UNIVERSITY OF CALGARY

A study of the *C. elegans teg-4* gene, using Tiling array analysis

by

Xuan Zhao

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF BIOLOGICAL SCIENCES

CALGARY, ALBERTA

JULY, 2011

© XUAN ZHAO 2011

UNIVERSITY OF
CALGARY

The author of this thesis has granted the University of Calgary a non-exclusive license to reproduce and distribute copies of this thesis to users of the University of Calgary Archives.

Copyright remains with the author.

Theses and dissertations available in the University of Calgary Institutional Repository are solely for the purpose of private study and research. They may not be copied or reproduced, except as permitted by copyright laws, without written authority of the copyright owner. Any commercial use or re-publication is strictly prohibited.

The original Partial Copyright License attesting to these terms and signed by the author of this thesis may be found in the original print version of the thesis, held by the University of Calgary Archives.

Please contact the University of Calgary Archives for further information:
E-mail: uarc@ucalgary.ca
Telephone: (403) 220-7271
Website: http://archives.ucalgary.ca

Abstract

teg-4, which encodes a protein homologous to the SAP130 splicing factor, was identified to be involved in the regulation of the proliferation/differentiation decision in the *C. elegans* germ line (Mantina et al., 2009). Tiling arrays were used in this work to identify the transcriptome changes in *teg-4* mutants. This work designed computational methods (using R language) for systematically analyzing tiling array data for the identification of splicing defects and differentially expressed genes. Data analyses identified 42 genes with altered expression levels and thousands of potential splicing defects in *teg-4* mutants. RNAi screens on the 42 genes identified 3 genes that can affect germline tumors. One of them, the gene C38D9.2, when mutated, showed strong tumor suppression abilities. Overall, this work developed new methods for analyzing tiling data, and provides more insights in understanding the mechanisms of *teg-4*'s regulation in the proliferation/differentiation decision.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor, Dr. Dave Hansen. His patience, encouragement and continuous support from the initial to the final level enabled me to develop an understanding of the subject.

My sincere thanks also go to my supervisory committee, Dr. John Cobb, Dr. Jeb Gaudet and Dr. Gordon Chua, who have provided valuable suggestions and encouragement. I would also like to thank Dr. Mark Bieda for agreeing to be part of my defence committee.

I warmly thank Dr. Mayi Arcellana-Panlilio and Dr. Xiuling Wang at the Southern Alberta Microarray Facility for their excellent work in performing the Tiling array experiment.

I would like to thank all the members in the Hansen lab, for their kindness, help, suggestions, and for sharing interesting life experiences with me.

I am deeply grateful to my husband, Xin Wang; not only for his tremendous love and care in my life, but also for the insightful academic discussion with him throughout my graduate study, from which I benefited so much.

Finally, I would like to thank my parents for their love and support.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definetion
5’ UTR	5’ untranslated region
cDNA	complementary DNA
cM	centiMorgans
DIC	differential interference contrast microscopy
DNA	deoxyribonucleic acid
DTC	distal tip cell
ECD	extracellular domain
EGF	Epidermal growth factor
EST	Expressed Sequence Tag
FDR	False Discovery Rate
gf	gain-of-function
GO	gene ontology
GSC	germ line stem cell
IGB	Integrated Genome Browser
N2	Bristol wild-type strain
NGM	nematode growth medium
NMD	nonsense-mediated decay
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PTC	premature termination codon
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
RNAi	RNA interference
ROI	Region of Interest
rRNA	Ribosomal RNA
SAP	Spliceosome-associated Protein
WT	wild-type

Chapter One: Introduction

1.1 Stem cells

Stem cells are a group of cells with amazing features. Unlike other cells that can only remain in their current state, stem cells have the potential of differentiating into virtually any cell type, yet still maintaining the ability of proliferating to produce more stem cells. Because of these intriguing characteristics, stem cells have the potential to be used in treating many neurodegenerative diseases such as Parkinson's disease (Tonnesen et al., 2011). However, in contrast to the rapid development in the application of stem cells in therapeutic areas, the mechanisms that regulate stem cell behaviour were largely unknown. Since stem cells are able to both proliferate and differentiate, the switch between these two behaviours is crucial for maintaining a balance between proliferation and differentiation. If the switch occurs too early, stem cells will differentiate prematurely without establishing a regenerating pool, and as a result, there will not be enough "ingredients" (constant renewing stem cells) from which cells of other types are made, causing irreversible loss of the differentiated tissue. In the opposite scenario, if the switch to differentiation comes too late or never occurs, stem cells will over proliferate with little or no differentiation at all, and this usually will lead to the formation of a tumor. The balance between proliferation and differentiation of stem cells is very delicate; any kind of imbalance will result in detrimental consequences. To regulate this dynamic balance, organisms utilize different strategies. One way to understand this intricate regulation is to use a well known model system.

1.2 Germline stem cells

Stem cells reside in both somatic tissues and the germ lines (Li and Xie, 2005). While somatic stem cells are essential for organogenesis and tissue regeneration, germline stem cells (GSCs) are only responsible for producing gametes for the creation of the next generation (Kimble and White, 1981; Lin and Spradling, 1993; Xie, 2008). GSCs undergo mitosis first to establish a pool of self-renewing stem cells; then they will cease the mitotic cell cycle and enter into the meiotic cell cycle to further differentiate into gametes. The decision between mitosis and meiosis is as crucial for GSCs to maintain the reproductive robustness as for other stem cell to maintain a balance between proliferation and differentiation. Studying the regulation of GSCs behaviour can greatly assist in the understanding of all other types of stem cells. The germ line is thus an appropriate model for studying stem cells. Specifically, this project used the *C. elegans* germ line as a model system.

1.3 The *C. elegans* germ line as a model system

1.3.1 C. elegans *as a model organism*

As a free-living soil dwelling transparent nematode, *C. elegans* was first applied in biological research by Sydney Brenner (Brenner, 1974), and ever since then it has been used extensively as a model organism for both genetic and developmental studies. *C. elegans* is a good organism for genetic studies due to the following characterstics. First of all, *C. elegans* has a rapid life cycle of only 3.5 days, within which it undergoes both embryogenesis and four larval stages (L1-L4) to reach adulthood. Secondly, *C. elegans* exists as both hermaphrodites, that are capable of self-fertilizing by producing both sperm

and oocytes, and males, which only generate sperm. These two features allow for performing both inbreeding (by using hermaphrodites) and crosses (by mating males with hermaphrodites), and researchers can obtain the desired progeny and experimental data in a short period of time (rapid life cycle).

C. elegans is also suitable for developmental studies. *C. elegans* is transparent and adults are approximately 1 mm in length, thus developmental processes can easily be monitored through the use of microscopy. Another beneficial and most interesting feature is that *C. elegans* is eutelic. Eutely guarantees a reliable and precise detection of any alteration on cell number and cell position caused by developmental abnormalities. Actually, the entire cell lineage of wild-type *C. elegans* from fertilized egg to adult has been determined (Sulston and Horvitz, 1977). All these are of great advantages for *C. elegans* being used as a model for development study.

Moreover, *C. elegans* is the first metazoan whose whole genome was sequenced (Consortium, 1998), and its genome is highly annotated (www.wormbase.org). This makes large scale bioinformatic studies possible. Many genes in *C. elegans* have homologues in humans (Lesney, 2001), meaning these genes are highly conserved and therefore studies on *C. elegans* can shed light on understanding many biological processes in humans.

1.3.2 The C. elegans *germ line*

The germ line of the *C. elegans* adult hermaphrodite contains two gonad arms (Figure1.1A). Each gonad arm is highly polarized, harbouring cells that undergo dynamic changes from distal to proximal regions (Figure 1.1B). In the distal most region of the

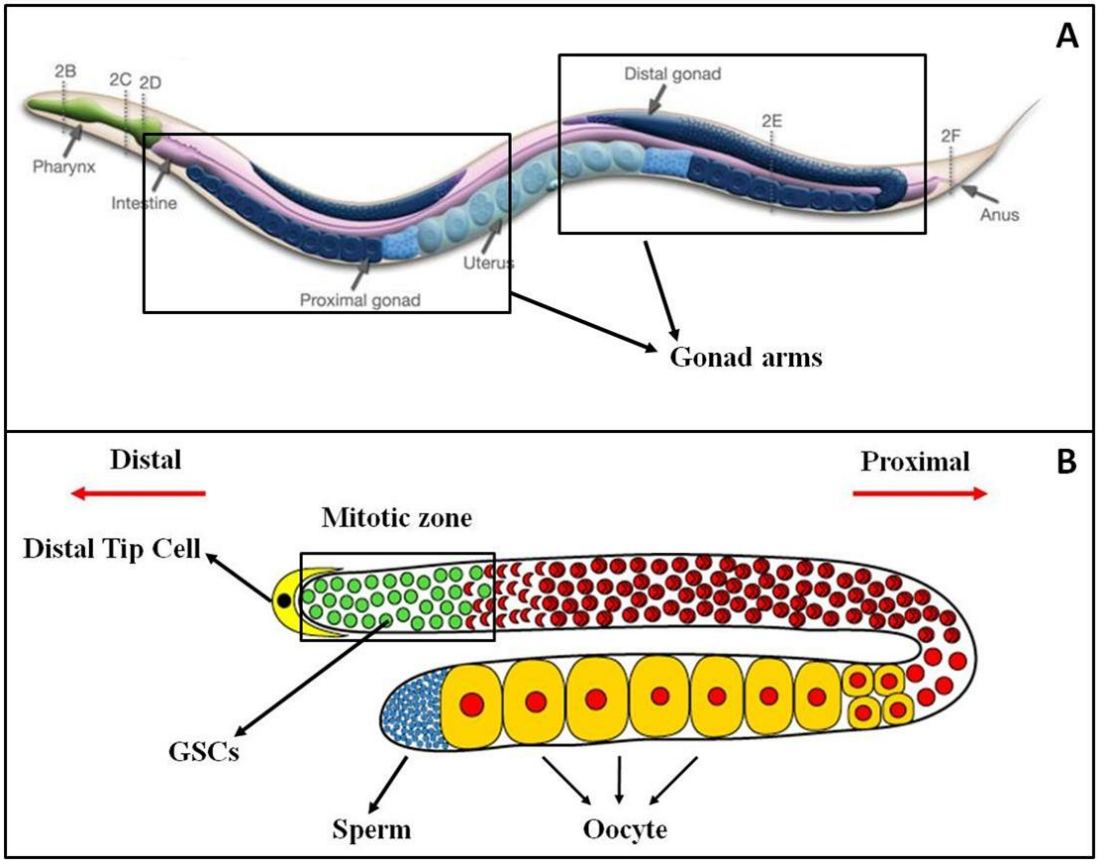


Figure 1.1. The *C. elegans* hermaphrodite germ line.

(A) A diagram structure of an adult hermaphrodite *C. elegans*. The major inner sections of *C. elegans* include pharynx, intestine, and germ line. Hermaphrodite animals have two gonad arms that are structurally the same. (B) A diagram of one gonad arm. The *C. elegans* germ line is highly polarized with a distal section and a proximal section. The distal end (top left) of each gonad arm is capped by a somatic distal tip cell (DTC). Germ cells undergo proliferation (green) in the mitotic zone. As cells move proximally, they will exit mitotic processes and switch into meiotic processes (red) to further differentiate to sperm (blue), or oocytess (yellow).

germ line resides cells that are mitotic and are continually proliferating to maintain a GSC pool (Lin, 1997). As more cells are generated, some of these cells will be pushed away from the mitotic zone and will migrate proximally; they will then exit mitosis and enter into meiosis, eventually differentiating into sperm or oocytes.

The maintenance of the germ cell mitotic population and the switch between mitosis and meiosis are highly regulated. Insufficient proliferation or premature entry into meiosis can lead to a decrease in the number of gametes, or in the most extreme circumstances, a nearly empty germ line with only a few sperm (Figure 1.2B). Excessive proliferation, without switching to meiosis, usually results in a tumorous germ line (Figure 1.2C). The underlying mechanism of precisely how the balance is maintained between mitosis and meiosis in the *C. elegans* germ line is still not fully understood. However, it is known that this regulation utilizes the highly conserved GLP-1/Notch signaling pathway (Seydoux and Schedl, 2001) followed by two redundant mRNA regulatory pathways (Kadyk and Kimble, 1998), which are discussed below.

1.4 Regulation of mitosis and meiosis in the *C. elegans* germ line

1.4.1 The role of DTC

The first breakthrough in understanding the regulation of the proliferation/differentiation decision in *C. elegans* germ line was in 1981, when Kimble and White conducted a series of experiments to investigate the effects of Distal Tip Cell (DTC) laser ablation at different developmental stages on germ cell status (Kimble and White, 1981). The DTC is a somatic cell that is located at the distal end of each gonad arm, and has long projections that cap and are in contact with the germ cells in the distal

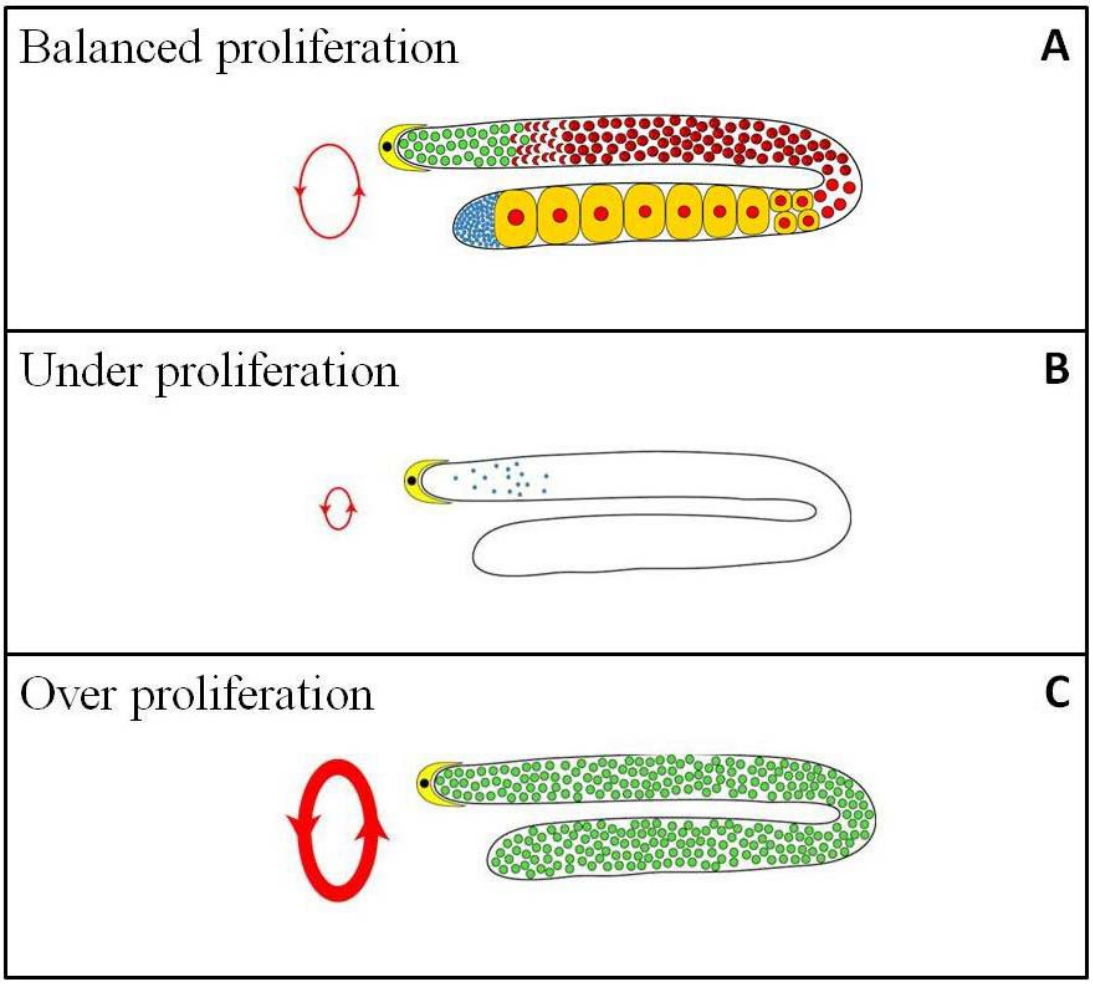


Figure 1.2. A diagram of wild-type and mutant gonad arms.

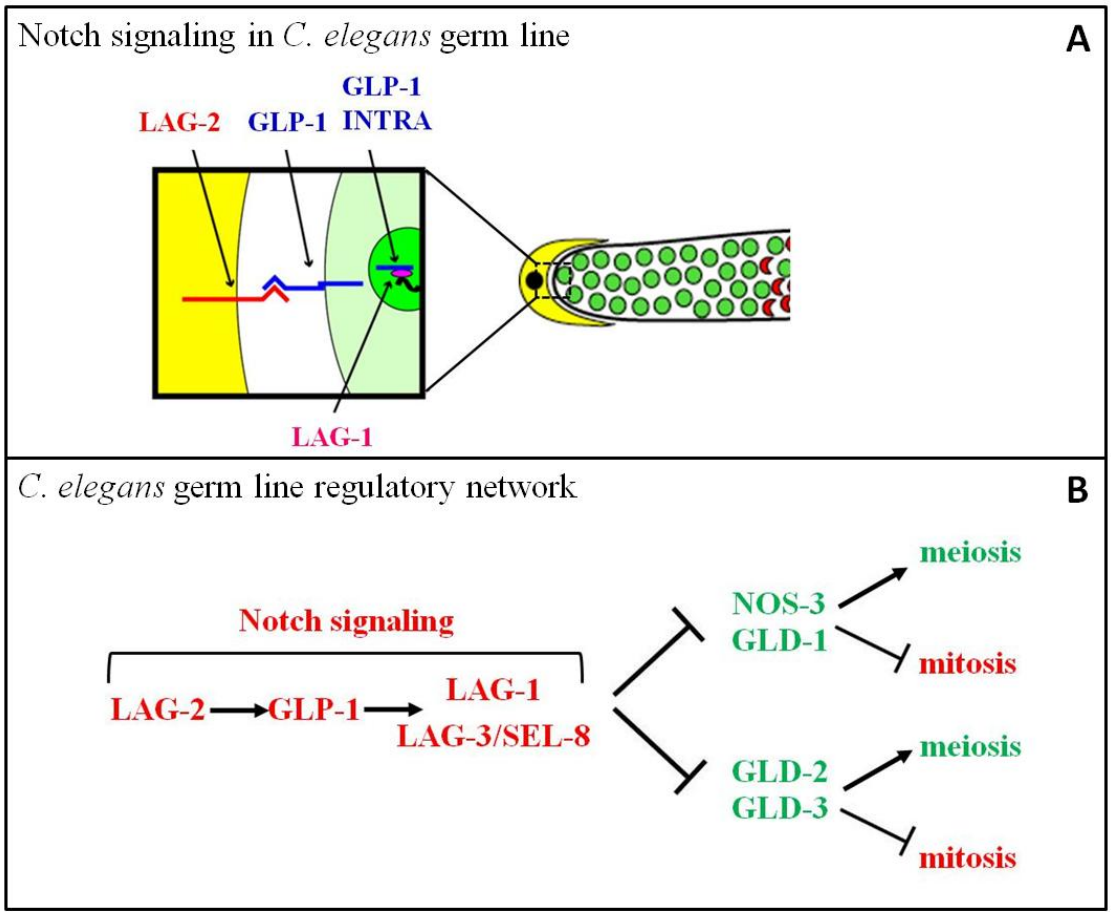
(A) A wild-type hermaphrodite germ line. Proliferative germ cells (green) only exist at the distal region, and cells that are located more proximally begin meiosis to undergo gametogenesis. (B) A "Glp" germ line. Germ cells enter into meiosis prematurely, and only a few sperm are produced (black dots). (C) A Tumorous germ line. Germ cells constitutively proliferate, without switching into meiosis.

mitotic region (Hall et al., 1999) (Figure1.1). In Kimble's and White's experiment, they noticed that after the removal of the DTC, cells adjacent to the DTC that normally mitotically divide all stopped proliferating, and entered into the meiotic cell cycle (Kimble and White, 1981). They also manipulated the position of the DTC and found that ectopic positioning of the DTC led to an alteration of germ cell organization polarity. More specifically, wherever the DTC was located, the surrounding germ cells were mitotic. Their findings suggested an important role for the DTC in controlling the decision between mitosis and meiosis for *C. elegans* germ cells as well as creating a niche suitable for cell proliferation.

1.4.2 GLP-1/Notch signaling pathway

It is now known that the underlying molecular mechanism, by which the DTC governs germ cell proliferation, is through a conserved signaling pathway, which in *C. elegans*, is known as the GLP-1/Notch signaling pathway (Figure 1.3). There are four fundamental components in this pathway: LAG-2 (Notch ligand), GLP-1 (Notch receptor), LAG-1 (transcription factor) and LAG-3/SEL-8 (transcription co-activator). LAG-2 is a conserved DSL (Delta-Serate-LAG-2) Notch ligand expressed in the DTC (Henderson et al., 1994). GLP-1, the Notch receptor, is a transmembrane protein that is expressed on the surface of the mitotic germ cells (Crittenden et al., 1994) adjacent to the DTC. LAG-2 has the ability to bind the GLP-1/Notch receptor (Henderson et al., 1994; Tax et al., 1994) when the DTC touches neighbouring germ cells. Upon binding between LAG-2 and GLP-1, a proteolytic cleavage event occurs within the GLP-1 receptor

Figure 1.3. Regulation of mitosis and meiosis at the *C. elegans* germ line.
(A) The core components of the GLP-1/Notch signaling pathway are shown. The DSL ligand (Delta/Serrate/**LAG-2**), the **GLP-1** Notch receptor and the **LAG-1** CSL transcription factor. The pathway is only active at the distal end of the gonad arm due to the restriction of the expression of the LAG-2 ligand to the membrane of the somatic distal tip cell (DTC). When the LAG-2 ligand binds to the GLP-1 receptor, the GLP-1 receptor is cleaved, to allow the intracellular portion of GLP-1 (**GLP-1 INTRA**), to translocate to the nucleus and bind to the LAG-1 transcription factor and SEL-8/LAG-3 co-activator. This complex transcribes genes that are necessary for promoting mitosis and inhibiting meiosis. (B) The regulatory network of the *C. elegans* germline proliferation/differentiation decision. Notch signaling is responsible for maintaining mitosis in distal region. Downstream of Notch signaling are two redundant pathways that promote meiosis (arrows) and inhibit mitosis (barred lines). NOS-3 and GLD-2 function in the same pathway, and GLD-2 and GLD-3 function in the other pathway. Figure adapted from Hansen and Schedl (2006) (Hansen and Schedl, 2006).


Notch signaling in C. elegans germ line
A
LAG-2
GLP-1
GLP-1
INTRA
LAG-1
C. elegans germ line regulatory network
B
Notch signaling
LAG-2
GLP-1
LAG-1
LAG-3/SEL-8
NOS-3
GLD-1
meiosis
mitosis
GLD-2
GLD-3
meiosis
mitosis

(Mumm and Kopan, 2000), causing its intracellular domain (Notch intracellular domain: NCID, or GLP-1 (INTRA)) to be translocated into the nucleus. In the nucleus, GLP-1 (INTRA) binds to LAG-1, a CSL (CBF-1/RBPJκ-Su(H)-LAG-1) transcription factor (Christensen et al., 1996). GLP-1 (INTRA) and LAG-1 recruit the transcription coactivator LAG-3/SEL-8 MAML to form a ternary complex (Doyle et al., 2000; Nam et al., 2006; Petcherski and Kimble, 2000; Wilson and Kovall, 2006), which then activate the transcription of downstream genes involved in cell proliferation (Figure 1.3A). Loss-of-function in any of the above components leads to compromised Notch signaling, resulting in phenotypes resembling DTC ablated gonads (Austin and Kimble, 1987; Christensen et al., 1996; Doyle et al., 2000; Henderson et al., 1994; Petcherski and Kimble, 2000; Tax et al., 1994). Conversely, the gain-of-function (gf) allele, *glp-1(oz112gf)*, causes constitutive cleavage of GLP-1, which causes the germ cells to proliferate continuously, regardless of the presence of the ligands, resulting in the formation of a germline tumor (Berry et al., 1997; Pepper et al., 2003). Generally, GLP-1/Notch signaling promotes proliferation of germ cells in the distal mitotic region due to cells being close to the DTC; as these cells move proximally, GLP-1 receptors no longer interact with LAG-2 on the DTC, thus no GLP-1 (INTRA) is released to nucleus, and no ternary complex is formed to turn on the transcription of downstream genes. Eventually, these proximally moving cells will exit mitosis and enter into meiosis (Hansen and Schedl, 2006).

1.4.3 Notch signaling in other systems

Besides promoting germ cell proliferation in the *C. elegans* germ line, Notch signaling is also involved in regulating stem cells in other systems. Like in the *C. elegans* germ line, Notch signaling functions similarly in the *Drosophila* ovary to establish and maintain the GSC niche in this organ. Expanded expression of Notch signaling can induce more cap cell formation, which then triggers more GSC self-renewal. In contrast, Notch signaling mutant ovaries have a decreased number of both cap cells and GSCs (Song et al., 2007). In addition to GSCs, Notch signaling is capable of regulating other types of stem cells. Notch signaling is a key regulator in maintaining mammalian neural stem cells (NSC) (Alexson et al., 2006; Hitoshi et al., 2002). Hematopoietic stem cell (HSC) self-renewal can be increased by Notch activation (Campbell et al., 2008; Weber and Calvi, 2010). HSCs transduced with the intracellular domain of Notch1 (ICN1) continue proliferating for more than 8 months, whereas untransduced control cells maintained this state for only 25 days (Varnum-Finney et al., 2000). Notch signaling is also tightly linked with tumorigenesis (Allenspach et al., 2002; Bolos et al., 2007). In the human glioma cell line SHG-44, an active form of the Notch receptor Notch1, was detected and over-expressed Notch1 induced the formation of cancer stem cell (CSC)-like colonies (Zhang et al., 2008).

1.4.4 Switch from mitosis to meiosis

Downstream of the GLP-1/Notch signaling pathway in the *C. elegans* germ line are two mRNA regulatory pathways: the GLD-1/NOS-3 pathway and GLD-2/GLD-3 pathways (Hansen and Schedl, 2006). These pathways function redundantly to promote

meiotic entry (Figure 1.3B). GLD-1 is a STAR (Signal transduction and activation of RNA) protein that contain a conserved KH RNA binding domain (Jones and Schedl, 1995; Vernet and Artzt, 1997), and acts as a translational repressor (Jan et al., 1999). Another component in the first pathway is NOS-3 (Hansen et al., 2004b), which belongs to the Nanos RNA-binding protein family (Kraemer et al., 1999). In the second pathway, GLD-2 harbours a catalytic domain of a cytoplasmic poly(A) polymerase (PAP) (Wang et al., 2002), but lacks a RRM-like region, which is critical for PAP RNA binding (Bard et al., 2000; Martin et al., 2000). Also within this pathway is GLD-3, a member of the Bicaudal-C family of RNA binding proteins (Eckmann et al., 2002). GLD-3 can bind to GLD-2 specifically (Wang et al., 2002). GLD-2 and GLD-3 can form a heterodimer that contains both an RNA binding domain (GLD-3) and a catalytic domain (GLD-2), thereby allowing GLD-2 to be targeted to specific mRNAs by GLD-3 and polyadenylate these targets (Wang et al., 2002). Based on the above information, it is hypothesized that these two pathways promote meiotic entry by repressing mitosis-promoting genes (GLD-1/NOS-3) and stabilizes meiosis-promoting RNAs (GLD-2/GLD-3 poly(A) polymerase) (Hansen and Schedl, 2006).

1.5 Screen to identify additional factors involved in mitosis and meiosis regulation

The core regulator of the balance between proliferation and differentiation in the *C. elegans* germ line is the mitosis-promoting GLP-1/Notch signaling pathway, and the GLD-1/NOS-3 and GLD-2/GLD-3 meiosis-promoting pathways. However, this key network itself is also highly regulated by many other components, directly or indirectly, and these additional players can help build a more detailed and complete picture for

deciphering the mechanism of maintaining the proliferation and differentiation balance. Genetic screens are powerful tools for identifying genes that have certain functions. Genetic screens are often performed by mutagenizing wild-type animals and then screening for mutant animals with a specific phenotype related to that gene function. By applying this approach *lag-1*, *lag-2* and *glp-1* were identified in a screen looking for sterile mutants that are defective in the proliferation/differentiation decision (Austin and Kimble, 1987; Lambie and Kimble, 1991).

However, this traditional way of screening has limitations since certain mutations may not show any obvious traits in a wild-type background, due to redundancy issues. This limitation makes it difficult to search for more genes involved in a regulatory pathway/network. An alternative screening technique is to performe an enhancer or suppressor screen using a sensitized genetic background, instead of wild-type background. Many additional players in the pathway regulating the proliferation and differentiation decision were identified using suppressor or enhancer screens. In one screen the *glp-1(oz112oz120)* allele was used as a genetic background. This allele shows a slight gain-of-function (gf) over-proliferation phenotype in the germ line when grown at 15°C—only 0.05% of the animals have gonads with late-onset over proliferation, the rest are phenotypically wild-type (Wilson-Berry, 1998). Because of this unique feature, mutagenesis on animals with this allele can allow for the identification of other mutations that can intensify this "subtle" phenotype. An enlarged proliferative zone in the germ line can result in a germline tumor in tested animals. This enhancer screen was performed in the laboratory of Dr. Tim Schedl (Washington University), and three *teg* (tumorous

enhancement of weak *glp-1(gf)*) genes, *teg-1*, *teg-2* and *teg-4*, were identified (Wilson-Berry, 1998). This thesis focuses on *teg-4*.

1.6 Initial characterization of *teg-4*

teg-4 has been partially characterized by Pallavi Mantina (Mantina et al., 2009), and this section briefly summarizes her major findings on *teg-4*.

1.6.1 Molecular identity of teg-4

A series of mapping approaches were employed, positioning *teg-4* on *C. elegans* chromosome I. and identified *teg-4* as being allelic to *tag-203*. The *teg-4(oz210)* allele, isolated from the genetic screen described above, contains a missense mutation, in which a guanine is mutated to adenine, causing a glycine to aspartic acid substitution. *teg-4* encodes a protein homologous to human SAP130, *Saccharomyces cerevisiae Rse1p*, and *Schizosaccharomyces pombe Prp12p* splicing factors (Chen et al., 1998; Das et al., 1999; Habara et al., 2001). All three proteins are subunits of the U2 snRNP-associated complex SF3b, which is required for binding of U2 snRNP to the branch site during pre-mRNA splicing (Jurica and Moore, 2003; Kramer, 1996; Nagai et al., 2001).

1.6.2 teg-4(oz210) *enhances the* glp-1(ar202gf) *over-proliferative germline phenotype*

In order to confirm the role of *teg-4(oz210)* as a tumor enhancer, a double mutant *teg-4(oz210)*; *glp-1(ar202gf)* was made. *glp-1(ar202gf)* is a weak gf allele, which has been widely used in studying the function of many other genes in the mitotic

(proliferation) vs. meiotic (differentiation) decision (Hansen et al., 2004a; Macdonald et al., 2008; Pepper et al., 2003).

The results revealed that in *teg-4(oz210)*; *glp-1(ar202gf)* double mutants, the size of the mitotic zone was significantly increased, compared with that in *glp-1(ar202gf)* animals (Figure 1.4) (Mantina et al., 2009). Moreover, at 15°C, a large portion (57%, n=146) of the double mutants displayed a Pro tumor (ectopic proliferation at the proximal end of the gonad arms) phenotype, whereas no *teg-4(oz210)* single mutants and only 21% (n=146) *glp-1(ar202gf)* single mutants showed this phenotype (Mantina et al., 2009).

1.6.3 Characterization of teg-4(oz210) *single mutant*

teg-4(oz210) single mutants did not display any tumorous germline phenotypes. Rather, they have a relatively smaller proliferative zone, when compared with wild-type animals (Mantina et al., 2009). *teg-4(oz210)* is a partial lf allele. While L4 animals treated with *teg-4*(RNAi) generate ~99% arrested embryos or larvae, *teg-4(oz210)* homozygotes are generally viable; but still 12.9% of the population (n=1952) die during embryogenesis. Among those that do reach adulthood, 15% (n=703), 49% (n=2771) and 67% (n=762) are still sterile, when grown under 15°C, 20°C and 25°C, respectively (Mantina et al., 2009). The reason for the sterility is likely due to the production of excessive sperm and the absence of proper oogenesis (Mog: masculinization of germ line (Graham and Kimble, 1993).

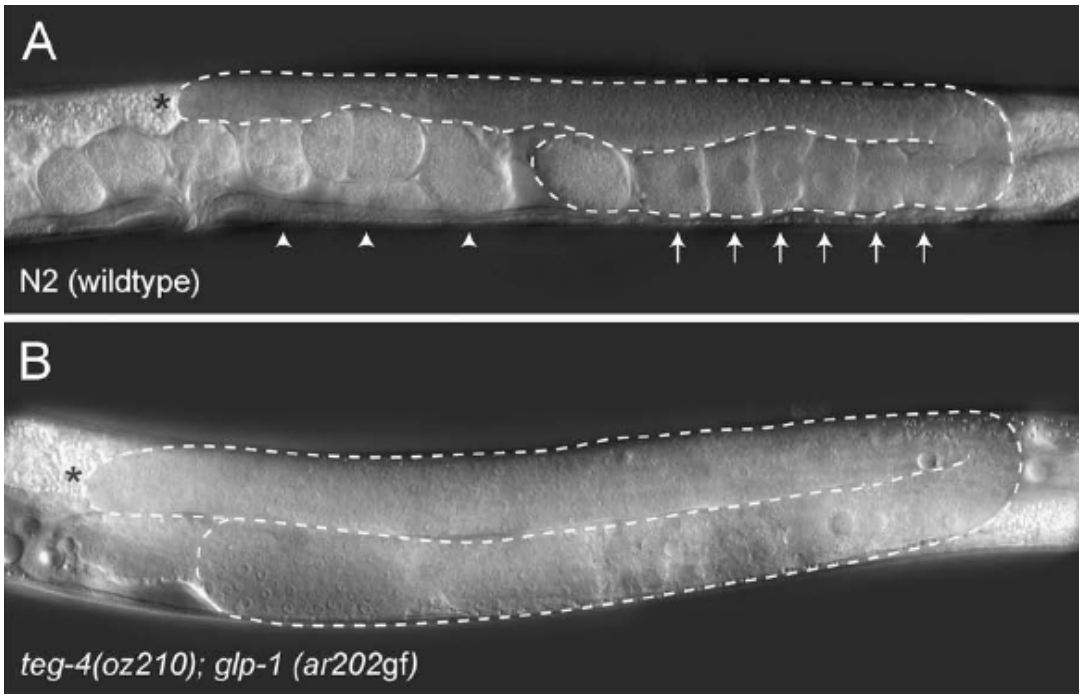


Figure 1.4. *teg-4(oz210)* causes a synthetic tumor with *glp-1(ar202gf)*.
Adult hermaphrodite animals one day past L4 were analyzed using differential interference contrast microscopy (DIC). In each image, only one gonad arm is shown, and is outlined with a dashed line. "*" indicates the distal end. (A) A wild-type (N2) gonad arm. Both oocytes (arrows) and embryos (arrowheads) were produced. (B) A *teg-4(oz210)*; *glp-1(ar202gf)* gonad arm. Gametogenesis did not occur; the germ line is full of proliferating cells. Figure adapted from Mantina *et al*. (Mantina et al., 2009).

1.7 Splicing and tumor formation

1.7.1 Pre-mRNA splicing

Splicing, the removal of non-coding sequences called introns from pre-mRNA, is an important means of RNA regulation in eukaryotes. Splicing processes are catalyzed by a mega-Dalton multicomponent RNA-protein complex called the spliceosome, which primarily consists of five spliceosomal small nuclear ribonucleoprotein complexes (U1, U2, U4, U5 and U6 snRNPs) (Brow, 2002; Jurica and Moore, 2003; Staley and Guthrie, 1998; Will and Luhrmann, 2010). As the spliceosome assembles across the intron, splicing occurs.

Pre-mRNAs contain conserved sequences that help the recognition of introns. For each intron, these conserved sequences are located at the 5' splice site, the 3' splice site and the branch site (Wachtel and Manley, 2009). After the nascent pre-mRNA is produced, the U1 snRNP and the U2 snRNP are recruited to the 5' site and the branch site, respectively, to form the pre-spliceosome (Complex A). To the pre-spliceosome is added the U4/U6.U5 tri-snRNP, forming the pre-catalytic spliceosome (Complex B). The U1 and U4 snRNPs are then released from Complex B, resulting in the activated spliceosome (Complex B*). Complex B* catalyzes the 1st step of splicing, in which the 5' site of the intron is cleaved, releasing a free exon I and an intron lariat-exon II intermediate. Concomitantly, Complex B* is converted into Complex C, which in turn catalyzes the 2nd step of splicing, in which the 3' site of the intron is cleaved and exon I and exon II are ligated (Karijolich and Yu, 2010; Wachtel and Manley, 2009; Will and Luhrmann, 2010) (Figure 1.5). There are other non-snRNP proteins participating in the

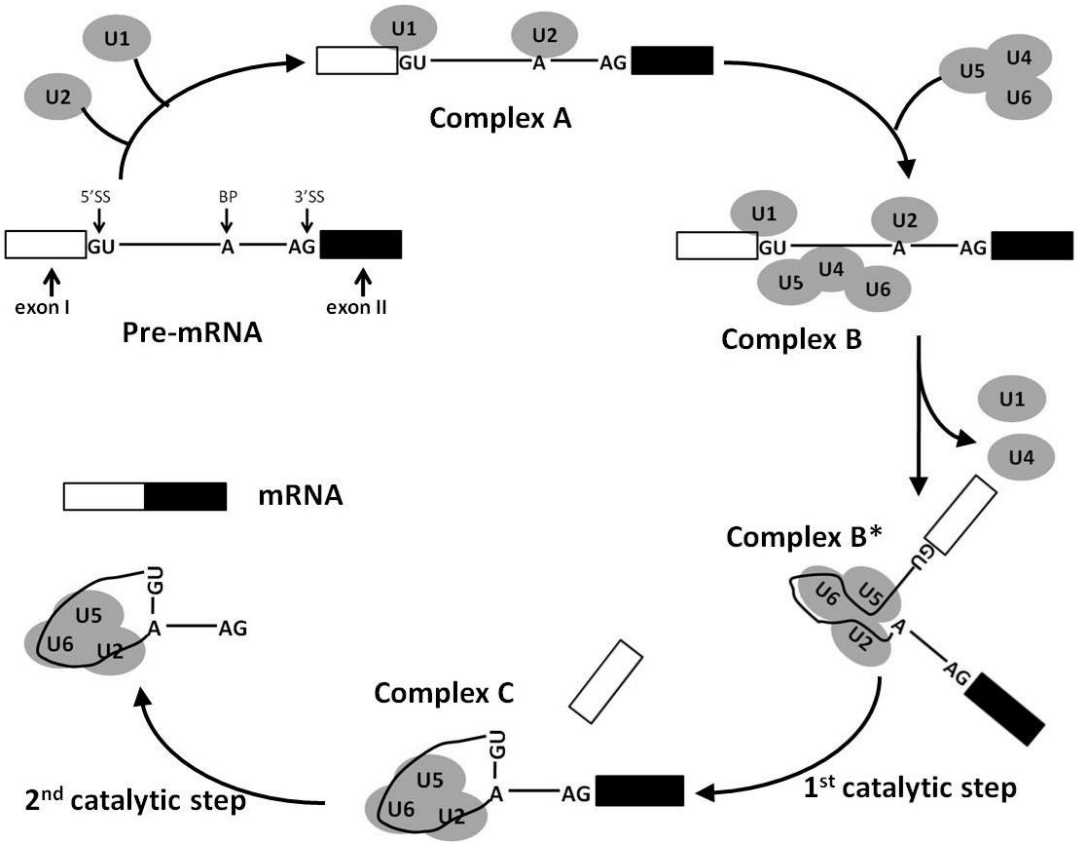


Figure 1.5. Major spliceosome assembly and pre-mRNA splicing.

Exons are represented in boxes and introns in solid lines. Ellipses represent five small nuclear ribonucleoproteins (U1, U2, U4, U5 and U6). The 5' splice site (5'SS), the 3' splice site (3'SS) and the branch point (BP) adenosine are indicated in the pre-mRNA. The conserved sequences at the 5' and 3' splice site and the branch site are shown.

assembly and activation of the spliceosome, and most of them belong to a family of DExD/H-box RNA helicases (Brow, 2002).

1.7.2 Alternative splicing

Alternative splicing, by which exons in a pre-mRNA are joined in multiple combinations, is a crucial mechanism utilized by metazoans for gene regulation. Alternative splicing greatly diversifies the transcriptome and the proteome, and it is the most plausible explanation for the paradox that the number of genes transcribed in multicellular organisms is disproportional to the phenotypic complexity they exhibit (Graveley, 2001; Modrek and Lee, 2002). The rate of alternative splicing correlates with the complexity of an organism, and humans have the highest rate of alternative splicing, known in which at least 60% of the multi-exon genes are alternatively spliced (Kim et al., 2007; Modrek and Lee, 2002).

There are four major types of alternative splicing: exon skipping, intron retention, alternative 5' splice site, and alternative 3' splice site (Ast, 2004) (Figure 1.6). Exon skipping, which accounts for at least one-third of the known alternative splicing events (Blencowe, 2006), is the most prevalent type of alternative splicing, and its prevalence increases gradually along the eukaryotic tree (Kim et al., 2008; Kim et al., 2007). Alternative 5' and 3' sites are the second most frequent types, accounting for ~25% of the known alternative splicing events (Blencowe, 2006). Intron retention is the rarest type, only accounting for 8% of the total splicing events (Ast, 2004).

Alternative splicing is functionally important; it not only increases genomic diversity, but also contributes to tissue specificity (Black, 2003; Graveley, 2001). In order

to fully elucidate the roles of alternative splicing, high-throughput approaches, which are powerful tools in the identification of genome-wide alternative splicing events, are utilized. Expressed sequenced tags (ESTs) used to be a major source of information for

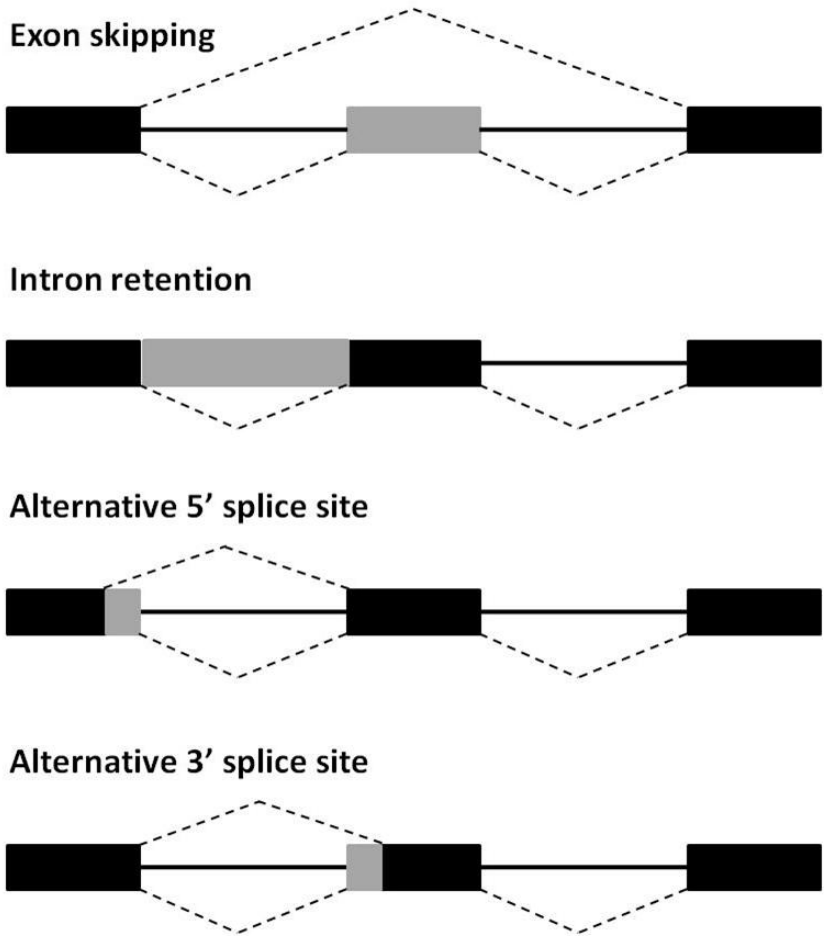


Figure 1.6. Major types of alternative splicing.
Exons are represented in boxes and introns in solid lines. In all four examples of alternative splicing, constitutive exons are shown in black and alternatively spliced regions in grey. Dashed lines indicate splicing activities.

the identification of alternative splicing events (Blencowe, 2006). Many findings on the identification of large-scale alternative splicing, such as the general alternative splicing rate in different organisms/tissues (i.e. normal tissues vs. cancerous tissues), as well as the distribution of different alternative splicing types, were discovered by employing EST analyses (Brett et al., 2002; Kim et al., 2008; Kim et al., 2007; Kim et al., 2004). However, one limitation of the EST analysis is its biased and insufficient coverage of the genome (Johnson et al., 2003; Pan et al., 2004). This limitation is overcome by two emerging technologies: custom microarrays and high-throughput sequencing. Both technologies facilitate a more thorough identification of global alternative splicing, and reveal many new alternative splicing events that were overlooked by EST analyses (Johnson et al., 2003; Pan et al., 2008a).

1.7.3 Splicing and tumorigenesis

Many diseases stem from defects in mRNA splicing, including cancer (Faustino and Cooper, 2003). Splicing can be linked with tumorigenesis in two ways. First, *cis*-acting mutations affecting the splicing process in oncogenes or tumor suppressor genes can be a direct cause of neoplasia. An increased or decreased level of a certain splicing variant in these tumor regulators often results in an enhancement (for oncogenes) or reduction (for tumor suppressors) of their original function, thus inducing tumor formation, progression or transformation. For example, a splicing mutation on the donor site on exon 5 of tumor suppressor p53 introduces a premature stop codon during translation, resulting in truncated p53 protein production (Schneider-Stock et al., 1997). mRNA of Wilms' tumor suppressor protein (WT1) has four isoforms, and they determine the protein location,

stability and binding affinity (Laity et al., 2000), thus affecting protein function. Oncogene KIT is a type III receptor tyrosine kinase, and Chen et al found that a deletion caused a new pre-mRNA 3' splice site (Chen et al., 2005b). This resulted in the production of aberrantly spliced KIT, which lacks a portion critical for autoinhibition, causing KIT to be constitutively accumulated, and eventually leads to the formation of a gastrointestinal stromal tumor.

Nevertheless, most cancer-related splicing alterations are not due to the nucleotide changes of the affected gene, but are the result of changes in trans-acting splicing factors (Srebrow and Kornblihtt, 2006; Wang and Cooper, 2007). In many cancer cells, splicing factors are either up-regulated or down-regulated (Grosso et al., 2008), and have profound significance. Slight up-regulation of the splicing factor SF2/ASF caused successful transformation of immortal rodent fibroblasts (Karni et al., 2007), and H37 protein (RNA binding protein, may participate in splicing process) showed tumor suppression properties by functioning in the cell cycle and apoptosis pathways (Oh et al., 2006).

Recently, several splicing factors have been found to be targets of antitumor drugs. Human splicing factor 3b (SF3b) is the target of several antitumor drugs: spliceostatin A, the methylated derivative of anticancer compound FR901464, specifically binds to SF3b subunits SAP155, SAP145 and SAP130 (Corrionero et al., 2011; Kaida et al., 2007); SAP130 is also the binding protein of the antitumor product pladienolide, and its derivative E7107 (Folco et al., 2011; Kotake et al., 2007); and SAP155 is the target of GEX1A, another antitumor natural product (Hasegawa et al., 2011). These drugs have been shown to inhibit the splicing of several genes by specifically interfering with SF3b.

1.8 Hypothesis and goals

According to the previous study by Mantina *et al*. (Mantina et al., 2009), *teg-4* is involved in regulating the proliferation/differentiation decision in the *C. elegans* germ line. However, how *teg-4* fits into the GLP-1/Notch signalling regulatory network is still unclear. Since *teg-4* is homologous to the splicing factor SAP130, a logical hypothesis is that *teg-4* is somehow involved in the the splicing process: The defects in the germline proliferation/differentiation balance in *teg-4* loss-of-function mutants are probably due to splicing abnormalities on downstream targets. Searching for *teg-4* targets is the primary goal of this thesis. Specifically, this project includes the following aims:

(1) Using tiling array analysis to identify the targets of *teg-4* (chapter 4);

(2) Using RNAi and genetic approaches to further investigate the functions of these targets (chapter 5).

Chapter Two: Materials and methods

2.1 General methods

Worms were maintained on nematode growth medium (NGM) plates seeded with *E. coli* OP50 (Brenner, 1974), and experiments were conducted at 20°C unless otherwise noted. All strains were derivatives of wild-type Brisol strain (N2).

2.2 Mutant strains construction

2.2.1 Generating teg-4 smg-2 *double mutants*

The *teg-4* and *smg-2* genes are both located on chromosome I, 18.28 cM away from each other. To make *teg-4 smg-2* double mutants, +/hT2g males were crossed with *teg-4(oz210)* L4 hermaphrodites. Green male progeny (*teg-4(oz210)*/hT2g) were then crossed with *smg-2(e2008)* L4 hermaphrodites and non-green male progeny (*teg-4(oz210)/smg-2(e2008)*) were selected to cross with +/hT2g males. The green hermaphrodite progeny from this cross consist of: 1) *smg-2(e2008)*/hT2g, 2) *teg-4(oz210)*/hT2g, 3) +/hT2g and 4) *teg-4(oz210) smg-2(e2008)*/hT2g.

Progeny were then put onto individual plates to allow them to self-fertilize (P generation), and their non-green progeny (F1 generation) were screened for recombinants (*teg-4(oz210) smg-2(e2008)*). Animals were primarily tested for the *smg-2* mutation, and this was done by running a reverse-transcriptase PCR on gene *rpl-12* (primers "rpl-12fwd" and "rpl-12rev" in Table D.5). In *smg-2* mutants, there is an alternative splice variant of *rpl-12* that does not get degraded as in wild-type animals and two bands will be seen instead of one. Once the presence of *smg-2(e2008)* was confirmed, animals were

then sequenced for the *teg-4(oz210)* mutation (primer "teg-4fwd" and "teg-4rev" in Table D.5), and *teg-4(oz210) smg-2(e2008)* double mutants were generated (Figure 2.1).

2.2.2 Generating teg-4; glp-1; C38D9.2 *triple mutants*

The *teg-4*, *glp-1* and C38D9.2 genes are located on chromosome I, chromosome III and chromosome V, respectively. The *teg-4(oz210)*; *glp-1(ar202gf)* strain was already constructed and balanced with hT2g: *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g. The allele *ok1853* contains a deletion in C38D9.2, and the strain *ok1853/ok1853* was obtained from CGC (www.cbs.umn.edu/CGC).

To make the *teg-4(oz210)*; *glp-1(ar202gf)*; C38D9.2*(ok1853)* triple mutant, +/hT2g males were first crossed with *ok1853/ok1853* L4 hermaphrodites. Green L4 male progeny (+/hT2g; +/hT2g; *ok1853*/+) were then crossed with L4 *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g hermaphrodites. Green L4 hermaphrodite progeny were picked onto individual plates to allow them to self-fertilize. The possible genotypes for these animals are: 1) *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g; *ok1853*/+, 2) *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g; +/+, 3) +/hT2g; +/hT2g; *ok1853*/+ and 4) +/hT2g; +/hT2g; +/+.

Screens were performed to identify only genotype 1) through a two-step procedure. In the first step, all plates were subject to PCR using primers named "ok1853fwd" and "ok1853rev" (Table D.5) to screen for the *ok1853* deletion, and genotype 2) and 4) were thereby excluded. In the second step the remaining plates were screened by looking for non-green tumorous animals, which would contain *teg-4(oz210)* and *glp-1(ar202gf)*, as a result, only plates that have genotype 1) remained.

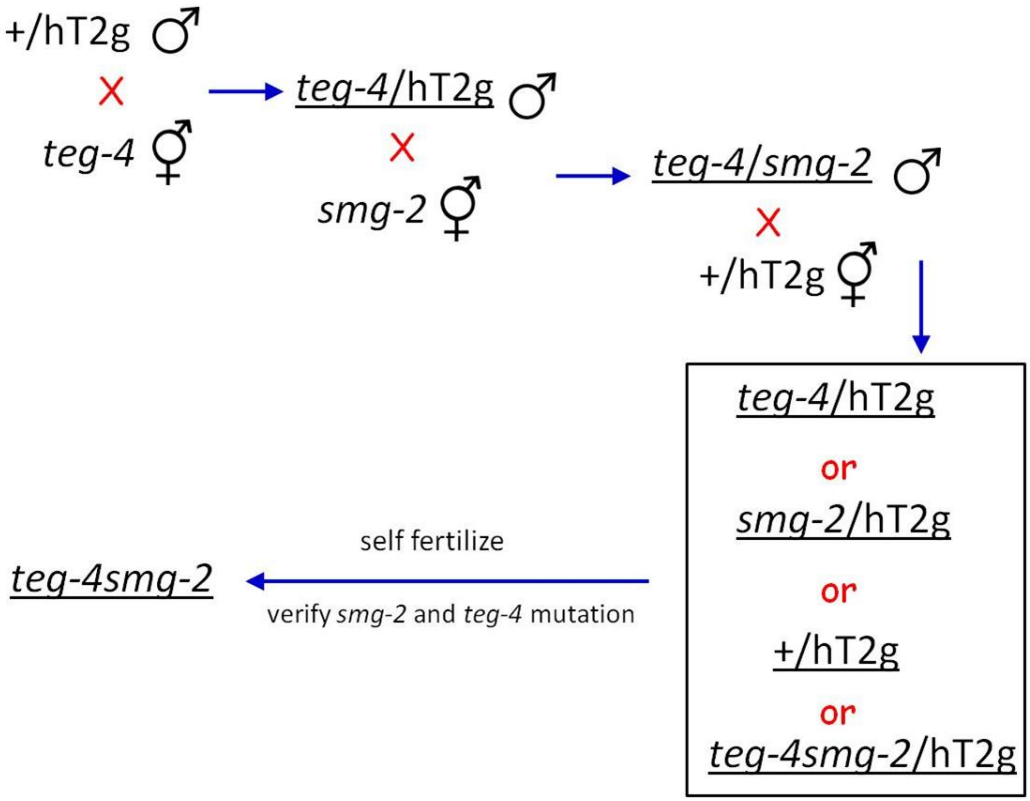


Figure 2.1. The construction of the *teg-4 smg-2* double mutant.
Hermaphrodite *teg-4* homozygotes were crossed with male +/hT2g animals, and green (hT2g has GFP expression in the pharynx) male progeny were all *teg-4*/hT2g, which were crossed with hermaphrodite *smg-2* homozygotes to generate non-green male progeny; *teg-4*/*smg-2*. Male *teg-4/smg-2* animals were crossed with hermaphrodite +/hT2g animals, and their hermaphrodite green progeny had four possible genotypes (boxed part), in which +/hT2g and *teg-4 smg-2*/hT2g are recombinants. Animals were put onto separate plates to allow them to self fertilize, and their non-green progeny were screened for both the *smg-2* and the *teg-4* mutation via PCR. Animals with both mutations are *teg-4 smg-2*.

Animals from *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g; *ok1853*/+ were picked individually onto different plates, and their progeny were screened for *ok1853*

homozygotes, which has the genotype *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g; *ok1853/ok1853*, the desired strain (Figure 2.2).

2.3 Whole worm lysis for PCR

Genomic DNA was isolated from worms through worm lysis (Hope, 1999). Standard procedures were as follows: 1-2 animals were added into 2.5 µL of worm lysis buffer (50 mM KCL, 10 mM Tris-HCl pH 8.2, 25 mM $MgCl_2$, 0.45% NP-40, 0.45% Tween-20, 0.01% Gelatin, 3 µL of proteinase K (20µg/µL)/mL) and freeze-cracked at -80°C for at least 10 min; each sample was then heated at 65°C for 60 min, followed by 95°C for 15 min. At the end of worm lysis, genomic DNA would be released into the solution, and are ready to undergo PCR.

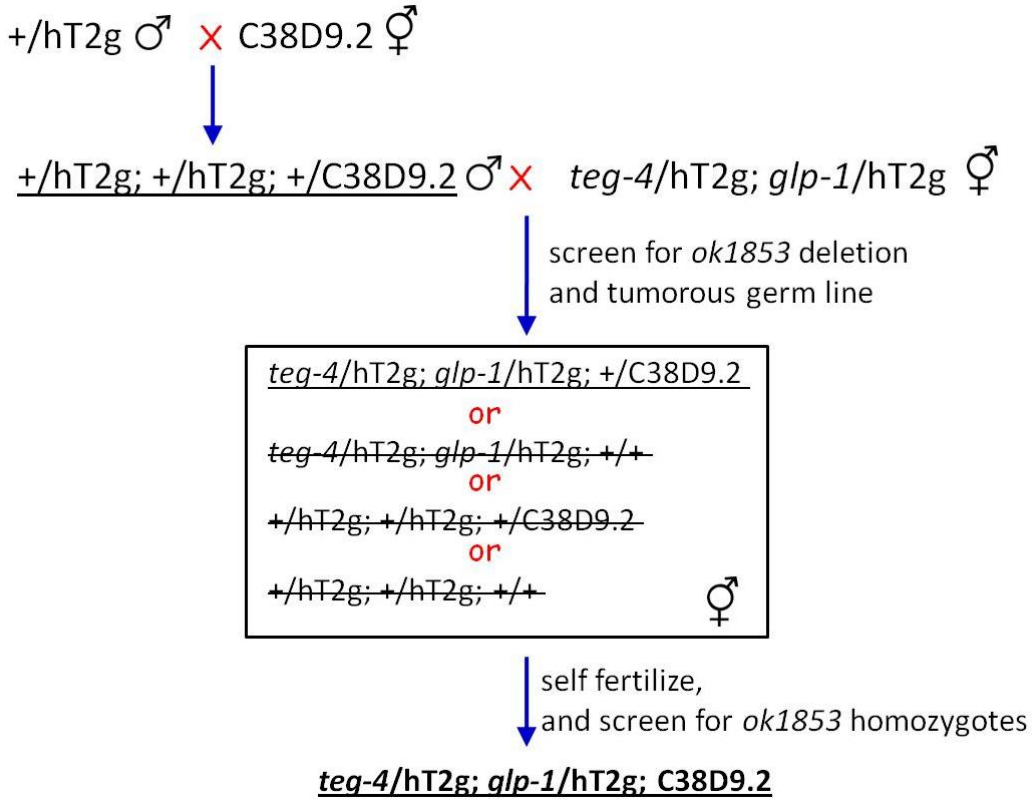
2.4 Tiling array experiments

2.4.1 Chip information

The GeneChip® *C. elegans* Tiling 1.0R Array from Affymetrix was used for this project. It is a single array that contains over 3.2 million perfect match/mismatch probes tiled through the entire non-repetitive *C. elegans* genome. Sequences for designing this array were selected from the WormBase web site, www.wormbase.org, release WS140, March 26, 2005. 25-mer probes were synthesized complementary on each corresponding sequence, and were tiled for an average resolution of 25 base pairs.

Figure 2.2. The construction of the *teg-4(oz210); glp-1(ar202gf);* C38D9.2*(ok1853)* triple mutant.

Hermaphrodite *ok1853* homozygotes were crossed with male +/hT2g, and green (hT2g has GFP expression in the pharynx) male progeny were all +/hT2g; +/hT2g*; +/ok1853* animals, which were crossed with hermaphrodite *teg-4/*hT2g*; ar202/*hT2g animals. The hermaphrodite green progeny have four possible genotypes (boxed part), and animals were put onto separate plates to allow them to self fertilize overnight. All of the parental animals on these plates were screened for the *ok1853* deletion through PCR. Plates with positive results (possible genotypes are *teg-4*/hT2g; *ar202*/hT2g; +/*ok1853* and +/hT2g; +/hT2g; +/*ok1853*) were kept, and their progeny were screened for the *teg-4(oz210)* and *glp-1(ar202gf)* mutation by looking for tumorous germline phenotypes. Animals that have this phenotype were the progeny of *teg-4(oz210)*/hT2g; *glp-1(ar202gf)*/hT2g; +/C38D9.2(*ok1853)* animals. Green progeny were put onto separate plates to allow them to self fertilize overnight and the parental animals were screened for *ok1853* homozygotes. Animals with the positive result are the *teg-4(oz210)/*hT2g*; glp-1(ar202gf)/*hT2g*;* C38D9.2*(ok1853)/* C38D9.2*(ok1853)*.


+/hT2g ♂ X C38D9.2 ♀
+/hT2g; +/hT2g; +/C38D9.2 ♂ X teg-4/hT2g; glp-1/hT2g ♀
screen for ok1853 deletion
and tumorous germ line
teg-4/hT2g; glp-1/hT2g; +/C38D9.2
or
teg-4/hT2g; glp-1/hT2g; +/+
or
+/hT2g; +/hT2g; +/C38D9.2
or
+/hT2g; +/hT2g; +/+
♀
self fertilize,
and screen for ok1853 homozygotes
teg-4/hT2g; glp-1/hT2g; C38D9.2

2.4.2 RNA isolation and purification

Four strains were used in this experiment: N2, *teg-4 (oz210)*, *smg-2(e2008)* and *teg-4(oz210) smg-2(e2008)*. RNA was extracted and purified from synchronized L4 worms. RNA isolation was performed using a Trizol based method (Appendix E); and isolated RNA samples were then treated with DNase to eliminate genomic DNA (Appendix E).

2.4.3 RNA quantity and quality determination

RNA quantity and quality were determined spectrophotometrically by using a NanoDrop (Thermo Scientific, DE). The ratio of sample absorbance at 260 nm and 280 nm (A260/A280) was used to assess the purity of RNA, and a ratio of ~2.0 is accepted as "pure" RNA.

A 0.7% electrophoresis agarose gel was also used to examine the RNA quality. RNA samples of good integrity will only display distinct 28S and 18S ribosomal RNA (rRNA) bands, and rRNA 28S/18S ratio should be around 2. A lack of these two ribosomal RNA bands, with only smear bands, indicates RNA has degraded. If additional bands were present, RNA samples were likely to be contaminated with genomic DNA.

RNA samples for the Tiling array experiments were required to be of both high quality and sufficient quantity. RNA needs to be intact and free of genomic DNA; and each experiment requires a total amount of at least 6 μg of RNA with a minimal concentration of 1μg/μl. RNA samples that failed to meet these standards were discarded and new samples were prepared.

2.4.4 Preparation of the RNA samples for the Tiling array experiment

Four strains were used for performing the Tiling array experiments; N2 (wild-type), *teg-4(oz210)*, *smg-2(e2008)*, and *teg-4(oz210) smg-2(e2008)*. The N2 and *teg-4(oz210)* strains were in the same experimental set, and the N2 strain was the control. The *smg-2(e2008)* and *teg-4(oz210) smg-2(e2008)* strains were in the other experiment set, and the *smg-2(e2008)* strain was the control. The gene *smg-2* is one of the components of the NMD (nonsense-mediated decay) pathway. The NMD pathway can recognize transcripts that contain premature termination codons (PTCs) and targets these transcripts for degradation (Isken and Maquat, 2008). In the *smg-2* mutant, the NMD pathway is disrupted. As a result, if splicing problems occurred in the *teg-4* mutant animals, the mis-spliced transcripts will continue to exist in the cell and can be detected, without being eliminated by the NMD pathway. Therefore, in order to avoid the influence of the NMD pathway on mis-spliced transcripts, strains *smg-2(e2008)*, and *teg-4(oz210) smg-2(e2008)* were used in the second Tiling array experiment.

RNA was extracted and purified from late L4 stage animals of each of the four strains described above. For the *smg-2(e2008)* and *teg-4(oz210) smg-2(e2008)* animals, three independent biological replicates of RNA samples were prepared. However, for the N2 and *teg-4(oz210)* animals, in the initial Tiling array experiment, only three technical replicates were used. Therefore, two more RNA samples (biological replicates) were isolated for each of them. Finally, when performing the second Tiling array experiment for the N2 and *teg-4(oz210)* animals, a total of five replicates of RNA samples were used (results obtained from this experiment were used for data analysis).

All RNA samples were provided to the Southern Alberta Microarray Facility (SAMF), where procedures regarding target preparation, target hybridization, fluidics station setup and probe array scanning were performed, by following the *GeneChip® Whole Transcript (WT) Double-Stranded Target Assay Manual (Affymetrix)*.

2.5 cDNA synthesis

cDNA was synthesized from purified RNA samples, using Roche Transcriptor Reverse Transcriptase (cat# 035312950021), following the protocol modified from the product manual (Appendix E). cDNA generated by using this method was only for the real-time qPCR assay.

2.6 Real time qPCR assay

2.6.1 qPCR experiment

A Bio-Rad iCycler thermal cycler and iQ™5 Optical System were used for performing real time quantitative PCR (qPCR), following the manufacturer's instruction (Bio-Rad). The experiment was a singleplex assay, using the DNA-binding dye SYBR Green I. Detailed procedures for sample preparation can be found in Appendix E.

Primary results were processed in iQ5 software. Primers were designed using a web-based program Primer3 (http://frodo.wi.mit.edu/primer3/), and an amplicon size of 75-200 bp was used to ensure a high amplification efficiency.

All designed primers were examined for their quality. Primers that were accepted for qPCR assays have to: 1) form no dimers and bind to the targets specifically; 2) have an efficiency of 90%-110%. Dimer formation can be detected by running a negative

control, in which no template was added. If no significant amplification was noticed after 35 cycles (C_q >35), it can be assumed that that primer dimers were not formed. Melt curves were used to test the primer specificities; single peak in a melt curve suggested primers bound to the target specifically, and multiple peaks indicated non-specific bindings. Primer efficiency was determined by running serial dilutions of a template. A standard curve generated by the results was used for calculating the primer efficiency. All primers used for qPCR assays are listed in Table D.3.

2.6.2 qPCR data analysis

Relative quantification was applied, and a reference gene was used as a normalizer. After C_q values were measured by iQ5 software, the Pfaffl method (Pfaffl, 2001) was used to determine the expression level of the target gene in the experimental sample relative to the control sample.

2.7 RNAi

RNAi knockdown of certain genes was performed by feeding worms with RNAi feeding vectors, obtained from the Ahringer RNAi bacterial library (Kamath and Ahringer, 2003). Specifically, worms were maintained on RNAi plates (see details in Appendix E) for three generations, and germline phenotypes for worms on each RNAi plate were examined using a dissection scope.

Chapter Three: Computational analysis: design

3.1 General description of computational analysis

3.1.1 Nomenclature

Raw data: original data obtained directly from SAMF.

Dataset: Data (raw or manipulated) from all replicates (arrays) in one sample. Four datasets were considered in the data analysis: *teg-4* dataset (includes 5 replicates), N2 dataset (includes 5 replicates), *teg-4 smg-2* dataset (includes 3 replicates) and *smg-2* dataset (includes 3 replicates).

3.1.2 Major principle of data analysis

The main purpose of the computational analysis aims to determine whether or not there are any splicing defects in the *teg-4* mutants, as well as whether or not some genes are expressed at different levels in *teg-4* mutants as compared to wild-type animals. In order to fulfill these two aims, it is necessary to know the expression levels of any desired region; then, by comparing these levels in different samples, both splicing defects in a certain gene and genes that are differently expressed can be identified. This is the fundamental principle of the entire computational analysis, as illustrated in Figure 3.1.

However, the original datasets only contain the information of the signal intensity of every single probe, as well as its position. Thus, individual probe signals were first transformed to “region signals”, after which statistical comparisons can be further performed.

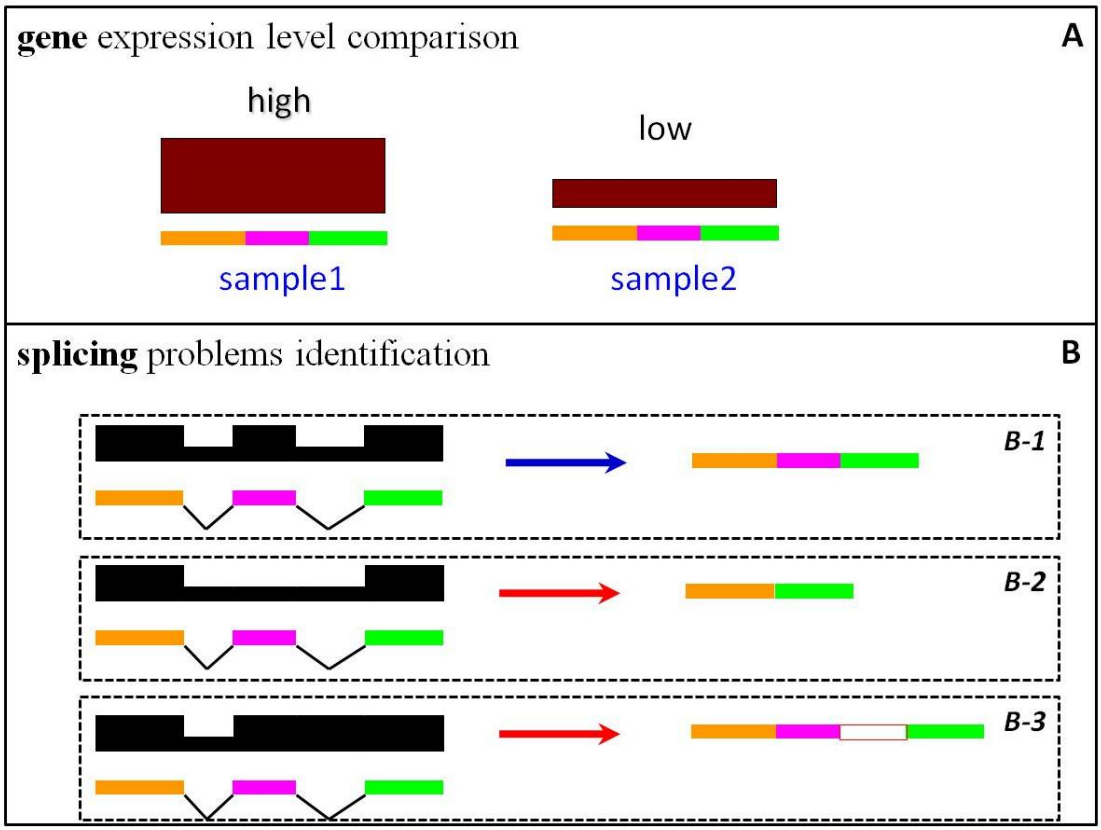


Figure 3.1. The core strategy of the computational analysis.

Signal intensities of whole genes can be used to identify genes with different expression levels (A). Signal intensities of exons and introns can be used for the identification of different manners of alternative splicing (B). If signals are high at exonic regions and low at intronic regions, there is no splicing defect (B-1). Unusually low signal at an exonic region indicates mis-splicing of the corresponding exon (B-2), and unusually high signal at an intronic region indicates retention of the corresponding intron (B-3).

Typically, three major procedures were included in this process: generation of required files (section 3.3, 3.4 and 3.5) — data pre-processing (section 3.5) — and statistical analysis (section 3.6). These tasks were accomplished by applying the R programming language, in which self-written R codes and compiled R packages (all R packages were downloaded from Bioconductor http://www.bioconductor.org/ unless otherwise noted) were used. Each of the above steps will be illustrated in detail in the following sections.

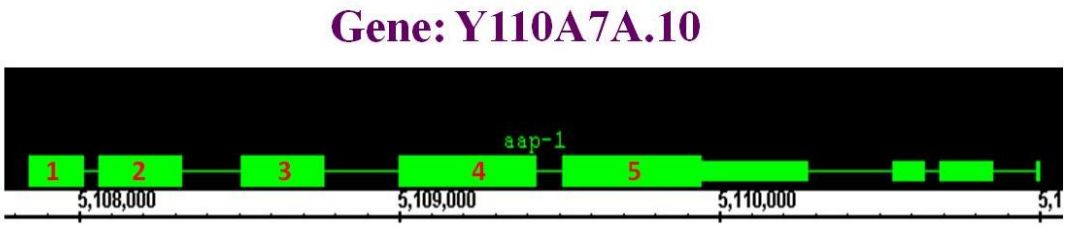
3.2 Quality assessment

Before any meaningful data manipulation was performed, an initial evaluation of the quality of the original data was necessary. When the reliability of the raw data was validated, further analysis could be carried out. Quality assessment include flaw-detection of original hybridization signal and detection of RNA degradation, R packages "affy" "affyPLM" and "simpleaffy" were employed in this process.

3.3 Annotation files generation

An annotation file is a table that defines important features of specific sequences; it includes the names, as well as the locations of the sequences. Figure 3.2 illustrates an example of a stereotypic annotation file.

The original annotation file was obtained from WormBase (WS170), and after a series of modifications, the original file was then transformed/subdivided into four different types of annotation files: 1) gene annotation file, 2) exon annotation file, 3)



sequence name	gene	chromosome	strand	start position	end position
Y110A7A.10.exon1	Y110A7A.10	I	1	5107847	5108019
Y110A7A.10.exon2	Y110A7A.10	I	1	5108067	5108326
Y110A7A.10.exon3	Y110A7A.10	I	1	5108508	5108773
Y110A7A.10.exon4	Y110A7A.10	I	1	5109001	5109432
Y110A7A.10.exon5	Y110A7A.10	I	1	5109513	5109950

Figure 3.2. An example of annotation file.

The top panel showed the structure of the gene Y110A7A.10, which includes five coding exons. The table below represents the format used for annotating this gene. Specifically, each coding exon within this gene was given a sequence name (1st column), a gene name (2nd column) and its position on the genome (chromosome [3rd column]: this gene is on chromosome I; strand [4th column]: "1" indicates the top strand and "-1" indicates the bottom strand; start position and end position [5th and 6th column]: their locations in base pairs.

intron annotation file and 4) exon/intron boundary annotation file. Each of them will be sequentially described in more detail.

3.3.1 Gene annotation file

Since only coding genes were of interest in this analysis, all pseudogenes were removed from the original file. Then, all coding genes were divided into two groups; genes without splicing variants, and genes with splicing variants. These two situations were treated differently (Figure 3.3A).

For every gene without splicing variants, the name and start/end positions of each coding exon within this gene were recorded (Figure 3.3B, left). For genes with splicing variants, only the exon regions that are common in all variants were selected (Figure 3.3B, right), and positions of these regions were recorded, in a format that is similar to that of the genes without splicing variants.

In the end, genes with and without splicing variants were combined to form the ultimate gene annotation files.

3.3.2 Exon annotation file

In creating exon annotation files, all genes were divided into two groups, genes with and genes without splicing variants (Figure 3.4A). For genes without splicing variants, all the coding exons were selected and their positions were recorded (Figure 3.4B). For genes with splicing variants, all coding regions within one gene were retained and renamed in an unrepeated fashion, as demonstrated in Figure 3.4C. For the sake of analysis simplicity, these two files were not put together, but were analysed separately.

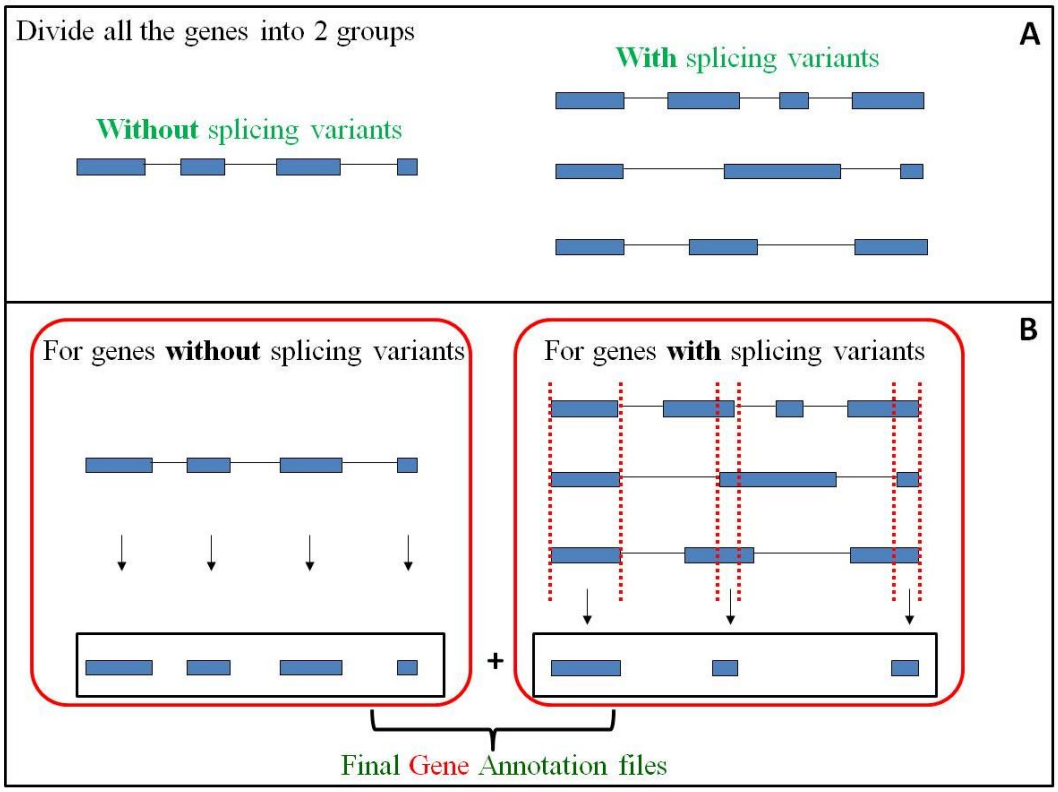


Figure 3.3. Making the Gene Annotation file

All coding genes were divided into two groups, 1) genes without splicing variants, and 2) genes with splicing variants (A). For genes without splicing variants, the way of annotating them is to record the information (see figure 3.2) of each coding exon (B, left box). For genes with splicing variants, the way of annotating them is to record the information (see figure 3.2) from coding regions that are only present in all variants (B, right box). Annotation information from both groups were put together to form the Final Gene Annotation file (B).

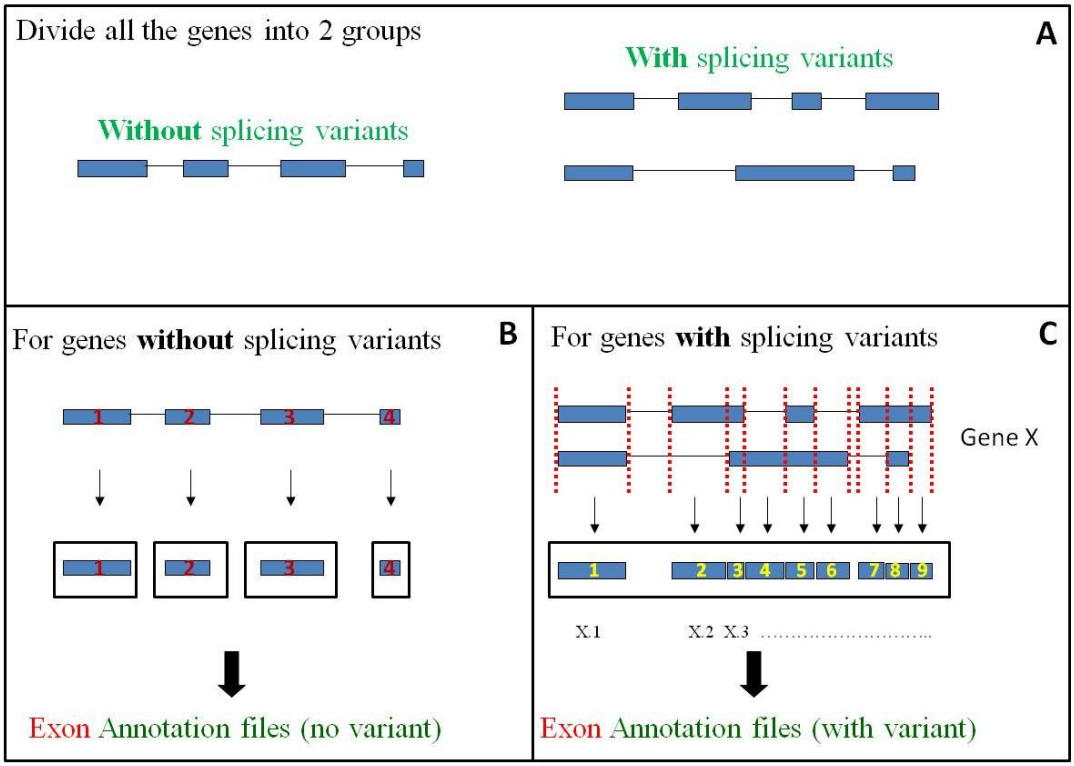


Figure 3.4. Making the Exon Annotation File.

All coding genes were divided into two groups, 1) genes without splicing variants, and 2) genes with splicing variants (A). For genes without splicing variants, the way of annotating the coding exons within them is to record the original information (see figure 3.2) for each of them, and this generates the Exon Annotation File (no variant) (B). For genes with splicing variants, all coding regions on this gene were kept, and renamed, in an unrepeated fashion. The information (see figure 3.2) from these newly named coding regions were recorded to generate the Exon Annotation File (with variants) (C).

3.3.3 Intron annotation file

In making intron annotation files, first, all genes were put into two groups, based on whether they have splicing variants (Figure 3.5A). For genes without splicing variants, the positions of all introns were recorded (Figure 3.5B). For genes with splicing variants, only these "pure intronic" regions were retained, as shown in (Figure 3.5C). These two situations were saved individually and were not analysed together.

3.3.4 Exon/intron boundary annotation file

The Exon/intron boundary defines the range between the start/end position of an exon -25 bp and start/end position of an exon +25 bp, as illustrated in (Figure 3.6A); and an exon/intron boundary annotation file records this information (Figure 3.6B).

3.4 Probe mapping

The original probe file was provided by Affymetrix, and contains the sequence of each 25 bp probe. First, a Python program entitled xMAN (extreme MApping of OligoNucleotide) (Li et al., 2008) was utilized to map these probes against the *C. elegans* genome (WS170) to their positions on each chromosomes, creating the primitive probe mapping file. Due to the fact that some of the probes have sequences that exist in multiple places in the entire genome, the signals from these probes are not trustworthy. Therefore, these probes were removed from the original file, leaving only the probes with unique sequences, and only the signals from these probes were considered.

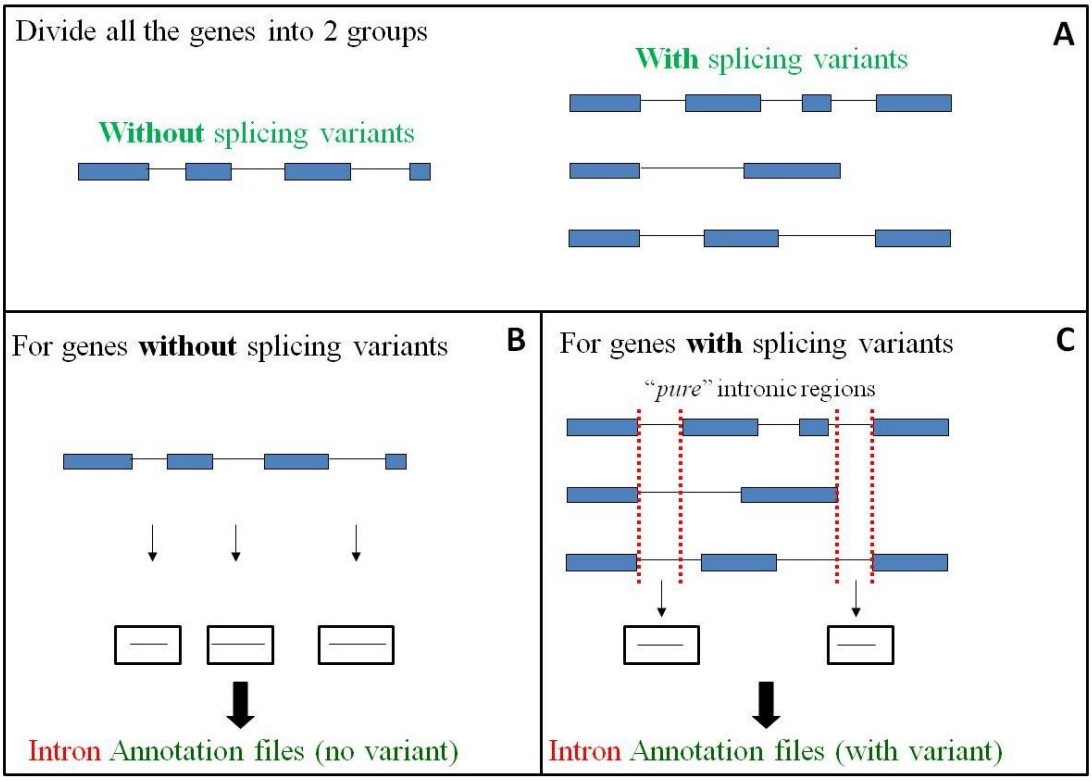


Figure 3.5. Making the Intron Annotation File.

All genes were divided into two groups, 1) genes without splicing variants, and 2) genes with splicing variants (A). For genes without splicing variants, the way of annotating the introns within them is to record the original information (see figure 3.2) for each of them, and this generates the Intron Annotation File (no variant) (B). For genes with splicing variants, only regions ("pure intronic regions") with absolutely no coding parts were kept, and information (see figure 3.2) from them were recorded to generate the Intron Annotation Files (with variants) (C).

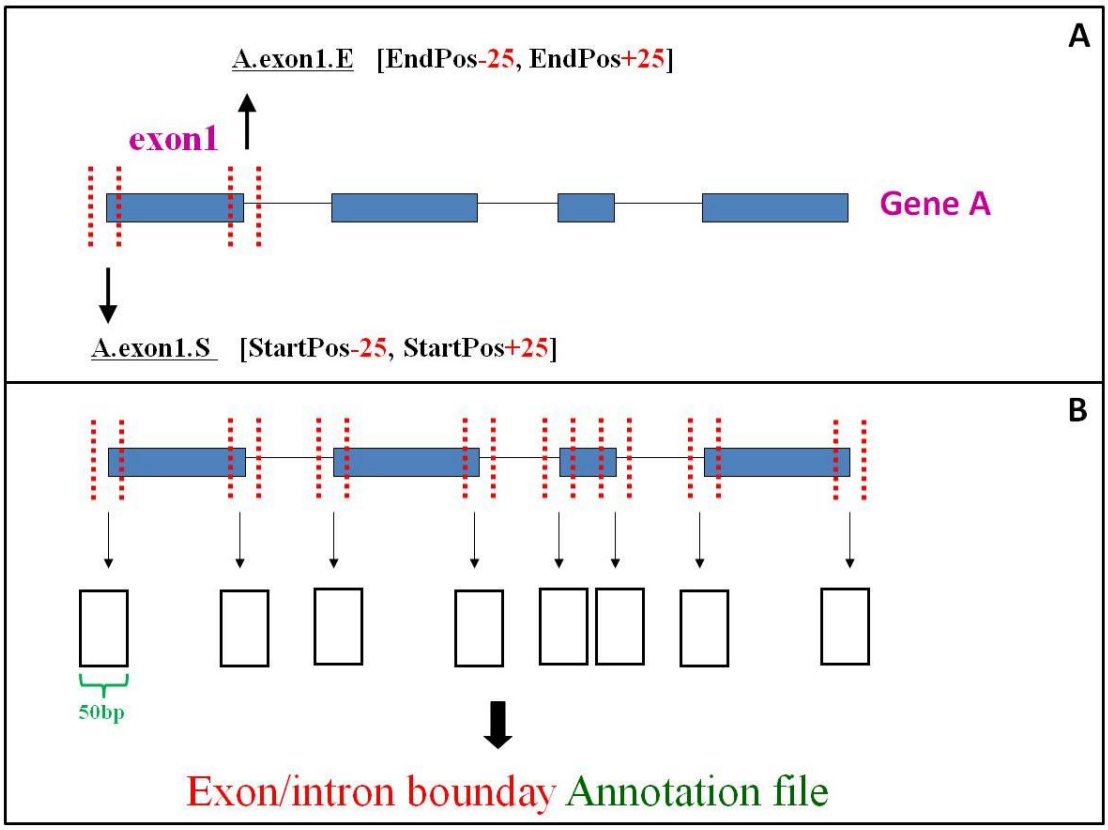


Figure 3.6. Making the Boundary Annotation File.

A boundary is defined as the range between the -25bp and +25bp at an exon/intron junction (A). Each boundary was given a name with this format: gene name. exon name [within which located the boundary]. S (or E) [S indicates this boundary is at the start of this exon, and E indicates this boundary is at the end of this exon] (A). By recording the information (see figure 3.2) of these boundaries, the Boundary Annotation File was generated (C).

3.5 Probe grouping

3.5.1 Probeset and ROI

As illustrated in section 3.1, the core strategy of this analysis is to properly convert individual probe signals into the expression pattern of a certain region. In doing so, instead of considering probe signals individually, signals from several probes need to be dealt with collectively. A group of probes is called a "probeset" and the range in which this probeset is contained is called an ROI (Range Of Interest). An ROI can be defined as any region in the genome; it can be a gene, an exon, an intron or even part of an exon. Figure 3.7 depicts the definition of probeset and ROI.

3.5.2 Grouping probes

So far, two types of files were available; annotation files and probe mapping files. Therefore, by cross referring the location information from these two files, every probe can be assigned to a specific range, and the probe grouping files were eventually generated. Since there are four kinds of annotation files (section 3.3), correspondingly, four types of probe grouping files were created: 1) gene probe file, 2) exon probe file, 3) intron probe file, 4) exon/intron boundary probe file. Each probe file defines different types of ROIs. As examples, in the gene probe file, an ROI corresponds to a gene, while in the exon probe file, an ROI corresponds to an exon. Figure 3.8 illustrates this process,

3.6 Data pre-processing

With the probe grouping files being generated, raw probe signals were ready to be converted into expression level of corresponding ROIs, i.e. gene, exon, intron, or

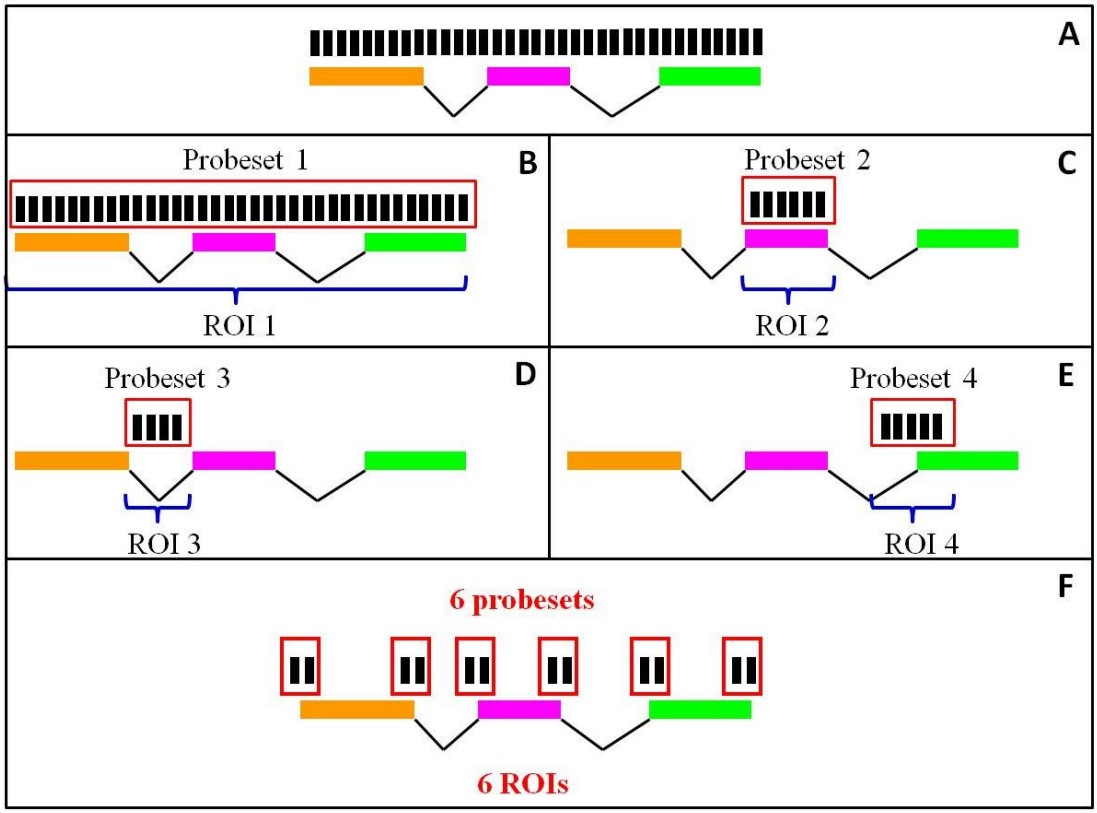


Figure 3.7. Definitions of "probeset" and "ROI (Region of Interest)".
Gene structures and probes (black bars) were first related to each other (A). Various types of probesets were utilized (B) all probes on this gene is a probeset, and this gene (including both exonic and intronic parts) is an ROI. (C) probes on the second exon is a probeset and this exon is an ROI; in (D), probes on the first intron is a probeset and this intron is an ROI; in (E), probes located on the junction of the last intron and exon is a probeset and this region is an ROI. If probeset is defined as the probes within the boundary region, this gene harbors six ROIs and all the probes within form six probesets (F).

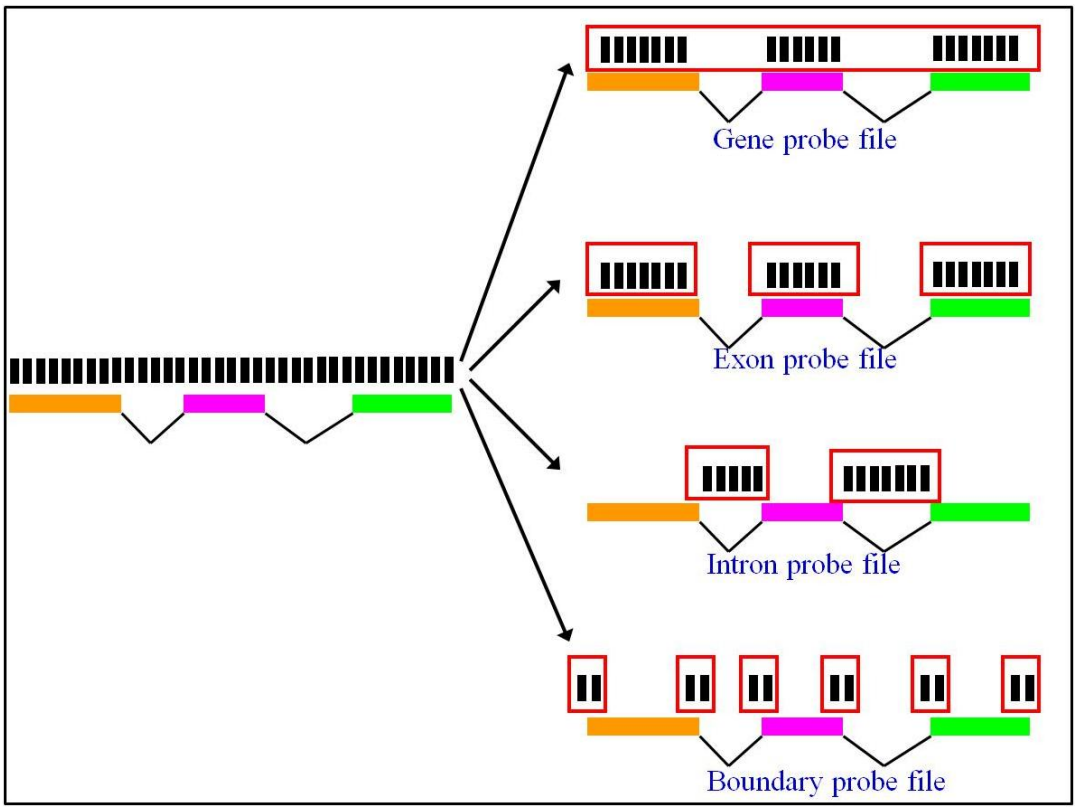


Figure 3.8. Making the probe grouping file.

Based on probe locations, probes (black bars) were eventually allocated into different groups.

boundary range. This process is called “data pre-processing”, in which three data manipulations were performed sequentially; background correction, quantile normalization and summarization.

Usually, the measured probe signals include signals from non-specific hybridization and the noise from the optical detection system (Huber et al., 2005). Therefore, in order to ensure the specificity of probe intensities, a background correction is necessary. Normalization aims at minimizing the sources of variations so that different measurements can be comparable to each other. Quantile normalization is often used when handling microarray data because of its fast speed and great performance in reducing the variances from different arrays (Bolstad et al., 2003). After quantile normalization, each array within the same comparing group will have the same distribution of intensities (Bolstad et al., 2003). Summarization is the last step of data pre-processing. It can transform individual probe signal intensities to the signals of the gene/exon/intron/boundary.

In R programming, a package named “gcrma” (see detailed supplemental information for using this package in Appendix B) was used to perform all three steps in the data pre-processing. After data pre-processing, the expression levels of different regions were created, and statistical comparisons between different datasets were performed.

3.7 Statistical analysis

Statistical analyses were performed to identify genes/exons/introns/boundaries that show different expression levels in *teg-4* mutants as compared with wild-type animals.

An R package named “limma” was employed for doing so. The final files generated by “limma” contain the expression information of all genes/exons/introns/boundaries in any comparing pair, and then different FDR (False Discovery Rate) values were applied to select regions that are considered to be differently expressed in comparing datasets.

The expression level of an exon contributes to the expression level of its corresponding gene. Therefore, for exons that were identified with different expression levels, it is possible that their corresponding genes were identified as well. The purpose of comparing exon expressions is to identify splicing problems (exon skipping). If the absolute expression levels of exons were used for making comparisons, many exons identified in this way are due to the differences in their gene expression levels, instead of splicing defects. Therefore, in order to avoid this situation, the relative expression levels of exons were calculated (exon expression/gene expression). The statistical analysis was performed on these ratios to identify the “true” exons that showed altered expression levels in different datasets.

3.8 Data visualization

The Integrated Genome Browser (IGB) (downloaded from http://bioviz.org/igb/download.shtml) was used for data visualization. IGB is a free open source, desktop graphic tool, which can be used to display genome-scale data (Nicol et al., 2009). Both the results from microarray experiments and the gene structure of all genes in a genome can be loaded into IGB for visualization at the same time (see examples of IGB images in Figure 3.2 and Figure 4.4). Therefore, all of the expression

data generated above (in which intensity values are represented in numbers) can be visualized for better understanding.

3.9 Additional methods for splicing defects identification

Splicing defects can occur in four manners: exon skipping, intron retention, mis-selection of 3' site, and mis-selection of 5' site. By comparing expression levels of exons and introns, exon skipping and intron retention can be determined; by comparing expression levels of boundary regions, mis-selection of 3' site and 5' sites can be identified. However, the unusual heterogeneity of tiling array signals poses great challenge for any computational analysis, and traditional analysis methods usually produce many false positive results, especially in searching for splicing defects (van Bakel et al., 2010). Therefore, in addition to the strategies described above, two unconventional methods were designed (section 3.10 and 3.11) to identify splicing defects.

3.10 Cluster analysis

The signals in a tiling array experiment are highly dynamic and uneven. Even if the overall expression level of one gene is significantly up-regulated in one sample, it does not necessarily mean that every single probe within that gene has a higher signal in that sample in the tiling array experiment. Individual probe signals are not always consistent with the gene expression levels, and in this situation, it is difficult to tell whether a region really has a different expression level in different samples, as shown in Figure 3.9. Keeping these considerations in mind, cluster analysis was utilized, aiming to recognize

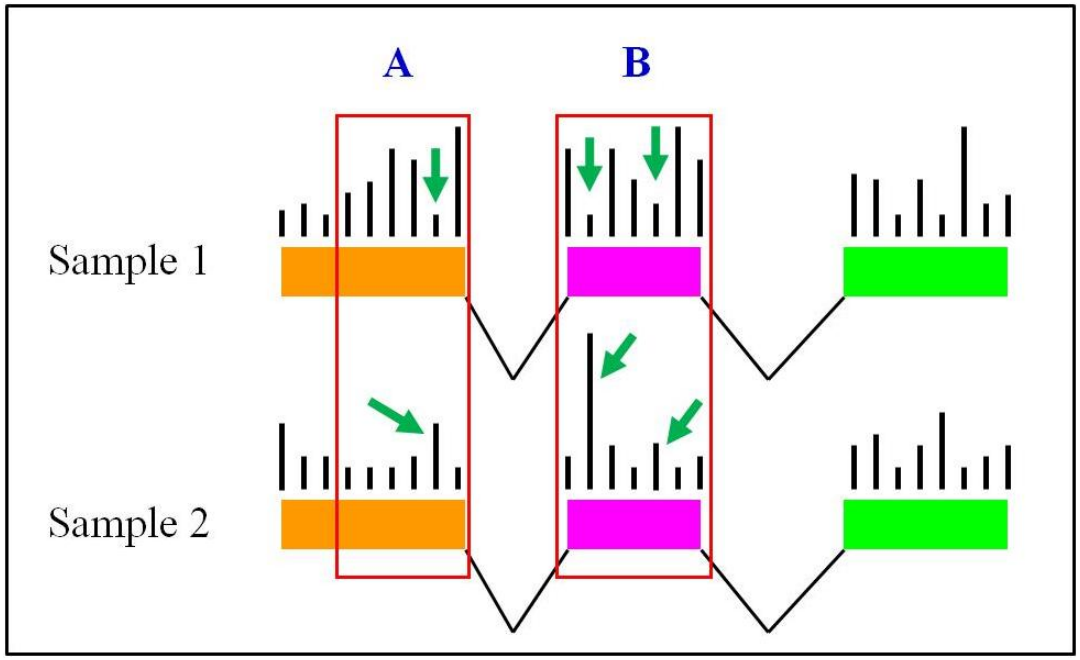


Figure 3.9. An example of the situation considered in "Cluster analysis".
Lines represent probe locations and the heights of them indicate the corresponding signal intensities. For region A and region B (boxed parts), although most of the probes have higher signals in the sample 1, there are still several probes that have higher signals in sample 2 (arrows).

these differently expressed regions, many of which may be due to splicing problems. Detailed procedures for performing such an analysis are described below.

3.10.1 Sorting probes

Based on the location information of every probe, probes that were located within inter-genic regions were first removed, ensuring that only genic probes were dealt with. Genic probes were further grouped into pure intronic (see section 3.3.3 for definition of "pure intronic") and non-intronic probes (the rest of genic probes) (Figure 3.10).

3.10.2 Normalization of probe signals

For each comparing pair (*teg-4* vs. N2 or *teg-4 smg-2* vs. *smg-2*), the signals of the probes in the two groups were normalized separately, generating two files with information for both probe locations and normalized probe signals (*i.e.* for *teg-4* vs. N2, normalized probe signals were generated for both intronic probes and non-intronic probes). After this, these two files were merged into one file (Figure 3.11), which was divided into the six different chromosomes. For example, *teg-4* vs. N2 pair was eventually broken into six parts, each of which contains normalized probe signals for both *teg-4* and N2 from only one chromosome (chrI to chrX).

3.10.3 Identifying probes that have different signals

For each chromosome, the R package "limma" was utilized. Probes that showed different ($p<0.05$) signal intensities in a comparing pair were retained. Afterwards, based on the relative signal intensities between the two comparing datasets (*i.e. teg-4* vs. N2),

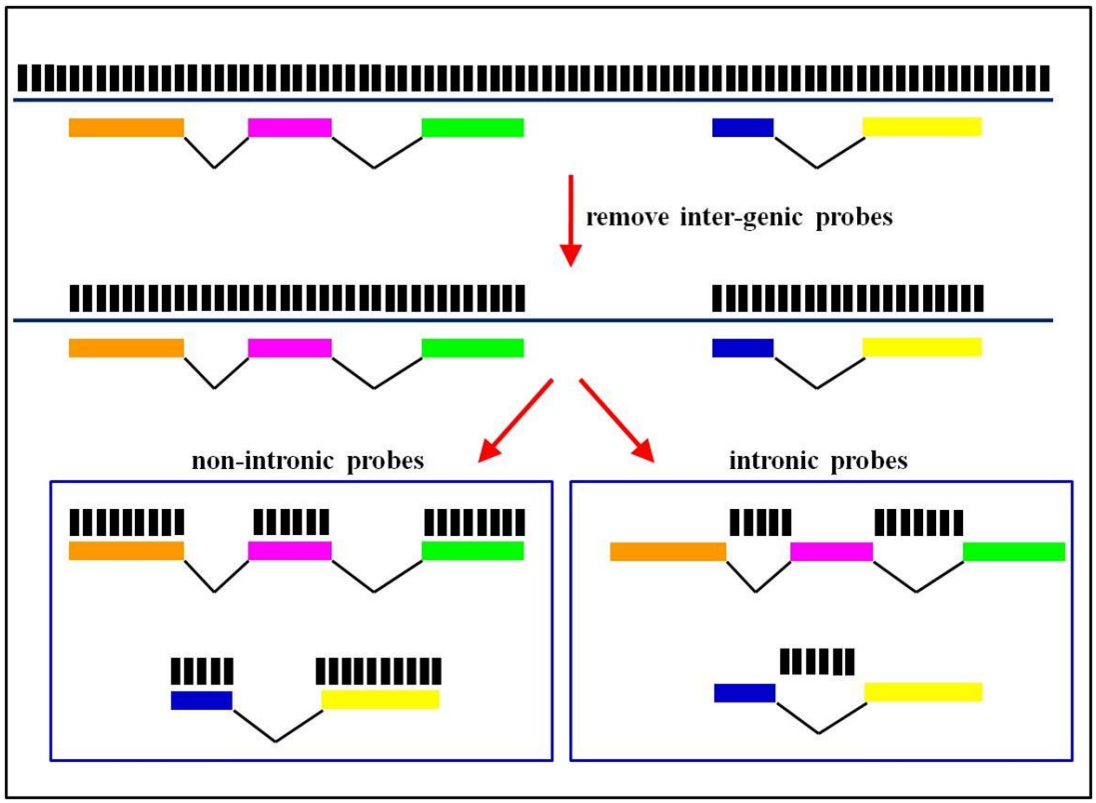


Figure 3.10. Sorting probes for the cluster analysis.

Tiling probes (black bars) were located across the genome, including the genic and inter-genic regions (top panel). Probes that are within inter-genic regions were removed, leaving only genic probes (middle panel). Genic probes were further divided into non-intronic probes (bottom panel, left box) and intronic probes (bottom panel, right box).

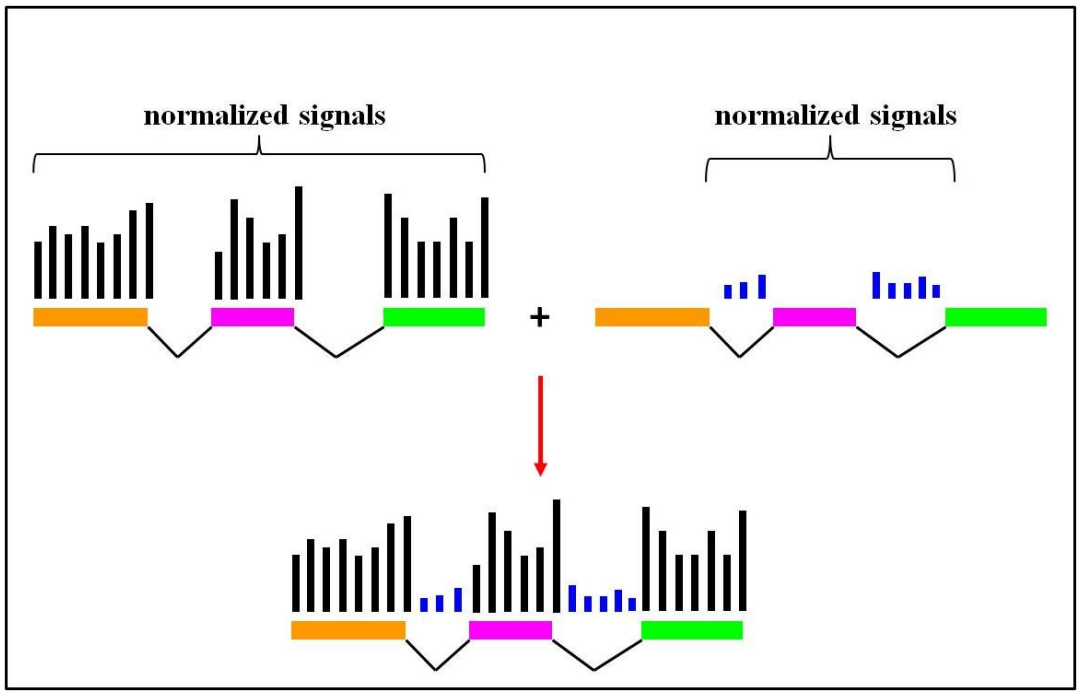


Figure 3.11. Non-intronic and intronic probes were combined for cluster analysis. Signal intensities of non-intronic probes (black lines, top panel, left) and intronic probes (blue lines, top panel, right) were normalized separately. Non-intronic and intronic probes with normalized signals were merged (bottom panel).

probes were further divided into two groups; probes in the first group all have signal intensities higher in one dataset (*i.e.* probes have higher value in N2), probes in the second group all have signal intensities higher in the other dataset (*i.e.* probes have higher value in *teg-4*), and they were stored in two separate files (Figure 3.12).

The sections above describe the preparation for the probes that was necessary to perform cluster analysis (described below).

3.10.4 Cluster creation

Generally speaking, creation of clusters is based on the probe location information from the two probe files for each chromosome described above. Two concepts were introduced here; maximum gap and minimal count. Maximum gap refers to the largest distance (in bp) between two adjacent probes (here distance=start position of former probe to the latter position of former probe) in a certain region. Minimal count refers to the total number of probes this region harbors. For a certain group of probes, if the maximum gap is 50 bp and minimal count is 2, then this group of probes can be considered as a cluster. Cluster locations can provide important information in splicing defect recognition, as illustrated in Figure 3.13. In the end, for each chromosome, one probe file was used to generate one cluster file.

3.10.5 Annotating clusters

Additional information was added to cluster files. First, cluster expression level was determined, with the help of R package “gcrma”. Second the number of probes contained in each cluster was also determined. Finally, clusters from all chromosomes were

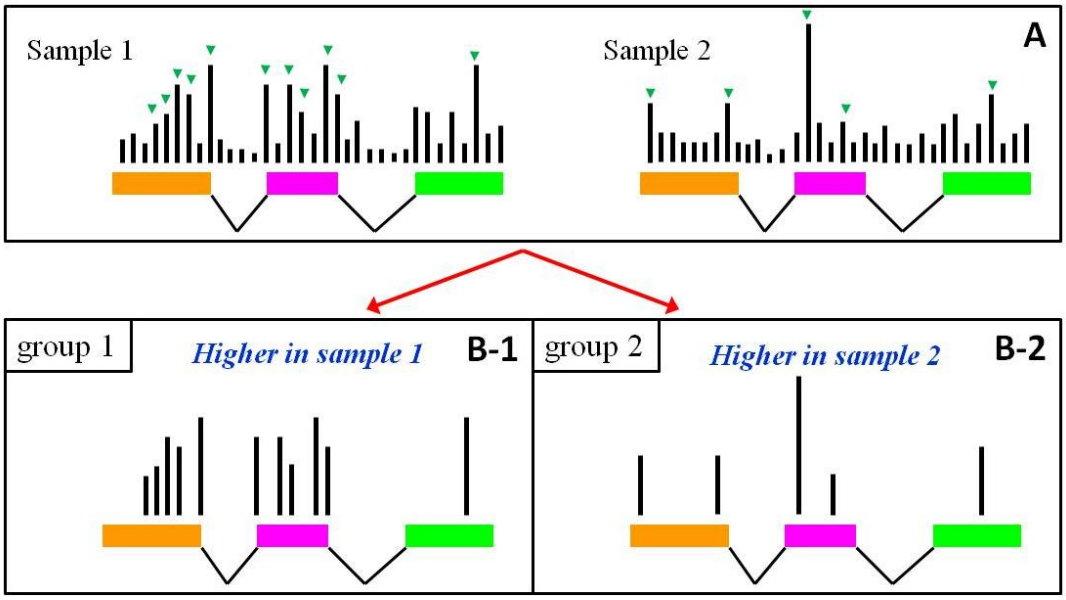
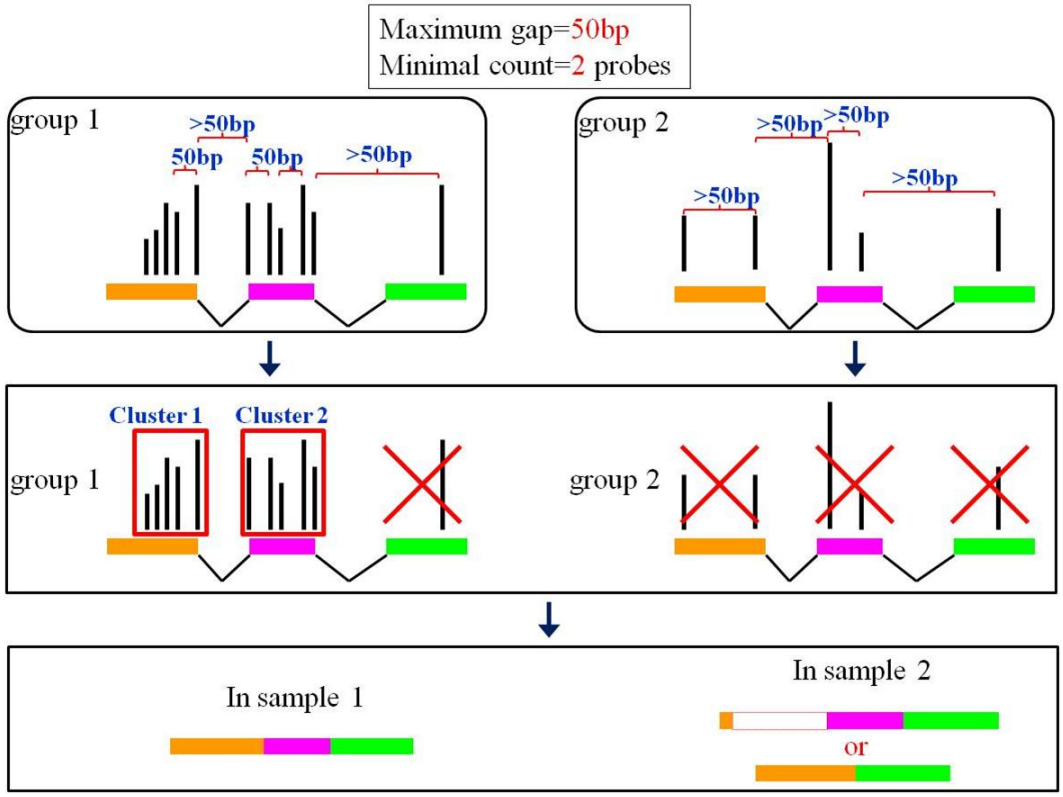


Figure 3.12. The probe grouping strategy for the cluster analysis.

Probe intensities (black lines) for the same gene from two different samples (Sample 1 and Sample 2) are presented in (A). According to the statistical analysis results, only probes that show significantly different signals from two samples were kept (A, green arrowheads). Based on the relative signal levels probes were divided into two groups (B-1 and B-2). In the first group, all probes have higher signals in Sample 1 (B-1), and in the second group, all probes have higher signals in Sample 2 (B-2).

Figure 3.13. Cluster identification.

A "cluster" is defined as a group of probes in which two criteria are satisfied: the maximum distance between two adjacent probes within a group [Maximum gap] is 50bp and the minimal number of probes [Minimal count] in this group is two. Distances between two adjacent probes were determined for both group 1 and group 2 (top panel) (see figure 3.12 for explanations of these two groups). Clusters were then identified by applying the two criteria. In this example, two clusters (Cluster 1 and Cluster 2) are identified in group 1 and none in group 2 (middle panel). Since both clusters are in group 1 (in which probe signals are higher in sample 1), it can be deduced that the expression levels of the regions that these clusters are higher in sample 1. Because both regions are exonic, it can be concluded that these two exonic parts were retained in sample 1, but were removed in sample 2 (bottom panel).


Maximum gap=50bp
Minimal count=2 probes
group 1
>50bp
50bp
50bp
>50bp
group 2
>50bp
>50bp
>50bp
>50bp
Cluster 1
Cluster 2
group 1
group 2
In sample 1
In sample 2
or

combined, and by cross-referencing cluster locations and annotation files, cluster identity were determined, *i.e.* exonic cluster, and the corresponding expression levels of these "identities" were also obtained. Since a very large number of candidates were expected to be identified through this approach, annotation information of the cluster files can help filter out a significant amount of insignificant noise (see details in Appendix C). As a consequence, the number of candidates subjected to further experimental verification can be largely reduced.

3.11 Single probe analysis

Orthodox ways of analysing microarray data always include normalization of the data from all arrays before performing statistical comparisons. But some research suggests that normalization, and other manners of data manipulation, might compromise the data, in terms of obscuring tiling array signal changes at gene structure boundaries (Gilbert and Rechtsteiner, 2009). This might not exert serious consequences for whole gene expression comparisons, but when tackling splicing defects, a dropped sensitivity of tiling array signals at boundaries may blur the distinction between exonic and intronic regions, making splicing errors more difficult to identify. Therefore, another attempt, which is completely different from ordinary analysing methods, was performed. In this method no normalization was conducted and raw probe intensity data was used directly for comparison purpose.

Examinations of ESTs (Expressed Sequence Tag) revealed that intron retention, as well as alternative 3' and 5' site selection, were frequently observed in cancerous tissues, whereas exon skipping was more rarely detected (Kim et al., 2008). This may be because

intron retention may introduce a premature stop codon, thus generating a truncated/dysfunctional protein. For this reason, this new approach only focuses on intron retention (entire or partial retention).

3.11.1 Extracting original probe signals

An R package "affy" was used to extract raw probe intensity data from each of the 16 arrays (N2 and *teg-4*: 5 arrays each, *smg-2* and *teg-4 smg-2*: 3 arrays each). Only intronic probes were kept, the rest of the probes were removed, resulting in 16 separate files with intronic probe intensities.

3.11.2 Probe selection

For each of the 16 intronic probe files created, all probes were ranked based on their intensities, and only probes with the highest 10% intensities were reserved. Next, probes that were present in all replicates (arrays) for each of the four datasets (*i.e.* probes that are present in all 5 replicates in N2 dataset) were identified. This was accomplished by simply comparing the probe locations among different arrays. As a result, each of the four datasets generated one file, in which only probes that were identified in all replicates were included.

Then, for files generated from *teg-4* and *teg-4 smg-2* datasets, probes that were also identified in their corresponding control groups were removed, leaving the probes that were identified only in the experimental datasets (*i.e.* in *teg-4* not in N2). Upon completing this, probes were further filtered by checking their identities. Specifically, only those probes that were mapped to boundary regions were retained and the rest were

deleted. Finally, two files were generated, one from *teg-4*/N2 and the other from *teg-4 smg-2*/*smg-2*. Figure 3.14 briefly summarizes the main idea of this approach.

3.12 Summary

Of all the data analysis approaches described, only the gene annotation/probe grouping files were used to identify differentially expressed genes. Exon/intron/boundary annotation/probe grouping files, cluster analysis and single probe analysis were all aimed at identifying splicing defects. By comparing exon/intron expression levels, splicing errors that affect an entire exon or intron could be identified; by comparing the expression of the probes near boundaries, and employing cluster and single probe approaches, minor/partial splicing defects could be identified.

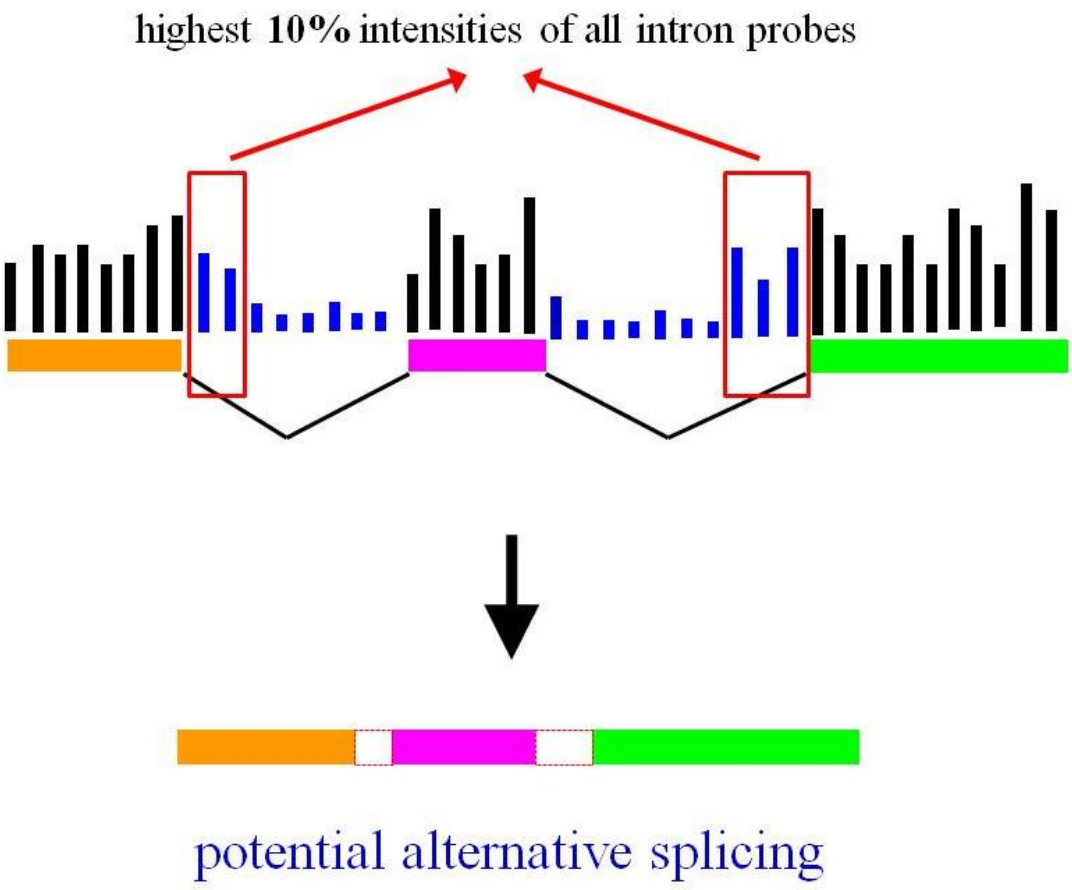


Figure 3.14. The major idea of how single probe analysis can be used for splicing defects (alternative 3' and 5' site selection) identification.

Only probes that are mapped to intronic regions (blue lines) were considered, and signal intensities of these probes were ranked increasingly. Probes with the top 10% signal intensities were further scrutinized. If some of these probes are located at boundary regions, it can be concluded that within these regions, partial retention of introns occurred.

Chapter Four: Computational analysis: results and verification

4.1 Quality assessment

Since carrying out a microarray experiment involves many steps, it is inevitable that some errors might be introduced during this process. Because enormous amount of information will be generated in the end, one small mistake at one single step is likely to have a significant impact in the final output. In some cases, the data is beyond correction with statistical methods and any conclusion drawn from them is questionable. Therefore, an initial evaluation of the data is essential and necessary for detecting possible quality issues, and a futile analysis on such an error-riddled raw data should be avoided.

4.1.1 Chip images

A first step in performing a quality assessment is by analyzing the images from the raw probe level data. In a good image, there should be no obvious unevenness in it. All 16 images produced from N2, *teg-4*, *smg-2* and *teg-4 smg-2* datasets show no uneveness (Figure 4.1C). For comparison, an example of a poor quality chip is provided, in Figure 4.1A, the large dim area at the left corner suggests this is a problematic chip.

The second kind of image is called a chip pseudo-image, and it is more visually informative in highlighting some subtle artifacts on a chip, which is very difficult to be observed in images of the raw probe intensities. The principle of a chip pseudo-image is illustrated elsewhere (Bolstad et al., 2005). All pseudo-images from the four datasets again suggested their high quality (Figure 4.1D); no major problems, such as the “circle” in the negative example (Figure 4.1B) were observed.

4.1.2 RNA degradation

High fidelity of microarray data requires input RNA of high quality, and degraded RNA can result in poor quality data. Although efforts were made during RNA sample preparation to ensure its quality before the microarray experimental procedures started, it is still not safe to assume that RNA degradation will not happen in the later process unless there is some way to test this.

Every individual probe on an Affymetrix GeneChip is numbered sequentially from its 5' end of the targeted transcript. If RNA degradation is significant, the probe intensities at the 3'end of a probeset will be *systematically* elevated when compared with the 5'end (Bolstad et al., 2005). By drawing an RNA degradation plot, any obvious advanced RNA degradation can be visualized. Figure 4.2A clearly represented a typical example of RNA degradation; at the 3'end, the intensities were greatly elevated. RNA degradation plots of all datasets showed no apparent systematic RNA degradation; and all curves share a similar pattern (Figure 4.2B and 4.2C), suggesting a high agreement between chips. Therefore, no serious RNA degradation occurred during the experiment and RNA for all chips were of a similar quality.

In summary, both images and RNA degradation plots validated the original data generated from all chips, and more in-depth analysis of these data, aimed to pull out more biologically meaningful information, can be performed.

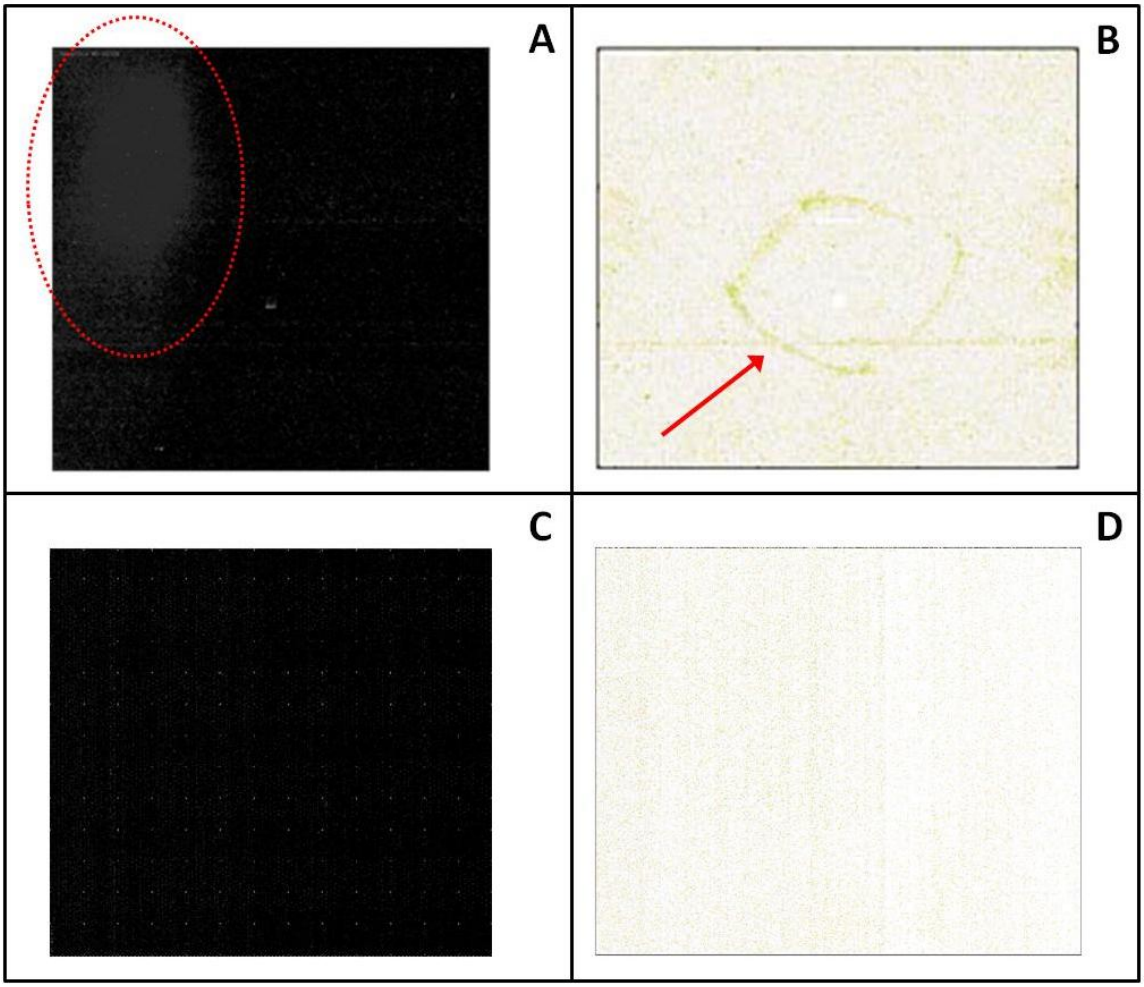
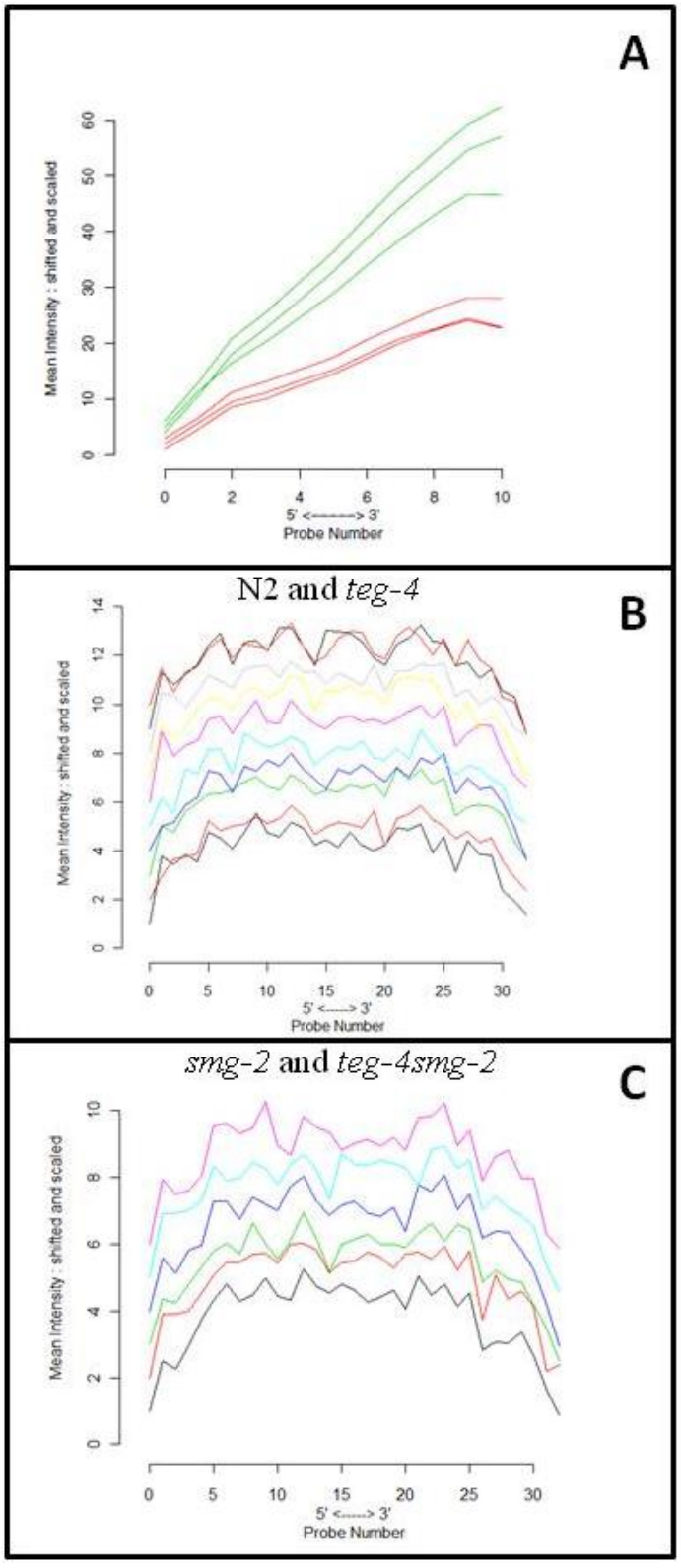


Figure 4.1. Images of chips with good (C and D) and bad (A and B) qualities.
(A) and (C): Images of raw probe intensities. (B) and (D): Chip pseudo-images. The dim area at the top-left corner (circled by dashed line) of (A) and the "circle" shape in (B) both indicate poor quality of these two chips. No obvious artifact is noticed in (C) and (D), suggesting good qualities of these two chips. (A) and (B) were adapted from (Bolstad et al., 2005), (C) and (D) are sample images of this work.

Figure 4.2. RNA degradation plots.

In (A), green lines represent the data of amplified RNA samples, and red lines represent the data of standard arrays. For green lines, probe intensities at the 3'end are systematically elevated when compared with red lines, indicating RNA degradation. Picture adapted from Bolstad et al. (Bolstad et al., 2005). In RNA degradation plots of all chips from N2 and *teg-4* samples (B) and *smg-2* and *teg-4 smg-2* (C), no systematic elevation of probe intensities at the 3'ends were observed. Plots within each comparing pair (N2 and *teg-4* or *smg-2* and *teg-4 smg-2*) have similar shapes, suggesting high agreement with each other.


A
Mean Intensity : shifted and scaled
0
10
20
30
40
50
60
0
2
4
6
8
10
5' <-------> 3'
Probe Number
N2 and teg-4
B
Mean Intensity : shifted and scaled
0
2
4
6
8
10
12
14
0
5
10
15
20
25
30
5' <------> 3'
Probe Number
smg-2 and teg-4smg-2
C
Mean Intensity : shifted and scaled
0
2
4
6
8
10
0
5
10
15
20
25
30
5' <------> 3'
Probe Number

4.2 Summary of annotation file information

Annotation files, as described in chapter two, are the basis of all data analysis. Four types of annotation files were generated; gene annotation files, exon annotation files, intron annotation files and boundary annotation files. A brief summary of these files is listed in Table 4.1.

Briefly, according to the annotation files, 20,012 coding genes, 126,865 exons, 105,177 introns and 244,415 boundary regions were included (Table 4.1). All of the analysis described below (including data pre-processing and statistical comparison) was performed within these genomic regions. Regions outside these annotations were not considered.

4.3 Summary of probe file information

Using xMAN, all probes were mapped against the WS170 *C. elegans* WormBase (www.wormbase.org) annotation file to their corresponding positions on each chromosome (including mitochondrial genome) to generate the preliminary probe file. After scrutinizing the preliminary probe file, it was found that 3,041,220 probes were mapped to 3,745,078 positions (Table 4.2). Apparently, many probes have sequences that are present in multiple positions on the genome; a total of 154,414 such probes were removed. From further analysis, 2,886,808 probes were retained, and all of them have a unique position in the genome. These probes were further grouped into different chromosomes (probes mapped to mitochondrial genome were not considered) and assigned to different genes/exons/introns etc. Therefore, all of the results and conclusions described below were drawn from only the information of these 2,886,806 probes.

Table 4.1 Summary of the information in annotation files.

chromosomes	No. genes	No. exons		No. introns		No. boundaries
		no variant	with variants	no variant	with variants	
ChrI	2854	13245	6030	11070	4849	37055
ChrII	3471	14994	5720	12404	4539	39748
ChrIII	2652	11351	6087	9299	4898	33518
ChrIV	3276	14486	5895	12085	4766	39243
ChrV	4986	22987	5167	19433	4096	54705
ChrX	2773	14638	6265	12559	5179	40146
Total	20012	91701	35164	76850	28327	244415

Table 4.2 Summary of probe file information.

No. total probes	No. total locations	No. removed probes	No. final probes	No. final locations
3041220	3745078	154414	2886806	2886806

Information corresponding to probes that were removed was excluded from any data analysis.

4.4 Data pre-processing

Figure 4.3 shows the data distribution before and after data pre-processing. It is obvious that the data distributions for the original datasets contained differences (Figure 4.3A). However, after data pre-processing, the distribution among different datasets within each category (gene/exon/intron/boundary) were quite similar (Figure 4.3B). These figures illustrate that the original data was successfully converted to comparable sets, and that statistical analysis using this data is warranted.

4.5 Identification of genes with different expression levels

24 genes showed significantly different (FDR<0.1) expression levels in the N2 vs. *teg-4* comparison (Table 4.3a), and 28 genes showed significantly different (FDR<0.1) expression levels in the *smg-2* and *teg-4 smg-2* comparison (Table 4.3b). Genes T09E11.2 and F21F8.4 were identified in both sets using a FDR of <0.1; however, T09E11.2 was up-regulated in both *teg-4* mutants (*teg-4* and *teg-4 smg-2*), but F21F8.4 displayed an opposite tendency: up-regulated in *teg-4 smg-2* but down-regulated in *teg-4*.

If N2 and *smg-2* animals and *teg-4* and *teg-4 smg-2* animals were referred as "*teg-4*$^{+}$" and "*teg-4*$^{-}$" animals, respectively, and the results from both sets were examined together (Table 4.3a and 4.3b), of the total 49 genes (F21F8.4 is not included), 23 showed higher expression in *teg-4*$^{-}$ animals (in red), and 26 showed higher expression in *teg-4*$^{+}$ animals (in black), with fold changes ranging from 1.24 to 45.23. Each of these

Figure 4.3. Boxplots of data before (A) and after (B) data pre-processing.
Each boxplot represents one replicate. In the N2 vs. *teg-4* (A, left panel; and B, upper panel), the first five boxplots were drawn from the N2 dataset, and the later five were drawn from the *teg-4* dataset. In the *smg-2* vs. *teg-4 smg-2* (A, right panel; and B, lower panel), the first three boxplots were drawn from the *smg-2* dataset, and the later three were from the *teg-4 smg-2* dataset. (A) The data distribution of raw probe intensities (before data pre-processing), and they display distinct distribution patterns. After data pre-processing (B), raw probe intensities were converted into expression levels of certain regions, and within each category (gene/exon/intron/boundary), all replicates share the similar pattern of data distribution in both N2 vs. *teg-4* and *smg-2* vs. *teg-4 smg-2*.

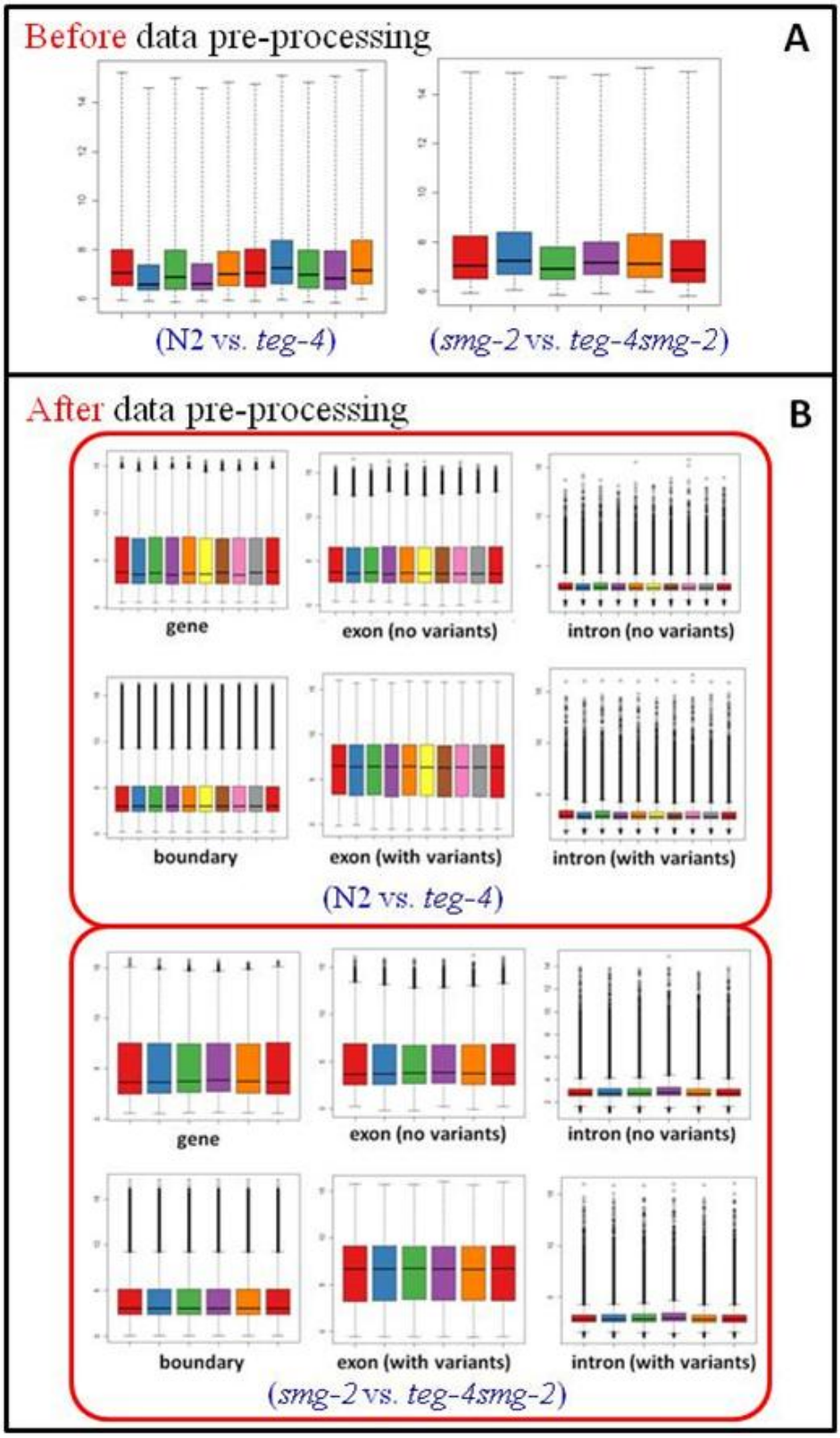

Before data pre-processing
A
(N2 vs. teg-4)
(smg-2 vs. teg-4smg-2)
After data pre-processing
B
gene
exon (no variants)
intron (no variants)
boundary
exon (with variants)
intron (with variants)
(N2 vs. teg-4)
gene
exon (no variants)
intron (no variants)
boundary
exon (with variants)
intron (with variants)
(smg-2 vs. teg-4smg-2)

Table 4.3a Genes with different expression levels in the N2 vs. *teg-4*.

gene ID	gene name	FDR	P.Value	N2 expression	*teg-4* expression	fold change
T09E11.2	***nhr-217***	**0.0092**	**1.09E-06**	**9.10**	**106.84**	**10.71**
F21F8.4	**.**	**0.0598**	3.30E-05	**7404.14**	**1570.42**	**4.94**
~~B0511.11~~	~~.~~	~~0.0319~~	~~1.14E-05~~	~~10.17~~	~~42.89~~	~~4.09~~
F55D10.1	*aman-1*	0.0333	1.36E-05	36.36	132.67	3.63
~~F28F8.7~~	~~.~~	~~0.0814~~	~~9.25E-05~~	~~6.63~~	~~22.01~~	~~3.19~~
T08A9.12	*spp-2*	0.0319	1.08E-05	612.51	196.32	3.16
K02B12.9	.	0.0604	5.23E-05	105.06	36.51	2.79
F49C12.7	.	0.0092	1.40E-06	137.01	49.67	2.75
F59D6.3	.	0.0203	5.18E-06	2708.56	1094.94	2.47
F26G1.10	.	0.0814	8.82E-05	52.25	23.56	2.18
T09F5.9	*clec-47*	0.0165	3.37E-06	8080.96	3739.99	2.15
K07E8.3	*sdz-24*	0.0648	5.95E-05	5220.52	2467.40	2.07
T20B3.1	.	0.0092	1.10E-06	468.26	234.43	2.01
M02H5.8	.	0.0814	9.54E-05	1605.39	881.13	1.85
C01G6.7	*acs-7*	0.0604	5.22E-05	421.04	234.06	1.80
ZK822.4	.	0.0604	4.44E-05	871.15	492.44	1.77
R09H10.3	.	0.0604	4.88E-05	2550.97	1513.46	1.69
VW02B12L.4	*adbp-1*	0.0598	3.91E-05	468.25	278.49	1.68
Y55F3AM.1	.	0.0652	6.31E-05	1272.33	772.09	1.66
Y57A10A.29	.	0.0598	3.85E-05	812.65	490.73	1.66
F52B11.2	.	0.0598	3.96E-05	2018.25	1298.87	1.56
F46E10.1	*acs-1*	0.0660	6.73E-05	2692.27	3946.21	1.47
F09B9.4	.	0.0500	2.30E-05	1768.37	1281.93	1.38
T21H3.1	.	0.0904	1.11E-04	24136.56	19439.06	1.24

Genes that are in red have higher expression in the *teg-4*, and genes that are in black have higher expression in the N2. Genes that have a strikethrough are the ones with low expression levels (<50) and are considered not expressed.

Table 4.3b Genes with different expression levels in the *smg-2* vs. *teg-4 smg-2*.

gene ID	*gene name*	FDR	P.Value	*smg-2* expression	*teg-4 smg-2* expression	fold change
T09E11.2	***nhr-217***	**0.0005**	**2.66E-08**	**9.50**	**432.15**	**45.23**
C03G5.12	*nspc-5*	0.0039	1.04E-06	5.62	231.91	39.60
B0379.3	*mut-16*	0.0037	6.92E-07	15.22	288.50	19.36
D1037.2	*smp-2*	0.0039	1.40E-06	25.73	437.46	17.39
F15D4.6	.	0.0037	7.46E-07	20.86	331.46	16.18
T28F2.5	*ccb-1*	0.0037	6.58E-07	1466.17	89.26	16.10
C38D9.2	.	0.0274	1.54E-05	56.47	737.15	13.06
C25G4.7	.	0.0534	3.81E-05	8.42	102.18	11.89
Y119C1B.1	.	0.0756	9.25E-05	6.35	76.85	11.15
Y73B3A.20	.	0.0630	7.06E-05	4249.85	469.92	9.72
F15D4.5	.	0.0194	8.89E-06	34.72	271.19	7.75
~~CE7X_3.2~~	.	~~0.0039~~	~~1.27E-06~~	~~33.06~~	~~4.63~~	~~7.07~~
B0511.4	*tag-344*	0.0619	5.24E-05	13.82	69.72	5.16
F59H6.3	.	0.0630	6.65E-05	3961.07	11971.41	3.06
~~Y113G7A.6~~	~~*ttx-1*~~	~~0.0619~~	~~5.53E-05~~	~~26.33~~	~~9.13~~	~~2.87~~
Y38E10A.10	*lips-16*	0.0619	5.99E-05	48.20	137.83	2.87
K10H10.2	.	0.0841	1.07E-04	435.60	153.51	2.86
~~Y39E4B.13~~	.	~~0.0274~~	~~1.68E-05~~	~~14.30~~	~~39.27~~	~~2.74~~
C30G12.2	.	0.0938	1.29E-04	1143.40	3076.54	2.66
K02E7.6	.	0.0619	5.79E-05	193.25	507.87	2.61
~~B0513.6~~	.	~~0.0965~~	~~1.38E-04~~	~~5.78~~	~~15.20~~	~~2.60~~
C15C6.3	.	0.0223	1.14E-05	763.39	1834.95	2.40
F14B6.6	.	0.0849	1.13E-04	112.67	271.23	2.39
F21F8.4	**.**	**0.0173**	**7.04E-06**	**909.07**	**2122.21**	**2.33**
T26H5.9	.	0.0386	2.56E-05	3313.32	1575.82	2.11
~~ZK1025.4~~	.	~~0.0570~~	~~4.36E-05~~	~~9.32~~	~~18.20~~	~~1.95~~
ZC410.5	.	0.0739	8.66E-05	736.71	438.27	1.68
T01H8.1	*rskn-1*	0.0630	6.97E-05	139.67	83.60	1.67

Genes that are in red have higher expression in the *teg-4 smg-2*, and genes that are in black have higher expression in the *smg-2*. Genes that have a strikethrough are the ones with low expression levels (<50) and are considered not expressed.

genes was visualized using IGB for all genotypes. Figure 4.4 displays two sample images of genes from Table 4.3b.

30 of these genes are of unknown or predicted functions. The remaining 19 genes are involved in many biological processes such as reproduction, metabolism, growth, etc. Some genes are multifunctional, while the others are only involved in one biological process. Figure 4.5 shows a general distribution of these functions.

4.6 Identification of splicing defects

As illustrated in chapter three, expression values from defined exonic and intronic regions were calculated to help determine the presence of major splicing problems (exon skipping and intron retention, or complete splicing), boundary expression values, as well as cluster and single probe analysis were employed to identify minor splicing errors (alternative 3' site and 5' site, or partial splicing)

Table 4.4 and Table 4.5 list the regions (exon or intron) with potential splicing defects, and most of them are fairly small in size (~50bp). Under FDR<0.1, only one in Table 4.4 (N2 vs. *teg-4*) and seven in Table 4.5 (*smg-2* vs. *teg-4 smg-2*) can meet this FDR criteria. Considering the small size of the splicing defects in these candidates, within each of them, intensity values from only one or two probes were used for analysis. As a result, there may not be sufficient information from which a powerful statistical analysis can be carried out, and the accuracy of the analysis is possibly weakened. For this reason, FDR<0.1 might be too stringent, and the cutoff was then increased to 0.2. The candidates in Table 4.4 and Table 4.5 are below this threshold.

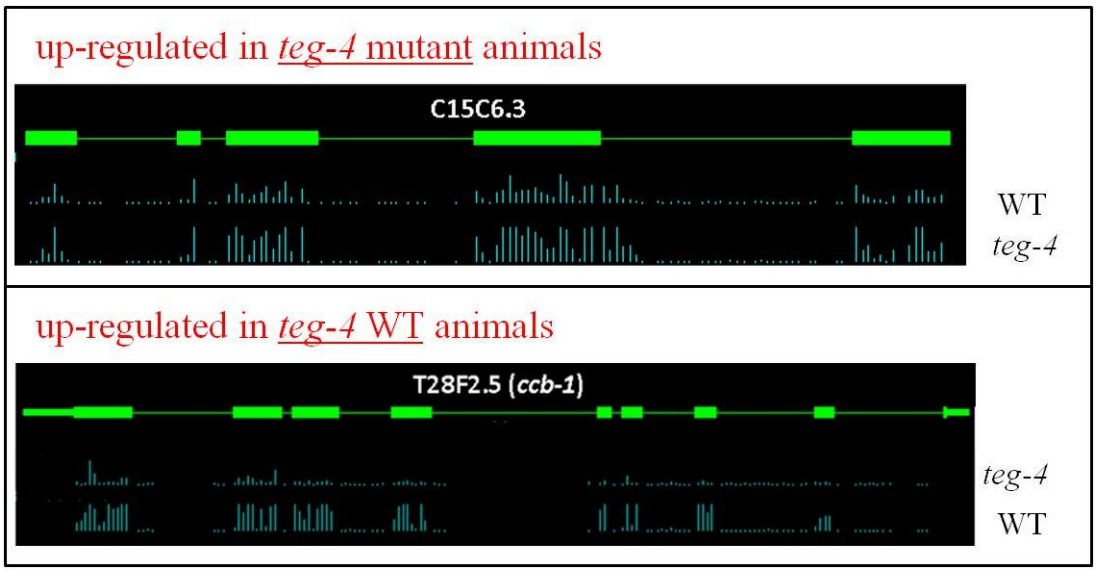


Figure 4.4. Sample IGB (Integrated Genome Browser) images for genes with different expression levels in *teg-4* mutants and *teg-4* wild-type (WT) animals. Gene structures are shown on the top for both upper and lower panels. Solid green rectangles are exons, and straight lines between exons are introns. Underneath the gene structures are the signal intensities of each sample, represented by the height of the green lines (probes). The expression level of the gene C15C6.3 is higher in *teg-4* mutants (upper section) and the expression level of the gene T28F2.5 is higher in *teg-4* WT (lower section) animals.

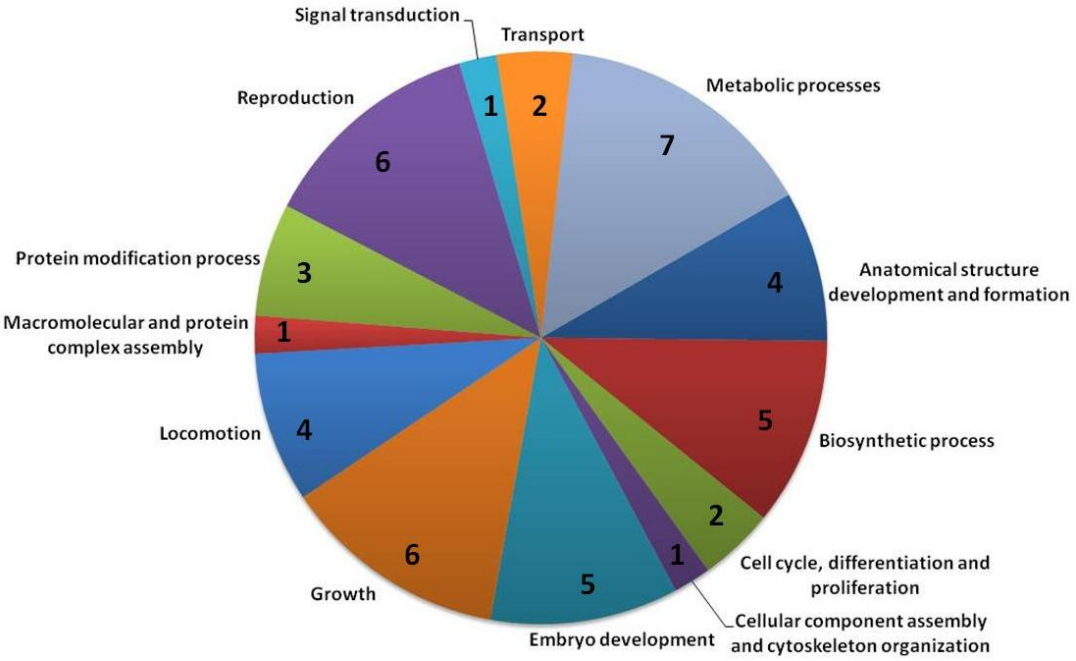


Figure 4.5. Distribution of GO terms of genes with different expression levels. 49 genes were identified to have different expression levels. 19 genes were mapped with 13 GO slim generic terms (biological processes), using AmiGO. Numbers indicate how many genes are in each GO slim term category.

Table 4.4 Splicing defects (exonic or intronic regions) identified in the N2 vs. *teg-4*.

region name	region identity	region size (bp)	FDR	N2*	*teg-4* *
K09A11.4.exon1	non-variant intron	47	0.0071	35.50	6.41
B0024.4.exon4	non-variant intron	65	0.0417	5.88	14.05
T22F3.10.exon8	non-variant intron	54	0.0417	3.24	26.53
C07A9.5.exon3	non-variant intron	56	0.1852	78.69	8.52
C05E4.9a.1.exon3	variant intron	68	0.0225	5.94	62.72
W05H7.4a.exon3	variant intron	51	0.0340	5.64	39.31
Y79H2A.3a.exon11	variant intron	747	0.1480	6.59	16.19
F32D1.9.1.exon4	variant intron	58	0.1961	17.25	5.25
F59A1.8.exon1	non-variant exon	50	0.0211	4.78	13.24
C46C11.1.28	variant exon	213	0.1995	398.16	120.11
F08G5.5.5	variant exon	134	0.1995	291.75	39.21
Y41D4A.4.7	variant exon	298	0.1995	1805.82	1679.03

* expression levels of these regions in the corresponding sample

Table 4.5 Splicing defects (exonic or intronic regions) identified in the *smg-2* vs. *teg-4 smg-2*.

* expression levels of these regions in the corresponding sample

The gene that is highlighted was tested using PCR

region name	region identity	region size (bp)	FDR	*smg-2* *	*teg-4 smg-2* *
Y34B4A.5.exon7	non variant intron	42	0.0749	5.42	33.68
ZK550.1.exon3	non variant intron	54	0.0749	76.96	3.96
C15H11.11.exon1	non variant intron	45	0.0749	4.97	61.35
R04F11.4.exon11	non variant intron	65	0.0749	6.32	23.50
Y38F1A.1.exon3	non variant intron	54	0.0749	19.59	6.13
T03F6.8.exon5	non variant intron	54	0.0749	4.12	65.63
F09A5.3.exon5	non variant intron	53	0.0768	6.26	29.91
T23F2.4.exon1	non variant intron	55	0.1299	82.96	5.69
K11H12.3.exon6 (5'UTR)	non variant intron	53	0.1299	9.28	21.51
R186.7.exon1	non variant intron	48	0.1545	23.31	3.88
M01A12.1.exon2	non variant intron	46	0.1545	27.65	7.03
ZK418.6.exon9	non variant intron	65	0.1725	17.70	5.08
F28F8.1.exon5	non variant intron	48	0.1725	6.35	227.81
C39B5.10.exon2	non variant intron	108	0.1725	11.08	4.60
F55H12.1.exon8	non variant intron	47	0.1725	26.56	6.83
Y8A9A.2.exon11	non variant intron	46	0.1725	5.66	70.12
Y106G6A.1.exon5	non variant intron	46	0.1725	19.63	5.32
Y18H1A.7.exon2	non variant intron	60	0.1725	3.33	40.10
K06H6.5.exon4	non variant intron	61	0.1725	43.28	8.11
C04A11.2.exon8	non variant intron	51	0.1736	27.50	6.14
T25D1.1.exon9	non variant intron	50	0.1736	32.53	5.64
ZC190.9.exon1	non variant intron	444	0.1736	4.37	76.44
Y39A1A.9.exon3	non variant intron	49	0.1736	32.99	6.86
Y5F2A.4.exon3	non variant intron	51	0.1736	4.36	45.92
B0454.6.exon2	non variant intron	48	0.1736	122.06	7.42
C16D6.3.exon1	non variant intron	50	0.1793	10.45	26.10
F23D12.2.exon5	non variant intron	45	0.1793	55.83	6.22
Y48G8AR.2.exon4	non variant intron	51	0.1793	23.14	3.92
F58G11.2.exon1	non variant intron	51	0.1793	100.54	7.67
F53A3.2.exon1	non variant intron	51	0.1899	49.89	5.87
Y67D8C.5.exon5	non variant intron	59	0.1927	5.01	8.90
ZK455.4.exon9	non variant intron	52	0.1927	123.89	10.18
T24E12.3.exon4	non variant intron	78	0.1927	5.89	31.38
F36A4.1.exon2	non variant intron	49	0.1959	434.76	1136.30
T21H8.5.exon2	non variant intron	56	0.1959	64.76	9.39
F53F1.2.exon2	non variant intron	50	0.1967	116.35	6.14
F10D2.3.exon6	non variant intron	47	0.1967	24.42	5.05
C54D1.1.exon4	non variant intron	72	0.1967	13.46	4.57
Y110A7A.11.exon6	non variant intron	49	0.1967	31.57	7.72
F52E1.13a.exon15	variant intron	54	0.1562	4.30	73.80
H22K11.4a.exon8	variant intron	58	0.1562	7.23	30.55
F02E9.9a.1.exon6	variant intron	45	0.1562	5.00	78.42
C05D11.11a.1.exon2	variant intron	48	0.1562	5.00	27.71
ZK546.14a.1.exon1	variant intron	53	0.1562	44.38	2.77
K11G12.1a.exon3	variant intron	47	0.1562	6.06	28.34
ZK6.11a.exon3	variant intron	47	0.1856	28.91	3.73
EEED8.1.1.exon3	variant intron	43	0.1856	37.20	3.05
ZK370.3a.exon9.1	variant intron	56	0.1856	5.06	74.29
T02G5.8.1.exon3	variant intron	50	0.1856	111.87	3.99
T19D12.4a.exon7	variant intron	57	0.1856	6.65	23.36
C07G1.5.1.exon6	variant intron	45	0.1856	7.19	44.97
F09E5.15.1.exon1	variant intron	64	0.1856	5.24	99.95
C06C3.1a.exon12	variant intron	43	0.1856	5.29	24.77
C46H11.11a.exon3	variant intron	82	0.1939	5.55	48.67

Most of the candidates in Table 4.4 and Table 4.5 are intronic regions, and only four exonic regions were identified in the N2 vs. teg-4 set, but none were identified in smg-2 vs. teg-4 smg-2. Correspondingly, the expression levels for these four exonic regions were much higher than those in the intronic regions.

Boundary expression, cluster and single probe analysis primarily deal with partial splicing problems, usually within the range of only two probes (50bp in length); hence the list of candidates for this category is much longer than that for the large splicing problems. Table 4.6 summarizes the major findings of these three situations, and a detailed list can be found in Table D.6. IGB images were also created for examples of potential splicing defects, as presented in Figure 4.6.

4.7 Background expression determination

Ideally, intronic probes should have no hybridization, resulting in their signals being zero. Therefore, generally speaking, intronic signals can be treated as background. In this analysis, background was determined by calculating the value that is higher than the intensities of 90% of intronic probes. Figure 4.7 shows an accumulated histogram of intronic expression in the N2 dataset, with over 90% of the intron probes having intensities below 50. In order to further validate this background value, PCR and gel electrophoresis were performed. The logic behind this is that if genes are truly expressed, there should be sufficient template in order to have visible bands on a 1.0% agarose gel.

Table 4.6 Summary of splicing defects identified in boundary, cluster, and single probe analyses.

a

boundary	single probe	cluster		
			N2 (*teg-4*)*	*teg-4* (N2)*
75	96	exon	1841	2428
		intron	313	453
		boundary	213	130
		total	2367	3011

b

boundary	single probe	cluster		
			smg-2 (*teg-4 smg-2*)*	*teg-4 smg-2* (*smg-2*)*
257	112	exon	873	903
		intron	223	266
		boundary	66	97
		total	1162	1266

[a] This table shows the number of splicing defects identified in the N2 vs. *teg-4*

[b] This table shows the number of splicing defects identified in the *smg-2* vs. *teg-4 smg-2*

* A (B): This means these clusters all have higher expression levels in sample A.

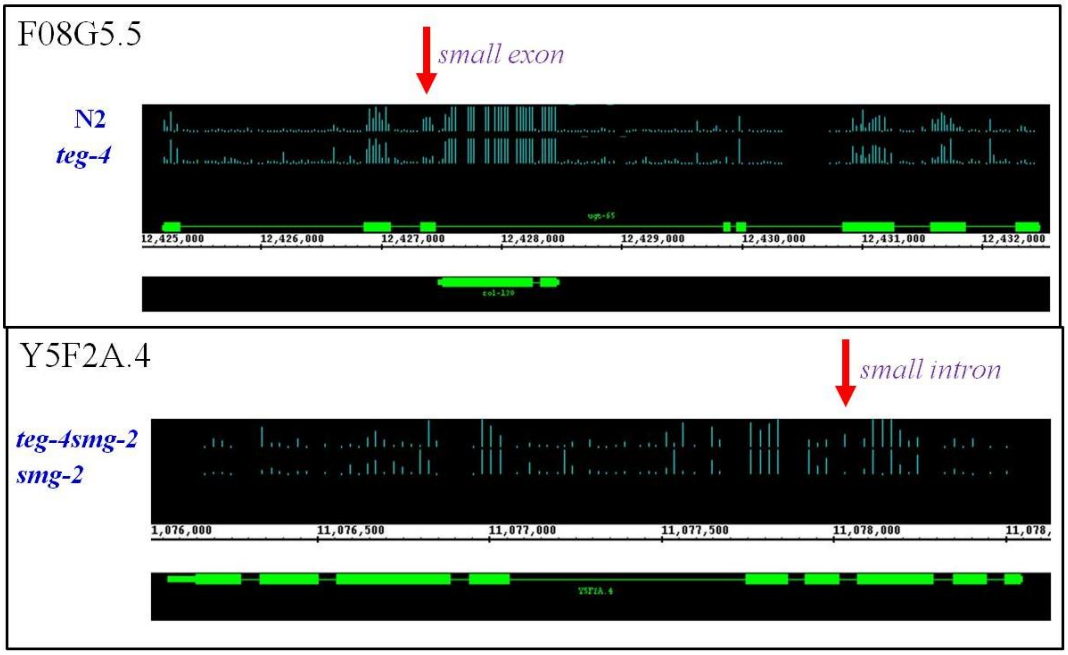


Figure 4.6. Sample IGB (Integrated Genome Browser) images for identified potential splicing defects.

Gene structures are shown on the bottom of each panel, and solid green rectangles are exons, and straight lines between are introns. Above the gene structures are the signal intensities of each sample, represented by the height of the green lines (probes). For the gene F08G5.5 (upper section), the third exon (small in size) is spliced out in the *teg-4* single mutant (red arrow, compare the overall levels of all probes within this exon between N2 and *teg-4*). For the gene Y5F2A.4 (lower section), the sixth intron (small in size) is retained in the *teg-4 smg-2* double mutant (red arrow, compare signal of the identified probe in this intron between *smg-2* and *teg-4 smg-2*).

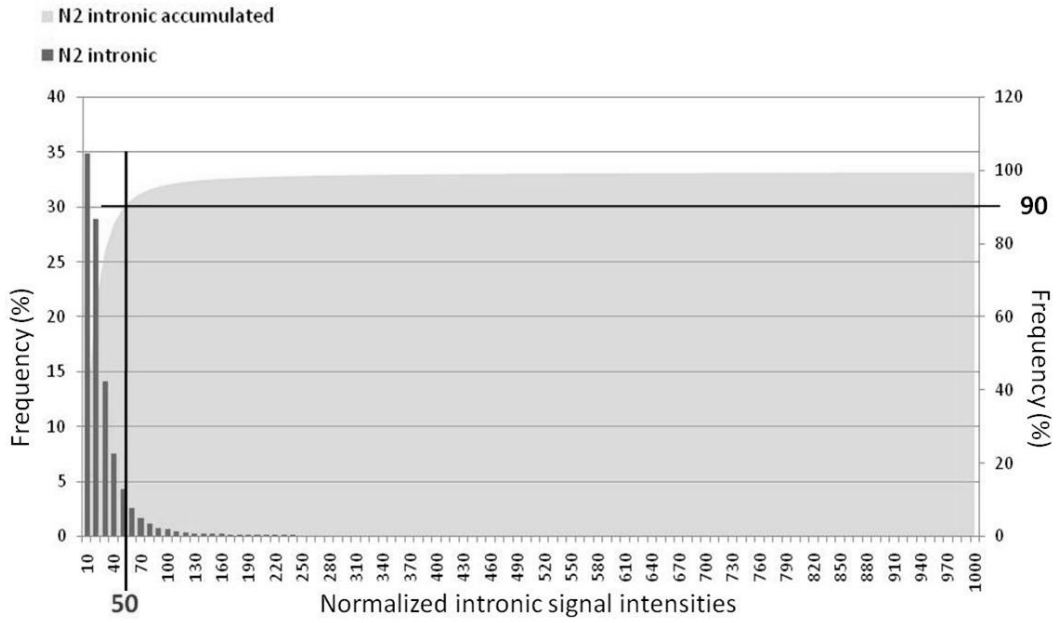


Figure 4.7. The distribution of the intensities of intronic probes.

Dark grey columns show the frequency of probes with intensities of a certain range (bin size=10). Light grey area shows the accumulated frequency of probes. 90% intronic probes have signal intensities <=50.

19 genes were chosen based on their computational expression levels, which are listed in Table 4.7 (All primers used for amplifying these genes are listed in Table D.1). Figure 4.8 showed the results of the gel electrophoresis. Genes with expression levels over 50 all have clear bands, but were only vaguely discernable or completely missing on the gel when under 50. This result is consistent with 50 as being the background level, and expression levels under 50 should be treated as un-expressed.

4.8 qPCR verification of genes with different expression levels

Table 4.3 listed the genes with differential expression; however, genes B0511.1 and F28F8.7 in Table 4.3a, and genes CE7X_3.2, Y113G7A.6, Y39E4B.13, B0513.6 and ZK1025.4 in Table 4.3b, all have expression levels less than the background (50), so these "un-expressed" genes were all deleted from the list of "genes with different expression" and will not be subjected to qPCR verification. Upon this, 18 genes showed higher expression in *teg-4* mutants (Table 4.3, in red) and 24 showed higher expression in *teg-4* wild-type animals (Table 4.3, in black).

4.8.1 Reference gene selection

Reference genes are those whose expression levels stay constant in different conditions, and they were frequently used when performing relative quantification in the qPCR experiment. Two ways of selecting reference genes were employed: 1) well-known housekeeping genes from other organisms, and 2) genes that have shown unchanged and relatively high expression level in computational analysis across all 4 datasets. By doing so, 4 candidates were chosen (Table 4.8), and they were all tested for expression levels

Table 4.7 Genes selected for doing background determination.

name	expression level (in N2)
C38D9.2	12
B0511.11	10
Y37B11A.3	19
C02C2.3	23
C35A5.2	20
F55D10.1	36
F58D5.4	49
C24G7.2	32
F58G1.7	31
C13B9.4	32
Y46C8AL.3	58
D1044.1	53
ZK550.2	70
Y74C9A.5	92
T08H4.1	84
T01G6.7	200
T24A11.2	221
Y75B8A.28	354
C26C6.5	165

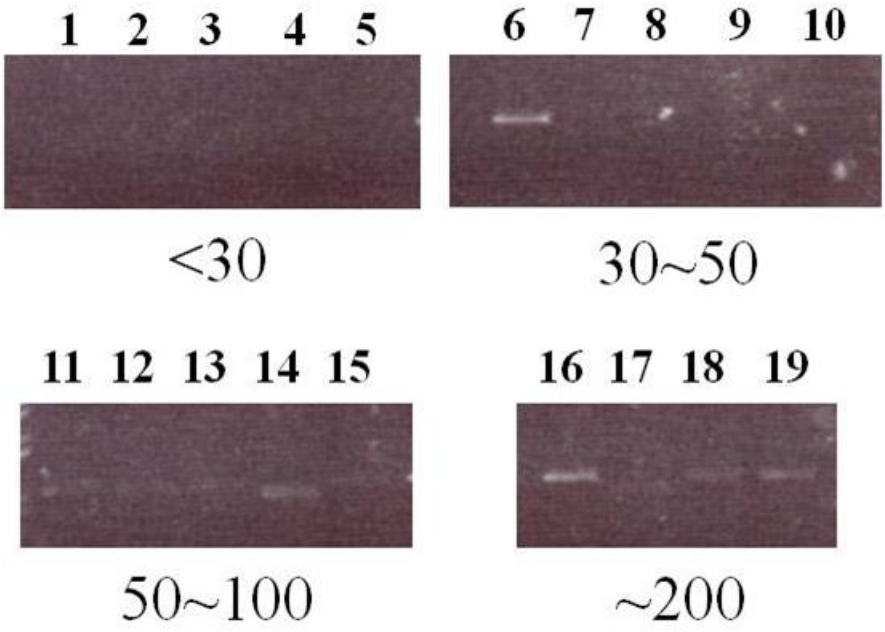


Figure 4.8. Agarose gel (1.0%) images of 19 genes (1 to 19).
Genes with predicted expression level less than 30 are not visible on the gel (top left). For genes with predicted expression levels between 30 and 50, only one (gene 6) shows visible band (top right). Genes with predicted expression levels over 50 all show visible bands (bottom, gene 11 to gene 19)

Table 4.8 Expression levels (calculated through computational analysis) of selected reference gene candidates.

Gene	computational expression level			
	N2	*teg-4*	*smg-2*	*teg-4 smg-2*
gpd-1	338	534	26	119
gpd-2	5037	4535	4725	4031
unc-15	3355	5129	5662	6148
unc-54	3737	6857	7021	7660

using qPCR (All primers used for amplifying these selected genes are listed in Table D.2).

Based on qPCR results, primers for amplifying *gpd-2* and *unc-54* formed primer dimers; and the expression level of housekeeping gene *gpd-1* was not the same in all four samples (Table 4.8, first row). Only gene *unc-15* remained consistently expressed in both predicted and experimental results (Table 4.8, data not shown); therefore, *unc-15* was selected as the reference gene for all qPCR experiments.

4.8.2 qPCR verification results

qPCR was first conducted on all 42 genes, but due to the requirement for high quality primers for qPCR, only half of them produced valid results. Figures 4.9A and 4.9B display the fold changes in both qPCR and computational results for all tested genes. Figure 4.9C plots the fold change values from both methods on a scatter chart. These results exhibit a high degree of similarity between the computational analysis and qPCR verification, validating the methods used in this analysis. Therefore, this strongly supports the computationally analyzed results.

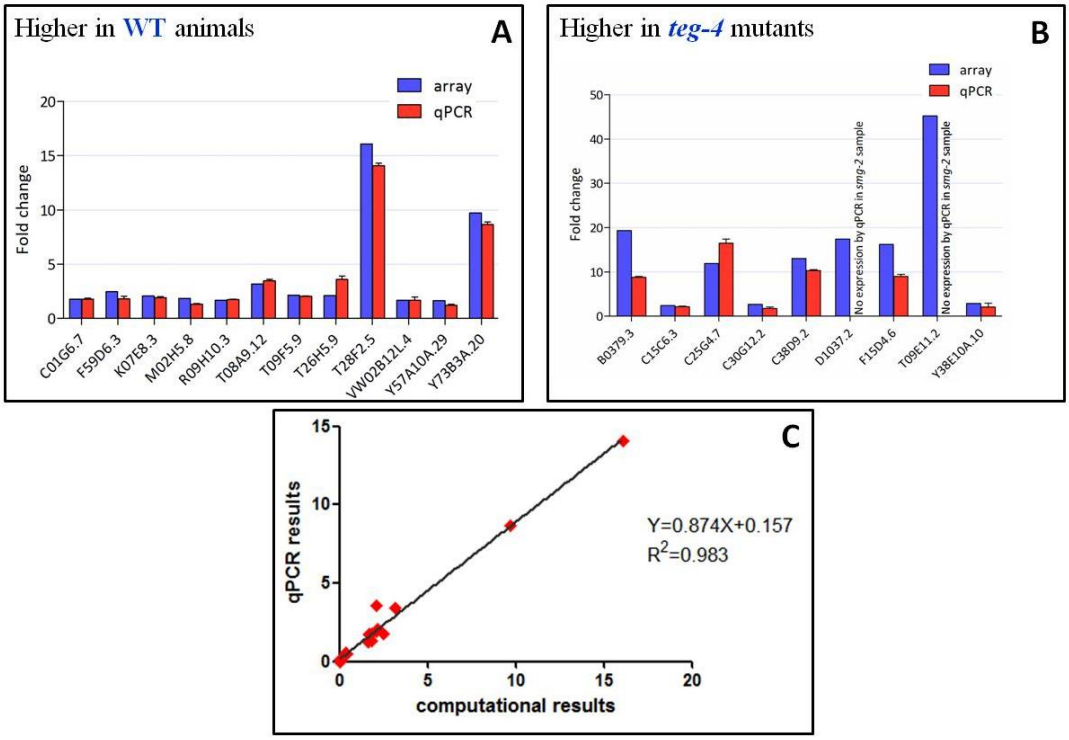
4.9 Verification of potential splicing defects

4.9.1 Candidate selection

Because of the heterogeneity of the tiling array, unless the splicing defect was one big exon/intron spliced/included entirely (which was not found), the splicing defect candidates all seem very subtle. It became quite difficult to select the “true” positive splicing defects solely through computational prediction. In other words, none of these

Figure 4.9. Comparison between the results of computational analysis and qPCR assays on differently expressed genes.

(A) Fold changes of genes that have higher expression levels in *teg-4* wild-type (WT) animals, including both N2 and *smg-2*. (B) Fold changes of genes that have higher expression levels in *teg-4* mutants, including both *teg-4* and *teg-4 smg-2*. Fold changes are represented as "higher expression level/lower expression level" for each gene included in (A) and (B) (names are listed under the X axis). Blue columns are the results from the arrays (tiling array analysis) and red columns are the results from the qPCR assays. For the gene D1037.2 and T09E11.2 in (B), their expression could not be detected in *smg-2* (*teg-4* WT) using qPCR assay. For all the genes included in (A) and (B), the value of "expression level in *teg-4* WT/ expression level in *teg-4* mutants" were calculated for both computational analysis and qPCR assay. Both results were plotted against each other (C, red dots). A trendline with linear regression was generated. The equation and R squared value are also displayed.


Higher in WT animals
A
array
qPCR
Fold change
20
15
10
5
0
C01G6.7
F59D6.3
K07E8.3
M02H5.8
R09H10.3
T08A9.12
T09F5.9
T26H5.9
T28F2.5
VW02B12L.4
Y57A10A.29
Y73B3A.20
Higher in teg-4 mutants
B
array
qPCR
Fold change
50
40
30
20
10
0
No expression by qPCR in smg-2 sample
No expression by qPCR in smg-2 sample
B0379.3
C15C6.3
C25G4.7
C30G12.2
C38D9.2
D1037.2
F15D4.6
T09E11.2
Y38E10A.10
C
qPCR results
computational results
Y=0.874X+0.157
R2=0.983
15
10
5
0
0
5
10
15
20

candidates really stood out. Using a series of filtering approaches (Appendix C), the list of candidates was narrowed down substantially. However, a large number of candidates remained and it was impractical to verify all of them. With the above stated considerations, the verification procedure involved randomly selecting certain candidates for verification. 45 candidates were selected for verification. Selected candidates are highlighted in grey in Table 4.4, Table 4.5 and Table D.6 (All primers used for identifying splicing defects are listed in Table D.4).

4.9.2 Splicing defect verification

Figure 4.10 summarizes the strategy for detecting splicing problems. Generally, if splicing errors were suspected to occur in a specific region, primers were designed on flanking exons. The amplicon size was limited to between 100 bp and 200 bp. This region was then PCR amplified (using cDNA) and visualized on an agarose gel (Figure 4.10A).

If splicing problems indeed occurred at a particular location, even though this potential problematic region is relatively small, a high percentage agarose gel (3.5%) can be used to differentiate between bands with only a small difference in size Figure 4.10B shows the gel image of two positive controls, each of which contains two DNA products, 20 bps different in length. Both bands are clearly discriminated. The resolution of the *C. elegans* Tiling Array is 25 bp. Therefore, theoretically, a splicing defect detected by tiling array can also be identified using this approach.

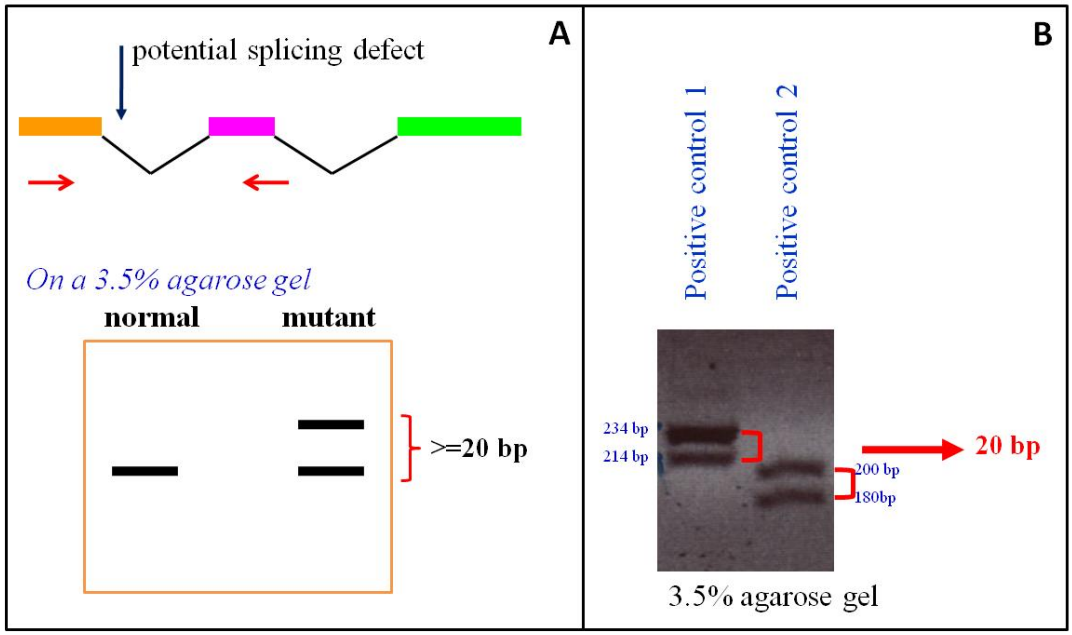


Figure 4.10. The strategy for splicing defects verification.

Primers were designed on flanking exons to amplify suspect regions with splicing defects (A, top). PCR products from normal (no splicing defects) and mutant (with splicing defects) animals have different band patterns on 3.5% agarose gel (A, bottom). Two positive controls (see Table D.5) were conducted to validate this strategy. Each of them contains two DNA products, 20bp difference in length. They are PCR amplified and visualized on 3.5% agarose gel (B).

Unfortunately, experimental verification did not confirm the predicted splicing defects in any of these 45 selected candidates. Figure 4.11 is a sample gel image of the result, on which the band pattern of each gene for all four strains appear to be the same.

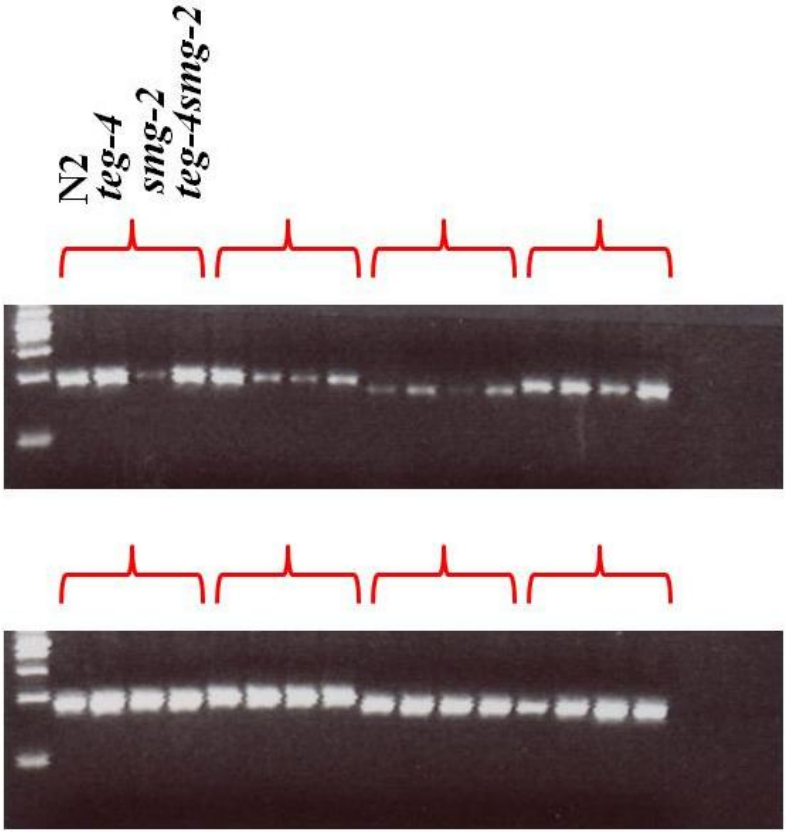


Figure 4.11. Results of verification of splicing defects from selected candidates. A total of 45 candidates with potential splicing defects were tested, using the strategy described in section 4.9.2 (Figure 4.9). Specifically, PCR amplification was performed for all four strains: N2, *teg-4*, *smg-2* and *teg-4 smg-2*. 3.5% agarose gel was used for visualizing the results.

Chapter Five: Functional analysis of potential targets

5.1 Introduction

According to the results from computational analysis (Chapter four), 42 genes showed different expression in *teg-4* mutants, but no confirmed splicing defects were identified. Therefore, this chapter will focus on further characterization of the differentially expressed genes, whose functions may involve regulation of the germline proliferation vs. differentiation balance.

5.2 RNAi screen

In order to study the functions of these candidate genes, an RNAi screen was performed. Due to the lack of certain vectors in the lab's RNAi library, RNAi knockdown was only conducted on 30 out of 42 genes. Two strains were used as sensitized genetic backgrounds; *glp-1(ar202gf)* and *teg-4(oz210); glp-1(ar202gf)*. *glp-1(ar202gf)* animals exhibit a slight over-proliferative germline phenotype and was used to search for genes that can enhance this phenotype. *teg-4(oz210); glp-1(ar202gf)* animals have tumorous germ lines, and they were used to identify genes that can suppress the tumor.

In the initial RNAi screen, the aim was to primarily identify candidates that show tumor enhancement or suppression; no phenotypic quantification was performed. RNAi of two genes stood out in this screen; RNAi of F14B6.6 caused tumor enhancement, whereas RNAi on C38D9.2 suppressed tumor formation.

In order to determine specifically to what extent these genes affect germline proliferation more in-depth, two more RNAi knockdown experiments of both genes were

conducted, with scoring of the germline phenotypes, as presented in Table 5.1. RNAi on F14B6.6 greatly increased the percentage of animals with tumorous germ lines from less than 27.8% in control animals (*glp-1(ar202gf)*) to 61.2% in RNAi knockdown animals (Table 5.1a). In *teg-4(oz210); glp-1(ar202gf)* animals subjected to C38D9.2 RNAi, the portion of animals with non-tumorous germ lines increased from 7.55% (in control animals) to 26.12% (in RNAi knockdown animals) (Table 5.1b, Figure 5.1). Subsequent experiments primarily focused on characterizing C38D9.2.

5.3 C38D9.2 RNAi on the *teg-4(oz210); glp-1(ar202gf)*

RNAi on C38D9.2 significantly suppressed the *teg-4(oz210)*; *glp-1(ar202gf)* germline tumor, to the degree that some animals began to produce several eggs (Table 5.2, "with embryos"), while 7.41% of the animals even looked completely wild-type (Table 5.2, "wild-type"), with fully functional germ line and no apparent tumor.

To further study this effect, non-tumorous animals were continually transferred to fresh RNAi plates, resulting in *teg-4(oz210); glp-1(ar202gf); C38D9.2 (RNAi)* animals being maintained. (*teg-4(oz210)*; *glp-1(ar202gf)* are normally maintained as heterozygotes). Germline phenotypes of these 'RNAi maintained animals' were scored (Table 5.3) and after several generation exposure to RNAi, the percentage of animals with non-tumorous germ line was boosted to 50.94%. *teg-4(oz210)*; *glp-1(ar202gf)* animals not exposed to C38D9.2 (RNAi) have less than 10% animals with non-tumorous germ lines (data not shown).

Table 5.1 Results of the RNAi of F14B6.6 (a, on *glp-1(ar202gf)* animals) and C38D9.2 (b, on *teg-4(oz210); glp-1(ar202gf)* animals).

a

	% of tumorous germ line	*n*
L4440 RNAi control	27.80%	394
F14B6.6 RNAi	61.20%	328

b

	% of non-tumorous germ line*	*n*
L4440 RNAi control	7.55%	583
C3BD9.2 RNAi	26.12%	513

* Non-tumorous germ line means embryos are clearly visible in the germ line

Table 5.2 Results of the C38D9.2 RNAi on the *teg-4(oz210); glp-1(ar202gf)*

	Germline phenotypes				
RNAi vector	% of tum[a]	% of wild-type[b]	% of with embryos[c]	% of total non-tum[d]	*n*
L4440	92.45%	1.37%	6.18%	7.55%	583
GFP	87.27%	2.27%	10.45%	12.73%	440
C38D9.2	73.88%	7.41%	18.71%	26.12%	513
F15D4.5	81.75%	4.56%	13.69%	18.25%	767

[a] tum: tumorous germ lines

[b] These germ lines are non-distinguishable with the wild-type germ lines

[c] Over-proliferation is still present in these germ lines, but embryos were also produced

[d] The total of the "% of wild-type" and the "% of with embryos"

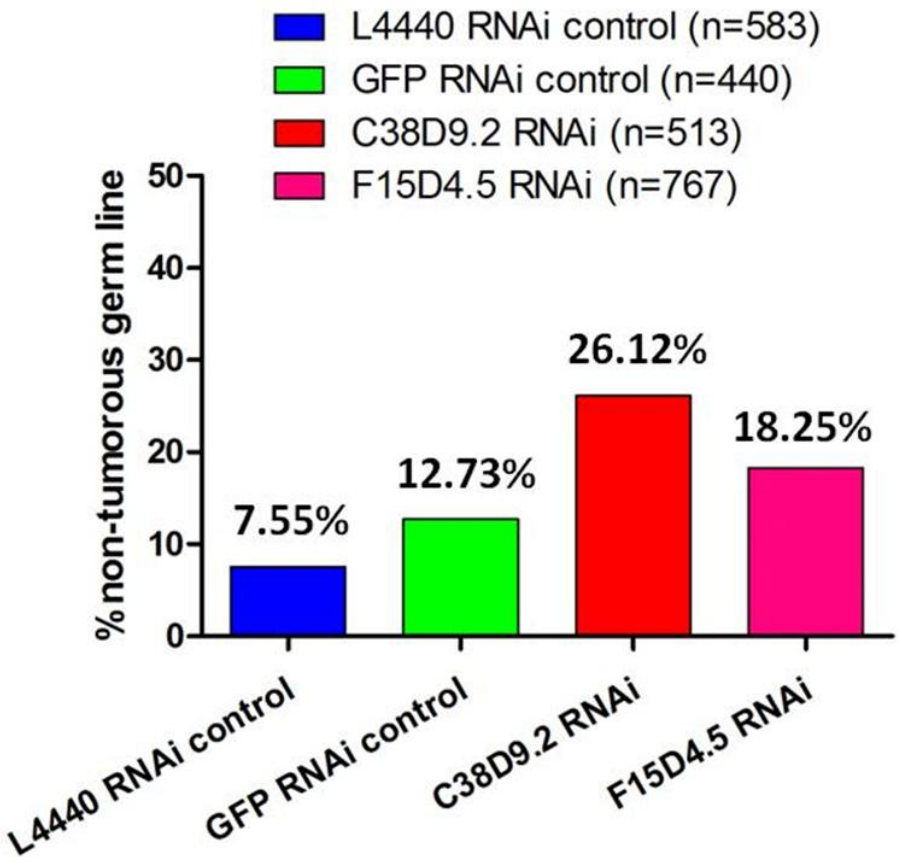


Figure 5.1. Results of RNAi on *teg-4(oz210); glp-1(ar202gf)* animals.
RNAi knockdown of C38D9.2 and F15D4.5 on *teg-4(oz210); glp-1(ar202gf)* animals was performed. Two controls were used by performing GFP (RNAi) and L4440 (RNAi). Percentage of non-tumorous (see Table 5.2 for detailed definition) animals were shown on the top of each column.

Table 5.3 Phenotypes of the *teg-4(oz210); glp-1(ar202gf); C38D9.2 (RNAi)* (*teg-4(oz210); glp-1(ar202gf)* animals were maintained on the C38D9.2 (RNAi) plates).

% of tum[a]	% of wild-type[b]	% of with embryos[c]	% of total non-tum[d]	*n*
49.06%	18.24%	32.70%	50.94%	159

[a] tum: tumorous germ lines

[b] These germ lines are non-distinguishable with the wild-type germ lines

[c] Over-proliferation is still present in these germ lines, but embryos were also produced

[d] The total of the "% of wild-type" and the "% of with embryos"

5.4 C38D9.2 RNAi on other tumors

Because TEG-4 is a potential splicing factor, it is plausible that C38D9.2 knockdown may affect other splicing factor related tumors. Two strains were selected for this purpose: *prp-17(oz273)*; *glp-1(oz264gf)* and *teg-1(oz230) unc-32(e189) glp-1(ar202gf)*. PRP-17 is a newly identified splicing factor in *C. elegans*, which can form synthetic tumors with *glp-1(oz264gf)* when mutated (Kerins et al., 2010); TEG-1 is another potential splicing factor that, when mutated, can form a tumor with *glp-1(ar202gf)* (Wang and Hansen, unpublished data). However, C38D9.2 RNAi did not suppress the tumors in either of these two strains (Table 5.4).

In order to determine if the expression level of C38D9.2 plays a role in tumor formation of other synthetic tumors, which have no obvious connection to splicing, C38D9.2 (RNAi) was performed on *gld-1(q485) gld-2(q497)* (Hansen et al., 2004b) and *rrf-1(pk1417)*; *puf-8(oz192)* ;*glp-1(oz264gf)* (Racher, 2010). Both strains have tumorous germ line. The results showed that C38D9.2 (RNAi) did not suppress the tumor in these two strains (data not shown).

5.5 C38D9.2 RNAi on N2 and *teg-4(oz210)* animals

In order to determine whether knocking down the expression of C38D9.2 can have an effect on the wild-type animals and *teg-4* single mutants, C38D9.2 RNAi was also performed on N2 and *teg-4(oz210)* animals as control experiments. However, no obvious phenotypic changes were observed in either N2 or *teg-4(oz210)* animals treated with C38D9.2 (RNAi) (data not shown).

Table 5.4 Results of the C38D9.2 (RNAi) on *teg-1(oz230) unc-32(e189) glp-1(ar202gf)* and *prp-17(oz273); glp-1(oz264gf)* animals

		Germline phenotypes		
Background genotype	RNAi vector	% of tum[a]	% of total non-tum[b]	*n*
teg-1(oz230) unc-32(e189) glp-1(ar202)	L4440	100.00%	0.00%	335
	GFP	100.00%	0.00%	258
	C38D9.2	100.00%	0.00%	428
	F15D4.5	100.00%	0.00%	238
prp-17(oz273); glp-1(oz264gf)	L4440	100.00%	0.00%	180
	GFP	100.00%	0.00%	131
	C38D9.2	100.00%	0.00%	223
	F15D4.5	100.00%	0.00%	220

[a] tum: tumorous germ lines

[b] non-tum: germ lines that have embryos

5.6 F15D4.5

Bioinformatic analyses indicate that C38D9.2 is not conserved across species and only exists in nematodes. However, C38D9.2 has a paralog, F15D4.5, which is 49% identical to C38D9.2 (Figure 5.2). F15D4.5 was also identified in the computational analysis as an up-regulated gene in *teg-4* mutants (Table 4.3), but was not identified during the initial RNAi screen.

To test whether F15D4.5 can act similarly to C38D9.2, RNAi knockdown of F15D4.5 was performed on *teg-4(oz210)*; *glp-1(ar202gf)*, *prp-17(oz273)*; *glp-1(oz264gf)* and *teg-1(oz230) unc-32(e189) glp-1(ar202gf)*. F15D4.5 did possess the ability to suppress the *teg-4(oz210)*; *glp-1(ar202gf)* tumor, although not as strong as that of C38D9.2. (only 18.25% of the animals are non-tumorous, Table 5.2, Figure 5.1). As with C38D9.2, F15D4.5 RNAi was unable to suppress the tumors in *prp-17(oz273)*; *glp-1(oz264gf)* and *teg-1(oz230) unc-32(e189) glp-1(ar202gf)* animals (Table 5.4).

5.7 *teg-4*, *glp-1* and C38D9.2 triple mutant

A deletion allele of C38D9.2 (*ok1853*) was available from the CGC (*C. elegans* Genetic Center), in which the second exon was deleted in-frame (Figure 5.3). A triple mutant *teg-4(oz210); glp-1(ar202gf); C38D9.2(ok1853)* was constructed to study whether this strain will show a similar germline phenotype as seen in the *teg-4(oz210); glp-1(ar202gf);* C38D9.2 (RNAi) animals. The result was very intriguing. At 20°C, 46.5% of the adult homozygotes are non-tumorous. 10.35% of the animals have wild-type looking germ lines (Table 5.5). They were indistinguishable from the heterozygotes (see section 2.2.2 for more strain information) at least under the dissection scope.

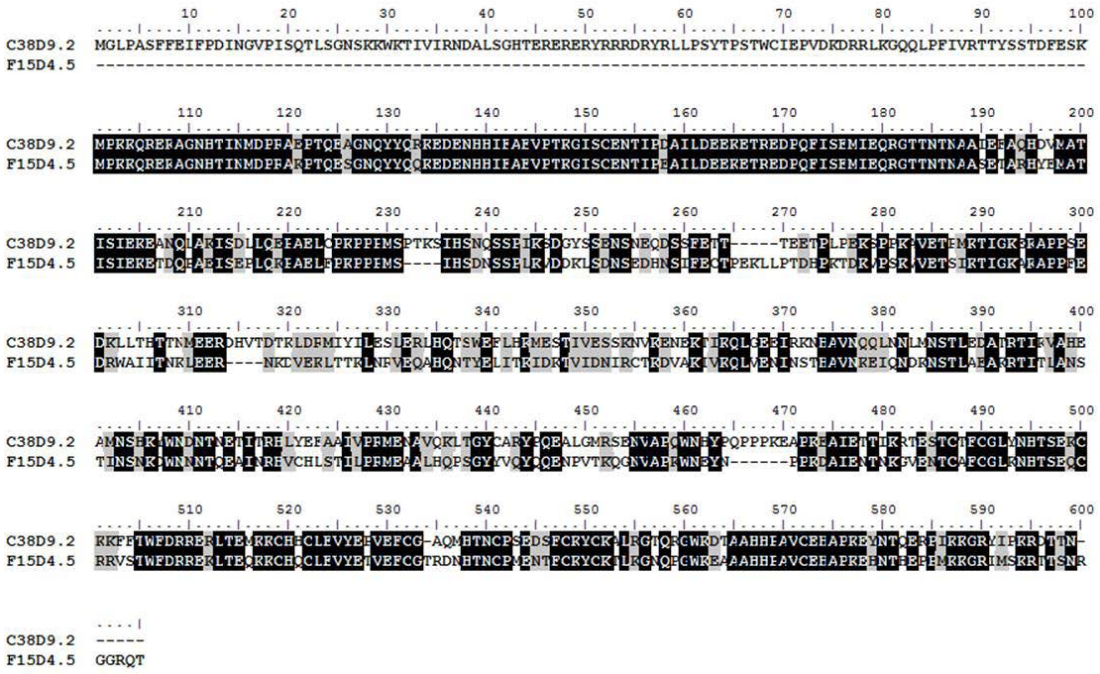


Figure 5.2. An alignment of the protein of C38D9.2 and F15D4.5.

The amino acid sequences of the gene C38D9.2 and F15D4.5 are aligned against each other. C38D9.2 and F15D4.5 are 49% identical (highlighted in black).

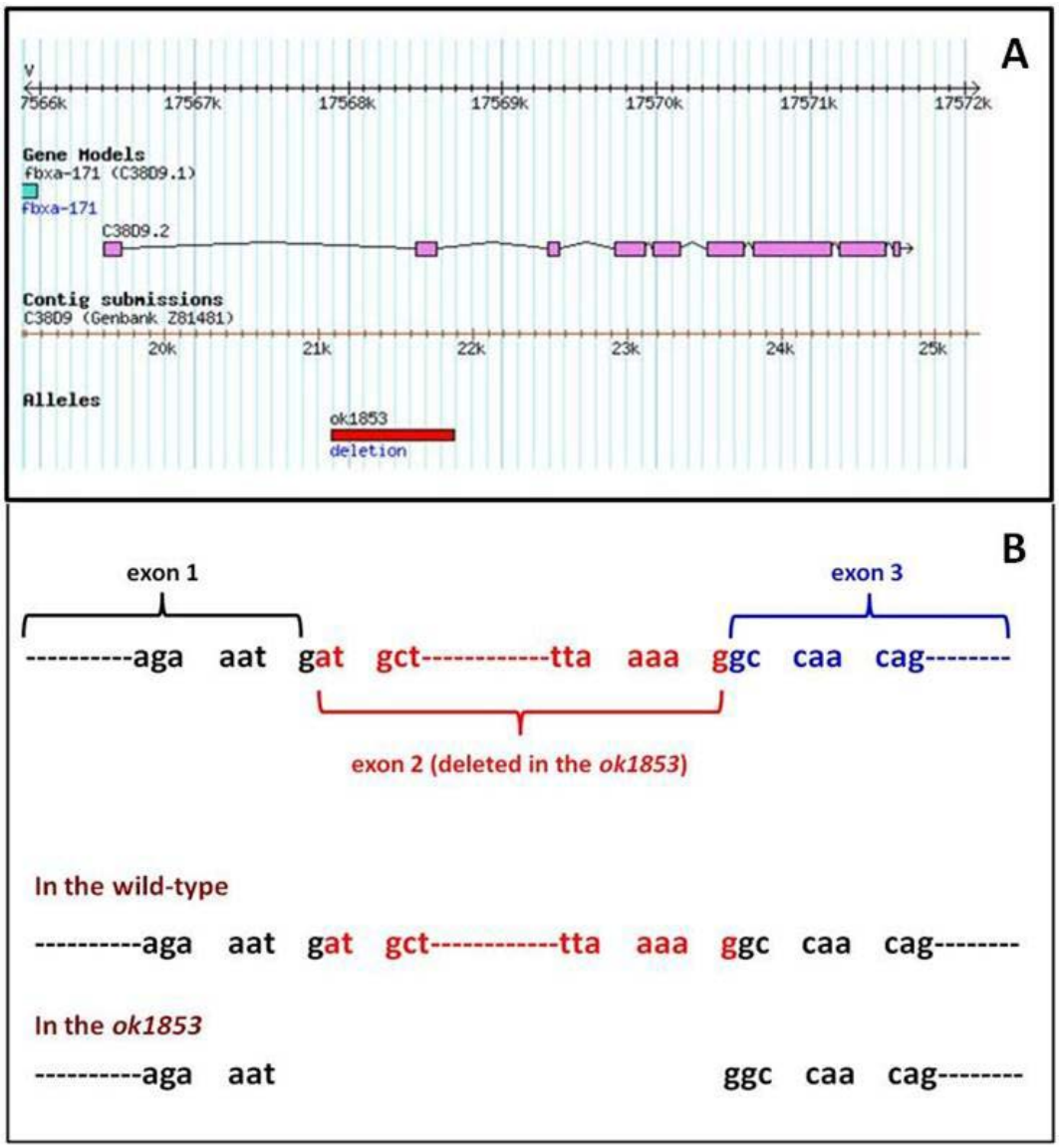


Figure 5.3. *ok1853* deletion does not cause a frame shift in the gene C38D9.2. (A) The gene structure of C38D9.2 and the location of the *ok1853* deletion. Picture adapted from http://www.wormbase.org/db/gene/gene?name=WBGene00008010;class=Gene. (B) The resulting sequence changes (cDNA) caused by the *ok1853* deletion. The second exon was removed in-frame.

Table 5.5 phenotypes of the triple mutant *teg-4(oz210); glp-1(ar202gf); C38D9.2(ok1853)*

% of tum[a]	% of wild-type[b]	% of with embryos[c]	% of total non-tum[d]	*n*
53.50%	10.35%	36.15%	46.50%	1043

[a] tum: tumorous germ lines

[b] These germ lines are non-distinguishable with the wild-type germ lines

[c] Over-proliferation is still present in these germ lines, but embryos were also produced

[d] The total of the "% of wild-type" and the "% of with embryos"

Chapter Six: Discussion and conclusion

The gene *teg-4*, which encodes a homolog of the human splicing factor SAP130, has been identified as a regulator of the balance between proliferation and differentiation in the *C. elegans* germ line (Mantina et al., 2009). This work further investigates how *teg-4* is involved in this regulation by searching for potential target genes. Through a series of computational analyses on tiling array data, and other genetic approaches, three genes whose overall expression levels were affected by a *teg-4* mutation have been shown to affect germline tumors.

6.1 Genes regulated by *teg-4*

Tiling array analysis identified 42 genes with different expression levels in a *teg-4* loss of function mutant (Table 4.3). Functional analysis by RNAi knockdown identified three of them as being involved in controlling germline proliferation; two with a suppression effect (C38D9.2 and F15D4.5), and the other, with an enhancement effect (F14B6.6).

6.1.1 C38D9.2

C38D9.2 RNAi showed strong suppression of the *teg-4(oz210); glp-1(ar202gf)* tumor; it did not act similarly on other splicing factor related tumors, nor did it repress the *glp-1(ar202gf)* tumor. This suggests that the interaction between *teg-4* and C38D9.2 is crucial for the tumor suppression and that the enhancement of *glp-1(ar202gf)* tumor is probably due, in part, to up-regulation of C38D9.2 in a *teg-4* mutant (Table 4.3).

Therefore, by knocking down C38D9.2 expression, *teg-4(oz210)* is unable to enhance *glp-1(ar202gf)* over-proliferation to the same extent (Figure 6.1).

Triple mutant *teg-4(oz210); glp-1(ar202gf); C38D9.2(ok1853)* animals also show a reduction in animals with a tumorous germ line, and the extent of tumor suppression was much higher than that of the C38D9.2(RNAi) on *teg-4(oz210); glp-1(ar202gf)*. This was unexpected, since in C38D9.2*(ok1853)* only a very small exon (129bp) is deleted, and does not cause a frame-shift. *C38D9.2(ok1853)* partially suppresses the tumorous phenotype of *teg-4(oz210); glp-1(ar202gf)*, indicating that this exon must contain sequences important for C38D9.2's function with respect to the regulation of proliferation in the germ line.

C38D9.2 has never been characterized in *C. elegans*, nor does it have any sequence homologous in species outside of nematodes, however, one previous study did identify that C38D9.2 can be regulated by two other genes; *deps-1* and *rde-3*. According to the genome wide microarray analysis of this study, C38D9.2 was up-regulated several hundred-fold in *deps-1* and *rde-3* mutants (Spike et al., 2008). DEPS-1 is a newly identified P-granule-associated protein, and *deps-1* mutants display defects in germ cell proliferation (Spike et al., 2008). *rde-3 (mut-2)* encodes a potential poly-A polymerase that is homologous to GLD-2; it has a role in the RNAi pathway (Chen et al., 2005a) and meiotic chromosome segregation (Hammond et al., 2001).

C38D9.2 (RNAi) and C38D9.2 (*ok1853*) can both suppress the *teg-4(oz210); glp-1(ar202gf)* tumor and *deps-1* and *rde-3*, which have germ line-related functions, can both regulate C38D9.2. Together, this suggests that C38D9.2 may have a role in regulating germ cell proliferation/differentiation. However, these results are very preliminary and it

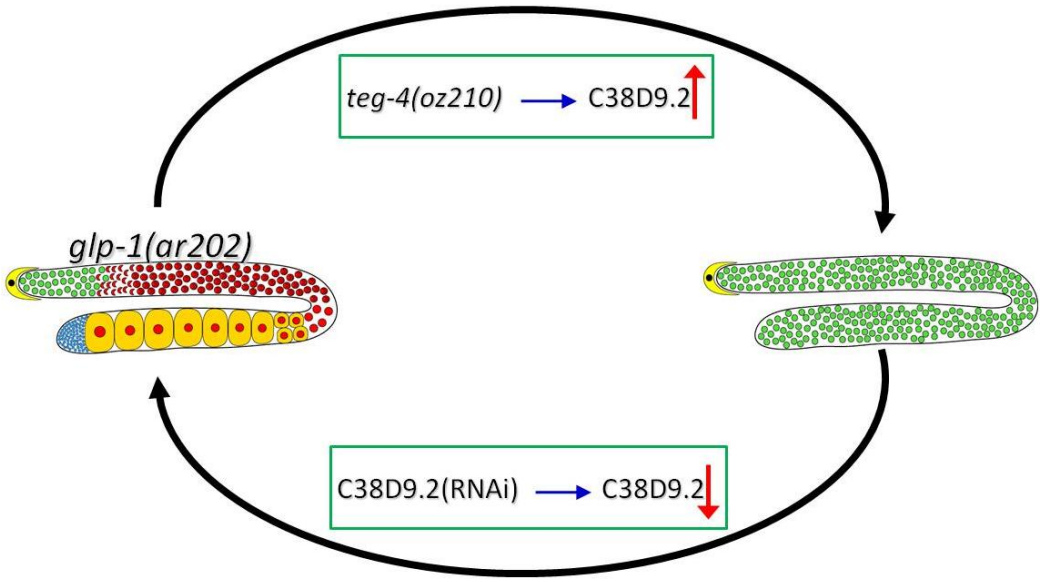


Figure 6.1. C38D9.2 (RNAi) suppresses the *teg-4(oz210); glp-1(ar202gf)* tumor. The enhancement of *glp-1(ar202gf)* over-proliferation by *teg-4(oz210)* is partially due to increased expression level of C38D9.2 in *teg-4(oz210)* animals. When knocking down C38D9.2 (C38D9.2 [RNAi]) in *teg-4(oz210); glp-1(ar202gf)* animals, the germline tumor is suppressed.

is too difficult to draw any conclusions at this time. Moreover, both tiling array analysis and a qPCR assay have confirmed that C38D9.2 is up-regulated in *teg-4* mutants, suggesting that *teg-4* normally functions to inhibit C38D9.2, but whether this inhibition is direct or indirect is unclear. The interaction between C38D9.2 and *teg-4* is undoubtedly important in tumor control, but the detailed mechanism is not understood, and it is likely that other components are involved in this process.

6.1.2 F14B6.6

Another interesting candidate identified by the tiling array and RNAi experiments is F14B6.6, which, like C38D9.2, is up-regulated in *teg-4* mutants (Table 4.3). However, RNAi knockdown of F14B6.6 and C38D9.2 exerted opposite effects on germline proliferation; F14B6.6 RNAi caused a tumor enhancement in *glp-1(ar202gf)*. Given that *teg-4(oz210)* can enhance *glp-1(ar202gf)* overproliferation, if *teg-4(oz210)* really caused an up-regulation of F14B6.6 (this was not verified by qPCR assay due to primer issues), a more likely consequence of F14B6.6 RNAi would be tumor repression instead of tumor enhancement, as illustrated in Figure 5.2A. Moreover, *glp-1(ar202gf)* itself can be a problem; even before F14B6.6 RNAi knockdown, the percentage of tumorous animals in this strain was already very high (27.8% in L4440 RNAi control, Table 5.1a). This makes the finding that reduced F14B6.6 level can enhance the *glp-1(ar202gf)* tumor not as profound of a result as that of the C38D9.2 RNAi experiment (in L4440 RNAi control, only 7.55% are non-tumorous, Table 5.1b). It is certainly possible that both the up-regulation of this gene in *teg-4(oz210)* and the tumor enhancement of the RNAi were

true; however, there must be other factors involved to explain this seemingly paradoxical result (Figure 6.2B).

F14B6.6 is a predicted galactosyltransferase, and glycosylation has important roles in regulating Notch receptors. The Notch receptor consists of three conserved domains: an extracellular domain (ECD), a transmemebrane domain and an intracellular domain. The ECD undergoes extensive glycosylation during Notch receptor synthesis, and this modification is essential for the proper interaction between the Notch receptor and its ligands (Fortini, 2009). The ECD contains 29-36 tandem epidermal growth factor-like (EGF) repeats, which are modified by three types of O-linked glycosylation: O-fucosylation, O-glucosylation and O-GlcNAcylation (Matsuura et al., 2008; Moloney et al., 2000; Whitworth et al., 2010). O-fucosylation has been the most extensively studied. Each EGF repeat has six cysteine residues (C_1 to C_6), and usually O-fucose is added to Ser/Thr (Ser: serine, Thr: threonine) between the second and the third cysteines (Shao and Haltiwanger, 2003). O-fucose can be elongated by sequential addition of three other sugar residues; N-acetylglucosamine (GlcNAc), galactose (Gal) and sialic acid (SA). However, not all EGF-like repeats are added with fully extended O-fucose (tetrasaccharides), and some contain only mono- or disaccharides (Moloney et al., 1997; Moloney et al., 2000). These residues are transferred by different glycosyltransferases (fucosyltransferase, galacosyltransferase, etc). Mutations in glycosyltransferase genes can change the glycosylation status and affect the proper function of Notch receptors, resulting in aberrant Notch signaling (Rampal et al., 2007).

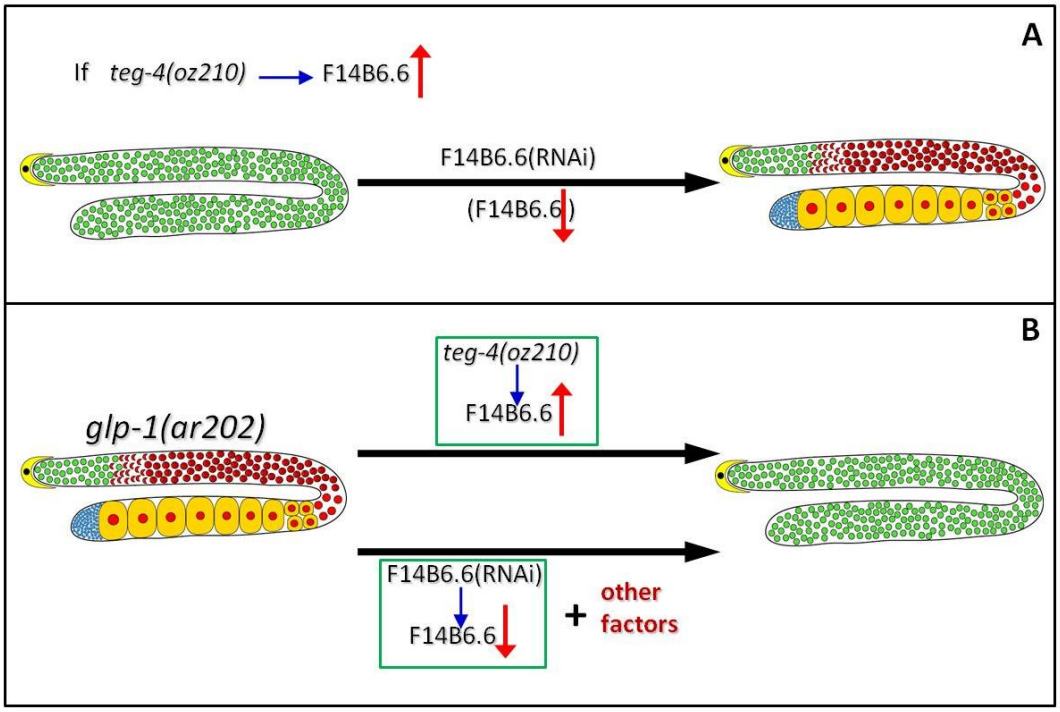


Figure 6.2. F14B6.6 (RNAi) enhances *glp-1(ar202gf)* over-proliferation.
If F14B6.6 expression level is increased in the *teg-4(oz210)*, knocking down F14B6.6 expression in the *teg-4(oz210); glp-1(ar202gf)* would more likely cause tumor suppression (A). In fact F14B6.6 (RNAi) in the *glp-1(ar202gf)* resulted in enhancement of the over-proliferation, thus more genes are involved to cause this effect (B).

6.2 Splicing and the control of the germ cell proliferation/differentiation balance

Because TEG-4 is a splicing factor, the initial goal of this work was to identify TEG-4's targets with splicing problems. However, no major splicing defects were identified. Although computational analysis predicted thousands of minor potential splicing defects, of all the tested 45 candidates, none could be confirmed.

6.2.1 Does teg-4(oz210) *cause splicing defects?*

Splicing is a fundamental process in living organisms. Significantly knocking down the function of some crucial splicing factors will cause serious splicing abnormalities of a wide spectrum of genes; major housekeeping functions will then be abrogated, and cell survival is threatened. RNAi of *teg-4* confirmed this. Significantly knocking down *teg-4* activity by RNAi leads to embryonic or larval (L1) arrest in both N2 and *glp-1(ar202gf)* backgrounds (Mantina et al., 2009). However, *teg-4(oz210)* displays a less severe phenotype; the *teg-4(oz210)* single mutant is largely fertile, and *teg-4(oz210); glp-1(ar202gf)* has tumorous germ line.

The *teg-4(oz210)* mutation only changes one amino acid of the TEG-4 protein (a glycine to an aspartic acid substitution). The exact molecular consequences caused by *oz210* (partial loss of function mutation) on *teg-4* is still largely unknown. Based on the results of this work, there are at least two possible explanations: 1) *teg-4(oz210)* reduces TEG-4's splicing efficiency and the splicing of one or more genes was affected, and that our techniques were not sensitive enough to detect them; or 2) *teg-4(oz210)* disrupts TEG-4's functions other then splicing, triggering other kinds of defects (such as

expression levels) on its targets. Both possibilities could lead to the tumorous germline phenotype seen in *teg-4(oz210); glp-1(ar202gf)*.

If *teg-4(oz210)* animals do have impaired splicing efficiency, it is plausible to conclude that one or more of its targets were not properly spliced. But those mis-splicing events must not be widespread, since even for *teg-4(oz210) smg-2(e2008)* animals, in which the NMD pathway is compromised, and presumably more mis-spliced transcripts could be accumulated, they are still viable. It is more likely that mis-splicing is restricted to certain genes that are more sensitive to splicing efficiency changes. These genes may have specific roles in germ cell regulation, and their disrupted splicing can induce the over-proliferation phenotype.

This work did not identify any major splicing defects. However, a great number of minor potential splicing defects were identified through tiling array analysis. Although mis-splicing was not verified in any of the selected 45 genes, and many of the rest are likely false positives, it remains possible that some might be true splicing errors, and they just have not been tested yet.

6.2.2 Other splicing factors indentified in C. elegans

In *C. elegans*, other splicing factors have also been identified as having functions in the regulation of germ line development.

mog genes

The *mog* genes are a group of genes that are essential for sex determination in the *C. elegans* germ line and includes six members; *mog-1* to *mog-6* (Gallegos et al., 1998; Graham and Kimble, 1993; Graham et al., 1993). Additionally, they are all synthetically

required with *gld-2* or *gld-3* for proper function of the mitosis/meiosis switch (Belfiore et al., 2004; Kasturi et al., 2010; Puoti and Kimble, 1999, 2000; Zanetti et al., 2011). All MOG genes are homologous to splicing factors. MOG-1, MOG-4 and MOG-5 are closely related to yeast splicing factors PRP16, PRP2 and PRP22, respectively (Puoti and Kimble, 2000). MOG-2 is a homolog of the vertebrate spliceosomal protein U2A' (Zanetti et al., 2011), MOG-3 shares conserved motifs with the yeast splicing factor cwc25 (Kasturi et al., 2010). As well, MOG-6 and human CYP-60 are 46% identical (Belfiore et al., 2004).

There are other splicing factors, in addition to *mog* genes that, when mutated, cause defects in the *C. elegans* germ line. Two of them are DDX-23 and PRP-17. They are homologous to the yeast splicing factors PRP28 and CDC40, respectively and both genes have been implicated in either germline sex determination or regulation of the mitosis/meiosis switch (Kerins et al., 2010; Konishi et al., 2008).

Mutations (loss of function or null) in the above splicing factors presumably cause splicing efficiency reduction, and lead to the abnormal germline phenotypes observed in the mutants. However, except for *mog-2*, no confirmed splicing defects in the rest of the mutants have been identified (Belfiore et al., 2004; Kasturi et al., 2010; Puoti and Kimble, 1999). A deletion (*q75*) in *mog-2* caused splicing problems in two genes: *ama-1* and *mog-1*, both of which retained an intron in *mog-2(q75)*. Failure of removing the second intron in *mog-1* resulted in a truncated protein that deletes conserved motifs required for MOG-1 function (Zanetti et al., 2011).

In the study of PRP-17 by Kerins *et al.* (Kerins et al., 2010), although no attempt was conducted for identifying downstream splicing defects caused by reduction of PRP-

17 splicing efficiency, the researchers did perform an RNAi screen of 114 *C. elegans* genes that code for orthologs of yeast and human splicing factors. They discovered that RNAi of 47 genes exhibited a tumor enhancement on different sensitized genetic backgrounds. These genes are not confined within a distinct splicing step, but function throughout the entire splicing process (Kerins et al., 2010). This suggested that the robustness of the general splicing process is essential for maintaining a proliferation/differentiation balance in the germ line, disruption in any splicing step can lead to germ cell over-proliferation (Kerins et al., 2010). They also performed another RNAi screen of several core RNA polymerase II and core ribosomal protein genes, but did not observe any germline related abnormalities. This indicated that the germline proliferation/differentiation phenotype was unlikely due to a change in general levels of gene expression, but were the result of splicing defects on certain genes (Kerins et al., 2010).

6.2.3 Do splicing factors have non-splicing functions?

Although the research on *mog-2* and PRP-17 strongly suggested that a reduction on splicing efficiency is the cause of mutated germline phenotype, splicing defects have been difficult to detect in splicing factor mutants that influence the proliferation/differentiation decision. It is possible that this specific phenotype is due to non-splicing causes. The RNA recognition and binding ability of splicing factors have enabled them to be multifunctional, and many of the processes they participate in are unrelated to splicing.

Auxiliary splicing factor U2AF65 has been reported to have functions related to transcription (Listerman et al., 2006) and apoptosis (Izquierdo, 2008). Using a yeast two-hybrid system, fourteen proteins have been identified that interact with U2AF65, four of which have functions other than splicing (Prigge et al., 2009). The splicing factor, SF2/ASF, has been confirmed to have a role in sumoylation. Its RNA recognition domain has made it both necessary and sufficient for the enhancement of this protein modification process (Pelisch et al., 2010).

TEG-4's human ortholog SAP130 is not only a component of SF3b (see section 1.6.1), but also a member of DDB1 (Damaged DNA-Binding Protein) family proteins (Das et al., 1999), a subunit of TFTC (TATA-binding protein [TBP]-free TBP-associated-factor [TAF_{II}] complex) (Brand et al., 2001). SAP130 can associate with the HAT (histone acetyltransferase) complex STAGA (SPT3-TAF_{II}-GCN5L acetylase), which has roles in chromatin modification and transcription-coupled processes (Martinez et al., 2001). However, the specific function of SAP130 within these complexes is still unknown. Recently, in addition to being identified as a target of several antitumor drugs (Folco et al., 2011; Kaida et al., 2007; Kotake et al., 2007), SAP130 has also been suggested to have a role in CRLs (Cullin-RING ubiquitin E3 ligases) mediated ubiquitination (Menon et al., 2008).

Although the significance of the non-splicing functions of splicing factors have not been fully determined, it is possible that they are involved in the regulation of the germline proliferation/differentiation decision, directly or indirectly.

6.3 Technologies for the detection of splicing defects

The detection of splicing defects has always been a challenge. Traditional methods, such as northern blot analysis, were only suitable when a few selected candidate-genes needed to be tested for different transcripts. However, such an approach cannot be applied to study the whole transcriptome. Recently, the advent of the genome-wide tiling array and high throughput RNA-sequencing (RNA-seq) technologies has enabled the large scale identification of alternative splicing.

This work utilized tiling array technology to study the *C. elegans* transcriptome. Tiling arrays use oligonucleotide probes that cover the entire non- repetitive genome. Although the relatively high resolution makes the recognition of mis-spliced introns/exons possible, the complexity of the tiling signal and the lack of well-developed software/program makes the tiling data very difficult to interpret and extract valuable information from, especially when performing customized analysis.

The identification of differently expressed genes was successful. On the contrary, the search for splicing defects was not as successful. Although thousands of potential mis-splicing events were identified, the results from the experimental verification suggested that the accuracy was not high enough. In the *C. elegans* genome, most introns are as small as 50 nucleotides (Blumenthal and Steward, 1997), and almost all introns identified in this study were 50 nucleotides or smaller. Since the tiling array probe is 25bp long, 50 nucleotides means that, at most, two probes (when perfectly aligned) are located within this region. This explains, at least partially, the high proportion of false positives.

High throughput RNA-seq is a newly emerging technology. It generates data at a single-base resolution, which provides great accuracy. Because of this, RNA-seq has also been used in surveying alternative splicing in several studies (Bainbridge et al., 2006; Pan et al., 2008b; Sultan et al., 2008). Making customized arrays is another option. For this work, if the probes that are specifically mapped to intronic or splice-junction regions can be specially designed, some of the sequences overlooked by using the tiling array can be identified.

However, each technology has its advantages and disadvantages. Tiling array provides great coverage, but its resolution is limited by the probe resolution (for *C. elegans* array, sequences less than 25bp in size cannot be detected). When facing low expression levels, tiling data also has a high false-positive rate (van Bakel et al., 2010). The data from RNA-seq is quite accurate, but to cover the whole transcriptome is very costly; most sequencing approaches were performed on certain cells/tissues. Customized design of probes requires specific considerations and can be time consuming. Perhaps the best way of performing a large-scale identification of splicing problems is by combining different approaches, so that the shortcomings of each technology will be complemented. Ramani *et al* used both RNA-seq and Tiling array (Ramani et al., 2009) or specially designed alternative splicing microarray (Ramani et al., 2011) to study the splicing changes during different developing stages in *C. elegans*. The generated results were of both high-coverage and high-accuracy.

6.4 Conclusion

Splicing factors have drawn great attention in cancer research because they are the targets of several antitumor drugs (see section 1.7). These drugs can impair the *in vivo* splicing of several genes by modulating the activity of splicing factors (Hasegawa et al., 2011). However, the molecular link between the mis-splicing of these genes and the drugs' antitumor properties is still missing. Whether splicing is involved in tumor regulation, or if other roles of splicing factors are more crucial, remains unclear. More studies are needed to unravel this mystery.

In *C. elegans*, a partial loss of function mutation in *teg-4* causes phenotypes similar to those seen in other splicing factor mutants; including over-proliferation in the germ line. Since mutations in many splicing factors show this same phenotype, splicing likely plays an important role in regulating proliferation in the germ line. However, the search for splicing defects in these mutants has not been successful. This work uses tiling arrays as a tool to search for mRNAs mis-spliced in *teg-4* mutants, but instead of discovering confirmed splicing defects, three genes with altered expression levels were identified. Therefore, the splicing defects are likely subtle, and the technique used could only identify downstream effects.

The formation of a tumor is a unique process. Whether splicing or non-splicing functions of splicing factors lead to this phenotype, the defects must be very specific and are unlikely to be large-scale alterations of general cellular process. Both the research on other splicing factors and this work confirmed this statement. If splicing defects are present in mutated animals, they are more likely restricted to a limited number of targets (which also explains why the search of these defects is so difficult).

As mentioned above, the over-proliferative germline phenotype has been observed in a large number of splicing factor mutants, including *teg-4*. This is not a coincidence. Therefore, although major splicing defects were not identified, it is likely splicing errors caused by the *teg-4* mutation resulted in downstream effects on C38D9.2. A simple model is proposed here (Figure 6.3). Mutation in the gene *teg-4* and other splicing factors leads to a reduction of general splicing efficiency, causing mis-splicing in certain genes (gene X). Gene X may participate in a non-splicing process, such as transcription. The splicing problems can disrupt the transcript of gene X (*i.e.* introduce a premature termination codon), and its function is compromised. This may result in changes of expression levels in the downstream genes (*i.e.* C38D9.2), which eventually triggers the over-proliferation in the germ line.

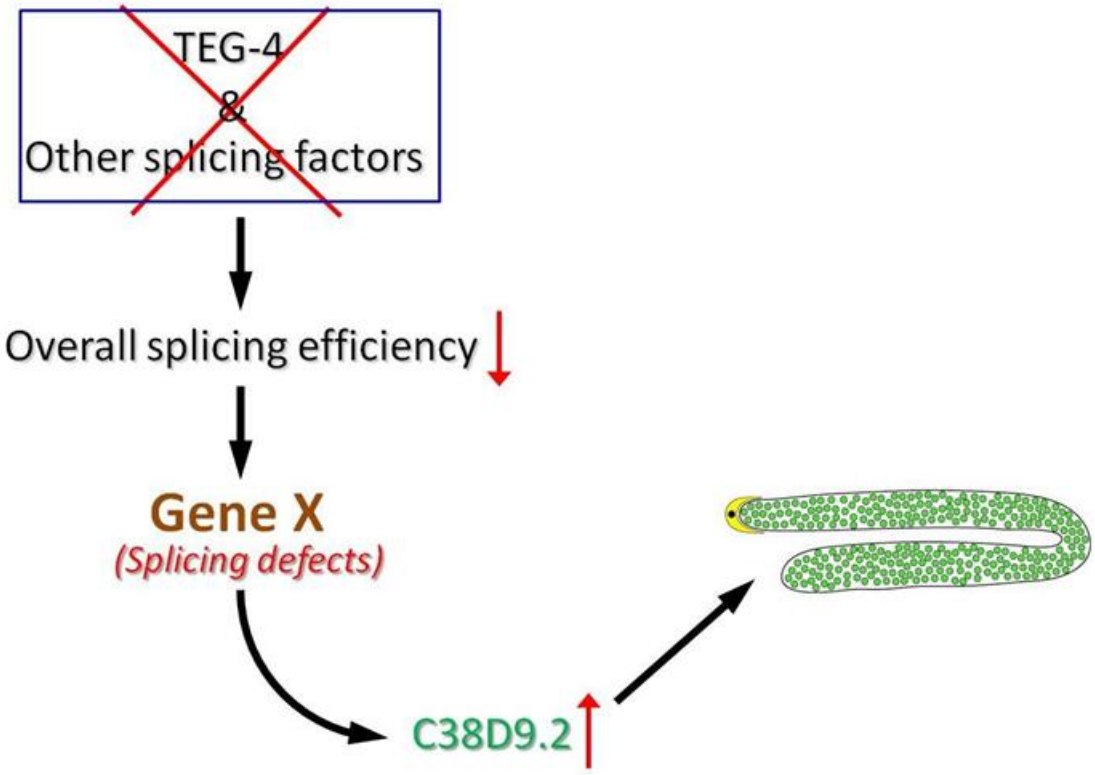


Figure 6.3. A proposed model for TEG-4 and other splicing factors to cause over-proliferation in the *C. elegans* germ line.

Mutations in the gene *teg-4*, or other splicing factors, can lead to an overall reduction of splicing efficiency. Compromised splicing efficiency induces splicing defects on certain genes (Gene X) that are more sensitive to the changes in splicing efficiency. These genes have non-splicing functions. Mis-spliced transcripts of Gene X disrupt its function. This changes the expression level of downstream targets (*i.e.* expression level of C38D9.2 is increased), resulting in germ line over-proliferation.

APPENDIX A: MAJOR SELF-WRITTEN R SCRIPTS

A.1. Gene annotation files generation (with splicing variants)

```
####the following scripts used Chromosome I for demonstration, and they can be applied
to the rest five chromosomes####
setwd("")  #set path for files to be read ("filename.txt")
#read files (Coding genes with variants)
V.ChrI<-read.table("filename.txt",header=T,sep="\t")

#extract information from different columns of "V.ChrI"
GeneName<-as.character(V.ChrI[["SequenceNameGene"]])
CodingStart<-V.ChrI[["CodingStartbp"]]
CodingEnd<-V.ChrI[["CodingEndbp"]]
SequenceNameGene<-V.ChrI[["SequenceNameGene"]]
SequenceNameTranscript<-V.ChrI[["SequenceNameTranscript"]]

GeneNameU<-GeneName[!duplicated(GeneName)]  #return a character vector with
                                            #unique elements

#######while loop A#######
#find the "start" and "end" positions for those common regions in all splicing variants
Final.Start<-c()
Final.End<-c()
```

```
i=1
while(i<=length(GeneNameU)){
Start<-CodingStart[SequenceNameGene==GeneNameU[i]]
End<-CodingEnd[SequenceNameGene==GeneNameU[i]]
Transcript<-
as.character(SequenceNameTranscript[SequenceNameGene==GeneNameU[i]])
levels<-levels(factor(Transcript))
x<-length(levels)

   #####while loop B#####
   Range<-c()
   j=1
   while (j<=length(Start)){
   Range<-c(Range,Start[j]:End[j])
   j=j+1
   }
   #####END of while loop B #####

StartU<-sort(Start[!duplicated(Start)],decreasing=FALSE)
EndU<-sort(End[!duplicated(End)],decreasing=FALSE)
Sort<-sort(Range)

a1<-is.element(Sort,StartU)
```

```
b1<-Sort[which(a1)]
a2<-is.element(Sort,EndU)
b2<-Sort[which(a2)]

   #####while loop C1#####
   k=1
   ha<-c()
   while(k<=length(StartU)){
   ha<-c(ha,length(which(b1==StartU[k])))
   k=k+1
   }
   #####END of while loop C1#####
F.Start<-StartU[which(ha==x)]

   #####while loop C2#####
   kk=1
   hb<-c()
   while(kk<=length(EndU)){
   hb<-c(hb,length(which(b2==EndU[kk])))
   kk=kk+1
   }
   #####END of while loop C2#####
F.End<-EndU[which(hb==x)]
```

```
Final.Start<-c(Final.Start,F.Start)

Final.End<-c(Final.End,F.End)

i=i+1

}

#######END of while loop A #######


FixedStart<-Final.Start

FixedEnd<-Final.End


##########Create the final table##########

#select the corresponding gene exons with the above start and end positions

TemTable<-data.frame(GeneName,CodingStart)

UTable<-V.ChrI[!duplicated(TemTable),]


SequenceNameGene<-UTable[["SequenceNameGene"]]

Strand<-UTable[["Strand"]]

ExonName<-UTable[["ExonName"]]

CodingStart<-UTable[["CodingStartbp"]]


p<-which(is.element(CodingStart,FixedStart))


SequenceNameGene<-SequenceNameGene[p]
```

```
Strand<-Strand[p]

ExonName<-ExonName[p]

NewCodingStart<-CodingStart[p]


#######while loop D#######

NewCodingEnd<-c()

n=1

while(n<=length(NewCodingStart)){

NewCodingEnd<-c(NewCodingEnd,FixedEnd[FixedStart==NewCodingStart[n]])

n=n+1

}

#######END of while loop D#######


FixedVariantTable<-

data.frame(SequenceNameGene,Strand,ExonName,NewCodingStart,NewCodingEnd)


#output

write.table(FixedVariantTable,file="FixedVariantChrI.txt",sep="\t")
```

A.2. Intron annotation files generation (with splicing variants)

```
#---------------------------------PART 1---------------------------------#
## extract location information of all exons and introns (with variants) from ##primary
annotation file WS170 (using Chromosome I for demonstration)
setwd("")   #set path for file to be read ("WS170ChrI.txt ")
####file reading (primary annotation file WS170)####
ChrI<-read.table("WS170ChrI.txt",header=T,sep="\t")

Table.intron<-
ChrI[,c("SequenceNameGene","SequenceNameTranscript","Strand","ExonName","Intro
nStartbp","IntronEndbp")]    #only extract intron related information
Table.exon<-
ChrI[,c("SequenceNameGene","SequenceNameTranscript","Strand","ExonName","Exon
Startbp","ExonEndbp")]    #only extract exon related information

#sequence name information from Table.intron
Name.intron<-as.character(Table.intron[["SequenceNameGene"]])
Transcript.intron<-as.character(Table.intron[["SequenceNameTranscript"]])
#sequence name information from Table.exon
Name.exon<-as.character(Table.exon[["SequenceNameGene"]])
Transcript.exon<-as.character(Table.exon[["SequenceNameTranscript"]])

x.intron<-Name.intron==Transcript.intron
```

```
VariantGeneTable.intron<-Table.intron[!x.intron,]
x.exon<-Name.exon==Transcript.exon
VariantGeneTable.exon<-Table.exon[!x.exon,]

IntronVariant<-
VariantGeneTable.intron[!is.na(VariantGeneTable.intron[["IntronStartbp"]]),c("Sequence
NameGene","Strand","ExonName","IntronStartbp","IntronEndbp")]
#output
write.table(IntronVariant,file="VariantIntronChrI.txt",sep="\t")  #include intron location
                                                                  # information
write.table(VariantGeneTable.exon,file="VariantExonChrI.txt",sep="\t")   #include exon
                                                                         #location
                                                                         #information
        #----------------------------END of PART 1----------------------------#
        #---------------------------------PART 2---------------------------------#
## create properly annotated intron (with variants) files(using the Chromosome I for
##demonstration)
setwd("")   #set path for files "VariantIntronChrI.txt " and "VariantExonChrI.txt"
####files reading####
VI.ChrI<-read.table("VariantIntronChrI.txt",header=T,sep="\t")
E.ChrI<-read.table("VariantExonChrI.txt",header=T,sep="\t")
TableV<-VI.ChrI
TableE<-E.ChrI
```

```
#define various vectors/dataframes
TemTable<-TableV[,c("SequenceNameGene","IntronStartbp","IntronEndbp")]
TableVN<-TableV[!duplicated(TemTable),]

GeneName<-as.character(TableVN[["SequenceNameGene"]])
GeneNameU<-GeneName[!duplicated(GeneName)]

IntronStart<-TableVN[["IntronStartbp"]]
IntronEnd<-TableVN[["IntronEndbp"]]
GeneIntron<-as.character(TableVN[["SequenceNameGene"]])
IntronName<-as.character(TableVN[["ExonName"]])

ExonStart<-TableE[["ExonStartbp"]]
ExonEnd<-TableE[["ExonEndbp"]]
GeneExon<-as.character(TableE[["SequenceNameGene"]])

##########Define function “FUNa” ##########
FUNa<-function(x,y){
  if(x<=length(GeneNameU)){
  StartE<-ExonStart[GeneExon==GeneNameU[x]]
  EndE<-ExonEnd[GeneExon==GeneNameU[x]]
        i=1
        RangeE<-c()
```

```
        while(i<=length(StartE)){
        RangeE<-c(RangeE,StartE[i]:EndE[i])
        i=i+1
        }
  StartI<-IntronStart[GeneIntron==GeneNameU[x]]
  EndI<-IntronEnd[GeneIntron==GeneNameU[x]]
  L<-length(StartI)
    if(y<=L){
    RangeI<-StartI[y]:EndI[y]
      if(!any(!is.element(RangeI,RangeE))){
      return(IntronName[GeneIntron==GeneNameU[x]][y])
      }else{
      return("needed")
      }
    }else{
    return("y should<=L")
    }
  }else{
  return("x should<=length(GeneNameU)")
  }
}
##########END of FUNa##########
j=1
```

```
Needed<-TableVN[1,]
Unwanted<-TableVN[1,]
while(j<=length(GeneNameU)){
StartI<-IntronStart[GeneIntron==GeneNameU[j]]
    k=1
    while(k<=length(StartI)){
        if(FUNa(j,k)=="needed"){
        Needed<-
rbind(Needed,TableVN[IntronName==(IntronName[GeneIntron==GeneNameU[j]][k]),])
        }else{
        Unwanted<-rbind(Unwanted,TableVN[IntronName==FUNa(j,k),])
        }
    k=k+1
    }
j=j+1
}
NeededIntronTable<-Needed[-1,]     # required for the following
UnwantedIntronTable<-Unwanted[-1,]   #not required for the following
#################################################################
#define various vectors/dataframes
NeededIntronS<-NeededIntronTable[["IntronStartbp"]]
NeededIntronE<-NeededIntronTable[["IntronEndbp"]]
Gene<-as.character(NeededIntronTable[["SequenceNameGene"]])
```

```
IName<-as.character(NeededIntronTable[["ExonName"]])
GeneU<-Gene[!duplicated(Gene)]

##########Define function “FUNb” ##########
FUNb<-function(x,y){
    if(x<=length(GeneU)){
    StartE<-ExonStart[GeneExon==GeneU[x]]
    EndE<-ExonEnd[GeneExon==GeneU[x]]
        i=1
        RangeE<-c()
        while(i<=length(StartE)){
        RangeE<-c(RangeE,StartE[i]:EndE[i])
        i=i+1
        }
    StartI<-NeededIntronS[Gene==GeneU[x]]
    EndI<-NeededIntronE[Gene==GeneU[x]]
    L<-length(StartI)
      if(y<=L){
      RangeI<-StartI[y]:EndI[y]
      a<-RangeI[!is.element(RangeI,RangeE)]
      b<-(a[-1]-a[-length(a)])!=1
      n<-which(b)
      Final<-sort(c(a[1],a[n],a[n+1],a[length(a)]),decreasing=FALSE)
```

```
      return(Final)
      }else{
      return("y should<=L")
      }
    }else{
    return("x should<=length(GeneU)")
    }
}
##########END of FUNb##########
ii=1
NIntronTable<-NeededIntronTable[,c(-4,-5)]
IStart<-c()
IEnd<-c()
S<-NIntronTable[1,]
NameMultiple<-c()
while(ii<=length(GeneU)){
StartI<-NeededIntronS[Gene==GeneU[ii]]
  kk=1
  while(kk<=length(StartI)){
        if(length(FUNb(ii,kk))==2){
        IStart<-c(IStart,FUNb(ii,kk)[1])
        IEnd<-c(IEnd,FUNb(ii,kk)[2])
        S<-rbind(S,NIntronTable[IName==(IName[Gene==GeneU[ii]][kk]),])
```

```
        Single<-cbind(S[-1,],IStart,IEnd)
        }else{
        NameMultiple<-c(NameMultiple,IName[Gene==GeneU[ii]][kk])
        }
  kk=kk+1
  }
ii=ii+1
}
Tem<-Single[,c(4,5)]
NoDuplicatedSingle<-Single[!duplicated(Tem),]     #required in the end
#############################################################################
jj=1
Multiple<-NeededIntronTable[1,]
while(jj<=length(NameMultiple)){
Multiple<-rbind(Multiple,NeededIntronTable[IName==NameMultiple[jj],])
jj=jj+1
}

MultipleTable<-Multiple[-1,]      #additional modifications are to be made
#############################################################################
GeneM<-as.character(MultipleTable[["SequenceNameGene"]])
GeneMU<-GeneM[!duplicated(GeneM)]
MIntronS<-MultipleTable[["IntronStartbp"]]
```

```
MIntronE<-MultipleTable[["IntronEndbp"]]
MIntronName<-as.character(MultipleTable[["ExonName"]])
StrandM<-MultipleTable[["Strand"]]

##########Define function “FUNcStart” and “FUNcEnd” ##########
#FUNcStart#
FUNcStart<-function(x){
  if(x<=length(MIntronName)){
  MStartExon<-ExonStart[GeneExon==GeneM[x]]
  MEndExon<-ExonEnd[GeneExon==GeneM[x]]
        i=1
        RangeE<-c()
        while(i<=length(MStartExon)){
        RangeE<-c(RangeE,MStartExon[i]:MEndExon[i])
        i=i+1
        }
  RangeI<-MIntronS[x]:MIntronE[x]
  a<-RangeI[!is.element(RangeI,RangeE)]
  b<-(a[-1]-a[-length(a)])!=1
  n<-which(b)
  Final<-sort(c(a[1],a[n],a[n+1],a[length(a)]),decreasing=FALSE)
  number<-1:length(Final)
  Start<-Final[number%%2==1]
```

```
  return(Start)
  }else{
  return("x should <=length(MIntronName)")
  }
}

#FUNcEnd#
FUNcEnd<-function(x){
  if(x<=length(MIntronName)){
  MStartExon<-ExonStart[GeneExon==GeneM[x]]
  MEndExon<-ExonEnd[GeneExon==GeneM[x]]
        i=1
        RangeE<-c()
        while(i<=length(MStartExon)){
        RangeE<-c(RangeE,MStartExon[i]:MEndExon[i])
        i=i+1
        }
  RangeI<-MIntronS[x]:MIntronE[x]
  a<-RangeI[!is.element(RangeI,RangeE)]
  b<-(a[-1]-a[-length(a)])!=1
  n<-which(b)
  Final<-sort(c(a[1],a[n],a[n+1],a[length(a)]),decreasing=FALSE)
  number<-1:length(Final)
```

```
  End<-Final[number%%2==0]
  return(End)
  }else{
  return("x should <=length(MIntronName)")
  }
}
##########END of  "FUNcStart" and "FUNcEnd" ##########

iii=1
MulGene<-c()
MulStrand<-c()
MulIntronName<-c()
MulIntronStart<-c()
MulIntronEnd<-c()

while(iii<=length(MIntronName)){
MulGene<-c(MulGene,rep(GeneM[iii],length(FUNcStart(iii))))
MulStrand<-c(MulStrand,rep(StrandM[iii],length(FUNcStart(iii))))
MulIntronName<-c(MulIntronName,rep(MIntronName[iii],length(FUNcStart(iii))))
MulIntronStart<-c(MulIntronStart,FUNcStart(iii))
MulIntronEnd<-c(MulIntronEnd,FUNcEnd(iii))
iii=iii+1
}
```

```
TemRepTable<-
cbind(MulGene,MulStrand,MulIntronName,MulIntronStart,MulIntronEnd)

##change the names of elements in vector "MulIntronName"
#function “NUMBER”#
NUMBER<-function(x){
  if(x<=length(MIntronName)){
  return(length(FUNcStart(x)))
  }else{
  return("x should <= length(MIntronName)")
  }
}

#function “Change”#
Change<-function(x){
  if(x<=length(MIntronName)){
  n=NUMBER(x)
  Mul<-MulIntronName[MulIntronName==MIntronName[x]]
     change<-c()
     for(m in 1:n){
     change<-c(change,paste(Mul[m],m,sep="."))
     }
```

```
  return(change)
  }else{
  return("x should <= length(MIntronName)")
  }
}

iiii=1
NewIntronName<-c()
while(iiii<=length(MIntronName)){
NewIntronName<-c(NewIntronName,Change(iiii))
iiii=iiii+1
}

MulTable<-cbind(MulGene,MulStrand,NewIntronName,MulIntronStart,MulIntronEnd)
## “MulTable” contains the correct Intron names and Start and End Region
#####################################################################
#combine MulTable and NoDuplicatedSingle
names(NoDuplicatedSingle)<-
c("SequenceNameGene","Strand","FixedExonName","FixedStart","FixedEnd")
colnames(MulTable)<-
c("SequenceNameGene","Strand","FixedExonName","FixedStart","FixedEnd")
Fixed<-rbind(NoDuplicatedSingle,MulTable)
```

```
Sequence<-as.character(Fixed[["SequenceNameGene"]])
StrandX<-Fixed[["Strand"]]
Intron<-as.character(Fixed[["FixedExonName"]])
StartX<-as.numeric(Fixed[["FixedStart"]])
EndX<-as.numeric(Fixed[["FixedEnd"]])

Order<-order(Sequence,StrandX,Intron,StartX,EndX)

SequenceNameGene<-Sequence[Order]
Strand<-StrandX[Order]
FixedExonName<-Intron[Order]
FixedStart<-StartX[Order]
FixedEnd<-EndX[Order]

FixedVariantIntron1<-
data.frame(SequenceNameGene,Strand,FixedExonName,FixedStart,FixedEnd)
Tem<-data.frame(FixedStart,FixedEnd)
FixedVariantIntronTable<-FixedVariantIntron1[!duplicated(Tem),]
#output
write.table(FixedVariantIntronTable,file="FixedVariantIntronChrI.txt",sep="\t")
```

#----------**END of PART 2**----------#

A.3. Exon annotation files generation (with splicing variants)

```
#---------------------------------PART 1---------------------------------#
##obtain original information of coding genes with splicing variants from WS170 (using
##Chromosome I for demonstration)
setwd("")   #set path for file to be read ("WS170ChrI.txt ")
####file reading (primary annotation file WS170)####
ChrI<-read.table("WS170ChrI.txt",header=T,sep="\t")

#remove NA from CodingStartbp and corresponding entries
noNAs<-ChrI[!is.na(ChrI["CodingStartbp"]),]

#generate genes with and without splicing variants
NameGene<-noNAs[["SequenceNameGene"]]
SequenceNameGene<-as.character(NameGene)

NameTranscript<-noNAs[["SequenceNameTranscript"]]
SequenceNameTranscript<-as.character(NameTranscript)

x<-SequenceNameGene==SequenceNameTranscript

VariantGeneTable<-noNAs[!x,]
NoVariantGeneTable<-noNAs[x,]
```

```
#output
write.table(NoVariantGeneTable,file="CodingNoVariantChrI.txt",sep="\t")
write.table(VariantGeneTable,file="CodingVariantChrI.txt",sep="\t")
#"CodingVariantChrI.txt" is required for PART 2
            #----------------------------END of PART 1----------------------------#
            #---------------------------------PART 2---------------------------------#
## create properly annotated exon (with variants) files(using the Chromosome I for
##demonstration)
setwd("")   #set path for files "CodingVariantChrI.txt " and "FixedVariantIntronChrI.txt"
####files reading####
V.ChrI<-read.table("CodingVariantChrI.txt",header=T,sep="\t")
FVI.ChrI<-read.table("FixedVariantIntronChrI.txt",header=T,sep="\t")
Table<-V.ChrI
TableF<-FVI.ChrI

#define various vectors/dataframes
GeneName<-as.character(Table[["SequenceNameGene"]])
GeneNameU<-GeneName[!duplicated(GeneName)]

CodingStart<-Table[["CodingStartbp"]]
CodingEnd<-Table[["CodingEndbp"]]
StrandE<-Table[["Strand"]]
```

```
##########Define function "GENE" ##########
GENE<-function(x){
    if(x<=length(GeneNameU)){
    Start<-CodingStart[GeneName==GeneNameU[x]]
    End<-CodingEnd[GeneName==GeneNameU[x]]
    StrandX<-StrandE[GeneName==GeneNameU[x]]

    StartU<-Start[!duplicated(Start)]
    EndU<-End[!duplicated(End)]

    StartAndEnd<-sort(c(StartU,EndU),decreasing=FALSE)

    NewStart<-StartAndEnd[-length(StartAndEnd)]

    EndTem1<-StartAndEnd[-1]-1
    EndTem2<-EndTem1[-length(EndTem1)]
    NewEnd<-c(EndTem2,(EndTem1[length(EndTem1)]+1))

    Gene<-rep(GeneNameU[x],length(NewStart))
    Strand<-rep(StrandX[1],length(NewStart))
      NewExonName<-c()
      i=1
      while(i<=length(NewStart)){
```

```
      NewExonName<-c(NewExonName,paste(GeneNameU[x],i))
      i=i+1
      }
    Table<-data.frame(Gene,Strand,NewExonName,NewStart,NewEnd)
    return(Table)
    }else{
    return("x should<=length(GeneNameU)")
    }
}
##########END of "GENE"##########
TableG<-GENE(1)[1,]
j=1
while(j<=length(GeneNameU)){
TableG<-rbind(TableG,GENE(j))
j=j+1
}

OriginalTable<-TableG[-1,]     ##intron regions are to be further removed from this table
#####################################################################
F.Start<-TableF[["FixedStart"]]
F.End<-TableF[["FixedEnd"]]

Range<-c()
```

```
k=1
while(k<=length(F.Start)){
Range<-c(Range,F.Start[k]:F.End[k])
k=k+1
}

O.Start<-OriginalTable[["NewStart"]]
O.End<-OriginalTable[["NewEnd"]]

##########Define function “REMOVE” ##########
REMOVE<-function(x){
  if(x<=length(O.Start)){
  RangeO<-O.Start[x]:O.End[x]
    if(any(is.element(RangeO,Range))){
    return("FALSE")
    }else{
    return("TRUE")
    }
  }else{
  return("x should<=length(O.Start)")
  }
}
##########END of  “REMOVE”##########
```

```
ii=1
n<-c()
while(ii<=length(O.Start)){
n<-c(n,REMOVE(ii))
ii=ii+1
}

Right<-which(n=="TRUE")
FinalTable<-OriginalTable[Right,]     ##this is the correct exon annotation,
                                      ##but it may contain UTR regions
                                      ##because the "FixedVariantIntron" files
                                      ##used for doing this include all the
                                      ##introns, not just introns between
                                      ##coding exons
#output
write.table(FinalTable,file="FixedExonVariantChrI.txt",sep="\t")
```

#----------END of PART 2----------#

A.4. Boundary annotation files generation

```
setwd("")   #set path for file to be read ("WS170ChrI.txt ")

####file reading (primary annotation file WS170)####

ChrI<-read.table("WS170ChrI.txt",header=T,sep="\t")

Chr<-ChrI


##order table based on "CodingStartbp"

Table<-Chr[,c(1,4,5,6,10,11)]

Start<-as.numeric(Table[["CodingStartbp"]])

Order1<-order(Start,decreasing=FALSE)

Table1<-Table[Order1,]


ExonStart<-as.numeric(Table1[["CodingStartbp"]])

ExonEnd<-as.numeric(Table1[["CodingEndbp"]])

removeID<-which((ExonEnd-ExonStart)<=26)

Table2<-Table1[(-removeID),]


###generate dataframe “Range"###

GeneName<-as.character(Table2[["SequenceNameGene"]])

ChrName<-as.character(Table2[["ChrName"]])

Strand<-as.character(Table2[["Strand"]])

ExonName<-as.character(Table2[["ExonName"]])

StartExon<-as.numeric(Table2[["CodingStartbp"]])
```

```
EndExon<-as.numeric(Table2[["CodingEndbp"]])

gene.name<-rep(GeneName,each=2)  #column1
chr<-rep(ChrName,each=2)  #column2
Strand<-rep(Strand,each=2)  #column3

Exon<-rep(ExonName,each=2)
paste<-rep(c("S","E"),length(ExonName))
RangeName<-paste(Exon,paste,sep=".")   #column4

i=1
R<-c()
while(i<=length(StartExon)){
R<-c(R,StartExon[i],EndExon[i])
i=i+1
}

StartR<-R-25   #column5
EndR<-R+25     #column6

RangeTem<-data.frame(gene.name,chr,Strand,RangeName,StartR,EndR)
SE<-data.frame(StartR,EndR)
Range<-RangeTem[!duplicated(SE),]
```

```
###dataframe "Range" generated###

############################################################

start<-as.numeric(Range[["StartR"]])

Order2<-order(start,decreasing=FALSE)

RangeO<-Range[Order2,]


StartR1<-as.numeric(RangeO[["StartR"]])

EndR1<-as.numeric(RangeO[["EndR"]])


diff<-StartR1[-1]-EndR1[-length(EndR1)]

ID1<-which(diff<=(-23))

ID2<-ID1+1

ID<-c(ID1,ID2)

IDs<-sort(ID)

IDfinal<-IDs[!duplicated(IDs)]


Boundary<-RangeO[(-IDfinal),]

#output

write.table(Boundary,file="Boundary chrI.txt",sep="\t")
```

A.5. Probe grouping files generation

```
setwd("")   #set path for files to be read

####files reading####

#read probe mapping files

ProbeI<-read.table("ProbeChrI.txt",header=T,sep="\t")

#read desired annotation files (use gene annotation file for demonstration)

GeneI<-read.table("GeneAnnotationChrI.txt",header=T,sep="\t")


Probe<-ProbeI

Gene<-GeneI


PositionS<-as.numeric(Probe[,5])

PositionE<-PositionS+24


Start<-Gene[["NewCodingStart"]]

End<-Gene[["NewCodingEnd"]]

GeneName<-as.character(Gene[["SequenceNameGene"]])


##########Define function “NAME” ##########

NAME<-function(x){

   if(x<=length(PositionS)){

   n<-which(PositionS[x]>=Start&PositionE[x]<=End)

      if(length(n)==1){
```

```
        return(GeneName[n])
        }else{
        return("NoName")
        }
    }else{
    return("x should<=length(PositionS)")
    }
}
##########END of "NAME"##########
i=1
NewName<-c()
while(i<=length(PositionS)){
NewName<-c(NewName,NAME(i))
i=i+1
}

m<-which(NewName!="NoName")
ProbeGeneName<-NewName[m]
ProbeNew<-Probe[m,]

GeneProbe<-cbind(ProbeNew,ProbeGeneName)
#output
write.table(GeneProbe,file="GeneProbeChrI.txt",sep="\t")
```

A.6. Create TPMAP files from probe grouping files

```
setwd("")   #set path for files to be read

####read probe grouping files####

Probe<-read.table("filename.txt",header=F)

#sort “Probe” based on sequence names

Name<-as.character(Probe[,3])

Order<-order(Name)

Table<-Probe[Order,]

#output  (format is to be changed)

write.table(Table,file="filename (for CDF).txt",quote=FALSE,sep="

",row.names=FALSE,col.names=FALSE)
```

##further modifications on "filename (for CDF).txt"##

--open file

--manually copy the following lines to the Header

```
  #seq_group_name Ce
  #version WS170
  #probeset_type tiling
```

--save and exit

A.7. Create tab separated sequence files from TPMAP files

```
setwd("")    #set path for files to be read
options(scipen=99)    #change scientific notation
####read TPMAP file####
probe<-read.table("filename (for CDF).txt",header=F)
#define vectors
ProbeSetName<-as.character(probe [,3])
ProbeX<- probe [,5]
ProbeY<- probe [,6]
Pos<- probe [,4]
ProbeInterrogationPosition<-Pos+12
ProbeSequence<-as.character(probe [,1])
TargetStrandedness<-rep("Antisense",length(ProbeSequence))

table<-
cbind(ProbeSetName,ProbeX,ProbeY,ProbeInterrogationPosition,ProbeSequence,Target
Strandedness)
#output
write.table(table,file="name_probe_tab.txt",quote=FALSE,sep="\t",row.names=FALSE)

##further modifications on "name_probe_tab.txt "##
--open file
--manually change the Header to the following:
```

Probe Set Name (1st column)

Probe X (2nd column)

Probe Y (3rd column)

Probe Interrogation Position (4th column)

Probe Sequence (5th column)

Target Strandedness (6th column)

--save and exit

A.8. Create BPMAP file from TPMAP file

```
library(affxparser)
setwd("")   #set path for file "filename (for CDF).txt"
tpmap2bpmap("filename (for CDF).txt","filename.bpmap",verbose=0)
```

A.9. Create CDF file from BPMAP file

```
library(affxparser)
library(aroma.affymetrix)
setwd("")   #set path for file "filename.bpmap"

##########Define function “bpmapCluster2Cdf” ##########
bpmapCluster2Cdf <-
function(filename, cdfName, nProbes=-1, gapDist=3000000, rows=NULL,
groupName="Ce", cols=NULL, field="fullname", verbose=10,
stringRemove="Ce:WS170;") {
  require("affxparser") || throw("Package not loaded: affxparser");
  require("R.utils") || throw("Package not loaded: R.utils");

  # Argument 'groupName':
  groupName <- Arguments$getCharacter(groupName);

  ##Function “BpmapUnit2df”##
  BpmapUnit2df <- function(u) {
    o <- order(u[["startpos"]]);
    mmx <- u[["mmx"]];
    mmy <- u[["mmy"]];
    mmx <- if (all(mmx == 0) | is.null(mmx)) 0 else mmx;
    mmy <- if (all(mmy == 0) | is.null(mmy)) 0 else mmy;
```

```
   data.frame(seqname=u$seqInfo[[field]],groupname=u$seqInfo$groupname,
u[c("pmx","pmy")], mmx=mmx, mmy=mmy,
            u[c("probeseq","strand","startpos","matchscore")],
stringsAsFactors=FALSE)[o,];
 }
 ##END of "BpmapUnit2df"##


 # - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 # Validating arguments
 # - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 # Argument 'filename':
 filename <- Arguments$getReadablePathname(filename);

 # Argument 'cdfName':
 cdfName <- Arguments$getCharacter(cdfName);

 # Argument 'verbose':
 verbose <- Arguments$getVerbose(verbose);

 verbose && enter(verbose, "Generating CDF from BPMAP");

 verbose && enter(verbose, "Reading BPMAP file");
```

```
verbose && cat(verbose, "Source pathname: ", filename);
bpmaplist <- readBpmap(filename, readMatchScore=TRUE);
verbose && exit(verbose);

verbose && enter(verbose, "Extracting X/Y locations");
bpmapdflist <- lapply(bpmaplist, FUN=BpmapUnit2df);
verbose && exit(verbose);

rm(bpmaplist); gc()  # save some space

if (is.null(rows)) {
  rows <- max(sapply(bpmapdflist, FUN=function(u) max(c(u$mmx,u$pmx))));
  verbose && cat(verbose, "NB: 'rows' of CDF are being set as ", rows,
      ". If this is not correct, stop now and specify 'rows' argument.");
}

if (is.null(cols)) {
  cols <- max(sapply(bpmapdflist,FUN=function(u) max(c(u$mmy,u$pmy))));
  verbose && cat(verbose, "NB: 'cols' of CDF are being set as ", cols,
      ". If this is not correct, stop now and specify 'col' argument.");
}

verbose && enter(verbose, "Creating structure for ", length(bpmapdflist), " units");
```

```
e <- vector("list", 1);
startps <- l <- ll <- vector("list", 50000);
count <- 0;
naValue <- as.character(NA);
nm <- rep(naValue, length(l));
for (ii in seq(along=bpmapdflist)) {
 name <- names(bpmapdflist)[ii];

 # Access ones
 bpmapdf <- bpmapdflist[[ii]];
 np <- nrow(bpmapdf);

 ch <- gsub(stringRemove, "", bpmapdf$seqname[1]);

 #if (all(bpmapdf$mmx == 0) & all(bpmapdf$startpos > 0)) {
 if (all(bpmapdf$startpos > 0) & bpmapdf$groupname[1]==groupName) {
  sp <- bpmapdf$startpos;
  d <- diff(sp);
  w <- whichVector(d > gapDist);
  ends <- c(w, np);
  starts <- c(1, w+1);
  k <- whichVector((ends-starts) > nProbes);
  verbose && cat(verbose, length(k), " ROIs for ", name, ".");
```

```
# Access ones
pmx <- bpmapdf$pmx;
pmy <- bpmapdf$pmy;
      mmx <- bpmapdf$mmx;
mmy <- bpmapdf$mmy;



for (jj in seq(along=k)) {
  w <- starts[k[jj]]:ends[k[jj]];
  np <- length(w);

                if (all(bpmapdf$mmx==0)) {
                 # PM only
      e[[1]] <- list(x=pmx[w], y=pmy[w], pbase=rep("A", np), tbase=rep("T", np),
                           atom=0:(np-1), indexpos=0:(np-1), groupdirection="sense",
natoms=np, ncellsperatom=1);
                } else {
                 # PM+MM
      e[[1]] <- list(x=c(pmx[w],mmx[w]), y=c(pmy[w],mmy[w]), pbase=rep("A", np*2),
tbase=rep(c("T","A"), each=np),
                           atom=rep(0:(np-1),2), indexpos=rep(0:(np-1),2),
groupdirection="sense", natoms=np, ncellsperatom=2);
```

```
                }
        names(e) <- paste(ch);
        na <- sum(unlist(sapply(e,FUN=function(u) u$natoms)));
        nc <- sum(unlist(sapply(e,FUN=function(u) u$natoms*u$ncellsperatom)));
        count <- count + 1;

        l[[count]] <- list(unittype=1, unitdirection=1, groups=e, natoms=na, ncells=nc,
ncellsperatom=nc/na, unitnumber=ii);
        startps[[count]] <- sp[w];
        nm[count] <- names(e);
        #if (verbose) { if (count %% 250 == 0) cat(verbose, ".") }
      } # for (jj ...)

     #if (verbose) cat("\n");
    } else {
     # keep all probes
     verbose && cat(verbose, "Skipping all ", np, " probes for ", name, ".");
     next;
    }
  } # for (ii ...)
  verbose && exit(verbose);
```

```
  l <- l[1:count];
  names(l) <- nm[1:count];

  verbose && enter(verbose, "Writing PPS file");
  startps <- startps[1:count];
  names(startps) <- names(l);
  saveObject(startps, file=sprintf("%s.pps", cdfName));
  rm(startps);
  verbose && exit(verbose);

  verbose && enter(verbose, "Writing CDF file");
  hdr <- list(probesets=length(l), qcprobesets=0, reference="", chiptype=cdfName,
filename=sprintf("%s.cdf", cdfName), nqcunits=0, nunits=length(l), rows=rows,
cols=cols, refseq="", nrows=rows, ncols=cols);
  verbose && cat(verbose, "Output pathname: ", hdr$filename);
  verbose && str(verbose, hdr);
  writeCdf(hdr$filename, cdfheader=hdr, cdf=l, cdfqc=NULL, overwrite=TRUE,
verbose=verbose);
  verbose && exit(verbose);

  res <- list(cdfList=l, cdfHeader=hdr);

  verbose && exit(verbose);
```

```
  invisible(res);
}
##########END of "BpmapCluster2Cdf"#########

bpmapFile<-"filename.bpmap"
chipType<-"filename"
bpmapCluster2Cdf(bpmapFile,chipType,rows=2560,cols=2560,verbose=-20)    #output
```

NOTE: function "bpmapCluster2Cdf" was originally obtained from http://www.aroma-project.org/node/42 (look for bpmapCluster2Cdf.R on this webpage). The original function was designed to make CDF files for analysing Human Promoter tiling array, therefore, modifications were made to adapt the analysis for *C. elegans* Tiling Array.

A.10. CDF package creation

```
library(makecdfenv)

setwd("")   #set path for file "filename.cdf"

make.cdf.package("filename.cdf",packagename="ce25bmrv02cdf",species="C_elegans")
```

A.11. Probe package creation

```
library("AnnotationDbi")
setwd("")   #set path for file "name_probe_tab.txt"
filename="name_probe_tab.txt"    #put corresponding file name at “name”
outdir<-getwd()
me<-"xuanzhao<xuazhao@ucalgary.ca>"
species<-"c.elegans"
makeProbePackage("ce25bmrv02",
datafile=gzfile(filename,open="r"),
outdir=outdir,
maintainer=me,
species=species,
version="0.0.1",
force=FALSE,
comparewithcdf=FALSE)
```

A.12. Cluster analysis

```
#probe intensities have already been statistically processed, and probes have been divided
#according to their relative signals in two comparing datasets (see detailed description in
#section 3.10.2 and 3.10.3). The following scripts are for identifying “clusters” from each
#probe category by applying the two criteria Max gap=50 and Min count=2 (section
#3.10.4). File “N2 higher than teg4I (p0.05).txt" is used for demonstration, in which all
#probes are on chromosome I, and have a significantly higher signal (p<0.05) in N2
#dataset.

setwd("")   #set path for file to be read
#read file
I<-read.table("N2 higher than teg4I (p0.05).txt",header=T)
L<-I

number<-as.numeric(L[,2])   #location of all probes
id<-1:length(number)
diff<-number[-1]-number[-length(number)]
#set criteria
x<-as.numeric(diff<=50)    # return a vector with elements 1 (diff<=50) and 0 (diff>50)

xdiff<-x[-1]-x[-length(x)]
start<-which(xdiff==1)
end<-which(xdiff==-1)
```

```
startid<-start+1

endid<-end+1


##########IMPORTANT CONSIDERATIONS##########

l.s<-length(startid)  #the size of vector "startid"

e.s<-length(endid)   #the size of vector "endid"

startid[1:5]

endid[1:5]

##Normally, l.s=e.s and "endid" have bigger elements.

##If one of the following three unusual situations happened, special modifications are

##needed to be made to vector "startid" and "endid".

  ##1)l.s>e.s. This would indicate the LAST two (or more) elements in the vector

  ##"number" meet the above criteria, the LAST element in vector "id" should be

  ##added to the END of "endid"

  ##2)l.s<e.s. This would indicate the FIRST two (or more) elements in the vector

  ##"number" meet the above criteria, the FIRST element in vector "id" should be

  ##added to the BEGINNING of "startdid"

  ##3)l.s=e.s, but "endid" have smaller elements. This would indicate both the FIRST

  ##two (or  more) and LAST two (or more) elements in the vector "number" meet the

  ##above  criteria, then the FIRST element in vector "id" should be added to the

  ##BEGINNING of "startdid" and the LAST element in vector "id" should be added to

  ##the END of "endid".

                ##############################################
```

```
name<-paste("I",1:l.s,sep=".")


table<-cbind(name,number[startid],number[endid])
colnames(table)=c("name","start","end")


#output
write.table(table,"cluster N2 higher than teg4I (0.05).txt",sep="\t")
```

A.13. Single probe analysis

```
# raw intensity data of intronic probes were obtained (section 3.11.1). The following
#scripts include identifying: probes with top 10% intensities, probes that are present in all
#replicates, probes that are only exist in experimental groups and probes that mapped to
#boundary regions (section 3.11.2).

#----------PART 1----------#
#identify probes with top 10% intensities, using one replicate from N2 dataset for
#demonstration
setwd("")  #set path for file to be read
#read file
Probe<-read.table("N2(1)intron intensity.txt",header=T)
int<-as.numeric(Probe[,3])     ##extract intensity data
order<-order(int,decreasing=TRUE)      ##rank the intensity data, decreasingly
Int.order<-Probe[order,]

top<-1:trunc(0.1*length(int))     ##intron probes with top 10% intensities
Int.top<-Int.order[top,]
#output
write.table(Int.top,file="top 10% intron probes (N2[1]).txt",sep="\t",row.names=F)
#repeat the above for all replicates in every dataset
#----------END of PART 1----------#
```

```
#---------------------------------PART 2---------------------------------#
#identify probes that are present in all replicates, using N2 dataset for demonstration
setwd("")  #set path for files to be read
#read files (files created in PART1)
N2.1<-read.table("top 10% intron probes (N2[1]).txt",header=T)
N2.2<-read.table("top 10% intron probes (N2[2]).txt",header=T)
N2.3<-read.table("top 10% intron probes (N2[3]).txt",header=T)
N2.4<-read.table("top 10% intron probes (N2[4]).txt",header=T)
N2.5<-read.table("top 10% intron probes (N2[5]).txt",header=T)

#consider both chromosome and location for each probe, generating a new vector with
#format: "chr.location"
N.in1<-paste(N2.1[["chr"]],N2.1[["location"]],sep=".")
N.in2<-paste(N2.2[["chr"]],N2.2[["location"]],sep=".")
N.in3<-paste(N2.3[["chr"]],N2.3[["location"]],sep=".")
N.in4<-paste(N2.4[["chr"]],N2.4[["location"]],sep=".")
N.in5<-paste(N2.5[["chr"]],N2.5[["location"]],sep=".")

combine<-c(N.in1,N.in2,N.in3,N.in4,N.in5)

du.2<-combine[duplicated(combine)]  ##elements that appeared at least 2 times
du.3<-du.2[duplicated(du.2)]  ##elements that appeared at least 3 times
du.4<-du.3[duplicated(du.3)]  ##elements that appeared at least 4 times
```

```
du.5<-du.4[duplicated(du.4)]      ##elements that appeared at least 5 times, desired one


rep.5<-N2.1[is.element(N.in,du.5),]     ##select these corresponding rows
chro<-as.character(rep.5[,1])
order<-order(chro)
rep.order<-rep.5[order,]      ##order these rows based on chromosomes


#output
write.table(rep.order,file="N2intron5.txt",sep="\t",row.names=F)
#repeat the above for the rest three datasets
```

#----------END of PART 2----------#

```
#---------------------------------PART 3---------------------------------#
#identify probes that are only exist in experimental groups, using N2 and teg-4 datasets
#for demonstration (teg-4: experimental group, N2: control group)
setwd("")   #set path for files to be read
#read files
teg4<-read.table("teg4intron5.txt",header=T)   ##read teg4 file created in PART2
N2<-read.table("N2intron5.txt",header=T)      ##read N2 file created in PART2

#consider both chromosome and location for each probe, generating a new vector with
#format: "chr.location"
t.in<-paste(teg4[["chr"]],teg4[["location"]],sep=".")   ##for teg-4
N.in<-paste(N2[["chr"]],N2[["location"]],sep=".")   ##for N2

MATCH1<-match(t.in,N.in)
ID1<-which(is.na(MATCH1))
teg4NOTN2<-teg4[ID1,]     ##identify probes that are present in teg-4 dataset only

#output
write.table(teg4NOTN2,file="intron teg4 not N2.txt",sep="\t",row.names=F)
#repeat the above for the teg-4smg-2 and smg-2 dataset

#----------------------------END of PART 3----------------------------#
```

```
#---------------------------------PART 4---------------------------------#
#identify probes that mapped to boundary regions, using N2 and teg-4 datasets for
#demonstration
setwd("")   #set path for files to be read
#read file created in PART3
teg4<-read.table("intron teg4 not N2.txt",header=T)
#read boundary probe files
BI<-read.table("BoundaryProbeChrI.txt",header=T)   ## for chrI
#then read the rest five boundary probe files:
  ##BII: chrII, BIII: chrIII, BIV: chrIV, BV: chrV, BX: chrX

##the following uses chromosome I for demonstration
chr<-teg4[["chr"]]
t.I<-teg4[chr=="chrI",]

loctI<-t.I[["location"]]
bouI<-as.numeric(BI[,5])        ##probe locations in boundary probe file

match1<-match(loctI,bouI)
ID1<-which(!is.na(match1))
I.teg4<-t.I[ID1,]           ##probes that are in boundary regions
loc<-I.teg4[["location"]]
```

```
GENE<-function(x){
 gene<-as.character(BI[bouI==loc[x],6])
 return(gene)
 }

gene<-c()
i=1
while(i<=length(ID1)){
gene<-c(gene,GENE(i))
i=i+1
}                              ##names of the genes that identified probes are in

teg4.bouI<-cbind(I.teg4,gene)    ##a dataframe with desired probes from chromosome I
#repeat the above scripts for each chromosome, and five more dataframes are created:
  ##teg4.bouII, teg4.bouIII, teg4.bouIV, teg4.bouV and teg4.bouX

#put information from all chromosomes together
teg4.intron.boun<-
rbind(teg4.bouI,teg4.bouII,teg4.bouIII,teg4.bouIV,teg4.bouV,teg4.bouX)
#output
write.table(teg4.intron.boun,file="teg4BoundaryHighIntrons.txt",sep="\t",row.names=F)
```

#----------END of PART 4----------#

APPENDIX B: SPECIAL CONSIDERATIONS FOR "GCRMA" PACKAGE

The R package "gcrma" was used for performing data pre-processing (section 3.6), and raw probe signals were converted into various regional signals. However, before this package could be used in R, two additional types of customized package were needed to be created and installed; 1) a CDF package, and 2) a probe package.

CDF package

A CDF package is made from a CDF file. CDF (Chip Definition File) files contain information on the design of the probes; which probes belong to the same probeset. In expression arrays, probesets correspond to genes. Therefore, CDF files are needed by many Bioconductor packages to analyze Affymetrix chips, to transform data from probe level to expression level. For some expression arrays, such as the *C. elegans* Genome Array, the CDF file is provided by Affymetrix. Bioconductor also provides the corresponding CDF packages (and other annotation files) that can be easily installed into R to create a corresponding CDF environment.

However, for tiling arrays, CDF files are not provided by Affymetrix. This is probably because probes in tiling arrays do not just correspond to the coding region of a gene, but also intronic and intergenic regions, as well as in the junctions of these regions. Consequently, the application of the current Bioconductor packages to analyze tiling arrays is limited, and customized CDF files are often needed for many R packages.

An R code on http://www.aroma-project.org/node/42 was found for making a CDF file from a BPMAP file (see "Create CDF file from BPMAP file" in APPENDIX A). Therefore, a method was designed to create customized CDF files. First, TPMAP files were created (see "Create TPMAP files from probe grouping files" in APPENDIX A),

and then BPMAP files were subsequently made (see "Create BPMAP files from TPMAP files" in APPENDIX A). Ultimately, CDF files will be created. Since all these files originated from the probe grouping files (section 3.5), each type of the probe grouping file will generate its corresponding CDF file.

All CDF files were made to CDF packages (see "CDF package creation" in APPENDIX A), and they all share the same name "ce25bmrv02cdf ", which is the name required by R packages when analysing the *C. elegans* Tiling array data in this project.

CDF packages that were generated were needed to be installed before they could be used. Only one package can be installed a time—depending on the type of analysis to be made. As an example, if the gene expression levels needed to be generated, the following procedures were followed:

1) Copy the corresponding CDF file for the gene to the directory "R/bin/"
2) open Command Prompt
3) change directory to "R/bin/" (type "cd \")
4) type "R CMD INSTALL ce25bmrv02cdf"

The package "ce25bmrv02cdf" would be successfully installed. If exon expression levels needed to be generated, the existing CDF package (for the gene) needed to be removed, and a new CDF package (for the exon) would be installed. To remove a CDF package, open R and run: *rm("ce25bmrv02cdf")*.

Probe package

The second type of customized package is the probe package. First, a tab separated sequence file needed to be made (see "Create tab separated sequence files from TPMAP files" in APPENDIX A), and a probe package can be generated (see "Probe package

creation" in APPENDIX A). Probe packages also needed installation and removal, in the same way as CDF packages, except that "ce25bmrv02cdf" were changed to "ce25bmrv02probe".

After the corresponding CDF and probe packages were both installed, the R package "gcrma" can properly function.

APPENDIX C: FILITERING METHODS FOR IDENTIFYING FALSE POSITIVE SPLICING DEFECTS

The list of potential splicing defects initially generated through tiling array analysis was very large, and it is not practical to experimentally verify all of them. However, some of the splicing defects can be removed without doing any experimental verification. Here I introduce several filtering methods that helped identify false positives.

1. Examine gene expression levels

There is no way that a splicing defect can be identified in a gene that is not expressed. Therefore, if a gene's expression level is below the background, all splicing defects identified on this gene are false positives (Figure C.1A).

2. Examine the expression level of regions with splicing defects

If a splicing problem caused retention of an intron, the expression level of this intron would increase. However, it is impossible that the expression level of an intron is even higher than the overall expression level of the gene in which this intron is located (Figure C.1B). Therefore, this splicing defect is not true.

If an exonic region is identified with splicing defects (exon skipping), in most cases, it is more likely that the expression level of this exon is lower in the mutant animals (*i.e.* *teg-4*), rather than in the wild-type animals (*i.e.* N2) (Figure C.1C). On the other hand, expression levels of the identified intronic regions are more likely higher in the mutant animals (Figure C.1D). Therefore, splicing defects with the opposite results were removed.

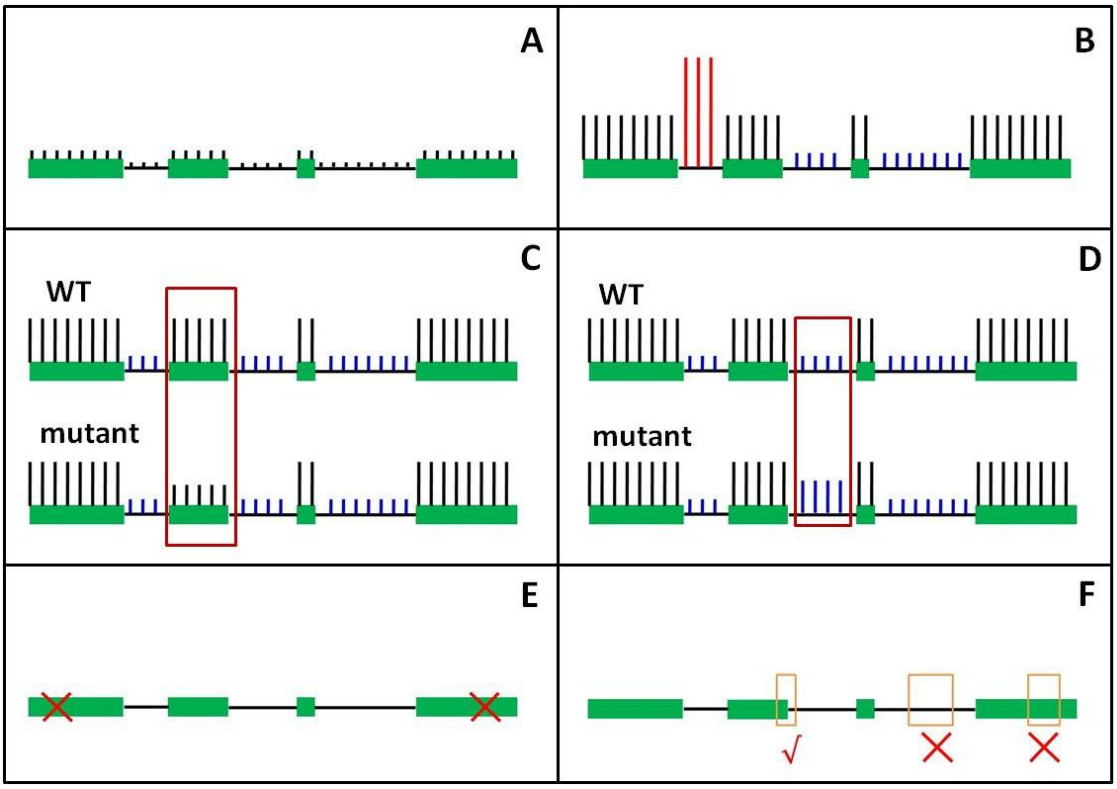


Figure C.1. Different methods for removing false splicing defects.

If a gene was not expressed (A), splicing defects were unlikely to be identified from it. If the calculated expression level of an intron (B, the second intron) is even higher than the overall expression level of the gene, this result is unlikely. The exon skipping (C) and the intron retention (D) should occur in the mutant rather than in the wild-type (WT) animals. Splicing problems unlikely occur at the end of a gene (E). In the cluster analysis, clusters that were identified in the boundary regions are more convincing (F).

3. Examine the location of a splicing defect

Splicing problems unlikely occur at the end of a gene. Therefore, based on the annotation information, if a potential splicing defect is located at the first exon or the last exon (Figure C.1E), this is likely a false positive result.

For the cluster analysis, identified clusters were located on different regions. However, unless an entire exon (not the exons at the end of a gene) or intron was identified as a cluster, only the clusters that are located at boundaries were most convincing (Figure C.1F). Therefore, the rest were removed.

APPENDIX D: ADDITIONAL TABLES

Table D.1 Primers used for background determination.

primer name*	primer sequence	amplicon size
C38D9.2fwd	ctacgaataccaacgcagca	862
C38D9.2rev	gcttcttttggagcctcctt	
B0511.11fwd	cgcaaacctcgatcaaatct	892
B0511.11rev	acagcctttgtctcctcatca	
Y37B11A.3fwd	ccttcagagaagacggttcg	853
Y37B11A.3rev	gtggagtctcgtcgatcaca	
C02C2.3fwd	cttactccgagcaccaaagtg	967
C02C2.3rev	ctgccacatagcaaagacga	
C35A5.2fwd	gccatcgtcttcgagtttct	995
C35A5.2rev	tttggcagctctccgtattt	
F55D10.1fwd	accaggctgctcatcaatct	855
F55D10.1rev	caggtttgcttcctttctcg	
F58D5.4fwd	ggaaacagcatttggacgat	920
F58D5.4rev	cggcgaaagatagtcaatcc	
C24G7.2fwd	tggtgaagtgagtcctccaa	932
C24G7.2rev	tgtgtccatcagtccattcc	
F58G1.7fwd	acacctgtccacgacatcaa	871
F58G1.7rev	aatagcacccgttgcagaac	
C13B9.4fwd	ccctgaacgagatccaaatg	991
C13B9.4rev	atccctagttggtgcaaacg	
Y46C8AL.3fwd	cttgggataacggaagtcca	991
Y46C8AL.3rev	ccggaagcatcaagattagc	
D1044.1fwd	gactccaaactcggaagcac	942
D1044.1rev	gtgttccctttcgtcagctc	
ZK550.2fwd	ctacggcctctggaatcaaa	948
ZK550.2rev	gactcgacaaatgggaaagc	
Y74C9A.5fwd	attggattggaacgtcgtgt	859
Y74C9A.5rev	ggaacattgaagcctggaaa	
T08H4.1fwd	atcgtgttccgaagttccag	979
T08H4.1rev	gagtgttcgatgggttggtt	
T01G6.7fwd	gtcaggatgtgggaatgacc	992
T01G6.7rev	cggatgtgagaactccaggt	
T24A11.2fwd	cgctgcaccatacttcttca	853
T24A11.2rev	ggctccagtgtattctttcgat	
Y75B8A.28fwd	tcgacccgaattggaactac	998
Y75B8A.28rev	ggcgtgaggaacaaataagc	
C26C6.5fwd	tctcaacaagcacagcaacc	987
C26C6.5rev	tcgcctctctatttgcctgt	

* The information in a primer name include the regions tested (i.e. the gene C38D9.2), and whether this is a forward (fwd) or reverse (rev) primer.

Table D.2 Primers used for reference gene selection.

primer name[a]	primer sequence	amplicon size
gpd-1fwd	tggcaccactggcgaaggtt	134
gpd-1rev	accgcgtccatctctccaca	
gpd-2fwd*	accatcgagaaggccaacgct	178
gpd-2rev*	tggcaagtggagcaaggcagt	
unc-15 fwd	tatgccctcgctcaacgcaagg	104
unc-15 rev	tcaacctcggcttgcttacggg	
unc-54 fwd*	tgcgttcccgttacgctgct	118
unc-54 rev*	cccatgaacatacgggcgca	

[a] The information in a primer name include the regions tested (*i.e.* the gene *gpd-1*), and whether this is a forward (fwd) or reverse (rev) primer.

* These primer pairs formed primer dimers.

Table D.3 Primers used for qPCR assays (for testing genes with different expression levels).

* The information in a primer name include the regions tested (*i.e.* the gene C01G6.7), and whether this is a forward (fwd) or reverse (rev) primer.

primer name*	primer sequence	amplicon size
C01G6.7fwd	tgtgctcacttcccgtctcgt	184
C01G6.7rev	cgctttggcgttggcatcgaa	
F59D6.3fwd	cccacttgcctacaactcgt	90
F59D6.3rev	gggatggtcgtttctgtttg	
K07E8.3fwd	tcatttcctgctcaccgact	109
K07E8.3rev	ataacaaggctggcccaact	
M02H5.8fwd	gcttctcgaaatggcctccgca	81
M02H5.8rev	tcgatggctccggaagaaagcc	
R09H10.3fwd	tttcggctcacgtcctcgaca	127
R09H10.3rev	acacgcccattgtcttgcgtg	
T08A9.12fwd	gatggatctgcggataagga	103
T08A9.12rev	gttcacactcctgtggagca	
T09F5.9fwd	gcttgggcagattcgactaa	109
T09F5.9rev	gtccagcactattccctgct	
T26H5.9fwd	gtgacgatgacccgcgaaga	115
T26H5.9rev	ggaggttcttgttctcgttg	
T28F2.5fwd	agccacacatccgccagtca	96
T28F2.5rev	atcatctgcgccggcgtgaa	
VW02B12L.4fwd	gggaagcgtcgagacgcagatt	126
VW02B12L.4rev	agaacattggcgactgcggc	
Y57A10A.29fwd	tgacacgcgcagaagttccc	150
Y57A10A.29rev	tcagctcgggctcgctcatt	
Y73B3A.20fwd	ccgccgccagaaaggagatt	84
Y73B3A.20rev	ttcccgccaagactcgcttc	
B0379.3fwd	cgatgactttgacgtggatg	73
B0379.3rev	gtccggtatcgtttctggtg	
C15C6.3fwd	gagcagaagcaggagtcgat	122
C15C6.3rev	catcctcatcggagctgatt	
C25G4.7fwd	gcaaacctgaaagtgttcaagaatccc	124
C25G4.7rev	agcgagtactgatggaccgtggt	
C30G12.2fwd	gcagaaggagatgaagcaag	98
C30G12.2rev	ggaccatcaaagcatcca	
C38D9.2fwd	tgcataccaactgcccaagcga	95
C38D9.2rev	atggtgtgccgcggtgtctt	
D1037.2fwd	tgccactgaagtcatccaga	80
D1037.2rev	cgacagcgaacaccaatatg	
F15D4.6fwd	tgccgagcttccgaatcagca	148
F15D4.6rev	ttggcggattggtatctctcga	

T09E11.2fwd	gccgcctgtgctgcattctt	90
T09E11.2rev	cttcctcgccttccggcaaa	
Y38E10A.10fwd	gttgctggagccaatagagg	95
Y38E10A.10rev	caccatagtccgttgacacg	

Table D.4 Primers used for splicing defects identification.

[a] The information in a primer name include the regions tested (*i.e.* the exon T02G5.8.1.exon3), and whether this is a forward (fwd) or reverse (rev) primer.

[b] The size of the PCR product in the wild-type animals.

primer name[a]	primer sequence	amplicon size[b]
T02G5.8.1.exon3fwd	acggactgaccgatgcttat	253bp
T02G5.8.1.exon3rev	agcgaggtgaacttgtcgaa	
T28F2.2.exon3.Sfwd	gtcgtggcttcttgaggaaa	116bp
T28F2.2.exon3.Srev	gtattggcttgtcgtgtcca	
K11G9.5.exon6.Sfwd	ttcagtgtcgctcaaactgg	258bp
K11G9.5.exon6.Srev	tcctgccacgttcagacata	
Y48G1A.1.exon6.Sfwd	ggagactgtatttgctgtgg	233bp
Y48G1A.1.exon6.Srev	catccttaaaatcgccgaaa	
C27C12.2.exon4.Efwd	aaaacgagcatccacaatcc	220bp
C27C12.2.exon4.Erev	gtttttggctcgaaatgtgc	
F26F2.7.exon12.Sfwd	atgggcttcattccgattc	238bp
F26F2.7.exon12.Srev	ttttctccaccattcgcttc	
F08B12.1.exon11.Efwd	tctcctggctctcattttcc	211bp
F08B12.1.exon11.Erev	ctggcacaccaagtttgaga	
F45E4.11.exon3.Sfwd	cggaatgtgggcagatagat	243bp
F45E4.11.exon3.Srev	tgacgtcctcccgaaataag	
F40G9.3.exon3.Sfwd	tccaccagacactccatacg	153bp
F40G9.3.exon3.Srev	agaatatcgaggcaaatgacg	
F32G8.4.exon3.Sfwd	ggtgtgaaattggagcacag	205bp
F32G8.4.exon3.Srev	attgctgcttggagactgct	
M05D6.7.exon3.Efwd	tcagaagcgacggaaaaact	279bp
M05D6.7.exon3.Erev	tgggctttcgtttctcaact	
T10B11.6.exon3.Efwd	ctggagcaaatccaaagcat	147bp
T10B11.6.exon3.Erev	ttgctctgaaaagtggtgaca	
Y105E8B.8a.exon6.Efwd	gccttgagaagcgagttttc	190bp
Y105E8B.8a.exon6.Erev	ggtccttcccatgaccattt	
Y105E8A.10a.exon8.Sfwd	gcttctcttcgagttgctat	205bp
Y105E8A.10a.exon8.Srev	ctgattgaaggaaccaata	
W09D6.6.exon4.Sfwd	ctgcattccactgaccgac	248bp
W09D6.6.exon4.Srev	cgacgcgaagttggagggaa	
M01E5.3.1.exon5.Sfwd	gctcctccagccataattca	158bp
M01E5.3.1.exon5.Srev	aacgcttcttgtctctgctg	
Y105E8B.1f.exon7.Efwd	gtcgcccgtaagctcgccat	121bp
Y105E8B.1f.exon7.Erev	tcaagttattaccaacgacg	
Y37E3.16.1.exon3.Sfwd	gtgctcctaatggcttcagc	200bp
Y37E3.16.1.exon3.Srev	gtatccaatagctcgccagt	
Y40B1B.6.1.exon2.Sfwd	agtatttggacgaagagata	187bp

Y40B1B.6.1.exon2.Srev	gtcgaatttctgagcagcag	
F29C12.1a.1.exon1.Efwd	cattggcagatcttccgtca	186bp
F29C12.1a.1.exon1.Erev	gtgcgaactgtggctgatta	
Y76A2B.1.exon6.Sfwd	cgactacggagaagcatcat	177bp
Y76A2B.1.exon6.Srev	gttcgaggcgcgtgtccacg	
F36H1.2a.exon23.Efwd	cttgatgcgaggaatgcacc	161bp
F36H1.2a.exon23.Erev	gttggatacgctcaaagtcg	
F52B11.3.1.exon2.Efwd	ctgcattgtctccgatgagg	193bp
F52B11.3.1.exon2.Erev	ggtaactccttgagccgaag	
B0250.5.exon3.Efwd	ggaacactgtgcatggattc	189bp
B0250.5.exon3.Erev	cgcttcggcacgcttgaaag	
ZC8.4a.exon12.Sfwd	cgtgtttcggaagacagtcg	187bp
ZC8.4a.exon12.Srev	cagatgattggaagcgacgg	
F25D7.3.exon7.Sfwd	gaggatatgaaagactcgat	162bp
F25D7.3.exon7.Srev	actataaatggtttcgattg	
ZK337.1a.1.exon15.Efwd	gagtggttaccgattcgatg	173bp
ZK337.1a.1.exon15.Erev	gatcggcaacttgataagtg	
Y37E3.16.1.exon9.Efwd	gtcttcgtgttcaccggttt	180bp
Y37E3.16.1.exon9.Erev	cgaatccaacacttttggct	
Y63D3A.6a.1.exon4.Sfwd	ctgccacaactttcggaatt	182bp
Y63D3A.6a.1.exon4.Srev	gtacattcgtgtggtgtcca	
Y37E3.16.1.exon1.Sfwd	gcttggttatccggaaaaga	164bp
Y37E3.16.1.exon1.Srev	acgacaagtttgagtacaga	
W05B5.3a.exon1.Sfwd	gtttgatgaatccacaaccg	165bp
W05B5.3a.exon1.Srev	gaattgtgccagctgatgtg	
F58D5.5.exon1.Efwd	cgccaacgaaaggctcaatc	180bp
F58D5.5.exon1.Erev	cgtcattgagatggatgatg	
C17D12.1a.exon7.Sfwd	ccactcaattatgcttgcaa	178bp
C17D12.1a.exon7.Srev	catcattgcaactttgctat	
Y6B3A.1a.exon10.Sfwd	ttgggttgatgttatgaaat	167bp
Y6B3A.1a.exon10.Srev	cgagagcatcttgcaaagag	
C47B2.9.exon1.Efwd	acgtggatgtgggcgatgat	150bp
C47B2.9.exon1.Erev	ctcggacttttccctgcatg	
ZC101.3.exon6.Sfwd	gccaagcttgataatatgga	158bp
ZC101.3.exon6.Srev	cgtggcaacgaaaatgttgg	
ZK675.1.1.exon3.Efwd	ctggatcaatatcttgtcgg	163bp
ZK675.1.1.exon3.Erev	gacttcttgagtagagagca	
F59E12.12.exon1.Sfwd	gacgagaaggaactgaatca	181bp
F59E12.12.exon1.Srev	caactgttctatgacgacag	

Y48C3A.14.exon5.Efwd	gagcctggattcactgcagt	170bp
Y48C3A.14.exon5.Erev	ataagtgtgatgagctcggc	
Y50D7A.7.1.exon4.Sfwd	tcgcgaagaagtcgatcagc	206bp
Y50D7A.7.1.exon4.Srev	tgcaaaggctgagaacctta	
R74.3a.exon1.Efwd	cttcttggcgatgatatggg	164bp
R74.3a.exon1.Erev	cagatcgcgcatcacatcct	
Y48G9A.4.exon7.Efwd	cggctggaacaaaagattgg	168bp
Y48G9A.4.exon7.Erev	tcgttgctctttgtgtggtc	
K08E7.5a.exon2.Efwd	agcgtccggagccgaatctg	180bp
K08E7.5a.exon2.Erev	atatttgtgatagtgacagg	
T07C12.7.1.exon2.Efwd	ggatgatctccttgaccaag	163bp
T07C12.7.1.exon2.Erev	cggaagtgcgagtttgacct	
C35C5.10.exon4.Efwd	aactggaatcttccttctta	145bp
C35C5.10.exon4.Erev	cgccaacgcgaaccaagctc	

Table D.5 Other primers used in this work.

primer name*	primer sequence	amplicon size
[a]unc15exon3fwd	gaggtcaccaaggaactcca	214bp
[a]unc15exon3rev1	tggcgatgactctctttcct	
[a]unc15exon3fwd	gaggtcaccaaggaactcca	234bp
[a]unc15exon3rev2	acggattctggtctccaact	
[a]T05H10.6exon2fwd	atgtctgacccaggaaccag	200bp
[a]T05H10.6exon2rev1	gggagaacaccgtctgatgt	
[a]T05H10.6exon2fwd	atgtctgacccaggaaccag	180bp
[a]T05H10.6exon2rev2	agcaatcttcaacgcttcat	
[b]ok1853fwd	ctgatataggtttctacgga	1393bp (in wt)
[b]ok1853rev	gtagtcacgcgtaaacaacg	594bp (in mutant)
[c]rpl-12fwd	cggagaagacatcgccaagg	212bp (in wt)
[c]rpl-12rev	gatggtgtcaacagtgaggtc	212bp and 320bp (in mutant)
[d]teg-4fwd	tgtgaaatccaataatcagc	878bp
[d']teg-4rev	gaaatcttaacccgccatta	

[a] Positive controls for testing the 20bp difference.

[b] For detecting the *ok1853* deletion.

[c] For confirming the *smg-2(e2008) mutation*.

[d] and [d'] For amplifying the genomic DNA of the gene *teg-4*.

[d'] For sequencing the *oz210* mutation in the gene *teg-4*.

Table D.6 Splicing defects identified

Names highlighted are the candidates that were tested (in all four tables below). In table c and table d, names boxed are present in both tables. These are the regions identified in both the N2 vs. *teg-4* and *smg-2* vs. *teg-4 smg-2* in the single probe analysis (section 3.11.2)

a: splicing defects at boundary regions (N2 vs. *teg-4*)

region name	p value	FDR	N2	*teg-4*
B0213.5.exon2.S	3.35E-05	0.1312	8.32	151.90
B0228.4a.exon14.S	1.21E-05	0.0698	7.29	217.22
B0238.13.exon13.S	5.45E-05	0.1710	449.03	18.39
B0511.11.exon1.E	5.79E-05	0.1765	11.10	210.06
B0564.7.1.exon1.E	1.10E-05	0.0695	5.37	144.21
C03A3.1a.exon2.S	4.25E-05	0.1415	6.34	78.56
C06B3.11.exon3.E	3.84E-07	0.0142	3.30	119.19
C09E10.2a.exon16.E	9.92E-07	0.0168	67.87	4.07
C15B12.5a.exon13.E	1.73E-05	0.0806	66.47	3.47
C16B8.1.1.exon6.E	2.55E-05	0.1055	120.32	8.14
C16C10.12.exon2.E	1.14E-05	0.0695	4.67	69.02
C24B9.2.exon5.E	1.50E-05	0.0761	3.75	28.30
C27A7.5a.exon5.E	7.78E-07	0.0167	7.86	71.15
C33A12.3.1.exon3.E	1.42E-05	0.0761	60.59	6.55
C34B2.11.exon2.E	3.85E-05	0.1362	7.64	49.72
C34F6.5.exon1.S	1.64E-05	0.0783	3.86	63.74
C38C3.7.exon4.S	3.36E-06	0.0335	5.41	94.68
C43E11.13.exon2.S	6.32E-06	0.0500	7.69	236.12
C48B6.4.exon6.E	9.30E-06	0.0618	5.13	76.81
F15D4.5.exon6.E	7.93E-07	0.0167	3.22	143.51
F26C11.1.exon7.S	6.93E-06	0.0520	4.18	69.96
F31E9.3.exon5.S	1.28E-05	0.0703	2.69	70.91
F33G12.3.1.exon2.S	6.43E-09	0.0014	3.47	95.05
F35C5.5a.exon1.S	5.56E-05	0.1718	12.18	110.36
F35C8.7b.exon2.E	2.34E-05	0.0990	63.96	7.67
F38G1.3.exon3.S	6.34E-05	0.1856	3.50	66.73
F41B4.1.exon1.S	2.04E-06	0.0224	7.22	162.00
F42G9.6b.exon7.E	3.54E-05	0.1312	99.29	4.54
F42G9.7.exon3.E	8.58E-06	0.0608	6.87	63.50
F43C9.1.exon2.E	1.89E-06	0.0218	5.59	126.79
F45C12.7.exon1.S	4.32E-08	0.0047	160.51	5.45
F47H4.11.exon2.S	4.45E-05	0.1437	55.19	7.25
F56A3.1.exon4.E	2.68E-06	0.0281	93.08	6.55
F56C3.9.exon4.E	4.16E-05	0.1405	5.16	55.13
F57F5.4a.1.exon5.E	7.94E-07	0.0167	2.87	76.93
F57G12.2.exon4.E	6.02E-05	0.1787	106.33	8.36
F59A1.10.exon1.E	2.29E-05	0.0986	180.14	10.64
K01A11.4.1.exon9.E	1.17E-05	0.0695	4.08	69.14

K05B2.4.exon5.S	1.56E-05	0.0761	5.17	104.84
M02G9.2.exon2.E	3.90E-06	0.0372	6.84	57.25
R03E9.2.exon12.E	5.30E-06	0.0466	6.00	206.57
R06B10.1.exon7.E	1.62E-06	0.0210	6.14	230.95
R07G3.2.exon5.E	1.52E-05	0.0761	4.58	28.32
R13H8.1b.exon2.E	1.39E-06	0.0203	6.42	204.11
R17.3.exon6.E	1.26E-05	0.0703	85.40	5.88
T01C4.2a.exon5.E	7.11E-06	0.0520	44.09	4.60
T01G5.2.exon2.E	2.12E-05	0.0951	5.69	154.87
T04C9.1a.exon1.S	4.44E-05	0.1437	9.92	183.60
T09B4.8.exon1.E	3.48E-05	0.1312	275.56	36.83
T09E11.2.exon1.E	9.13E-07	0.0167	7.56	532.25
T09E11.2.exon1.S	4.82E-05	0.1535	19.35	353.40
T10B11.6.exon3.E	3.68E-05	0.1323	226.80	10.15
T14B4.1.exon4.S	4.70E-07	0.0147	4.27	85.97
W02H5.2.exon2.S	6.38E-06	0.0500	6.80	59.11
W04E12.4.exon1.S	3.88E-07	0.0142	9.20	204.38
W08E12.8.exon3.S	8.82E-07	0.0167	3.35	76.82
Y37H2A.5.exon3.E	1.55E-06	0.0210	9.49	159.41
Y41D4A.4.1.exon3.E	9.15E-06	0.0618	14.10	468.30
Y41D4B.10.exon2.E	3.39E-05	0.1312	59.38	6.67
Y42G9A.4a.exon2.E	4.98E-06	0.0455	60.37	4.91
Y44E3A.2.exon1.S	4.13E-05	0.1405	5.40	108.83
Y48G8AR.1.exon1.E	3.58E-05	0.1312	46.82	8.52
Y56A3A.4.exon2.E	7.52E-08	0.0055	73.04	2.61
Y59C2A.3.exon6.E	1.56E-05	0.0761	6.80	32.55
Y69A2AR.18a.1.exon2.S	5.96E-05	0.1787	4.93	25.20
Y73B6A.1.exon2.S	1.44E-07	0.0079	2.17	47.92
Y73B6A.4.exon1.S	3.23E-05	0.1312	5.86	62.37
Y77E11A.1.1.exon2.E	4.03E-05	0.1405	30.85	6.53
Y79H2A.3a.exon6.S	2.22E-05	0.0974	6.74	62.58
Y80D3A.8.exon8.S	1.76E-06	0.0215	5.60	96.74
ZK355.2a.exon3.E	3.59E-05	0.1312	4.46	48.65
ZK381.1.exon4.E	1.25E-06	0.0196	10.76	134.93
ZK524.2a.exon22.E	5.68E-06	0.0479	9.09	123.50
ZK757.4a.exon1.S	1.12E-05	0.0695	92.53	7.71
ZK858.5.exon2.E	1.88E-05	0.0858	128.05	7.81

b: splicing defects at boundary regions (*smg-2* vs. *teg-4 smg-2*)

region name	p value	FDR	*smg-2*	*teg-4 smg-2*
AC3.2.exon2.E	8.63E-05	0.1606	4.84	143.18
B0024.14a.exon5.E	5.25E-05	0.1585	49.34	2.63
B0198.2.exon2.S	7.39E-05	0.1606	116.35	4.27
B0207.9.exon1.S	2.14E-04	0.1974	76.56	6.05
B0280.12a.exon12.S	2.12E-04	0.1974	69.01	7.36
B0302.1a.exon15.E	1.53E-04	0.1840	94.76	5.77
B0379.3a.exon5.E	2.15E-05	0.1399	6.55	550.16
B0412.1a.exon6.E	1.79E-04	0.1914	53.50	4.56
B0416.7a.exon3.E	1.80E-04	0.1914	116.98	6.93
B0454.9.exon3.E	1.88E-04	0.1914	91.56	5.38
B0457.1a.exon6.S	2.28E-04	0.1984	55.80	6.37
B0457.4.exon2.S	1.15E-05	0.1399	102.02	4.06
B0464.3.1.exon1.E	1.70E-04	0.1906	113.13	2.84
B0491.4.exon5.E	1.51E-04	0.1834	43.99	7.17
B0513.4.exon1.E	6.69E-05	0.1606	5.69	314.04
B0546.3.exon3.S	7.07E-05	0.1606	51.97	4.03
B0546.4b.exon3.S	6.33E-05	0.1606	7.60	77.28
C01C10.3.exon4.S	2.28E-04	0.1984	67.18	3.97
C01F1.6.exon3.S	9.90E-05	0.1707	4.39	67.51
C02B10.3.1.exon2.S	1.48E-04	0.1810	60.71	3.90
C02B8.1.1.exon2.E	2.28E-04	0.1984	5.44	175.25
C04E12.4.exon7.E	5.68E-05	0.1606	4.22	47.99
C05D12.1.exon10.E	3.66E-05	0.1477	80.93	4.47
C06G3.8.exon2.S	9.37E-05	0.1707	3.47	63.02
C06H5.8.exon1.S	4.38E-05	0.1566	5.41	283.28
C08F1.6.exon1.E	2.11E-04	0.1974	7.43	86.89
C09D4.5.1.exon4.S	1.58E-05	0.1399	3.13	106.68
C10A4.3.exon2.S	1.70E-04	0.1906	64.06	4.34
C14C10.7.exon3.E	8.17E-06	0.1232	117.74	4.25
C17B7.9.exon2.E	1.88E-04	0.1914	57.70	5.34
C17E7.11.exon4.S	9.83E-05	0.1707	6.14	86.20
C17G1.7.1.exon1.S	1.91E-04	0.1920	60.58	5.19
C17H12.8.exon6.S	9.41E-05	0.1707	53.24	4.40
C23H4.7.exon6.S	1.81E-05	0.1399	3.17	94.20
C27C12.2.exon4.E	3.08E-05	0.1477	89.12	3.57
C27F2.10.1.exon8.E	1.45E-04	0.1805	5.12	37.53
C28C12.7a.1.exon4.S	2.57E-05	0.1416	4.56	78.19
C29F4.2.exon4.S	2.22E-04	0.1984	56.98	4.50

C32F10.4.1.exon3.E	6.81E-05	0.1606	59.70	6.91
C33A12.4.exon1.S	1.04E-04	0.1744	5.79	78.01
C33B4.3a.exon3.S	1.88E-04	0.1914	235.48	36.07
C33C12.4.exon3.S	3.67E-05	0.1477	3.46	95.90
C34E10.5.1.exon5.E	1.25E-04	0.1744	4.27	38.79
C34F6.3.exon1.S	1.55E-04	0.1851	7.37	277.55
C35A5.7.exon9.E	1.23E-04	0.1744	4.46	144.87
C35B1.8.exon2.E	1.33E-04	0.1752	83.51	4.09
C35D10.11.1.exon1.S	1.22E-04	0.1744	5.05	129.09
C44C10.5.exon1.S	1.75E-04	0.1914	55.45	3.29
C46H3.2a.exon11.S	2.15E-04	0.1978	52.57	4.42
C48D1.5.exon1.E	8.40E-05	0.1606	78.58	7.59
C51E3.1.exon3.E	1.21E-04	0.1744	40.98	4.27
C53A5.2.exon5.E	5.16E-06	0.1192	7.87	222.13
C54D1.1.exon22.E	1.28E-05	0.1399	148.63	6.84
D1037.2.exon1.E	1.57E-04	0.1851	568.76	11.81
D1037.2.exon11.S	5.97E-06	0.1192	9.49	1050.65
D1086.10.1.exon2.E	1.84E-04	0.1914	9.09	185.79
D2030.4.1.exon3.S	7.89E-05	0.1606	50.16	4.77
E02A10.3.exon3.S	2.06E-04	0.1974	83.60	5.01
E03H4.4.exon6.S	5.49E-05	0.1585	9.04	93.24
E04F6.10.exon1.E	1.78E-04	0.1914	23.52	3.61
F07B10.1.exon5.S	1.32E-04	0.1752	6.50	95.08
F08B12.1.exon11.E	6.42E-05	0.1606	324.92	5.13
F08D12.12.1.exon4.E	6.30E-05	0.1606	126.04	7.30
F09C8.2.1.exon4.E	7.58E-05	0.1606	47.05	3.66
F09D1.1.exon3.E	9.62E-05	0.1707	43.76	5.88
F11A10.8.exon1.S	8.59E-05	0.1606	5.82	58.70
F11A6.1a.exon1.S	2.29E-04	0.1984	11.74	252.07
F11C1.6a.1.exon11.E	2.29E-06	0.1192	209.95	4.28
F11C7.6a.exon1.S	1.97E-04	0.1944	5.57	58.66
F12B6.1.exon9.S	1.39E-05	0.1399	52.14	3.08
F14B8.6.exon9.S	5.44E-05	0.1585	96.50	3.48
F15D3.8.exon4.E	2.32E-04	0.1984	31.46	6.25
F15D4.5.exon6.E	3.41E-05	0.1477	3.40	248.20
F15G9.4a.exon60.S	2.32E-04	0.1984	74.78	7.12
F17H10.1.1.exon10.E	7.96E-05	0.1606	174.59	5.23
F19B10.4.exon6.S	1.67E-04	0.1906	8.04	139.45
F20G4.3.exon4.S	2.03E-04	0.1974	6.53	43.72
F26A1.11.exon3.S	4.69E-05	0.1583	1.98	51.94
F26C11.3.exon20.S	5.85E-06	0.1192	142.26	2.08

F26D2.13.exon2.S	1.49E-05	0.1399	3.23	314.06
F26F2.7.exon12.S	3.11E-05	0.1477	84.38	4.14
F29F11.6.1.exon2.E	2.05E-05	0.1399	78.85	5.64
F29G6.1.exon3.E	1.39E-04	0.1783	94.13	4.61
F31F6.4a.exon2.S	1.20E-04	0.1744	6.12	153.95
F32G8.4.exon3.S	8.02E-05	0.1606	5.74	272.12
F33C8.1a.exon18.S	6.44E-05	0.1606	70.00	3.97
F36D3.1.1.exon3.E	7.11E-05	0.1606	62.93	4.04
F37A4.2.1.exon1.S	6.24E-05	0.1606	6.05	63.26
F37B12.3.exon4.S	1.25E-04	0.1744	165.56	6.09
F38B6.6.exon3.S	1.20E-04	0.1744	28.79	4.00
F38B6.6.exon8.E	1.69E-04	0.1906	22.91	3.88
F38E11.7.exon4.S	2.13E-04	0.1974	4.63	28.89
F39B1.1.exon3.E	1.14E-04	0.1744	81.50	5.72
F39C12.2a.exon2.S	1.70E-04	0.1906	7.11	73.54
F39E9.3.exon4.E	8.05E-05	0.1606	124.30	7.99
F40G9.3.exon3.S	7.63E-05	0.1606	4.42	98.03
F43G9.6.exon17.S	7.85E-06	0.1232	3.10	116.72
F45E4.11.exon3.S	7.26E-05	0.1606	6.10	96.67
F45E6.2.exon8.E	1.85E-04	0.1914	64.55	6.32
F46A8.4.exon1.S	2.23E-04	0.1984	65.64	3.44
F46A9.4.exon2.S	1.19E-04	0.1744	41.96	4.07
F46C3.3.exon28.E	2.24E-04	0.1984	24.96	4.71
F46F3.4.exon1.S	1.87E-04	0.1914	4.61	105.43
F48B9.8.exon1.E	1.21E-04	0.1744	2.27	20.03
F48G7.7.exon1.S	5.40E-05	0.1585	43.91	4.59
F49E11.1b.exon2.E	1.72E-04	0.1914	81.31	6.18
F52B11.3.1.exon1.E	4.42E-05	0.1566	57.98	2.55
F52D10.1.exon3.E	4.87E-05	0.1585	79.71	4.19
F53B6.6.exon9.E	5.42E-05	0.1585	6.35	109.05
F56C4.1.exon1.S	1.20E-04	0.1744	29.08	2.18
F56D2.1.1.exon9.S	2.07E-04	0.1974	3.02	70.62
F56D6.1.exon2.S	1.60E-05	0.1399	1.86	76.93
F56F11.5.exon6.S	2.89E-05	0.1442	4.00	43.25
F56F3.1.exon4.S	4.64E-06	0.1192	2.82	105.66
F56H6.6.exon7.S	1.77E-04	0.1914	6.56	99.91
F57B10.8.exon3.E	1.82E-04	0.1914	7.97	125.72
F58B6.3b.exon7.E	5.48E-05	0.1585	4.49	53.90
F59F5.8.exon4.S	4.57E-05	0.1568	7.80	104.57
H14N18.1c.exon1.E	1.41E-04	0.1783	100.14	8.70
H23L24.4.exon2.S	1.93E-04	0.1920	76.56	3.75

K01A11.4.1.exon11.E	1.79E-04	0.1914	3.56	162.46
K02F2.1a.exon2.E	1.67E-04	0.1906	61.87	4.72
K03A1.5.exon9.E	2.13E-04	0.1974	6.71	133.30
K03D7.4.exon3.S	2.00E-04	0.1956	3.33	33.22
K03D7.7.exon3.S	8.48E-05	0.1606	3.04	41.85
K04F10.6a.exon2.E	2.81E-05	0.1442	28.90	2.37
K06H6.1.exon1.S	7.64E-05	0.1606	139.34	5.58
K08B4.6.exon3.S	8.33E-05	0.1606	30.97	4.55
K08C9.2.exon5.E	9.87E-05	0.1707	3.64	54.63
K08E7.5a.exon10.S	1.32E-04	0.1752	67.28	3.86
K08F8.6.exon12.E	1.32E-04	0.1752	5.07	63.45
K08H2.8.exon2.S	1.35E-04	0.1754	8.43	99.40
K09A9.4.1.exon9.E	5.91E-05	0.1606	6.86	172.29
K10E9.1.exon4.E	1.78E-04	0.1914	66.39	3.13
K10G9.3.exon1.E	1.69E-05	0.1399	4.15	120.07
K11D12.10a.exon8.E	2.24E-04	0.1984	4.31	38.32
K11G9.5.exon6.S	4.32E-06	0.1192	111.53	2.56
M03C11.7.1.exon5.E	1.05E-05	0.1355	1.98	56.59
M05D6.7.exon3.E	1.55E-04	0.1851	116.35	8.90
M153.4.exon4.S	1.05E-04	0.1744	3.43	36.61
M18.1.exon2.S	1.83E-04	0.1914	4.03	77.96
M7.3.exon10.S	2.19E-04	0.1984	5.59	35.23
R03E9.3a.exon17.S	1.24E-04	0.1744	6.94	45.15
R05D7.2.exon2.E	2.83E-05	0.1442	3.21	57.84
R07A4.3.exon4.E	1.87E-05	0.1399	92.10	2.41
R07C12.1.exon4.S	1.99E-04	0.1956	31.71	3.76
R07C3.14.exon2.E	1.22E-04	0.1744	2.78	54.29
R08F11.3.exon3.E	3.42E-05	0.1477	5.95	176.05
R106.2.exon8.S	1.59E-05	0.1399	54.24	3.46
R10E4.3.exon1.E	6.68E-05	0.1606	83.45	3.18
R13A5.9.exon2.E	9.89E-05	0.1707	4.53	65.19
R31.1.exon5.S	1.84E-04	0.1914	75.56	5.49
R53.5.exon1.E	1.57E-04	0.1851	123.53	5.66
R90.1.exon11.S	1.18E-04	0.1744	6.25	88.89
T01B7.9.exon1.E	2.10E-04	0.1974	3.06	22.86
T01D3.2.exon2.E	1.30E-04	0.1752	4.50	51.94
T01D3.4.exon2.E	9.92E-06	0.1355	83.79	2.05
T01H8.1a.exon6.S	1.46E-04	0.1810	219.09	7.33
T02E1.8.exon5.E	1.38E-04	0.1777	63.24	4.60
T02G5.3.exon3.E	6.95E-05	0.1606	4.18	106.27
T04C9.6a.1.exon7.E	1.35E-04	0.1754	4.34	92.14

T04H1.4.exon5.E	5.13E-05	0.1585	5.35	59.62
T05E12.2.exon4.S	1.66E-04	0.1906	36.13	3.95
T05G5.8.exon1.E	3.33E-05	0.1477	2.21	80.43
T06C10.4.1.exon1.S	3.48E-05	0.1477	72.72	2.52
T09E11.10.exon5.S	7.72E-05	0.1606	2.74	60.57
T09E11.2.exon1.E	4.70E-07	0.0628	5.75	1762.16
T09E11.2.exon5.E	4.14E-06	0.1192	5.94	323.39
T11F9.4.exon4.S	1.21E-04	0.1744	4.87	42.75
T13G4.3.exon15.S	1.84E-04	0.1914	107.64	5.60
T13H2.5a.exon10.E	7.09E-05	0.1606	130.17	5.33
T17H7.4a.1.exon2.S	8.73E-05	0.1610	20.31	2.20
T20B5.3.exon13.S	5.42E-05	0.1585	104.16	6.63
T21D12.7.exon1.E	1.10E-04	0.1744	31.72	4.83
T22B7.3.exon2.S	5.37E-05	0.1585	5.35	116.95
T22D2.1.exon15.E	1.61E-04	0.1891	37.44	5.74
T23B7.3.exon3.E	1.68E-04	0.1906	157.19	5.98
T24H10.7a.exon5.S	2.12E-04	0.1974	28.55	3.66
T25E12.4a.exon6.E	3.92E-05	0.1483	280.25	6.21
T26H2.9.exon6.E	1.18E-04	0.1744	96.53	6.14
T27A10.3a.1.exon6.S	6.91E-05	0.1606	97.16	7.74
T27E9.4a.1.exon10.S	2.29E-05	0.1399	5.90	107.93
T28D6.7.exon1.E	1.47E-04	0.1810	66.49	4.95
T28F12.2a.1.exon7.S	3.80E-05	0.1477	4.76	46.56
T28F2.2.exon3.S	3.36E-06	0.1192	4.79	159.10
T28F2.5.exon4.S	2.26E-04	0.1984	86.45	5.16
VW02B12L.1.1.exon6.E	7.25E-05	0.1606	4.89	48.19
VZK822L.2.exon3.E	3.56E-05	0.1477	102.03	7.41
W02B3.4.exon2.E	2.32E-04	0.1984	62.47	7.01
W03D2.4.1.exon1.E	2.39E-05	0.1399	2.69	42.85
W03D8.6.1.exon27.E	7.29E-05	0.1606	130.88	12.20
W04A4.2.exon2.S	2.21E-05	0.1399	2.61	116.28
W04E12.7.exon3.S	8.02E-05	0.1606	3.88	46.49
W04G3.6d.exon9.S	6.70E-05	0.1606	3.24	93.28
W06A11.3.exon1.S	1.13E-04	0.1744	3.93	23.95
W06A7.3a.exon1.S	6.26E-05	0.1606	4.01	70.17
W06D12.5.exon6.S	2.22E-04	0.1984	35.94	4.60
W10C8.2.exon1.E	4.75E-06	0.1192	2.45	90.72
Y113G7A.4a.exon1.S	1.41E-04	0.1783	3.06	131.67
Y119C1B.8a.exon2.E	8.63E-05	0.1606	65.89	3.69
Y18H1A.6.exon4.E	2.01E-04	0.1963	90.36	3.33
Y24D9A.6.exon2.E	1.51E-04	0.1834	63.98	5.16

Y32G9A.6.exon5.S	1.67E-04	0.1906	4.52	45.29
Y32G9A.8.exon4.S	2.31E-04	0.1984	3.79	39.29
Y34D9B.1a.exon3.E	6.04E-05	0.1606	4.12	54.64
Y38A10A.6.exon5.E	1.43E-04	0.1796	3.87	75.86
Y38C9A.2.1.exon1.E	8.37E-06	0.1232	214.94	6.94
Y38E10A.20.exon4.E	1.18E-04	0.1744	6.31	94.96
Y39B6A.47.exon13.S	4.01E-05	0.1493	197.49	2852.60
Y39D8B.1.exon2.E	2.30E-04	0.1984	5.48	57.44
Y39G10AL.3.1.exon7.E	1.56E-05	0.1399	6.76	107.29
Y41D4B.18.exon1.S	2.10E-04	0.1974	184.53	6.51
Y45F10A.3.exon1.S	1.25E-04	0.1744	6.62	75.69
Y46G5A.15.exon10.E	4.91E-05	0.1585	163.94	3.18
Y47D3A.6a.exon10.E	1.73E-05	0.1399	5.79	78.83
Y48E1B.13c.exon3.E	1.08E-04	0.1744	2.91	55.77
Y48E1B.14a.1.exon7.E	2.17E-04	0.1984	6.87	54.48
Y48G10A.2.exon7.E	8.36E-05	0.1606	33.97	3.31
Y48G1A.1.exon6.S	8.42E-06	0.1232	97.07	4.23
Y48G1A.6.exon1.S	1.09E-04	0.1744	2.94	146.18
Y4C6B.1.exon2.E	1.24E-04	0.1744	52.89	5.00
Y50E8A.10.exon4.S	3.75E-05	0.1477	54.79	4.01
Y51H4A.4.exon3.E	1.41E-04	0.1783	4.59	49.22
Y51H7C.9.exon10.S	2.65E-05	0.1416	42.15	2.73
Y53F4B.12.exon6.S	5.73E-05	0.1606	114.58	2.76
Y53F4B.21.exon4.E	2.42E-05	0.1399	4.05	70.51
Y53G8AR.2a.exon3.S	8.36E-05	0.1606	148.08	3.63
Y54F10AL.1a.exon7.E	1.93E-04	0.1920	76.37	5.88
Y54G2A.28.exon1.E	1.03E-04	0.1744	3.82	103.64
Y55F3AM.3a.exon2.E	1.24E-04	0.1744	9.58	141.98
Y57A10A.28.exon2.S	9.56E-05	0.1707	4.72	100.25
Y57E12AL.2.exon2.E	1.78E-04	0.1914	60.34	5.27
Y62H9A.10.exon11.S	7.75E-05	0.1606	4.78	53.58
Y64G10A.7.exon4.E	4.53E-05	0.1568	60.55	3.90
Y66A7A.6.1.exon8.S	2.59E-05	0.1416	2.21	122.94
Y68A4A.7.exon3.E	1.30E-04	0.1752	5.28	68.38
Y69A2AR.19.exon1.E	4.40E-05	0.1566	73.81	4.63
Y71A12B.8.exon11.S	1.26E-04	0.1746	25.27	3.36
Y71D11A.5.exon10.S	2.31E-05	0.1399	6.84	102.77
Y71G12B.2.exon1.E	7.36E-05	0.1606	5.02	130.14
Y73B3A.20.exon3.E	5.73E-07	0.0628	3126.12	5.74
Y73F8A.24.exon1.E	3.84E-05	0.1477	80.22	5.15
Y79H2A.11.exon1.S	1.23E-04	0.1744	4.06	55.05

Y9C9A.2.exon4.E	1.32E-04	0.1752	6.38	42.06
ZC21.6b.exon1.S	2.02E-05	0.1399	5.14	165.70
ZC250.1.exon1.E	2.08E-05	0.1399	5.22	72.63
ZC434.2.1.exon4.E	2.22E-04	0.1984	5.27	67.14
ZC53.1.exon5.S	2.14E-04	0.1974	3.83	47.99
ZC581.7.exon7.E	1.23E-04	0.1744	3.43	115.00
ZK1127.5.exon2.S	2.11E-04	0.1974	9.93	81.21
ZK1193.2.exon17.E	1.26E-04	0.1744	3.59	27.77
ZK20.4.1.exon3.E	3.46E-05	0.1477	3.30	51.09
ZK265.7.1.exon1.E	9.95E-05	0.1707	127.15	5.99
ZK455.8a.exon8.E	1.92E-04	0.1920	44.14	5.88
ZK488.5.exon7.E	1.83E-04	0.1914	562.51	15.04
ZK488.6.exon3.S	1.92E-04	0.1920	75.78	5.97
ZK909.2f.exon1.S	1.31E-04	0.1752	46.32	3.87

c: splicing defects identified using single probe analysis (N2 vs. *teg-4*)

name	chr	location	intensity (in *teg-4*)
B0250.5.exon3.E	chrV	20489664	324
B0379.6.exon1.E	chrI	10097526	370
C04G6.2.exon4.E	chrII	5094018	325
C11H1.9a.exon7.E	chrX	14318589	305
C12C8.1.exon3.E	chrI	9321245	571
C18F10.8.exon1.S	chrIII	6267880	293
C34E10.9.exon4.E	chrIII	5262913	638
C41G7.6.exon1.E	chrI	9527366	795
C50D2.6.exon3.S	chrII	85339	274
C55F2.2.exon3.E	chrIV	7896470	458
F10F2.8.exon4.E	chrIII	4636487	343
F12F3.1a.1.exon1.S	chrV	6141564	338
F12F3.1b.exon2.S	chrV	6141512	250
F13E9.11.exon2.S	chrIV	10866321	270
F23B2.5a.1.exon6.E	chrIV	9145958	667
F23B2.5a.1.exon6.S	chrIV	9145882	572
F29C12.1a.1.exon1.E	chrII	13108088	245
F31C3.2a.exon7.S	chrI	15038745	416
F35D11.10.exon6.E	chrII	4620209	227
F36A4.1.exon2.E	chrIV	4272751	589
F36H1.2a.exon23.E	chrIV	11046491	332
F47G4.4.1.exon8.S	chrI	14064467	216
F48F5.2.exon4.E	chrV	20441211	366
F52B11.3.1.exon2.E	chrIV	14094491	266
F55C7.2.exon1.E	chrI	4005471	217
F56A6.1a.exon1.S	chrI	618188	233
F56B3.6.exon3.E	chrIV	790732	344
F59A7.3.exon3.S	chrV	2026481	277
K02F2.6.exon7.S	chrI	6820754	506
K07E8.9.exon3.S	chrII	661138	353
M01E5.3.1.exon5.S	chrI	13282435	658
M01E5.5a.exon5.E	chrI	13296835	306
M60.5.exon6.E	chrX	8245327	318
R02E4.3.exon2.S	chrX	4101773	201
R06B10.7.exon1.E	chrIII	1001378	205
R07B1.9.exon7.S	chrX	9878489	499
R07E3.2.exon3.S	chrX	10336483	541

R13H4.3.exon1.E	chrV	11833610	260
T01A4.3.exon2.E	chrI	4465370	607
T01G1.2.exon1.S	chrIV	11354385	298
T05A1.2.exon2.E	chrIV	9564359	1182
T14G8.4.exon7.S	chrX	12862284	234
T23G7.1.exon11.S	chrII	9172017	365
T25B9.10.1.exon8.E	chrIV	10771297	437
T28D6.5b.exon11.E	chrIII	11340953	719
T28D9.4a.exon6.E	chrII	6477929	846
VF13D12L.1.1.exon1.E	chrII	11700146	499
W02A11.3.1.exon1.S	chrI	12738084	562
W02B8.1.exon2.S	chrII	13900462	436
W05B2.2.exon3.S	chrIII	10966084	247
W05B5.3a.exon3.E	chrI	13059085	441
W07A8.5.exon4.S	chrV	20730846	239
W07E11.2.exon3.S	chrX	10092306	471
W07E11.3a.exon2.S	chrX	10094549	415
W09D6.6.exon4.S	chrIII	11097517	279
W09G3.1a.exon3.S	chrI	13829182	715
Y105E8A.10a.exon8.S	chrI	14429956	326
Y105E8B.1f.exon7.E	chrI	14623685	287
Y105E8B.8a.exon6.E	chrI	14671334	505
Y116F11B.8.exon4.E	chrV	19855824	496
Y34D9A.2.exon1.E	chrI	1062735	244
Y37E3.16.1.exon3.S	chrI	2117124	279
Y38H8A.1.exon3.E	chrIV	13485364	221
Y39A1A.22.exon1.S	chrIII	10705782	422
Y40B1B.6.1.exon2.S	chrI	13438882	192
Y40B1B.6.1.exon3.S	chrI	13440785	484
Y42G9A.3a.exon5.E	chrIII	6147885	238
Y45G5AM.1a.1.exon1.S	chrV	4187015	271
Y47D3A.22.exon4.E	chrIII	11288937	650
Y47G6A.15b.exon2.E	chrI	3474409	387
Y48C3A.4.exon5.S	chrII	13280028	208
Y52B11B.1.exon1.E	chrI	11123897	218
Y52B11B.1.exon1.S	chrI	11123735	320
Y54E2A.7.exon3.S	chrII	14772065	272
Y54E5B.1a.exon12.E	chrI	14814943	330
Y57G11C.18.exon1.S	chrIV	14842707	517

Y57G11C.24a.1.exon3.E	chrIV	14886796	235
Y58G8A.1.exon4.E	chrV	269273	237
Y65B4A.4.exon3.E	chrI	622191	214
Y76A2B.1.exon6.S	chrIII	13516560	501
Y87G2A.10.1.exon1.S	chrI	13588712	502
ZC101.3.exon1.S	chrII	14678394	450
ZC376.4.exon2.S	chrV	14178247	303
ZC8.4a.exon12.S	chrX	4989875	490
ZC8.4a.exon18.S	chrX	4992575	368
ZC84.3.exon1.E	chrIII	9190963	275
ZC84.3.exon9.S	chrIII	9187893	189
ZK105.1.exon2.S	chrV	7713981	1476
ZK256.1a.1.exon11.E	chrI	13023258	371
ZK256.1a.1.exon3.S	chrI	13012196	192
ZK256.1a.1.exon6.S	chrI	13015138	323
ZK337.1a.1.exon1.S	chrI	14994758	489
ZK337.1a.1.exon14.S	chrI	14999147	492
ZK370.8.exon4.S	chrIII	8753842	451
ZK370.8.exon7.S	chrIII	8752925	283
ZK930.7.exon4.E	chrII	11915246	368

d: splicing defects identified using single probe analysis (*smg-2* vs. *teg-4 smg-2*)

name	chr	location	intensity (in *teg-4 smg-2*)
B0212.1.exon1.E	chrIV	3572642	271
B0250.2.exon5.S	chrV	20472083	198
B0379.6.exon1.E	chrI	10097552	191
BE10.1.exon5.S	chrIII	12787547	227
C09G5.3.exon2.E	chrII	10704425	242
C11H1.9a.exon3.E	chrX	14319904	233
C17D12.1a.exon7.S	chrI	11615015	192
C23H3.7.exon3.E	chrII	54679	244
C30D11.1.exon14.E	chrIII	4144941	248
C32F10.4.1.exon1.E	chrI	5809984	264
C35C5.10.exon4.E	chrX	11558172	329
C41C4.5.1.exon7.S	chrII	8120491	260
C47B2.9.exon1.E	chrI	12983223	191
C50E3.6.exon2.S	chrV	7590046	188
C53B4.5.exon1.S	chrIV	8978903	292
F10C2.3.exon7.S	chrV	12022917	204
F10F2.4.exon3.S	chrIII	4624062	381
F10F2.7.exon3.S	chrIII	4629146	211
F12F3.1c.exon1.E	chrV	6140844	217
F15E6.3.1.exon2.S	chrIV	4295316	203
F15G9.5.exon2.E	chrX	9734347	245
F23B2.5a.1.exon1.S	chrIV	9144691	193
F23B2.5a.1.exon2.S	chrIV	9144790	388
F23B2.5a.1.exon6.E	chrIV	9145958	305
F23B2.5a.1.exon6.S	chrIV	9145882	572
F25D7.3.exon7.S	chrI	10413325	310
F32A7.4.exon6.E	chrI	14842771	208
F32E10.2.exon2.S	chrIV	7576427	214
F35D11.3.1.exon2.E	chrII	4609175	217
F35D11.9.exon3.S	chrII	4619145	278
F36A2.4.exon3.S	chrI	8810964	196
F37F2.3.exon3.E	chrI	1451522	233
F39D8.1a.exon2.E	chrX	15413152	219
F47F2.2.exon1.E	chrX	3887528	190
F48F5.1.exon2.E	chrV	20437342	215
F48F5.2.exon3.E	chrV	20441471	195
F48F5.2.exon4.E	chrV	20441211	286

F54C8.1.exon2.S	chrIII	9440060	252
F56B3.6.exon1.E	chrIV	788268	217
F58D5.5.exon1.E	chrI	12041404	194
F58E2.5.exon1.E	chrIV	3445154	195
F59C6.8.exon1.S	chrI	10509080	198
F59E12.12.exon1.S	chrII	5652313	241
H08M01.1.exon9.S	chrIV	13834503	190
H12C20.6a.exon3.S	chrV	11571034	358
K02B12.9.exon3.S	chrI	8521858	392
K02F2.6.exon7.S	chrI	6820754	261
K04C2.5.exon2.E	chrIII	6891217	275
K08E7.5a.exon2.E	chrIV	12573963	199
K09E4.1.exon1.S	chrII	14137766	207
K10C9.7.exon2.E	chrV	1067252	313
M01E5.2.1.exon4.E	chrI	13278842	266
M03A8.3.exon6.S	chrX	6802418	249
R02F2.8.exon6.E	chrIII	5493429	384
R74.3a.exon1.E	chrIII	4194205	248
T01A4.3.exon2.E	chrI	4465370	248
T01G9.3.exon2.S	chrI	8282321	388
T07C12.7.1.exon2.E	chrV	9951487	290
T07F10.5.exon2.S	chrV	12865877	360
T10B10.1.exon1.E	chrX	15176368	215
T19E7.1.exon2.S	chrIV	5654511	414
T23G7.1.exon10.E	chrII	9171952	208
T23G7.1.exon11.S	chrII	9172017	309
T27C10.7.exon1.E	chrI	10878292	295
T28D6.5a.exon1.E	chrIII	11334814	218
T28D9.4a.exon6.E	chrII	6477929	230
VM106R.1.exon1.E	chrII	10829298	446
W05B5.3a.exon1.S	chrI	13061803	203
W07E11.2.exon2.E	chrX	10092720	436
W07E11.3a.exon1.E	chrX	10095073	192
W07E11.3a.exon2.S	chrX	10094549	267
W07E11.3b.exon2.E	chrX	10094920	216
W07E11.4.exon1.S	chrX	10087778	333
W09D6.6.exon4.S	chrIII	11097517	217
Y105E8A.10a.exon8.S	chrI	14429956	348
Y105E8A.7a.exon7.E	chrI	14396465	210

Y105E8B.2a.exon1.S	chrI	14642493	233
Y105E8B.8a.exon6.E	chrI	14671334	318
Y111B2A.27.exon1.E	chrIII	12532154	225
Y22D7AL.11.exon3.S	chrIII	1633206	256
Y34D9A.2.exon3.S	chrI	1064657	706
Y37E3.16.1.exon1.S	chrI	2118459	204
Y37E3.16.1.exon9.E	chrI	2113910	265
Y37E3.16.1.exon9.S	chrI	2113782	338
Y38F1A.4.exon1.S	chrII	12968843	262
Y38H8A.1.exon1.S	chrIV	13483500	191
Y39A1A.22.exon1.S	chrIII	10705782	249
Y45F10D.4.exon4.E	chrIV	13784570	304
Y47G6A.15b.exon4.S	chrI	3473067	200
Y48C3A.14.exon5.E	chrII	13394577	201
Y48C3A.4.exon4.E	chrII	13281393	278
Y48G9A.4.exon7.E	chrIII	2141773	208
Y50D7A.7.1.exon4.S	chrIII	296265	268
Y52B11A.5.exon2.E	chrI	11016465	260
Y53G8AR.7a.exon10.E	chrIII	3289028	288
Y54G2A.15.exon3.S	chrIV	2843594	207
Y54G2A.15.exon4.S	chrIV	2843436	246
Y54G2A.15.exon5.S	chrIV	2843266	328
Y57G11C.18.exon4.E	chrIV	14840678	195
Y63D3A.6a.1.exon4.S	chrI	14110937	222
Y69A2AR.22.exon4.S	chrIV	2504663	217
Y6B3A.1a.exon10.S	chrI	13616262	192
Y73B3A.6.exon2.S	chrX	127020	250
Y76A2A.1.exon2.S	chrIII	13460314	282
Y87G2A.10.1.exon1.E	chrI	13588792	190
Y87G2A.6.1.exon2.S	chrI	13570026	271
ZC101.3.exon6.S	chrII	14680123	388
ZK256.1a.1.exon3.S	chrI	13012196	192
ZK337.1a.1.exon15.E	chrI	15001034	293
ZK370.8.exon9.S	chrIII	8752208	215
ZK675.1.1.exon3.E	chrII	7893586	260
ZK930.7.exon4.E	chrII	11915246	233

APPENDIX E: ADDITIONAL METHODS

RNA isolation from *C. elegans*

1. Flush plate with 1ml of PBS, and transfer to a 1.5 ml epi-tube. Wash with PBS twice to get rid of the bacteria (centrifuge 1 min @ 1000 rpm).
2. Add 200 µL of Trizol into the epi-tube and vortex vigorously for 1min.
3. Freeze-thaw sample twice.
4. Add additional 200 µL of Trizol, vortex vigorously and incubate at room temperature for 5 minutes.
5. Centrifuge @ 14,000rpm for 10 minutes (at 4°C) to remove insoluble material.
6. Transfer liquid to a fresh 1.5 ml epi-tube and add 100 µL chloroform.
7. Invert/vortex for 15 seconds, and incubate at room temperature for 3 minutes.
8. Centrifuge @ 14,000rpm for 15 minutes (at 4°C) to separate phases.
9. Remove the upper aqueous phase to a fresh 1.5 ml tube, and add 500 µL isopropanol and mix.
10. Incubate at room temperature for 10 minutes, and Centrifuge @ 14,000rpm for 15 minutes (at 4°C) to precipitate and recover RNA.
11. Carefully remove aqueous away from the RNA pellet.
12. Wash the pellet with 100 µL of 75% ethanol (in nuclease-free H_2O) twice, and dissolve the pellet in 50 µL nuclease-free H_2O.

RNA purification

1. DNase I digestion.

   Add the followings to an epi-tube, mix well (gently) and incubate at 37°C for 30 minutes.

   50 µL total RNA + 5.7 µL reaction buffer (10×) +1.0 µL Dnase I (10 units/ µL)

2. Phenol/ Chloroform (1:1) extraction.
   1) Add 40 µL Phenol/ Chloroform (1:1), vortex 30 seconds.
   2) 10 minutes on ice.
   3) Add 50 µL nuclease-free H_2O.
   4) Spin at 14,000rpm (at 4°C) for 5 minutes.
   5) Collect the upper phase.
3. Ethanol precipitation.
   1) Add 20 µL of 3M NaOAc, 250 µL of 100% ethanol, place sample at –80°C for 1 hour.
   2) Spin at 14,000rpm (at 4°C) for 10 minutes and remove the supernatant.
   3) Wash the pellet with 100 µL of 75% ethanol (in nuclease-free H_2O) twice, and dissolve the pellet in 30 µL nuclease-free H_2O.

cDNA synthesis

Add the followings to an epi-tube, mix well (gently) and incubate at 65°C for 10 minutes.

5 µL Oligo dT (10 pmol/µL) +1 µL random primer (1.6 µg/µL) + 1 µL RNA + 6 µL nuclease-free H_2O

Place on ice immediately after incubation.

Add the followings: *4 µL Reverse Transcriptase Buffer (5×) + 2 µL dNTP (10mM, pre-mixed) + 0.5 µL Reverse Transcriptase (Roche, cat# 03531295001) + 0.5 µL Rnase inhibitor*

Incubate at 25°C for 10 minutes and then 55°C for 30 minutes.

2X SYBR Green PCR mix recipe

Per 10ml:

$MgCl_2$ (50Mm, made in nuclease-free H_2O): 1.2 ml

KCl (1M, made in nuclease-free H_2O): 1.6 ml

Tris-HCl (1M, made in nuclease-free H_2O): 0.4 ml

dNTP: 40 µL each

Tween: 100 µL

Glycerol (100%): 1ml

Fluorescein (20 µM): 10 µL

SYBR Green I (1/100, diluted in DMSO): 160 µL

nuclease-free H_2O: 5.37 ml

For each qPCR reaction (25 µL), add 12.5 µL 2X SYBR Green PCR mix.

Making RNAi plates

Per 3 liters:

$NaHPO_4$: 18 grams

KH_2PO_4: 9 grams

KH_4Cl: 3 grams

Casamino Acids: 15 grams

Agar: 60 grams

Autoclave with stir bar and then add:

3 ml of 1M $CaCl_2$

3 ml of 1M $MgSO_4$

3 ml of 5mg/ml cholesterol

Add the following only after the temperature has reached ~46°C:

30 ml of β-20% lactose (sterile filtered)

3.0 ml of 25mg/ml carbenicillin

References

Alexson, T.O., Hitoshi, S., Coles, B.L., Bernstein, A., and van der Kooy, D. (2006). Notch signaling is required to maintain all neural stem cell populations--irrespective of spatial or temporal niche. Dev Neurosci *28*, 34-48.
Allenspach, E.J., Maillard, I., Aster, J.C., and Pear, W.S. (2002). Notch signaling in cancer. Cancer Biol Ther *1*, 466-476.
Ast, G. (2004). How did alternative splicing evolve? Nat Rev Genet *5*, 773-782.
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.
Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V.*, et al.* (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics *7*, 246.
Bard, J., Zhelkovsky, A.M., Helmling, S., Earnest, T.N., Moore, C.L., and Bohm, A. (2000). Structure of yeast poly(A) polymerase alone and in complex with 3'-dATP. Science *289*, 1346-1349.
Belfiore, M., Pugnale, P., Saudan, Z., and Puoti, A. (2004). Roles of the C. elegans cyclophilin-like protein MOG-6 in MEP-1 binding and germline fates. Development *131*, 2935-2945.
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.
Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*, 291-336.
Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell *126*, 37-47.
Blumenthal, T., and Steward, K. (1997). RNA Processing and Gene Structure.
Bolos, V., Grego-Bessa, J., and de la Pompa, J.L. (2007). Notch signaling in development and cancer. Endocr Rev *28*, 339-363.
Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A., and Speed, T.P. (2005). Quality Assessment of Affymetrix GeneChip Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor (New York, Springer Science+Business Media, Inc.), pp. 33-47.
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.
Brand, M., Moggs, J.G., Oulad-Abdelghani, M., Lejeune, F., Dilworth, F.J., Stevenin, J., Almouzni, G., and Tora, L. (2001). UV-damaged DNA-binding protein in the TFTC complex links DNA damage recognition to nucleosome acetylation. Embo J *20*, 3187-3196.
Brenner, S. (1974). The genetics of Caenorhabditis elegans. Genetics *77*, 71-94.
Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat Genet *30*, 29-30.

Brow, D.A. (2002). Allosteric cascade of spliceosome activation. Annu Rev Genet *36*, 333-360.
Campbell, C., Risueno, R.M., Salati, S., Guezguez, B., and Bhatia, M. (2008). Signal control of hematopoietic stem cell fate: Wnt, Notch, and Hedgehog as the usual suspects. Curr Opin Hematol *15*, 319-325.
Chen, C.C., Simard, M.J., Tabara, H., Brownell, D.R., McCollough, J.A., and Mello, C.C. (2005a). A member of the polymerase beta nucleotidyltransferase superfamily is required for RNA interference in C. elegans. Curr Biol *15*, 378-383.
Chen, E.J., Frand, A.R., Chitouras, E., and Kaiser, C.A. (1998). A link between secretion and pre-mRNA processing defects in Saccharomyces cerevisiae and the identification of a novel splicing gene, RSE1. Mol Cell Biol *18*, 7139-7146.
Chen, L.L., Sabripour, M., Wu, E.F., Prieto, V.G., Fuller, G.N., and Frazier, M.L. (2005b). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. Oncogene *24*, 4271-4280.
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.
Consortium, C.e.S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018.
Corrionero, A., Minana, B., and Valcarcel, J. (2011). Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A. Genes Dev *25*, 445-459.
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.
Das, B.K., Xia, L., Palandjian, L., Gozani, O., Chyung, Y., and Reed, R. (1999). Characterization of a protein complex containing spliceosomal proteins SAPs 49, 130, 145, and 155. Mol Cell Biol *19*, 6796-6802.
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.
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.
Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419-437.
Folco, E.G., Coil, K.E., and Reed, R. (2011). The anti-tumor drug E7107 reveals an essential role for SF3b in remodeling U2 snRNP to expose the branch point-binding region. Genes Dev *25*, 440-444.
Fortini, M.E. (2009). Notch signaling: the core pathway and its posttranslational regulation. Dev Cell *16*, 633-647.
Gallegos, M., Ahringer, J., Crittenden, S., and Kimble, J. (1998). Repression by the 3' UTR of fem-3, a sex-determining gene, relies on a ubiquitous mog-dependent control in Caenorhabditis elegans. Embo J *17*, 6337-6347.
Gilbert, D., and Rechtsteiner, A. (2009). Comments on sequence normalization of tiling array expression. Bioinformatics *25*, 2171-2173.

Graham, P.L., and Kimble, J. (1993). The mog-1 gene is required for the switch from spermatogenesis to oogenesis in Caenorhabditis elegans. Genetics *133*, 919-931.
Graham, P.L., Schedl, T., and Kimble, J. (1993). More mog genes that influence the switch from spermatogenesis to oogenesis in the hermaphrodite germ line of Caenorhabditis elegans. Dev Genet *14*, 471-484.
Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet *17*, 100-107.
Habara, Y., Urushiyama, S., Shibuya, T., Ohshima, Y., and Tani, T. (2001). Mutation in the prp12+ gene encoding a homolog of SAP130/SF3b130 causes differential inhibition of pre-mRNA splicing and arrest of cell-cycle progression in Schizosaccharomyces pombe. Rna *7*, 671-681.
Hall, D.H., Winfrey, V.P., Blaeuer, G., Hoffman, L.H., Furuta, T., Rose, K.L., Hobert, O., and Greenstein, D. (1999). Ultrastructural features of the adult hermaphrodite gonad of Caenorhabditis elegans: relations between the germ line and soma. Dev Biol *212*, 101-123.
Hammond, S.M., Caudy, A.A., and Hannon, G.J. (2001). Post-transcriptional gene silencing by double-stranded RNA. Nat Rev Genet *2*, 110-119.
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.
Hansen, D., and Schedl, T. (2006). The regulatory network controlling the proliferation-meiotic entry decision in the Caenorhabditis elegans germ line. Curr Top Dev Biol *76*, 185-215.
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.
Hasegawa, M., Miura, T., Kuzuya, K., Inoue, A., Won Ki, S., Horinouchi, S., Yoshida, T., Kunoh, T., Koseki, K., Mino, K.*, et al.* (2011). Identification of SAP155 as the Target of GEX1A (Herboxidiene), an Antitumor Natural Product. ACS Chem Biol *6*, 229-233.
Henderson, S.T., Gao, D., Lambie, E.J., and Kimble, J. (1994). lag-2 may encode a signaling ligand for the GLP-1 and LIN-12 receptors of C. elegans. Development *120*, 2913-2924.
Hitoshi, S., Alexson, T., Tropepe, V., Donoviel, D., Elia, A.J., Nye, J.S., Conlon, R.A., Mak, T.W., Bernstein, A., and van der Kooy, D. (2002). Notch pathway molecules are essential for the maintenance, but not the generation, of mammalian neural stem cells. Genes Dev *16*, 846-858.
Hope, I.A. (1999). C. elegans : a practical approach (Oxford ; New York, Oxford University Press).
Huber, W., Irizarry, R.A., and Gentleman, R. (2005). Preprocessing Overview. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor (New York, Springer Science+Business Media, Inc.), pp. 3-12.
Isken, O., and Maquat, L.E. (2008). The multiple lives of NMD factors: balancing roles in gene and genome regulation. Nat Rev Genet *9*, 699-712.
Izquierdo, J.M. (2008). Fas splicing regulation during early apoptosis is linked to caspase-mediated cleavage of U2AF65. Mol Biol Cell *19*, 3299-3307.

Jan, E., Motzny, C.K., Graves, L.E., and Goodwin, E.B. (1999). The STAR protein, GLD-1, is a translational regulator of sexual identity in Caenorhabditis elegans. EMBO J *18*, 258.
Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141-2144.
Jones, A.R., and Schedl, T. (1995). Mutations in gld-1, a female germ cell-specific tumor suppressor gene in Caenorhabditis elegans, affect a conserved domain also found in Src-associated protein Sam68. Genes Dev *9*, 1491-1504.
Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell *12*, 5-14.
Kadyk, L.C., and Kimble, J. (1998). Genetic regulation of entry into meiosis in Caenorhabditis elegans. Development *125*, 1803-1813.
Kaida, D., Motoyoshi, H., Tashiro, E., Nojima, T., Hagiwara, M., Ishigami, K., Watanabe, H., Kitahara, T., Yoshida, T., Nakajima, H*., et al.* (2007). Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. Nat Chem Biol *3*, 576-583.
Kamath, R.S., and Ahringer, J. (2003). Genome-wide RNAi screening in Caenorhabditis elegans. Methods (San Diego, Calif *30*, 313-321.
Karijolich, J., and Yu, Y.T. (2010). Spliceosomal snRNA modifications and their function. RNA Biol *7*, 192-204.
Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol *14*, 185-193.
Kasturi, P., Zanetti, S., Passannante, M., Saudan, Z., Muller, F., and Puoti, A. (2010). The C. elegans sex determination protein MOG-3 functions in meiosis and binds to the CSL co-repressor CIR-1. Dev Biol *344*, 593-602.
Kerins, J.A., Hanazawa, M., Dorsett, M., and Schedl, T. (2010). PRP-17 and the pre-mRNA splicing pathway are preferentially required for the proliferation versus meiotic development decision and germline sex determination in Caenorhabditis elegans. Dev Dyn *239*, 1555-1572.
Kim, E., Goren, A., and Ast, G. (2008). Insights into the connection between cancer and alternative splicing. Trends Genet *24*, 7-10.
Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Res *35*, 125-131.
Kim, H., Klein, R., Majewski, J., and Ott, J. (2004). Estimating rates of alternative splicing in mammals and invertebrates. Nat Genet *36*, 915-916; author reply 916-917.
Kimble, J.E., and White, J.G. (1981). On the control of germ cell development in Caenorhabditis elegans. Dev Biol *81*, 208-219.
Konishi, T., Uodome, N., and Sugimoto, A. (2008). The Caenorhabditis elegans DDX-23, a homolog of yeast splicing factor PRP28, is required for the sperm-oocyte switch and differentiation of various cell types. Dev Dyn *237*, 2367-2377.

Kotake, Y., Sagane, K., Owa, T., Mimori-Kiyosue, Y., Shimizu, H., Uesugi, M., Ishihama, Y., Iwata, M., and Mizui, Y. (2007). Splicing factor SF3b as a target of the antitumor natural product pladienolide. Nat Chem Biol *3*, 570-575.
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.
Kramer, A. (1996). The structure and function of proteins involved in mammalian pre-mRNA splicing. Annu Rev Biochem *65*, 367-409.
Laity, J.H., Chung, J., Dyson, H.J., and Wright, P.E. (2000). Alternative splicing of Wilms' tumor suppressor protein modulates DNA binding activity through isoform-specific DNA-induced conformational changes. Biochemistry *39*, 5341-5348.
Lambie, E.J., and Kimble, J. (1991). Two homologous regulatory genes, lin-12 and glp-1, have overlapping functions. Development *112*, 231-240.
Lesney, M.S. (2001). Ecce homology: A primer on comparative genomics. Modern Drug Discovery *4*, 26-38.
Li, L., and Xie, T. (2005). Stem cell niche: structure and function. Annu Rev Cell Dev Biol *21*, 605-631.
Li, W., Carroll, J.S., Brown, M., and Liu, S. (2008). xMAN: extreme MApping of OligoNucleotides. BMC Genomics *9 Suppl 1*, S20.
Lin, H. (1997). The tao of stem cells in the germline. Annu Rev Genet *31*, 455-491.
Lin, H., and Spradling, A.C. (1993). Germline stem cell division and egg chamber development in transplanted Drosophila germaria. Dev Biol *159*, 140-152.
Listerman, I., Sapra, A.K., and Neugebauer, K.M. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat Struct Mol Biol *13*, 815-822.
Macdonald, L.D., Knox, A., and Hansen, D. (2008). Proteasomal regulation of the proliferation vs. meiotic entry decision in the Caenorhabditis elegans germ line. Genetics *180*, 905-920.
Mantina, P., Macdonald, L., Kulaga, A., Zhao, L., and Hansen, D. (2009). A mutation in teg-4, which encodes a protein homologous to the SAP130 pre-mRNA splicing factor, disrupts the balance between proliferation and differentiation in the C. elegans germ line. Mech Dev *126*, 417-429.
Martin, G., Keller, W., and Doublie, S. (2000). Crystal structure of mammalian poly(A) polymerase in complex with an analog of ATP. EMBO J *19*, 4193-4203.
Martinez, E., Palhan, V.B., Tjernberg, A., Lymar, E.S., Gamper, A.M., Kundu, T.K., Chait, B.T., and Roeder, R.G. (2001). Human STAGA complex is a chromatin-acetylating transcription coactivator that interacts with pre-mRNA splicing and DNA damage-binding factors in vivo. Mol Cell Biol *21*, 6782-6795.
Matsuura, A., Ito, M., Sakaidani, Y., Kondo, T., Murakami, K., Furukawa, K., Nadano, D., Matsuda, T., and Okajima, T. (2008). O-linked N-acetylglucosamine is present on the extracellular domain of notch receptors. J Biol Chem *283*, 35486-35495.
Menon, S., Tsuge, T., Dohmae, N., Takio, K., and Wei, N. (2008). Association of SAP130/SF3b-3 with Cullin-RING ubiquitin ligase complexes and its regulation by the COP9 signalosome. BMC Biochem *9*, 1.

Modrek, B., and Lee, C. (2002). A genomic view of alternative splicing. Nat Genet *30*, 13-19.
Moloney, D.J., Lin, A.I., and Haltiwanger, R.S. (1997). The O-linked fucose glycosylation pathway. Evidence for protein-specific elongation of o-linked fucose in Chinese hamster ovary cells. J Biol Chem *272*, 19046-19050.
Moloney, D.J., Shair, L.H., Lu, F.M., Xia, J., Locke, R., Matta, K.L., and Haltiwanger, R.S. (2000). Mammalian Notch1 is modified with two unusual forms of O-linked glycosylation found on epidermal growth factor-like modules. J Biol Chem *275*, 9604-9611.
Mumm, J.S., and Kopan, R. (2000). Notch signaling: from the outside in. Dev Biol *228*, 151-165.
Nagai, K., Muto, Y., Pomeranz Krummel, D.A., Kambach, C., Ignjatovic, T., Walke, S., and Kuglstatter, A. (2001). Structure and assembly of the spliceosomal snRNPs. Novartis Medal Lecture. Biochem Soc Trans *29*, 15-26.
Nam, Y., Sliz, P., Song, L., Aster, J.C., and Blacklow, S.C. (2006). Structural basis for cooperativity in recruitment of MAML coactivators to Notch transcription complexes. Cell *124*, 973-983.
Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics *25*, 2730-2731.
Oh, J.J., Razfar, A., Delgado, I., Reed, R.A., Malkina, A., Boctor, B., and Slamon, D.J. (2006). 3p21.3 tumor suppressor gene H37/Luca15/RBM5 inhibits growth of human lung cancer cells through cell cycle arrest and apoptosis. Cancer Res *66*, 3419-3427.
Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008a). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.
Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008b). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.
Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D.*, et al.* (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell *16*, 929-941.
Pelisch, F., Gerez, J., Druker, J., Schor, I.E., Munoz, M.J., Risso, G., Petrillo, E., Westman, B.J., Lamond, A.I., Arzt, E.*, et al.* (2010). The serine/arginine-rich protein SF2/ASF regulates protein sumoylation. Proc Natl Acad Sci U S A *107*, 16119-16124.
Pepper, A.S., Killian, D.J., and Hubbard, E.J. (2003). Genetic analysis of Caenorhabditis elegans glp-1 mutants suggests receptor interaction or competition. Genetics *163*, 115-132.
Petcherski, A.G., and Kimble, J. (2000). LAG-3 is a putative transcriptional activator in the C. elegans Notch pathway. Nature *405*, 364-368.
Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res *29*, e45.

Prigge, J.R., Iverson, S.V., Siders, A.M., and Schmidt, E.E. (2009). Interactome for auxiliary splicing factor U2AF(65) suggests diverse roles. Biochim Biophys Acta *1789*, 487-492.
Puoti, A., and Kimble, J. (1999). The Caenorhabditis elegans sex determination gene mog-1 encodes a member of the DEAH-Box protein family. Mol Cell Biol *19*, 2189-2197.
Puoti, A., and Kimble, J. (2000). The hermaphrodite sperm/oocyte switch requires the Caenorhabditis elegans homologs of PRP2 and PRP22. Proc Natl Acad Sci U S A *97*, 3276-3281.
Racher, H. (2010). Characterization of puf-8's role as a negative regulator of proliferation in the C. elegans germ line. In Biological Sciences (Calgary, University of Calgary), pp. 280.
Ramani, A.K., Calarco, J.A., Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M.*, et al.* (2011). Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Genome Res *21*, 342-348.
Ramani, A.K., Nelson, A.C., Kapranov, P., Bell, I., Gingeras, T.R., and Fraser, A.G. (2009). High resolution transcriptome maps for wild-type and nonsense-mediated decay-defective Caenorhabditis elegans. Genome Biol *10*, R101.
Rampal, R., Luther, K.B., and Haltiwanger, R.S. (2007). Notch signaling in normal and disease States: possible therapies related to glycosylation. Curr Mol Med *7*, 427-445.
Schneider-Stock, R., Oda, Y., and Roessner, A. (1997). New splicing mutation in exon 5-6 of the p53-tumor suppressor gene in a malignant schwannoma. Hum Mutat *9*, 91-94.
Seydoux, G., and Schedl, T. (2001). The germline in C. elegans: origins, proliferation, and silencing. Int Rev Cytol *203*, 139-185.
Shao, L., and Haltiwanger, R.S. (2003). O-fucose modifications of epidermal growth factor-like repeats and thrombospondin type 1 repeats: unusual modifications in unusual places. Cell Mol Life Sci *60*, 241-250.
Song, X., Call, G.B., Kirilly, D., and Xie, T. (2007). Notch signaling controls germline stem cell niche formation in the Drosophila ovary. Development *134*, 1071-1080.
Spike, C.A., Bader, J., Reinke, V., and Strome, S. (2008). DEPS-1 promotes P-granule assembly and RNA interference in C. elegans germ cells. Development *135*, 983-993.
Srebrow, A., and Kornblihtt, A.R. (2006). The connection between splicing and cancer. J Cell Sci *119*, 2635-2641.
Staley, J.P., and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell *92*, 315-326.
Sulston, J.E., and Horvitz, H.R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol *56*, 110-156.
Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D.*, et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956-960.
Tax, F.E., Yeargers, J.J., and Thomas, J.H. (1994). Sequence of C. elegans lag-2 reveals a cell-signalling domain shared with Delta and Serrate of Drosophila. Nature *368*, 150-154.
Tonnesen, J., Parish, C.L., Sorensen, A.T., Andersson, A., Lundberg, C., Deisseroth, K., Arenas, E., Lindvall, O., and Kokaia, M. (2011). Functional integration of grafted neural

stem cell-derived dopaminergic neurons monitored by optogenetics in an in vitro Parkinson model. PLoS One *6*, e17560.
van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol *8*, e1000371.
Varnum-Finney, B., Xu, L., Brashem-Stein, C., Nourigat, C., Flowers, D., Bakkour, S., Pear, W.S., and Bernstein, I.D. (2000). Pluripotent, cytokine-dependent, hematopoietic stem cells are immortalized by constitutive Notch1 signaling. Nat Med *6*, 1278-1281.
Vernet, C., and Artzt, K. (1997). STAR, a gene family involved in signal transduction and activation of RNA. Trends Genet *13*, 479-484.
Wachtel, C., and Manley, J.L. (2009). Splicing of mRNA precursors: the role of RNAs and proteins in catalysis. Mol Biosyst *5*, 311-316.
Wang, G.S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet *8*, 749-761.
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.
Weber, J.M., and Calvi, L.M. (2010). Notch signaling and the bone marrow hematopoietic stem cell niche. Bone *46*, 281-285.
Whitworth, G.E., Zandberg, W.F., Clark, T., and Vocadlo, D.J. (2010). Mammalian Notch is modified by D-Xyl-alpha1-3-D-Xyl-alpha1-3-D-Glc-beta1-O-Ser: implementation of a method to study O-glucosylation. Glycobiology *20*, 287-299.
Will, C.L., and Luhrmann, R. (2010). Spliceosome Structure and Function. Cold Spring Harb Perspect Biol.
Wilson-Berry, L. (1998). Regulation of the Mitotic/Meiotic Cell Fate Decision in Caenorhabditis elegans. In Biology and Biomedical Sciences (St. Louis, Washington Univerisity), pp. 327.
Wilson, J.J., and Kovall, R.A. (2006). Crystal structure of the CSL-Notch-Mastermind ternary complex bound to DNA. Cell *124*, 985-996.
Xie, T. (2008). Germline stem cell niches.
Zanetti, S., Meola, M., Bochud, A., and Puoti, A. (2011). Role of the C. elegans U2 snRNP protein MOG-2 in sex determination, meiosis, and splice site selection. Dev Biol.
Zhang, X.P., Zheng, G., Zou, L., Liu, H.L., Hou, L.H., Zhou, P., Yin, D.D., Zheng, Q.J., Liang, L., Zhang, S.Z.*, et al.* (2008). Notch activation promotes cell proliferation and the formation of neural stem cell-like colonies in human glioma cells. Mol Cell Biochem *307*, 101-108.