-12, homologous genes required for cell interactions during development of C. elegans. *Cell*, **58**, 565-571.

Azzalin, C. M., Reichenbach, P., Khoriauli, L., Giulotto, E. and Lingner, J. (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science 318, 798–801.

Banan, A., Fields, J.Z., Zhang, Y. and Keshavarzian, A. (2001) Key role of PKC and Ca2+ in EGF protection of microtubules and intestinal barrier against oxidants. *Am J Physiol Gastrointest Liver Physiol*, **280**, G828-843.

Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185-198.

Behm-Ansmant, I., Gatfield, D., Rehwinkel, J., Hilgers, V. and Izaurralde, E. (2007a) A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *Embo J*, **26**, 1591-1601.

Behm-Ansmant, I., Kashima, I., Rehwinkel, J., Sauliere, J., Wittkopp, N. and Izaurralde, E. (2007b) mRNA quality control: an ancient machinery recognizes and degrades mRNAs with nonsense codons. *FEBS Lett*, **581**, 2845-2853.

Bender, A.M., Kirienko, N.V., Olson, S.K., Esko, J.D. and Fay, D.S. (2007) lin-35/Rb and the CoREST ortholog spr-1 coordinately regulate vulval morphogenesis and gonad development in C. elegans. *Dev Biol*, **302**, 448-462.

Berry, L.W., Westlund, B. and Schedl, T. (1997) Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors. *Development*, **124**, 925-936.

Bertani, G. (1951). Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. *J Bacteriol*, **62**, 293-300.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M.

(2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242-2246.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. and Kim, S.K. (2002) A global analysis of Caenorhabditis elegans operons. *Nature*, **417**, 851-854.

Boisvert, M.E. and Simard, M.J. (2008) RNAi pathway in C. elegans: the argonautes and collaborators. *Curr Top Microbiol Immunol*, **320**, 21-36.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

Booth, E.O., Van Driessche, N., Zhuchenko, O., Kuspa, A. and Shaulsky, G. (2005) Microarray phenotyping in Dictyostelium reveals a regulon of chemotaxis genes. *Bioinformatics*, **21**, 4371-4377.

Brenner, S. (1974) The genetics of Caenorhabditis elegans. *Genetics*, **77**, 71-94.

Buhler, M., Steiner, S., Mohn, F., Paillusson, A. and Muhlemann, O. (2006) EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length. *Nat Struct Mol Biol*, **13**, 462-464.

C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012-2018.

Cali, B.M., Kuchma, S.L., Latham, J. and Anderson, P. (1999) smg-7 is required for mRNA surveillance in Caenorhabditis elegans. *Genetics*, **151**, 605-616.

Calvi, L.M., Adams, G.B., Weibrecht, K.W., Weber, J.M., Olson, D.P., Knight, M.C., Martin, R.P., Schipani, E., Divieti, P., Bringhurst, F.R., Milner, L.A., Kronenberg, H.M. and Scadden, D.T. (2003) Osteoblastic cells regulate the haematopoietic stem cell niche. *Nature*, **425**, 841-846.

Chang, C., Hopper, N.A. and Sternberg, P.W. (2000) Caenorhabditis elegans SOS-1 is necessary for multiple RAS-mediated developmental signals. *Embo J*, **19**, 3283-3294.

Chang, J.C. and Kan, Y.W. (1979) beta 0 thalassemia, a nonsense mutation in man. *Proc Natl Acad Sci U S A*, **76**, 2886-2889.

Chang, Y.F., Imam, J.S. and Wilkinson, M.F. (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*, **76**, 51-74.

Chen, N. and Greenwald, I. (2004) The lateral signal for LIN-12/Notch in C. elegans vulval development comprises redundant secreted and transmembrane DSL proteins. *Dev Cell*, **6**, 183-192.

Chiu, S.Y., Serin, G., Ohara, O. and Maquat, L.E. (2003) Characterization of human Smg5/7a: a protein with similarities to Caenorhabditis elegans SMG5 and SMG7 that functions in the dephosphorylation of Upf1. *Rna*, **9**, 77-87.

Christensen, S., Kodoyianni, V., Bosenberg, M., Friedman, L. and Kimble, J. (1996) lag-1, a gene required for lin-12 and glp-1 signaling in Caenorhabditis elegans, is homologous to human CBF1 and Drosophila Su(H). *Development*, **122**, 1373-1383.

Church, D.L., Guan, K.L. and Lambie, E.J. (1995) Three genes of the MAP kinase cascade, mek-2, mpk-1/sur-1 and let-60 ras, are required for meiotic cell cycle progression in Caenorhabditis elegans. *Development*, **121**, 2525-2535.

Crittenden, S.L., Bernstein, D.S., Bachorik, J.L., Thompson, B.E., Gallegos, M., Petcherski, A.G., Moulder, G., Barstead, R., Wickens, M. and Kimble, J. (2002) A conserved RNA-binding protein controls germline stem cells in Caenorhabditis elegans. *Nature*, **417**, 660-663.

Crittenden, S.L. and Kimble, J. (2008) Analysis of the C. elegans germline stem cell region. *Methods Mol Biol*, **450**, 27-44.

Crittenden, S.L., Troemel, E.R., Evans, T.C. and Kimble, J. (1994) GLP-1 is localized to the mitotic region of the C. elegans germ line. *Development*, **120**, 2901-2911.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.

Crotty, T., Cai, J., Sakane, F., Taketomi, A., Prescott, S.M. and Topham, M.K. (2006) Diacylglycerol kinase delta regulates protein kinase C and epidermal growth factor receptor signaling. *Proc Natl Acad Sci U S A*, **103**, 15485-15490.

Cui, Y., Hagan, K.W., Zhang, S. and Peltz, S.W. (1995) Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. *Genes Dev*, **9**, 423-436.

Cui, M., Chen, J., Myers, T.R., Hwang, B.J., Sternberg, P.W., Greenwald, I., Han, M. (2006) SynMuv genes redulndantly inhibit *lin-3/EGF* expression to prevent inappropriate vulval induction in *C. elegans*. *Dev Cell*. **10**, 667-672.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W. and Steinmetz, L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A*, **103**, 5320-5325.

Doyle, T.G., Wen, C. and Greenwald, I. (2000) SEL-8, a nuclear protein required for LIN-12 and GLP-1 signaling in Caenorhabditis elegans. *Proc Natl Acad Sci U S A*, **97**, 7877-7881.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet*, **21**, 10-14.

Eberle, A.B., Stalder, L., Mathys, H., Orozco, R.Z. and Muhlemann, O. (2008) Posttranscriptional gene regulation by spatial rearrangement of the 3' untranslated region. *PLoS Biol*, **6**, e92.

Eckmann, C.R., Crittenden, S.L., Suh, N. and Kimble, J. (2004) GLD-3 and control of the mitosis/meiosis decision in the germline of Caenorhabditis elegans. *Genetics*, **168**, 147-160.

Eckmann, C.R., Kraemer, B., Wickens, M. and Kimble, J. (2002) GLD-3, a bicaudal-C homolog that inhibits FBF to control germline sex determination in C. elegans. *Dev Cell*, **3**, 697-710.

Eisenmann, D.M., Maloof, J.N., Simske, J.S., Kenyon, C. and Kim, S.K. (1998) The beta-catenin homolog BAR-1 and LET-60 Ras coordinately regulate the Hox gene lin-39 during Caenorhabditis elegans vulval development. *Development*, **125**, 3667-3680.

Ercan, S., Giresi, P.G., Whittle, C.M., Zhang, X., Green, R.D. and Lieb, J.D. (2007) X chromosome repression by localization of the C. elegans dosage compensation machinery to sites of transcription initiation. *Nat Genet*, **39**, 403-408.

Ferguson, E.L. and Horvitz, H.R. (1985) Identification and characterization of 22 genes that affect the vulval cell lineages of the nematode Caenorhabditis elegans. *Genetics*, **110**, 17-72.

Ferguson, E.L. and Horvitz, H.R. (1989) The multivulva phenotype of certain Caenorhabditis elegans mutants results from defects in two functionally redundant pathways. *Genetics*, **123**, 109-121.

Ferguson, E.L., Sternberg, P.W. and Horvitz, H.R. (1987) A genetic pathway for the specification of the vulval cell lineages of Caenorhabditis elegans. *Nature*, **326**, 259-267.

Fiegler, H., Carr, P., Douglas, E.J., Burford, D.C., Hunt, S., Scott, C.E., Smith, J., Vetrie, D., Gorman, P., Tomlinson, I.P. and Carter, N.P. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36**, 361-374.

Filipowicz, W. (2005) RNAi: the nuts and bolts of the RISC machine. *Cell*, **122**, 17-20.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806-811.

Fitzgerald, K. and Greenwald, I. (1995) Interchangeability of Caenorhabditis elegans DSL proteins and intrinsic signalling activity of their extracellular domains in vivo. *Development*, **121**, 4275-4282.

Francis, R., Maine, E. and Schedl, T. (1995) Analysis of the multiple roles of gld-1 in germline development: interactions with the sex determination cascade and the glp-1 signaling pathway. *Genetics*, **139**, 607-630.

Fraser, A.G., Kamath, R.S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M. and Ahringer, J. (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature*, **408**, 325-330.

Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S., Maziere, P., Grocock, R.J., Freilich, S., Thornton, J. and Enright, A.J. (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol*, **3**, 2032-2042.

Fribourg, S., Gatfield, D., Izaurralde, E. and Conti, E. (2003) A novel mode of RBD-protein recognition in the Y14-Mago complex. *Nat Struct Biol*, **10**, 433-439.

Gaiano, N. and Fishell, G. (2002) The role of notch in promoting glial and neural stem cell fates. *Annu Rev Neurosci*, **25**, 471-490.

Gardner, L.B. (2008) Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol*, **28**, 3729-3741.

Gatfield, D. and Izaurralde, E. (2004) Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila. *Nature*, **429**, 575-578.

Gatfield, D., Unterholzner, L., Ciccarelli, F.D., Bork, P. and Izaurralde, E. (2003) Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. *Embo J*, **22**, 3960-3970.

Gehen, S. C., Staversky, R. J., Bambara, R. A., Keng, P. C. & O'Reilly, M. A. (2008) hSMG-1 and ATM sequentially and independently regulate the G(1) checkpoint during oxidative stress. *Oncogene*, **27,** 4065-4074.

Gehring, N.H., Neu-Yilik, G., Schell, T., Hentze, M.W. and Kulozik, A.E. (2003) Y14 and hUpf3b form an NMD-activating complex. *Mol Cell*, **11**, 939-949.

Greenwald, I.S., Sternberg, P.W. and Horvitz, H.R. (1983) The lin-12 locus specifies cell fates in Caenorhabditis elegans. *Cell*, **34**, 435-444.

Grimson, A., O'Connor, S., Newman, C.L. and Anderson, P. (2004) SMG-1 is a phosphatidylinositol kinase-related protein kinase required for nonsense-mediated mRNA Decay in Caenorhabditis elegans. *Mol Cell Biol*, **24**, 7483-7490.

Guan, Q., Zheng, W., Tang, S., Liu, X., Zinkel, R.A., Tsui, K.W., Yandell, B.S. and Culbertson, M.R. (2006) Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast. *PLoS Genet*, **2**, e203.

Han, M., Aroian, R.V. and Sternberg, P.W. (1990) The let-60 locus controls the switch between vulval and nonvulval cell fates in Caenorhabditis elegans. *Genetics*, **126**, 899-913.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H. and Shiu, S.H. (2007) A large number of novel coding small open reading frames in the intergenic regions of

the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res*, **17**, 632-640.

Hannon, G.J. (2002) RNA interference. *Nature*, **418**, 244-251.

Hansen, D., Hubbard, E.J. and Schedl, T. (2004a) Multi-pathway control of the proliferation versus meiotic development decision in the Caenorhabditis elegans germline. *Dev Biol*, **268**, 342-357.

Hansen, D., Wilson-Berry, L., Dang, T. and Schedl, T. (2004b) Control of the proliferation versus meiotic development decision in the C. elegans germline through regulation of GLD-1 protein accumulation. *Development*, **131**, 93-104.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S.E. and Guigo, R. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol*, **7 Suppl 1**, S4 1-9.

He, F., Li, X., Spatrick, P., Casillo, R., Dong, S. and Jacobson, A. (2003) Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. *Mol Cell*, **12**, 1439-1452.

Hedgecock, E.M. and Herman, R.K. (1995) The ncl-1 gene and genetic mosaics of Caenorhabditis elegans. *Genetics*, **141**, 989-1006.

Henderson, S.T., Gao, D., Christensen, S. and Kimble, J. (1997) Functional domains of LAG-2, a putative signaling ligand for LIN-12 and GLP-1 receptors in Caenorhabditis elegans. *Mol Biol Cell*, **8**, 1751-1762.

Heo, J.S. and Han, H.J. (2006) PKC and MAPKs pathways mediate EGF-induced stimulation of 2-deoxyglucose uptake in mouse embryonic stem cells. *Cell Physiol Biochem*, **17**, 145-158.

Herman, R.K. and Hedgecock, E.M. (1990) Limitation of the size of the vulval primordium of Caenorhabditis elegans by lin-15 expression in surrounding hypodermis. *Nature*, **348**, 169-171.

Hodgkin, J., Papp, A., Pulak, R., Ambros, V. and Anderson, P. (1989) A new kind of informational suppression in the nematode Caenorhabditis elegans. *Genetics*, **123**, 301-313.

Horak, C.E. and Snyder, M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol*, **350**, 469-483.

Horvitz, H.R. and Sulston, J.E. (1980) Isolation and genetic characterization of cell-lineage mutants of the nematode Caenorhabditis elegans. *Genetics*, **96**, 435-454.

Hsu, V., Zobel, C.L., Lambie, E.J., Schedl, T. and Kornfeld, K. (2002) Caenorhabditis elegans lin-45 raf is essential for larval viability, fertility and the induction of vulval cell fates. *Genetics*, **160**, 481-492.

Huang, L.S., Tzou, P. and Sternberg, P.W. (1994) The lin-15 locus encodes two negative regulators of Caenorhabditis elegans vulval development. *Mol Biol Cell*, **5**, 395-411.

Hubbard, E.J., Dong, Q. and Greenwald, I. (1996) Evidence for physical and functional association between EMB-5 and LIN-12 in Caenorhabditis elegans. *Science*, **273**, 112-115.

Hubbard, E.J. and Greenstein, D. (2005) Introduction to the germ line. *WormBook*, 1-4.

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H. and Linsley, P.S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*, **19**, 342-347.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.

Ishida, S., Shigemoto-Mogami, Y., Kagechika, H., Shudo, K., Ozawa, S., Sawada, J., Ohno, Y. and Inoue, K. (2003) Clinically potential subclasses of retinoid synergists revealed by gene expression profiling. *Mol Cancer Ther*, **2**, 49-58.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V. and Kim, S.K. (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in Caenorhabditis elegans. *Proc Natl Acad Sci U S A*, **98**, 218-223.

Joshua-Tor, L. (2006) The Argonautes. *Cold Spring Harb Symp Quant Biol*, **71**, 67-72.

Kadyk, L.C. and Kimble, J. (1998) Genetic regulation of entry into meiosis in Caenorhabditis elegans. *Development*, **125**, 1803-1813.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P. and Ahringer, J. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, **421**, 231-237.

Kamath, R.S., Martinez-Campos, M., Zipperlen, P., Fraser, A.G. and Ahringer, J. (2001) Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in Caenorhabditis elegans. *Genome Biol*, **2**, RESEARCH0002.

Katz, W.S., Hill, R.J., Clandinin, T.R. and Sternberg, P.W. (1995) Different levels of the C. elegans growth factor LIN-3 promote distinct vulval precursor fates. *Cell*, **82**, 297-307.

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

Keene, J.D., Komisarow, J.M. and Friedersdorf, M.B. (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc*, **1**, 302-307.

Khajavi, M., Inoue, K. and Lupski, J.R. (2006) Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. *Eur J Hum Genet*, **14**, 1074-1081.

Kim, Y. K., Furic, L., Desgroseillers, L. and Maquat, L. E. (2005) Mammalian Staufen1 recruits Upf1 to specific mRNA 3UTRs so as to elicit mRNA decay. *Cell*, **120**, 195–208.

Kimble, J. (1981) Alterations in cell lineage following laser ablation of cells in the somatic gonad of Caenorhabditis elegans. *Dev Biol*, **87**, 286-300.

Kimble, J. and Hirsh, D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Dev Biol*, **70**, 396-417.

Kimble, J.E. and White, J.G. (1981) On the control of germ cell development in Caenorhabditis elegans. *Dev Biol*, **81**, 208-219.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D. and Dunham, I. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res*, **17**, 691-707.

Kozak, M. (1984) Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. *Nature*, **308**, 241-246.

Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283-292.

Kozak, M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol*, **196**, 947-950.

Kraemer, B., Crittenden, S., Gallegos, M., Moulder, G., Barstead, R., Kimble, J. and Wickens, M. (1999) NANOS-3 and FBF proteins physically interact to control the sperm-oocyte switch in Caenorhabditis elegans. *Curr Biol*, **9**, 1009-1018.

Lambie, E.J. and Kimble, J. (1991) Two homologous regulatory genes, lin-12 and glp-1, have overlapping functions. *Development*, **112**, 231-240.

Lamont, L.B., Crittenden, S.L., Bernstein, D., Wickens, M. and Kimble, J. (2004) FBF-1 and FBF-2 regulate the size of the mitotic region in the C. elegans germline. *Dev Cell*, **7**, 697-707.

Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C. and Brenner, S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926-929.

Lee, B.S. and Culbertson, M.R. (1995) Identification of an additional gene required for eukaryotic nonsense mRNA turnover. *Proc Natl Acad Sci U S A*, **92**, 10354-10358.

Lee, M.H. and Schedl, T. (2004) Translation repression by GLD-1 protects its mRNA targets from nonsense-mediated mRNA decay in C. elegans. *Genes Dev*, **18**, 1047-1059.

Lee, M.Y., Lee, S.H., Kim, Y.H., Heo, J.S., Park, S.H., Lee, J.H. and Han, H.J. (2006a) Effect of EGF on [3H]-thymidine incorporation and cell cycle regulatory proteins in primary cultured chicken hepatocytes: Involvement of Ca2+/PKC and MAPKs. *J Cell Biochem*, **99**, 1677-1687.

Lee, M.Y., Park, S.H., Lee, Y.J., Heo, J.S., Lee, J.H. and Han, H.J. (2006b) EGF-induced inhibition of glucose transport is mediated by PKC and MAPK signal pathways in primary cultured chicken hepatocytes. *Am J Physiol Gastrointest Liver Physiol*, **291**, G744-750.

Leeds, P., Peltz, S.W., Jacobson, A. and Culbertson, M.R. (1991) The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes Dev*, **5**, 2303-2314.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A.G. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, **38**, 896-903.

Lejeune, F., Li, X. and Maquat, L.E. (2003) Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell*, **12**, 675-687.

Lelivelt, M.J. and Culbertson, M.R. (1999) Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome. *Mol Cell Biol*, **19**, 6710-6719.

Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*, **100**, 189-192.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat Genet*, **21**, 20-24.

Longman, D., Johnstone, I.L. and Caceres, J.F. (2000) Functional characterization of SR and SR-related genes in Caenorhabditis elegans. *Embo J*, **19**, 1625-1637.

Longman, D., Plasterk, R.H., Johnstone, I.L. and Caceres, J.F. (2007) Mechanistic insights and identification of two novel factors in the C. elegans NMD pathway. *Genes Dev*, **21**, 1075-1085.

Losson, R. and Lacroute, F. (1979) Interference of nonsense mutations with eukaryotic messenger RNA stability. *Proc Natl Acad Sci U S A*, **76**, 5134-5137.

Lykke-Andersen, J., Shu, M.D. and Steitz, J.A. (2001) Communication of the position of exon-exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Science*, **293**, 1836-1839.

Lyne, R., Burns, G., Mata, J., Penkett, C.J., Rustici, G., Chen, D., Langford, C., Vetrie, D. and Bahler, J. (2003) Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, **4**, 27.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A. and Gingeras, T.R. (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet*, **38**, 1151-1158.

Mango, S.E. (2001) Stop making nonSense: the C. elegans smg genes. *Trends Genet*, **17**, 646-653.

Marin, V.A. and Evans, T.C. (2003) Translational repression of a C. elegans Notch mRNA by the STAR/KH domain protein GLD-1. *Development*, **130**, 2623-2632.

Matzke, M.A. and Birchler, J.A. (2005) RNAi-mediated pathways in the nucleus. *Nat Rev Genet*, **6**, 24-35.

Medghalchi, S.M., Frischmeyer, P.A., Mendell, J.T., Kelly, A.G., Lawler, A.M. and Dietz, H.C. (2001) Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum Mol Genet*, **10**, 99-105.

Mellor, H. and Parker, P.J. (1998) The extended protein kinase C superfamily. *Biochem J*, **332 (Pt 2)**, 281-292.

Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F. and Dietz, H.C. (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet*, **36**, 1073-1078.

Merida, I., Avila-Flores, A. and Merino, E. (2008) Diacylglycerol kinases: at the hub of cell signalling. *Biochem J*, **409**, 1-18.

Mitchell, P. and Tollervey, D. (2003) An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3'-->5' degradation. *Mol Cell*, **11**, 1405-1413.

Morrison, M., Harris, K.S. and Roth, M.B. (1997) smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. *Proc Natl Acad Sci U S A*, **94**, 9782-9785.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344-1349.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M, Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev*, **21**, 708-718.

Nishizuka, Y. (1984) The role of protein kinase C in cell surface signal transduction and tumour promotion. *Nature*, **308**, 693-698.

Ohmachi, M., Rocheleau, C.E., Church, D., Lambie, E., Schedl, T. and Sundaram, M.V. (2002) C. elegans ksr-1 and ksr-2 have both unique and redundant functions and are required for MPK-1 ERK phosphorylation. *Curr Biol*, **12**, 427-433.

Ohnishi, T., Yamashita, A., Kashima, I., Schell, T., Anders, K.R., Grimson, A., Hachiya, T., Hentze, M.W., Anderson, P. and Ohno, S. (2003) Phosphorylation of hUPF1 induces formation of mRNA surveillance complexes containing hSMG-5 and hSMG-7. *Mol Cell*, **12**, 1187-1200.

Olivas, W. and Parker, R. (2000) The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *Embo J*, **19**, 6602-6611.

Oliveira, V., Romanow, W.J., Geisen, C., Otterness, D.M., Mercurio, F., Wang, H.G., Dalton, W.S. and Abraham, R.T. (2008) A protective role for the human SMG-1 kinase against tumor necrosis factor-alpha-induced apoptosis. *J Biol Chem*, **283**, 13174-84.

Page, M.F., Carr, B., Anders, K.R., Grimson, A. and Anderson, P. (1999) SMG-2 is a phosphorylated protein required for mRNA surveillance in Caenorhabditis elegans and related to Upf1p of yeast. *Mol Cell Biol*, **19**, 5943-5951.

Pepper, A.S., Lo, T.W., Killian, D.J., Hall, D.H. and Hubbard, E.J. (2003) The establishment of Caenorhabditis elegans germline pattern is controlled by overlapping proximal and distal somatic gonad signals. *Dev Biol*, **259**, 336-350.

Petcherski, A.G. and Kimble, J. (2000) LAG-3 is a putative transcriptional activator in the C. elegans Notch pathway. *Nature*, **405**, 364-368.

Petricoin, E.F., 3rd, Hackett, J.L., Lesko, L.J., Puri, R.K., Gutman, S.I., Chumakov, K., Woodcock, J., Feigal, D.W., Jr., Zoon, K.C. and Sistare, F.D. (2002) Medical applications of microarray technologies: a regulatory science perspective. *Nat Genet*, **32 Suppl**, 474-479.

Poulin, G., Dong, Y., Fraser, A.G., Hopper, N.A. and Ahringer, J. (2005) Chromatin regulation and sumoylation in the inhibition of Ras-induced vulval development in Caenorhabditis elegans. *Embo J*, **24**, 2613-2623.

Qiao, L., Lissemore, J.L., Shu, P., Smardon, A., Gelber, M.B. and Maine, E.M. (1995) Enhancers of glp-1, a gene required for cell-signaling in Caenorhabditis elegans, define a set of genes required for germline development. *Genetics*, **141**, 551-569.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444-454.

Rehwinkel, J., Letunic, I., Raes, J., Bork, P. and Izaurralde, E. (2005) Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *Rna*, **11**, 1530-1544.

Reinke, V. (2002) Functional exploration of the C. elegans genome using DNA microarrays. *Nat Genet*, **32 Suppl**, 541-546.

Reinke, V., Gil, I.S., Ward, S. and Kazmer, K. (2004) Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development*, **131**, 311-323.

Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J., Davis, E.B., Scherer, S., Ward, S. and Kim, S.K. (2000) A global profile of germline gene expression in C. elegans. *Mol Cell*, **6**, 605-616.

Reinke, V. and White, K.P. (2002) Developmental genomic approaches in model organisms. *Annu Rev Genomics Hum Genet*, **3**, 153-178.

Rodriguez-Gabriel, M.A., Watt, S., Bahler, J. and Russell, P. (2006) Upf1, an RNA helicase required for nonsense-mediated mRNA decay, modulates the transcriptional response to oxidative stress in fission yeast. *Mol Cell Biol*, **26**, 6347-6356.

Rual, J.F., Ceron, J., Koreth, J., Hao, T., Nicot, A.S., Hirozane-Kishikawa, T., Vandenhaute, J., Orkin, S.H., Hill, D.E., van den Heuvel, S. and Vidal, M. (2004) Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library. *Genome Res*, **14**, 2162-2168.

Ruiz-Echevarria, M.J. and Peltz, S.W. (2000) The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. *Cell*, **101**, 741-751.

Ryder, S.P., Frater, L.A., Abramovitz, D.L., Goodwin, E.B. and Williamson, J.R. (2004) RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. *Nat Struct Mol Biol*, **11**, 20-28.

Saltzman AL, Kim YK, Pan Q, Fagnani MM, Maquat LE, Blencowe BJ. 2008. Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol Cell Biol*, **28**, 4320-4330.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097-6100.

Schroeter, E.H., Kisslinger, J.A. and Kopan, R. (1998) Notch-1 signalling requires ligand-induced proteolytic release of intracellular domain. *Nature*, **393**, 382-386.

Shaye, D.D. and Greenwald, I. (2002) Endocytosis-mediated downregulation of LIN-12/Notch upon Ras activation in Caenorhabditis elegans. *Nature*, **420**, 686-690.

Simske, J.S. and Kim, S.K. (1995) Sequential signalling during Caenorhabditis elegans vulval induction. *Nature*, **375**, 142-146.

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

Sriraman, V., Modi, S.R., Bodenburg, Y., Denner, L.A. and Urban, R.J. (2008) Identification of ERK and JNK as signaling mediators on protein kinase C activation in cultured granulosa cells. *Mol Cell Endocrinol*.

Sternberg, P.W. (1988) Lateral inhibition during vulval induction in Caenorhabditis elegans. *Nature*, **335**, 551-554.

Sternberg, P.W. and Horvitz, H.R. (1986) Pattern formation during vulval development in C. elegans. *Cell*, **44**, 761-772.

Stiernagle, T. (2006) Maintenance of *C. elegans*. *Wormbook*, **11**, 1-11.

Sulston, J.E. and Horvitz, H.R. (1977) Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol*, **56**, 110-156.

Sulston, J.E., Schierenberg, E., White, J.G. and Thomson, J.N. (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol*, **100**, 64-119.

Sulston, J.E. and White, J.G. (1980) Regulation and cell autonomy during postembryonic development of Caenorhabditis elegans. *Dev Biol*, **78**, 577-597.

Sundaram, M. and Greenwald, I. (1993) Suppressors of a lin-12 hypomorph define genes that interact with both lin-12 and glp-1 in Caenorhabditis elegans. *Genetics*, **135**, 765-783.

Tabara, H., Grishok, A. and Mello, C.C. (1998) RNAi in C. elegans: soaking in the genome sequence. *Science*, **282**, 430-431.

Tan, P.B., Lackner, M.R. and Kim, S.K. (1998) MAP kinase signaling specificity mediated by the LIN-1 Ets/LIN-31 WH transcription factor complex during C. elegans vulval induction. *Cell*, **93**, 569-580.

Tarraga, J., Medina, I., Carbonell, J., Huerta-Cepas, J., Minguez, P., Alloza, E., Al-Shahrour, F., Vegas-Azcarate, S., Goetz, S., Escobar, P., Garcia-Garcia, F., Conesa, A., Montaner, D. and Dopazo, J. (2008) GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res*, **36**, W308-314.

The FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2006) The Transcriptional Landscape of the Mammalian Genome. *Science*, **309**, 1559-1563.

Timmons, L., Court, D.L. and Fire, A. (2001) Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. *Gene*, **263**, 103-112.

Timmons, L. and Fire, A. (1998) Specific interference by ingested dsRNA. *Nature*, **395**, 854.

Wang, L., Eckmann, C.R., Kadyk, L.C., Wickens, M. and Kimble, J. (2002) A regulatory cytoplasmic poly(A) polymerase in Caenorhabditis elegans. *Nature*, **419**, 312-316.

Watt, F.M. and Hogan, B.L. (2000) Out of Eden: stem cells and their niches. *Science*, **287**, 1427-1430.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bahler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239-1243.

Winston, W.M., Molodowitch, C. and Hunter, C.P. (2002) Systemic RNAi in C. elegans requires the putative transmembrane protein SID-1. *Science*, **295**, 2456-2459.

Winston, W.M., Sutherlin, M., Wright, A.J., Feinberg, E.H. and Hunter, C.P. (2007) Caenorhabditis elegans SID-2 is required for environmental RNA interference. *Proc Natl Acad Sci U S A*, **104**, 10565-10570.

Wreden, C., Verrotti, A.C., Schisa, J.A., Lieberfarb, M.E. and Strickland, S. (1997) Nanos and pumilio establish embryonic polarity in Drosophila by promoting posterior deadenylation of hunchback mRNA. *Development*, **124**, 3015-3023.

Wu, Y. and Han, M. (1994) Suppression of activated Let-60 ras protein defines a role of Caenorhabditis elegans Sur-1 MAP kinase in vulval differentiation. *Genes Dev*, **8**, 147-159.

Wultsch, T., Chourbaji, S., Fritzen, S., Kittelt, S., Grunblatt, E., Gerlach, M., Gutknecht, L., Chizat, F., Golfier, G., Schmitt, A., Gass, P., Lesch, K.P. and Reif, A. (2007) Behavioural and expressional phenotyping of nitric oxide synthase-I knockdown animals. *J Neural Transm Suppl*, 69-85.

Yamashita, A., Kashima, I. and Ohno, S. (2005) The role of SMG-1 in nonsense-mediated mRNA decay. *Biochim Biophys Acta*, **1754**, 305-315.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30**, e15.

Yochem, J. and Greenwald, I. (1989) glp-1 and lin-12, genes implicated in distinct cell-cell interactions in C. elegans, encode similar transmembrane proteins. *Cell*, **58**, 553-563.

Yochem, J., Weston, K. and Greenwald, I. (1988) The Caenorhabditis elegans lin-12 gene encodes a transmembrane protein with overall similarity to Drosophila Notch. *Nature*, **335**, 547-550.

Yoo, A.S., Bais, C. and Greenwald, I. (2004) Crosstalk between the EGFR and LIN-12/Notch pathways in C. elegans vulval development. *Science*, **303**, 663-666.

Zamore, P.D. and Haley, B. (2005) Ribo-gnome: the big world of small RNAs. *Science*, **309**, 1519-1524.

Zhang, B., Gallegos, M., Puoti, A., Durkin, E., Fields, S., Kimble, J. and Wickens, M.P. (1997) A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line. *Nature*, **390**, 477-484.

Zien, A., Gebhard, P.M., Fundel, K. and Aigner, T. (2007) Phenotyping of chondrocytes in vivo and in vitro using cDNA array technology. *Clin Orthop Relat Res*, **460**, 226-233.