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.

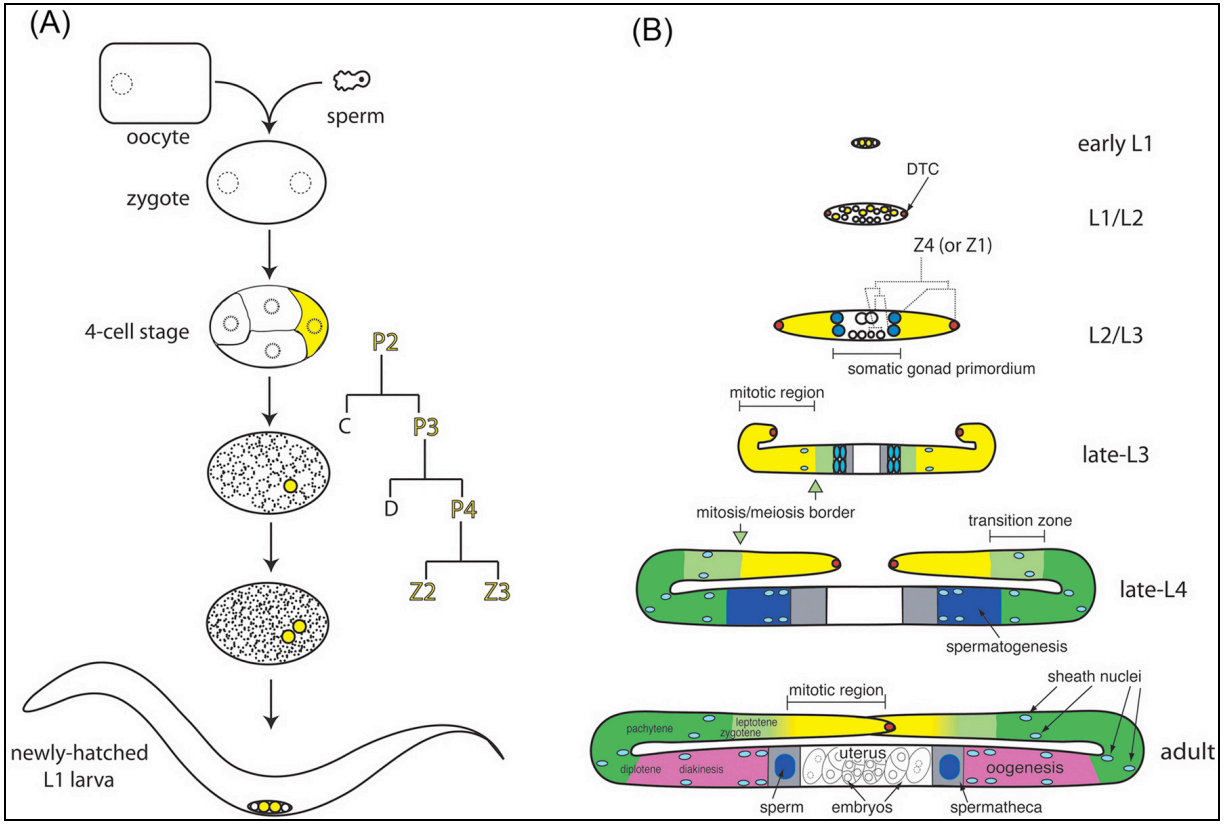

Figure 1.1. Cartoon representation of gonadogenesis. (A) Fertilization and the embryonic germ line: Fertilization of oocyte by sperm leads to embryonic development. Germline lineages are in yellow. (B) Post-embryonic hermaphrodite gonad development: Germline colour scheme: yellow = mitotic region, light 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.

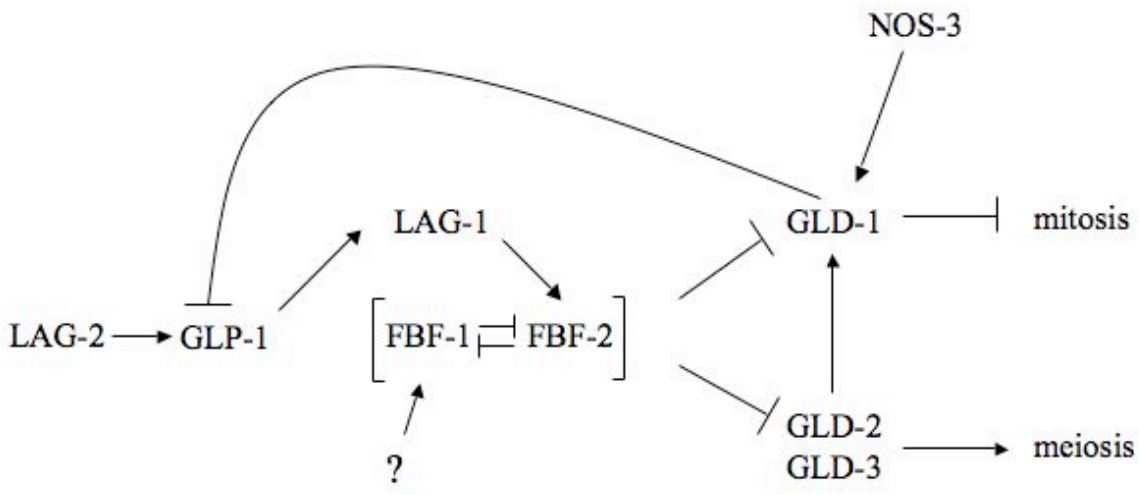

Figure 1.2. Regulation of the mitosis/meiosis decision by the interplay of pro- and anti-meiotic factors. Notch signalling activates anti-meiotic factors but is in turn suppressed by GLD-1 permitting entry into meiosis, stimulated by GLD-2 activation of pro-meiotic targets. This circuitry provides only 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.

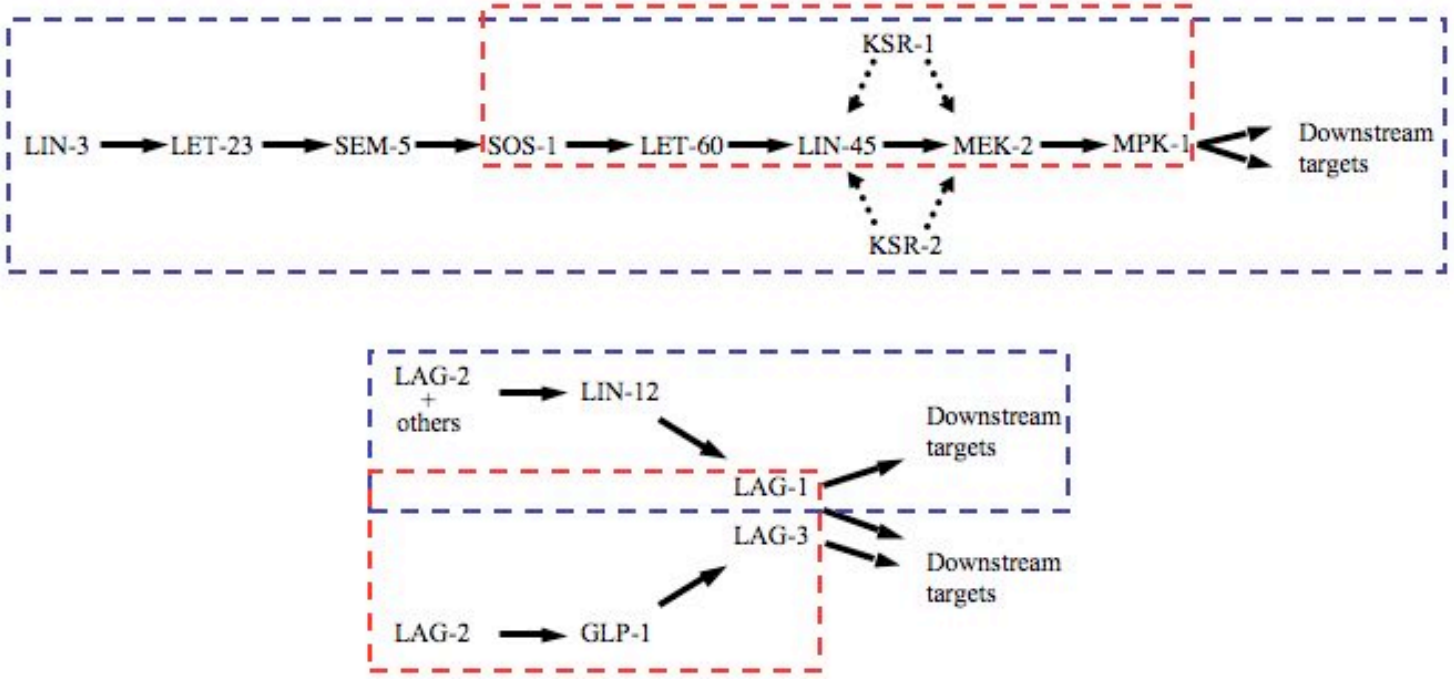

Figure 1.3. The canonical EGF/ras/MAPK and Notch signalling pathways as they are known to act in the vulva and germline. The EGF/ras/MAPK signalling pathway is shown at the top and Notch at the bottom. Components known to act in the vulva are outlined in blue and the germline in red. The classic model 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.

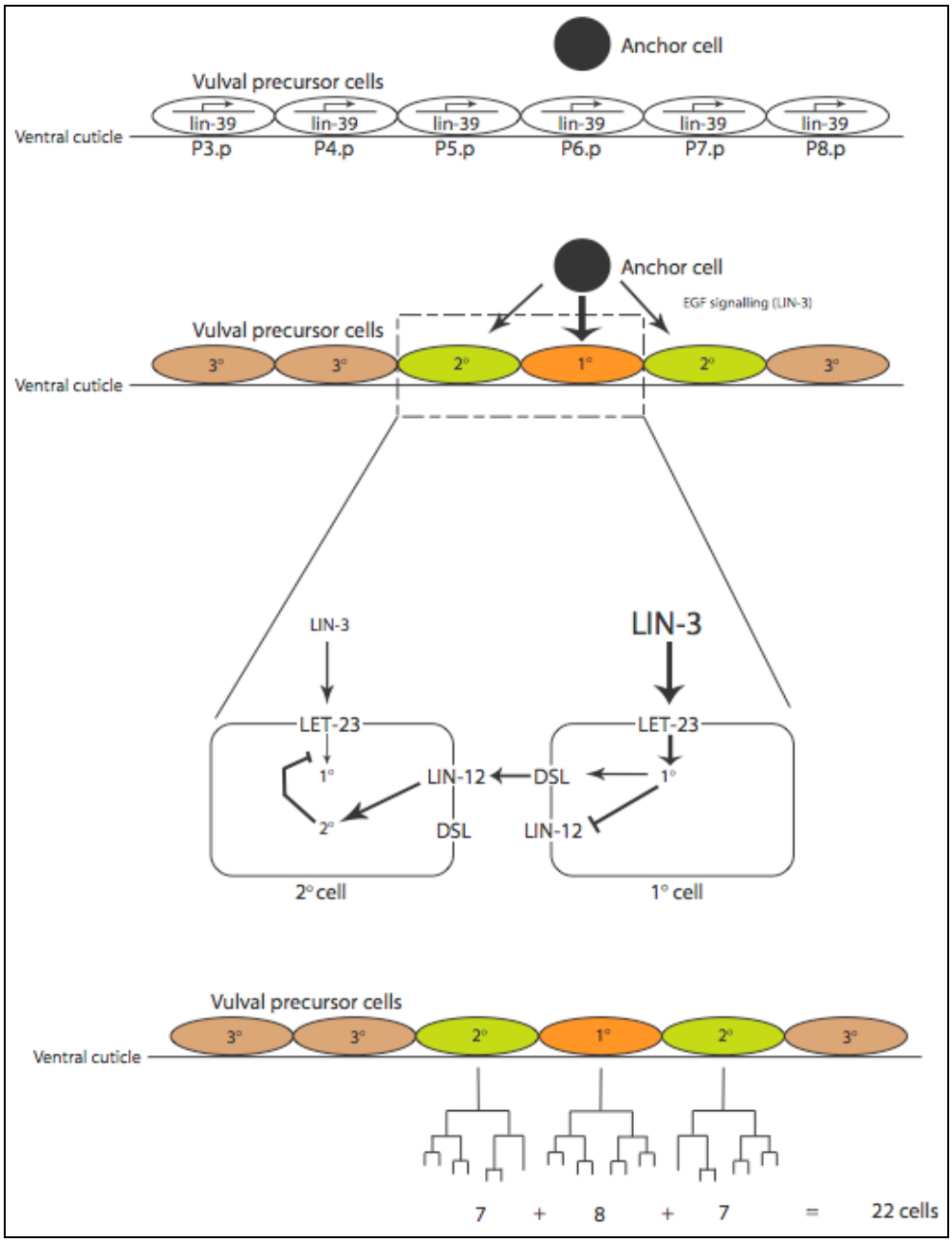

Figure 1.4. Vulval specification and lineage. Expression of *lin-39* imparts the potential on six cells (P3.p-P8.p) along the ventral axis of the worm to adopt vulval fates. EGF signalling from the anchor cell, part of the somatic gonad, leads to the specification of these cells as either 1^{o}, 2^{o} 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.

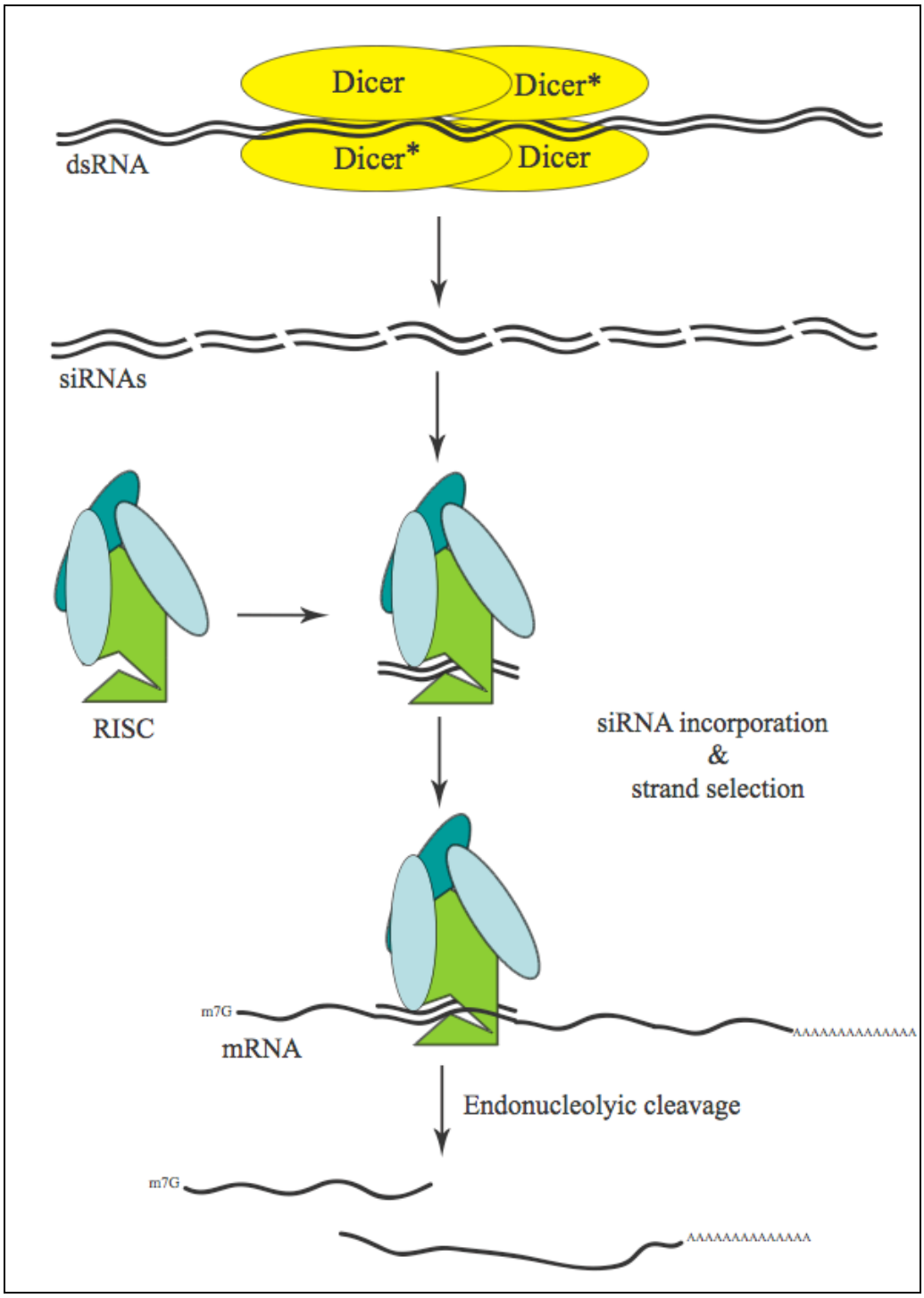

Figure 1.5. Mechanism of RNAi gene silencing. Two Dicer homo-dimers associate in anti-parallel orientation to cleave dsRNA. Only one catalytic centre in each Dicer homo-dimer is active (*). Active catalytic domains are spaced by ~22nt giving (siRNAs) of that length. siRNAs are incorporated into the RISC 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.

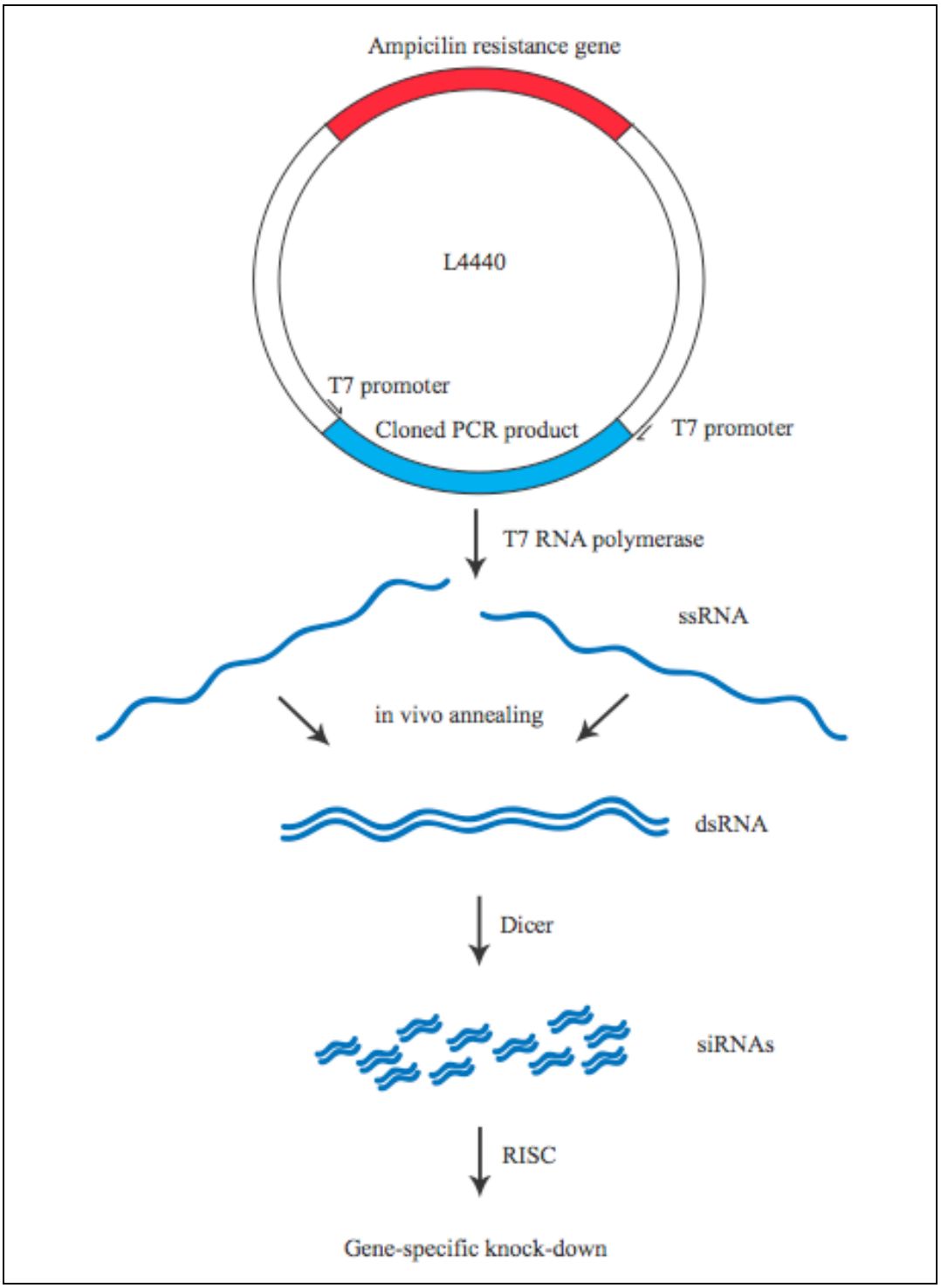

Figure 1.6. L4440 RNA interference feeding vector. A PCR product homologous to a target gene of interest is cloned between inverted T7 promoter sites. The vector is then transformed into an *Eschericia coli* strain expressing T7 RNA polymerase (HT115(DE3)), resulting in transcription of anti-parallel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

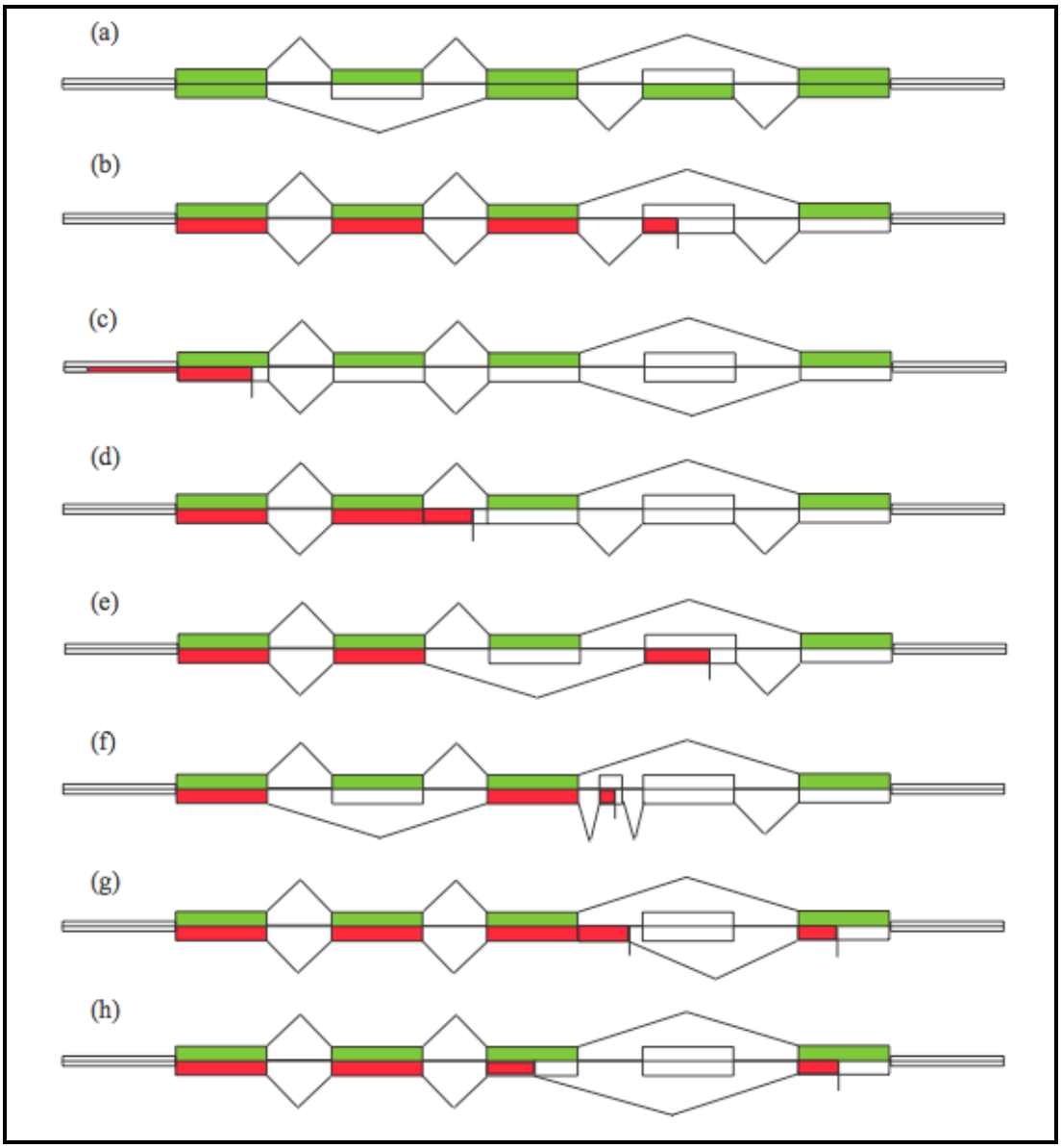

Figure 1.7. The recognized post-transcriptional causes of NMD targeting. Numerous translational and splicing events can lead to NMD targeting. The true coding ORF of the transcript not ending in a PTC is shown in green. The ORF ending in a detected PTC leading to NMD targeting is shown in red. (a) A 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.

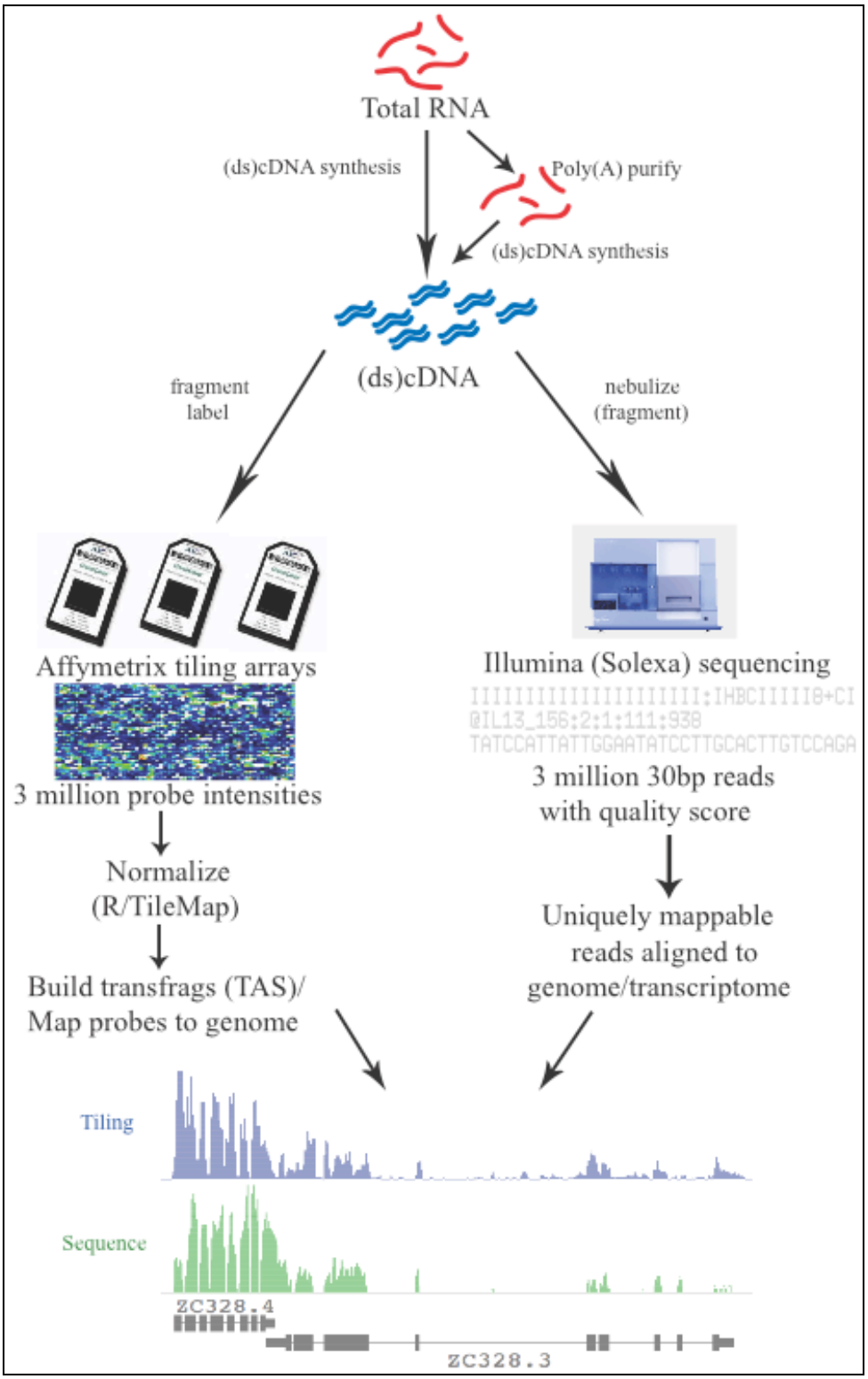

Figure 1.8. Technical flow-through of Affymetrix tiling array and Illumina sequencing technologies. The two independent technologies provide analogous and complementary datasets, tiling arrays of 35bp resolution and high depth, the sequencing of 1bp resolution but lower depth.