Gene Identification in *Caenorhabditis elegans* through Next Generation Sequencing

by

Tammy Wong

B.Sc. (Hons.), Simon Fraser University, 2010

Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2012

Approval

Name: **Tammy Wong**

Degree: **Master Science (Molecular Biology and Biochemistry)**

Title of Thesis: **Gene Identification in *Caenorhabditis elegans* through Next Generation Sequencing**

Examining Committee:

Chair: Edgar C. Young, Associate Professor

Jack Chen
Senior Supervisor
Associate Professor

David L. Baillie
Supervisor
Professor

Nancy Hawkins
Supervisor
Associate Professor

Fiona S. L. Brinkman
Internal Examiner
Professor

Date Defended/Approved: August 14, 2012

Partial Copyright Licence

Abstract

The emergence of Next Generation DNA Sequencing (NGS) technologies has made cloning of genes much more efficient by dramatically reducing the cost, labour and time required. In this work, I have applied a NGS and devised a bioinformatics program pipeline to detect genomics alterations in the mutant strains of the nematode *Caenorhabditis elegans*. In the first project, I have detected a mutation in *tba-5,* validated through complementation tests, and proposed a model elucidating the role of TBA-5 in cilia structure. In the second project, the variant detection pipeline is used to isolate potential candidate genes corresponding to various *dsh-2(or302)* suppressors, namely *Sup327*, *Sup245* and *Sup305*. Additionally, I have applied the bioinformatics pipeline to confirm that the current *C. briggsae* reference genome harbours thousands of assembly errors, many of which affect the correct prediction of gene models.

Keywords: next generation sequencing; *Caenorhabditis elegans; Caenorhabditis briggsae; dyf-10; tba-5; dsh-2*

I would like to dedicate this work to my mentors, co-workers and family who have provided tremendous help and support.

Acknowledgements

I would like to specially thank my senior supervisor Dr. J. N. Chen and committee members Dr. N. C. Hawkins and Dr. D. L. Baillie for their encouragement, advice and comments on my work. I would like to thank Kyla Hingwing for the following: construction of the mutant strain, and suppressor strains, mapping of the suppressors, confirmation of variations in the suppressor strain A00842. I also thank Kyla for teaching me *C. elegans* genetics and worm handling, and for supplying me with the strains *dyf-10(e1383)* and *dpy-5(e61)dyf-10(e1383)*. I would like to thank Dr. Maja Tarailo-Graovac for her advice and supervision on many experimental techniques such as outcrossing, generation of the double mutant and generation of the *tba-5(tm4200)* cDNA library. Dr. Maja Tarailo-Graovac also helped me sequence and analyze the PCR product generated from the cDNA library. I would like to thank Jun Wang for training me how to handle worms, perform DiO dye filling assays, freeze strains and use both the GFP dissecting and GFP compound microscopes. I would like to thank Christian Frech helping me troubleshoot computational issues. I would like to thank Bora Uyar and Jeff Chu for detecting InDels from the transcriptome data. I would like to thank Caleb Choo for his help in comparing the similarity between the alpha and beta tubulins of *Caenorhabditis elegans*. I would like to thank Dr. Limin Hao for the strain *tba-5(qj14)*, Shohei Mitani at the National BioResource Project for the strain *tba-5(tm4200)*, and Caenorhabditis Genetic Center for the strains MT2179 *nDf25/unc-13(e1091)lin-11(n566)I* and *RB1545 tba-9(ok1858)X*. I would like to thank Ata Roudgar and Martin Siegert at Westgrid for providing me with technical support and setting up the Joffre server for our lab use. I would like to thank Shannon Ho Sui in the Brinkman lab for sharing her experiences about MAQ and repetitive genomes. I would finally like to thank the Chen lab in general, for their inputs into my work including suggestions, support and much more.

Table of Contents

List of Tables

List of Figures

List of Acronyms or Glossary

BAM format	Binary Alignment/MAP format
BAQ	Base Alignment Quality
BaseQual	Minimum Base Quality (VarScan)
BB	Basal Body of the cilia
BBS	Bardet-Biedl Syndrome
bp	Base Pair
BWA	Burrows-Wheeler Alignment
CB	Hawaiian strain
cDNA	complementary DNA
CGC	*Caenorhabditis* Genetic Center
Con	Consensus base
COTRASIF	Conservation-aided transcription factor binding site finder
Cov	Minimum Read Coverage
cPAP	cytoplasmic poly(A) polymerase
CQUAL	Consensus Quality
DEL	Deletion
Depth	Read depth
DIC	Differential Interference Contrast
DiO	dioctadecyloxacarbocyanine percchlorate
Dpy	Dumpy phenotype, characterized by the general physical shortening of the worms
DS	Distal Segment of the cilia
Dsh	Dishevelled
dsRNA	Double stranded RNA
Dyf	Dye Filling Defective phenotype
EMS	Ethyl methanesulfonate
EST	Expression Sequence Tags
EtBr	Ethidium Bromide
EXP	Expression Pattern
FN	False negatives
FP	False positives
Freq	Variant frequency
GBrowse	Generic Genome browser
GFP	Green Fluorescent Protein
GO	Gene Ontology
HMM	Hidden Markov Model
HOM	Homolog
HSV	Herpes simplex virus
IFT	Intra-flagellar Transport

InDel	Insertions and Deletions
INS	Insertions
JSRDs	Joubert Syndrome Related Disorders
LB	Lysis Buffer
MAPK	Mitogen-activated kinase
MAQ	Mapping and Assembly with Quality
MaxCov	Maximum Coverage
MinCov	Minimum Coverage
MIS	Missense
MKS	Meckel Syndrome
MS	Middle Segment of the cilia
NBRP	National BioResource Project
NGS	Next Generation Sequencing
NON	Nonsense
OMRF	Oklahoma Medical Research Foundation
OSTs	Open Reading Frame Sequence Tags
PCR	Polymerase Chain Reaction
PID	Protein sequence level similarities or Global Percentage identity
PKD	Polycystic Kidney Disease
Porc	Porcupine protein
Ref	Reference base
RMIS	Radical Missense
RNAi	RNA interference
RNA-seq	RNA Sequencing
RT	Room Temperature
SAGE	Serial Analysis of Gene Expression
SAM	Sequence Alignment/Map
SEAL	Sequence Alignment evaluation suite
SND	Single Nucleotide Difference
SNP	Single Nucleotide Polymorphism
SNPQ	SNP quality
SPL	Splice Junction mutation
STDFQ	Standard FASTQ
TZ	Transition Zone of the cilia
UncVul	phenotype characterized by worms that are small, kinky, paralyzed and vulvaless. The phenotype results in "bags" of worms.
UTR	Untranslated Region
VarFreq	Minimum Variant Frequency
WestGrid	Western Canada Research Grid
WGS	Whole-genome shotgun sequencing
Wnt	Wingless in Drosophila
WT	Wild Type

1. General Introduction

1.1. Strength of *C. elegans* as a model organism

Caenorhabditis elegans is a great model organism for genetic studies. *C. elegans* was originally introduced as a model organism by Sydney Brenner in 1963. The organism has many advantageous properties. For example, the generation time is only three days at 20°C and they are very inexpensive to maintain compared to other organisms, such as mice. *C. elegans* are easy to propagate, have large brood size, can recover after starvation, and stocks can be frozen. In addition, there are a wide range of tools available for geneticists. The hermaphroditic nature of the organism allows for efficient generation of recessive mutants, and the presence of males allows the transfer of genes and genetic markers. Mutations can be easily introduced through various methods, and with the advantage of its short generation time, large brood size and ability to self fertilize, mutants can easily be isolated, facilitating forward genetic studies (Brenner, 1974; reviewed in Jorgensen and Mango, 2002). There are plenty of morphological markers for single nucleotide polymorphism (SNP) mapping. There are also plenty of balancers that allow many recessive sterile and lethal mutations to be detected and propagated. RNA interference (RNAi) used for the knockdown of gene function is also available in *C. elegans* for reverse genetics; by injecting double stranded RNA (dsRNA) into the organism, the corresponding mRNA will be degraded and therefore the gene involved will be knocked down (Ahringer, 2006). Furthermore, *C. elegans* has its genome sequenced (*C. elegans* Sequencing Consortium, 1998).

1.2. Classical approach to gene cloning

The traditional forward genetic approach in *C. elegans* involves random mutagenesis using biological agents or chemical mutagens, isolation of mutants with the defective phenotype and characterization of molecular lesions. For a simple F2

mutagenesis screen, mutagenized worms are grown for two generations to produce homozygous mutants. Homozygous worms that show a mutant phenotype of interest are then plated separately and observed to see if the phenotype is transmitted to the next generation. Typically, a screen of 12,000 haploid genomes using standard concentration of mutagen would require approximately two weeks to complete and is expected to recover six mutations in a particular gene. Once it is known which biological process the phenotype is associated with, a simple or modifier screen can be used to identify more genes associated with the biological pathway. The simple screen involves screening for additional mutants with the same phenotype. The modifier screen involves mutagenizing a strain with the mutant phenotype of interest, and screening for animals with second- site mutations that either enhance (enhancer screen) or suppress (suppressor screen) the original mutant phenotype. In addition, there are other types of screens that are utilized in classical gene cloning. Selection screens, for example drug selection, helps eliminate worms with irrelevant genotype. However, not all mutations result in a highly visible phenotype that can be screened. For example, there are mutations that can only be assayed under high magnification (microscope screens), that are unobservable in the first generation of homozygotes (multigenerational screens), that require the use of green fluorescent proteins to locate the movement of cells (green fluorescent protein screens), or that requires the use of laser to ablate certain cells (laser ablation screens). Furthermore, lethal mutations screens are extremely difficult to work with. First, lethal mutations are difficult to isolate and maintain as heterozygotes. Second, it is difficult and tedious to sort through many lethal mutations to identify those that are defective for a particular biological process; identifying the correct lethal mutations typically involves a secondary screen for other interesting mutant phenotypes (reviewed in Jorgensen and Mango, 2002).

The traditional forward genetic approach also utilizes mapping with genetic and/or SNP markers and Sanger sequencing to identify the mutant genes, which often takes significant effort and many years to complete (Fay and Bender, 2006; Hobert, 2010). In *C. elegans*, the relative gene density and limited recombinant frequencies make traditional mapping with genetic and/or SNP marker very time consuming (Sarin *et al.*, 2008). Two-factor mapping involves measurement of the recombinational distance between a gene and a known marker. The first step is to construct a double mutant that

contains the gene of interest (*mut-1*) and a visible marker (*vis-1*). Double mutants are crossed with WT males to generate F1 animals that are heterozygous for both genes The F1 progeny are then self-fertilized. The F2 self-progeny broods are then scored for recombinant progenies with phenotypes 'mutant non-marker' and 'marker non-mutant' and the linkage calculated. If the *vis-1* is close to *mut-1*, there will be few progenies with recombination between *vis-1* and *mut-1* (Hodgkin, 1999; Davis and Hammarlund, 2006). Similar to two-factor mapping, three-factor mapping relies on recombination events between the mutant gene and two visible markers to determine the relative location of the mutant gene between the two markers (Davis and Hammarlund, 2006). In SNP mapping, SNPs between the wild type (WT) N2 Bristol strain and the closely related CB4856 Hawaiian strain are used as genetic markers. Since there are no associated phenotypes with SNPs, mutant phenotypes that are easily masked by conventional marker mutations can be mapped. Furthermore, there is a dense collection of SNPs that can potentially provide single-gene resolution, unlike other markers. SNP mapping is completed in two steps. In the first step, rough mapping of the mutation using two-factor mapping identifies the chromosome and the rough position of the mutation of interest. In the second step, the goal is to place the mutation between two SNPs using interval mapping (Davis *et al*., 2005). After cloning, the gene harbouring the mutation of interest is usually sequenced using Sanger sequencing, also known as chain terminator sequencing. To sequence a specific region of interest a mix consisting of primer that is complementary to the template at the region of interest, DNA polymerase for DNA extension, the four dexoynucleotide bases and a low concentration of one of the chain terminating nucleotide are mixed together. The reaction is repeated using each of the four chain terminating nucleotides. Since the chain terminating nucleotide is present in low quantities, it results in a series of different sized DNA fragments that are terminated at certain positions where the particular terminating nucleotide is used. The fragments from each reaction are then subjected to gel electrophoresis and the sequence determined (Hong, 1982).

1.3. Next generation sequencing

The advent of next-generation sequencing (NGS), in contrast, is a time saving and cost reducing strategy that can be utilized in forward genetics. NGS identifies

sequence variants throughout a genome before and after mutagenesis. By narrowing down the list of candidate mutations that are responsible for the mutant phenotype of interest through outcrossing or a rough mapping of the mutation, researches can quickly narrow down their search to a few variants that is contained in a specific sequence interval. Compared to the lengthy and labour intensive procedures associated with multi-year cloning based methods that also lack in throughput, scalability and resolution, NGS offers unprecedented speed and ease, taking approximately a week to sequence a whole genome (Hobert, 2010). NGS costs are significantly lower compared to personnel and reagent costs of traditional multi-year mapping based cloning projects, and will likely continue to drop as technology advances (Sarin *et al.*, 2008). Over the past few years competing companies marketing NGS, for example Illumina, have decreased the costs of generating raw reads for human whole genome sequencing by many orders of magnitude compared to the costs of generating the draft of Craig Venter's genome. Craig Venter's genome was sequenced with ABI's Sanger-CE instruments, at a cost of about $10 million (Niedringhaus *et al.*, 2011; Levy *et al.*, 2007). Now the cost of sequencing the whole genome of a *C. elegans* strain is only a few thousand dollars (Sarin *et al.*, 2008; Hobert, 2010). The cost of sequencing will continue to decrease as NIH/NHGRI continue to fund groups to improve NGS with an ambitious goal of bringing down the cost of a human genome to under $1000 (Niedringhaus *et al.*, 2011). In addition, the high throughput of many NGS platforms, for example Illumina's Genome Analyzer, allows rapid sequencing and data acquisition (Medvedev *et al.*, 2009; Metzker, 2010; Sarin *et al*., 2008).

To date, NGS has been utilized in the cloning of many genes. For example, B9D1 has been identified as a novel Meckel Syndrome (MKS) gene in human through exon-enriched NGS of a set of ciliopathy genes in many MKS pedigrees. MKS is characterized by renal cysts, hepatic fibrosis, polydactyly, and central nervous system defects, which were thought to have been due to defects in primary cilia. By enriching and sequencing exons of a set of known ciliopathy genes in several MKS pedigrees, Hopp and colleagues identified a change in a splice-donor site of B9D1. B9D1 has been found to be structurally similar to MKS1 and is shown to be important in ciliogenesis. It was found that B9D1 localizes to the basal body (BB) of primary cilia in mammals, and depletion of B9D1 led to a decrease in the number of cilia (Hopp *et al.*, 2011). In

another example, mutations in human TMEM237 identified through NGS have been associated with Joubert syndrome related disorders (JSRDs). JSRDs is characterized by a broad and variable range of phenotypes typically seen in other ciliopathies. Huang and colleagues performed homozygosity mapping and NGS of DNA samples from individuals afflicted with JSRDs. They found a nonsense mutation in the gene TMEM237. TMEM237 was found to localize to the transition zone (TZ) and loss of TMEM237 results in defective ciliogenesis and deregulation of Wnt signaling. In *C. elegans*, TMEM237 has been found to interact with *nphp-4* to control BB-TZ anchoring to the membrane and ciliogenesis (Huang *et al.*, 2011). In yet another example, whole genome sequencing using NGS identified the molecular identity of *lys-12* in *C. elegans* as the gene R07B5.9*,* which is involved in neuronal fate decision (Sarin *et al.*, 2008) (Figure 1.1).

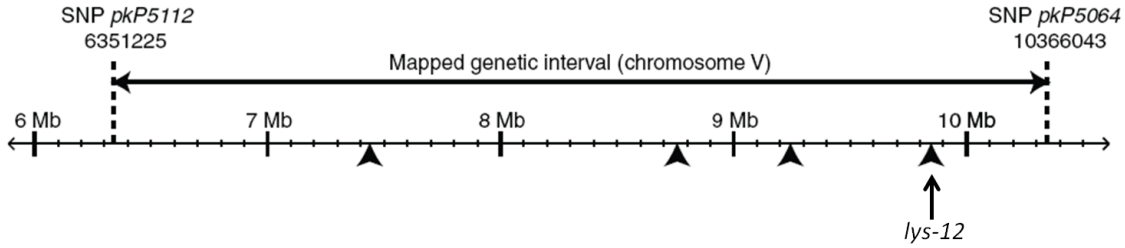


Figure 1.1 Physical location of variations detected by Sarin et al (2008).

Sarin *et al.* mapped *lsy-12* to a 4Mb interval, between SNP markers pkP5112 and pkP5064 on chromosome V. Their pipeline detected four exonic mutations in the 4 Mb region, marked by arrow heads. Complementation tests and mapping showed that the nonsense mutation near the 10 Mb interval was the only mutation that occurred in the five other strains that contains mutant alleles of *lys-12*. The mutation occurs in R07B5.8 and injecting the full length R07B5.8 into *lys-12* mutant rescues the mutant phenotype. As a result, Sarin *et al.* validated R07B5.8 as the gene *lys-12*. Figure modified from Sarin *et al.*, 2008.

Sarin *et al.* applied NGS towards identifying the molecular location of an unknown mutant gene by looking at molecular lesions and hence validated a quick alternative method to the traditional forward genetic approach. In the study, mutations in the *C. elegans* genome were induced by ethyl methanesulfonate (EMS) treatment. EMS typically introduces point mutations by guanine alkylation, where the alkylated guanine may be mistakenly paired with thymine in subsequent DNA replication; as a result, a G/C pair can be mutated into a A/T pair (O'Neil, 2006). Paired-end Illumina sequencing technology was used for DNA sequencing of the mutated strain, which generated 4.35

Gb of paired 35-mer sequence reads in a short period of a week. The reads were then mapped onto the reference genome using alignment tools such MAQ (Mapping and Assembly with Quality); the average coverage of the reads over the entire genome was approximately 28X. Overall, the researchers were able to find and validate mutations that were caused by EMS mutagenesis. They were also able to determine the exact variation that was responsible for the phenotype caused by the mutant gene. In addition, Sarin *et al*. found that the frequency of (1) identified variations due to sequencing error, (2) true variations due to EMS treatment and (3) variations due to differences between the starting strain and the reference genome was 21.6%, 18.9% and 56.8% respectively (Sarin *et al.*, 2008).

NGS is a powerful tool for gene cloning, but there are some bioinformatic challenges. Although the read length produced by platforms such as Illumina/Solexa is increasing, it is still shorter than reads obtained from the traditional Sanger sequencing; Illumina/Solexa produces ~75-100 bp reads compared to Sanger's > 800 bp reads. Shorter reads is a disadvantage for *de novo* sequencing since smaller reads make smaller overlaps. In *de novo* sequencing, short read lengths can lead to a higher number of gaps and regions where no reads align that results in poorer data quality. Although smaller reads are effective when mapped back to a reference genome, not all genomes have been previously assembled. There is also a tendency for higher rate of errors for shorter reads, whether it is due to repeated sequence regions, or sequencing errors (Kato, 2009; www.illumina.com). Many short read alignment tools have been developed to circumvent the difficulties associated with short reads: for example, using gapped versus ungapped alignment, trimming reads or taking into account the quality value of each base (Yu *et al.*, 2012). Another strategy for overcoming the limitations of short reads is paired-end sequencing where both ends of a DNA fragment are sequenced. Apart from containing sequence information, paired-end reads also contain long range positional information since the distance between each paired reads are known. Paired-end sequencing allows for a more precise alignment of reads across regions containing repetitive sequence and produce longer contigs for *de novo* sequencing compared to single-end reads; the longer the insert size between the paried-end reads, the more uniform the sequencing coverage (www. illumina.com). Compared to Roche 454, Illumina/Solexa does not have as high of a sequencing error rate in

homopolymer regions. A homopolymer region is defined as having three or more identical consecutive bases. However, Illumina/Solexa sequencing platforms tend to have higher sequencing error rates at 3' ends of reads and at regions in the genome that are associated with GGC motifs. In addition, different tiles of the sequencing plate tend to produce reads of different quality (reviewed in Luo *et al.*, 2012). In Illumina's Solexa sequencing approach, fragmented DNA samples are ligated to adaptors that bind to linker molecules on the surface of the flow cell for amplification. Each flow cell contains eight lanes where independent samples can be sequenced simultaneously. Each lane is further divided into hundreds of tiles, where four images are taken for each tile corresponding to the four base dyes (reviewed in Dolan and Denver, 2008). Overall, it was found that there is approximately 0.5% sequencing error per base in the raw reads, which were randomly distributed, but the frequency of single-base errors decreased with higher coverage (reviewed in Luo *et al.*, 2012).

Similar to the research completed by Sarin *et al.,* a goal of my project is to identify the physical location of the various EMS-mutated genes by identifying putative mutations using NGS while altering the method to circumvent the high frequency of sequencing error and strain variations. First, to increase the economic value and test the efficiency of this new alternative method, this project will simultaneously identify more than one uncloned gene in the *C. elegans strains*. Second, there will be a higher read coverage ranging from approximately 56X-100X over the entire genome. A higher read coverage will constitute a decrease in the frequency of variations due to sequencing errors and, as a result, reduce the amount of Sanger re-sequencing needed to verify the mutations. Finally, I will be comparing each variation from the mutated strains to the variations detected in two WT strains: Hobert and Horvitz. These WT strains have been mapped to the same reference genome. Since the Hobert and Horvitz strains are WT for the genes of interest, variations that are shared between the mutated and WT strains cannot be responsible for the mutant phenotype. This is an alternative method that will eliminate variations caused by strain differences and therefore dramatically narrow down the list of possible candidates, making cloning faster.

2. Bioinformatic and Experimental Analysis of *tba-5*

2.1. Background

Caenorhabditis elegans can sense a range of chemical cues through their chemosensory cilia which consist of the amphid and phasmid neurons. The chemical cues are important for locating food and mate, avoiding noxious conditions, and for proper development. In *C. elegans*, the dendritic ends of sensory neurons are connected to non-motile cilia which are exposed to the environment through channels composed of socket and sheath cells. Each of the eleven pairs of chemosensory neurons has dedicated receptors that are able to detect certain sets of stimuli in the surrounding environment (White *et al.*, 1986; Starich *et al.*, 1995; Bargmann, 2006).

In humans, defects in proteins involved with cilia are the basis of many diseases, collectively known as ciliopathies. Many such ciliopathies include human polycystic kidney disease (PKD), Bardet-Biedl syndrome (BBS) and Meckel syndrome (MKS). PKD is the most commonly inherited disease in the United States and is characterized by extensive cystic enlargement of the kidneys. Although none of the human proteins disrupted in PKD are associated directly with the assembly of cilia in humans, most of them are present on the cilia of other model organisms such as *Drosophila melanogaster* and *C. elegans*. For example, polycystin 1, polycystin 2, fibrocystin, nephrocystin, and inversin are all disrupted in human PKD and are found to be involved with cilia in *C. elegans* and *D. melanogaster*. Interestingly it was found that human PKD proteins function in kidney development and homeostasis as well as mechanosensation. BBS on the other hand is rare genetic disorder, but is characterized by several common phenotypes such as obesity, polydactyly and retinal dystrophy. Many of the BBS proteins are essential in cilia assembly and function or are associated with centrioles, which can influence cilia function as well as cell division in many different organisms. For example, BBS1 and BBS4 are important for the formation of cilia in mouse olfactory

epithelium. BBS5 has been found to be located in the basal body in mouse and worm. BBS8 has been found to be localized to centrosomes and basal bodies. Defective centrosomes may explain some of the phenotypic defects observed in BBS patients that are not easily explained by ciliary defects; the centrosome is important in many different cellular processes such as protein degradation, vesticular transport and axonal guidance (reviewed in Pan *et al.*, 2005; reviewed in Forsythe and Beales, 2012). MKS is a lethal disease characterized by renal cysts, hepatic fibrosis, polydactyly, and central nervous system defects. It has been found that there is overlap in genes that are mutated in BBS and MKS, as several BBS genes have also been found in fetuses with Meckel-like phenotypes. Cloning of MKS genes has also indicated that the encoded proteins are involved with sensory cilia (reviewed in Badano *et al.*, 2005).

One of the simplest methods to assay the structural integrity of the sensory cilia in *C. elegans* is to test for dye filling. Hydrophobic fluorescent dyes such as DiO, DiI, DiD and FITC acts as neuronal stains by preferentially filling amphid and phasmid neurons via their exposed ciliated endings in wild type (WT) animals. As such, *C. elegans* mutants that are defective in chemotaxis are also abnormal in the DiO filling of certain amphid and phasmid channel neurons (Figure 2.1), a mutant phenotype that is termed Dyf (Hedgecock *et al.*, 1985; Bargmann, 2006; Inglis, 2007). There are six pairs of neurons in the head that are normally dye filled in WT animals – ASK, ADL, ASI, AWB, ASH and ASJ. In addition, PHA and PHB are two pairs of neurons in the tails that are normally dye filled (Starich *et al.*, 1995).

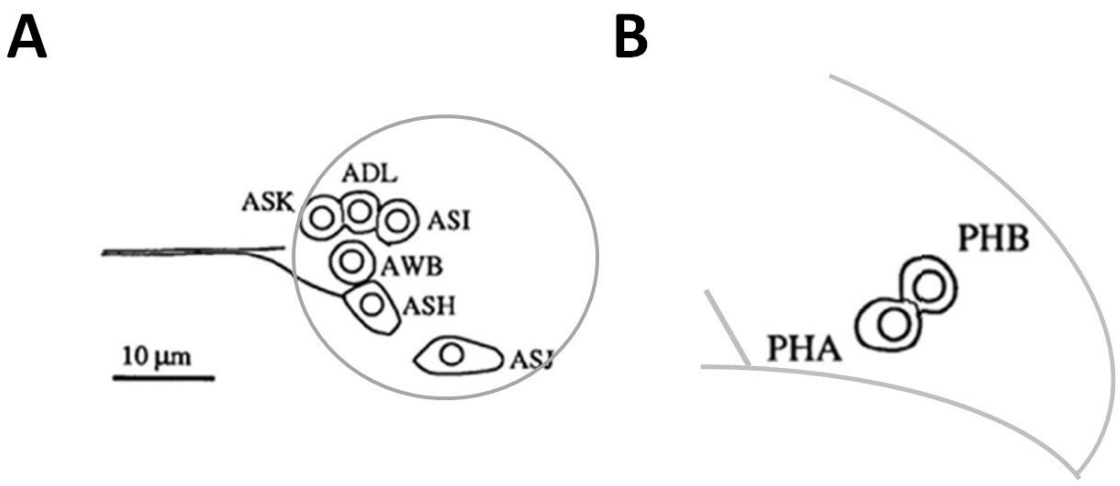


Figure 2.1. Schematic of neuronal cell bodies that have been stained with DiO in the amphid and phasmid neurons

(A) ASK, ADL, ASI, AWB, ASH and ASJ are the amphid neuronal cell bodies in the head that will fill with dye in a DiO dye filling assay. The grey circle indicate the approximate location of the anterior region of the pharyngeal bulb. (B) PHA and PHB are the phasmid neuronal cell bodies in the tail that will fill with dye in a DiO dye filling assay. The grey lines outline the tail of the worm and the approximate location of the anus. Anterior is left and dorsal is up. ASI and AWB both fill weakly with DiO. Figure modified from Starich *et al.*, 1995.

The mutated gene of interest in this project is *dyf-10,* which confers a recessive, dye-filling defective (Dyf) phenotype indicative of defects in the cilia structure important for chemosensory functions. Starich and colleagues in the Riddle lab obtained *dyf-10(e1383)* through an ethyl methanesulfonate (EMS) mutagenesis screen on the background strain SP1709. They selected for mutants that cannot dye fill, but have minimal effect on the viability, fertility or movement of the worm. *e1383* was one of the mutations that was discovered at the time of the study and since the mutant displays a Dyf phenotype, the gene was named *dyf-10(e1383)*. From three-factor mapping, Starich and colleagues have mapped *dyf-10(e1383)* on chromosome I between *dpy-5* and *unc-13* (Table 2.1). In the three-factor mapping, a triple mutant consisting of *dyf-10(e1383)* and two flanking markers*, dpy-5* and *unc-13,* are constructed. *dpy-5* mutants display a phenotype where the body is physically shorter than WT animals. *unc-13* mutants display a uncoordinated mutant phenotype where the animal's movement deviates from the normal smooth sinuous movement of WT worms on the agar surface (Brenner, 1974). 15/15 Dpy non-Unc recombinant worms with genotype *dpy-5 dyf-10* + were

observed and 12/14 Unc non-Dpy recombinant worms with genotype + + *unc-13* were observed (Starich *et al.,* 1995). They also completed deficiency mapping. In deficiency mapping, the mutation that is linked to a marker is crossed to the balanced deficiency. Progeny that contains both the mutation and the deficiency are observed for complementation (Fay, 2006). It was found that *dyf-10(e1383)* failed to complement with the deficiency hDf8. This means that *dyf-10(e1383)/hDf8* animals displayed a Dyf phenotype which indicates that *dyf-10(e1383)* is located in the *hDf8* region on chromosome I (Starich *et al.,* 1995). However, as communicated by Dr. D. L. Baillie, even though the left break-point of hDf8 has been defined, the right break point has not been defined. Based on the three-factor mapping data, mutations that correspond to *dyf-10(e1383)* will be searched between *dpy-5* and *unc-13,* which is a region of approximately two million bp (Table 2.1).

Table 2.1. Predicted Genomic Location of dyf-10(e1383) on Chromosome I

Gene Name	*dpy-5*	*dyf-10*	*unc-13*
Rearrangement: hDf8	-	+	-
Genetic Location (cM)	-0.00±0.002	1.65±0.046	2.07±0.004
Physical Location (bp)	5431981-5433112	unknown	7416520-7455962
Approximate location of *dyf-10*		1.98 million bp region	

Note: *dyf-10* has been mapped between *dpy-5* and *unc-13*. Deficiency mapping with hDf8 also resulted in no complementation (Starich, *et al.*, 1995).

It is hypothesized that *dyf-10* affects the intra-flagellar transport (IFT) system, as with most other ciliary genes. The sensory cilia, or the ciliary axoneme, is divided into three domains, namely the transition zone (TZ), middle segment (MS) and distal segment (DS), which have vastly different morphologies. Both the TZ and MS are made up of the canonical nine microtubule doublets without a central pair, whereas the DS is made up of nine microtubule singlets without a central pair. The microtubules, for example in the amphid neurons, form a bundle that protrudes into the amphid channel in the sheath cell and becomes enveloped at the distal end by the socket cells and surrounding cuticle. The assembly, maintenance and functions of cilia are the responsibilities of the IFT system; it involves the bi-directional movement of IFT particles from the base to the tip of the axoneme by microtubule based kinesin and dynein motors (Figure 2.2). IFT particles are multimeric protein complexes that are proposed to deliver

assembly precursors, for example structural precursors such as tubulins and signal transduction components, to the tips of the axoneme. The IFT components are organized into three main complexes: IFT-A, IFT-B and BBSome complex. In anterograde IFT, the two kinesin-2 motors, Kinesin-II and OSM-3, acts together in the MS to traffick the IFT machinery to the tip of the cilia. The IFT machinery consists of the IFT-A, IFT-B, BBS proteins, dyenin and other cargo. Dynein is transported in an inactivated form during anterograde IFT to the ciliar tip. Past the MS, OSM-3 is solely responsible for transport of the remaining IFT machinery. Following rearrangement at the cilia tip activated Dynein aids in the recycling, or retrograde transport, of the IFT machinery that includes inactivated OSM-3 and other ciliary components back to the basal body. At or near the basal body/transitional fibers, the IFT components rearrange once again to assemble the new IFT machinery and anterograde IFT transport takes place once again (Reviewed in Inglis *et al.*, 2009; reviewed in Badano *et al.*, 2006).

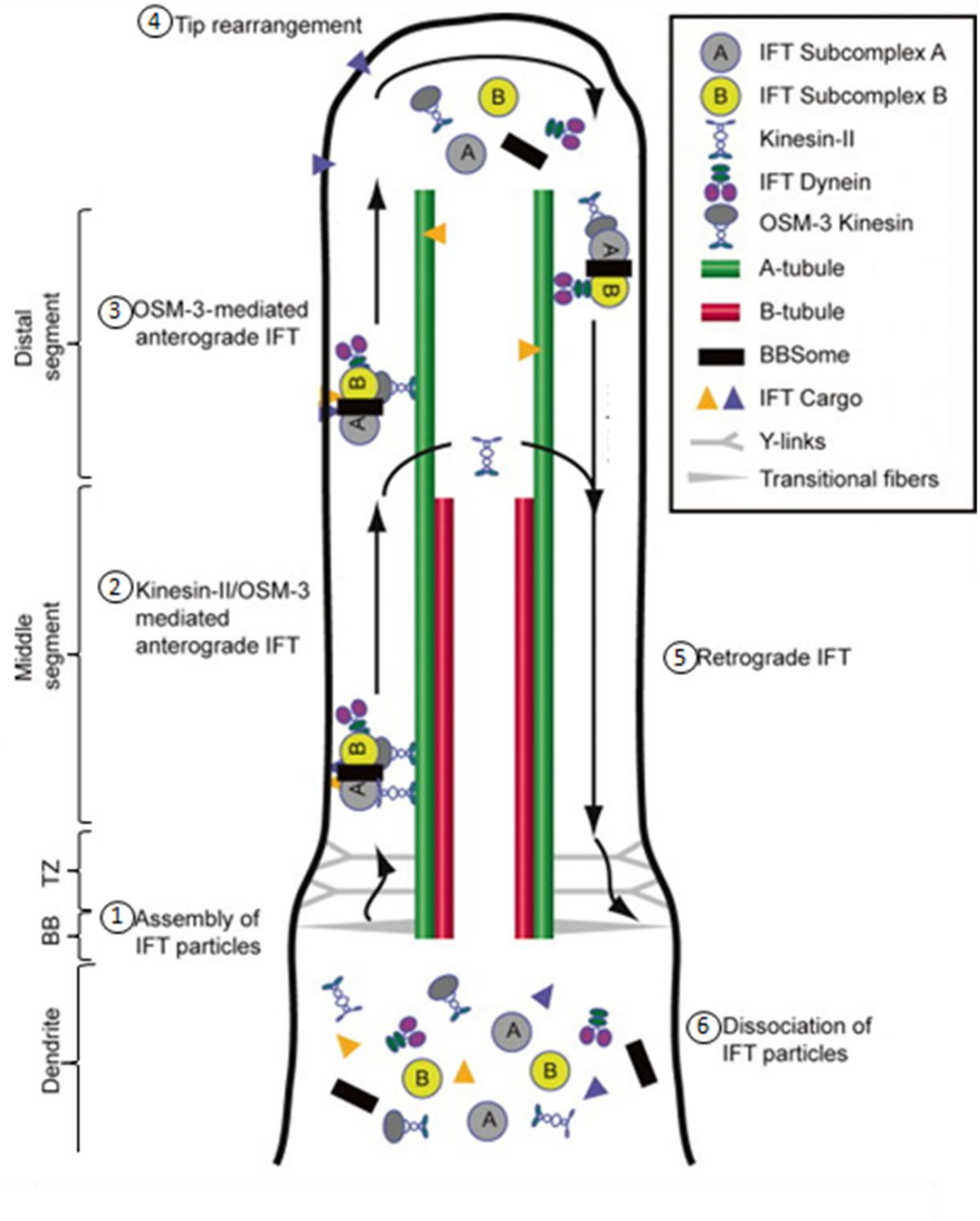

4 Tip rearrangement
IFT Subcomplex A
IFT Subcomplex B
Kinesin-II
IFT Dynein
OSM-3 Kinesin
A-tubule
B-tubule
BBSome
IFT Cargo
Y-links
Transitional fibers
Distal segment
3 OSM-3-mediated anterograde IFT
Middle segment
2 Kinesin-II/OSM-3 mediated anterograde IFT
5 Retrograde IFT
TZ
BB
1 Assembly of IFT particles
Dendrite
6 Dissociation of IFT particles

Figure 2.2. Summary of IFT anterograde and retrograde transport

(1) The IFT machinery for anterograde transport is assembled at the basal body (BB). The IFT machinery includes IFT-A, IFT-B, BBS proteins, inactivated dyenin and cargo that is (2) trafficked towards the tip of the cilia by two kinesin-2 motors: Kinesin II and OSM-3. (3) Past the middle segment (MS), OSM-3 is solely responsible for transport of the remaining IFT machinery. (4) Once at the tip, the IFT machinery is rearranged and (5) activated dyenin directs the retrograde transport of the IFT machinery and inactivated OSM-3 back to the BB. (6) Once the retrograde IFT machinery reaches the BB, rearrangement of the IFT machinery occurs and anterograde IFT transport starts anew. Figure adapted from Inglis *et al.* 2009.

My study, as well as the study done by Hao and colleagues, has determined that *dyf-10* is the ciliary tubulin *tba-5* (Hao *et al.*, 2011). *tba-5* currently has three alleles (Figure 2.3). *dyf-10(e1383)* is a C>T missense mutation in a highly conserved residue causing a P360L amino acid change that results in a Dyf phenotype. *tba-5(qj14)* is characterized by a missense C>T mutation in another highly conserved residue in the first exon causing a A19V amino acid change that also results in a Dyf phenotype. Hao and colleagues determined that *qj14* mutants had intact MS and that the mutation interferes with protofilament-protofilament interaction. They found that A19V is located in the region of the protein that is important for proper lateral interactions between adjacent microtubule protofilaments. Dye filling experiments also confirmed *qj14*'s role in destabilizing microtubules, as the dye filling of *qj14* is temperature dependent. Particularly, *qj14* is a cold-sensitive mutant where the gene is functional at higher temperature and the mutant phenotype gets stronger at colder temperatures. This was not observed in *e1383* mutants; dye filling was constant at different temperatures (Hao *et al.*, 2011). The third allele, *tba-5(tm4200),* is a null allele that features a 291 bp deletion that deletes part of an intron and a part of the second last exon (Figure 2.3). *tm4200* deletes 105 bp of the 5' end of the second last exon, as a result, the deletion is inframe. *tm4200* does not display dye filling defects, has normal cilium morphology as assessed by IFT markers and has similar rates of IFT along the MS as WT worms (Hao *et al.*, 2011). Although *tba-5* was not previously predicted to be expressed in ciliated neurons (Hurd *et al.*, 2010), a recent study confirmed *tba-5's* expression in ciliated neurons (Hao *et al.,* 2011); TBA-5::GFP injected into *tba-5(qj14)* worms shows expression in amphid and phasmid sensory neurons. Hao and colleagues also completed a complementation test by crossing *qj14* males with *e1383* hermaphrodites. They found that the two alleles did not complement each other: 0 out of 14 worms dye filled in the amphid and phasmid neurons. In other words, *tba-5(qj14)* and *dyf-10(e1383)*

are alleles of the same gene (Hao *et al.*, 2011). Their study provided some evidence that *tba-5* is *dyf-10,* however, they did not confirm that the *tm4200* allele also fails to complement with both *qj14* and *e1383*. Furthermore, they only dye filled a small number of worms for their complementation test. This project will further confirm that *dyf-10* is the tubulin gene *tba-5* through a three-way complementation tests between the three different alleles. Hao and colleagues also surmise that *dyf-10(e1383)* is a 'recessive gain-of-function' mutation. Through dye filling experiments, this project will also determine whether *dyf-10(e1383)* is indeed a 'recessive gain-of-function' mutation.

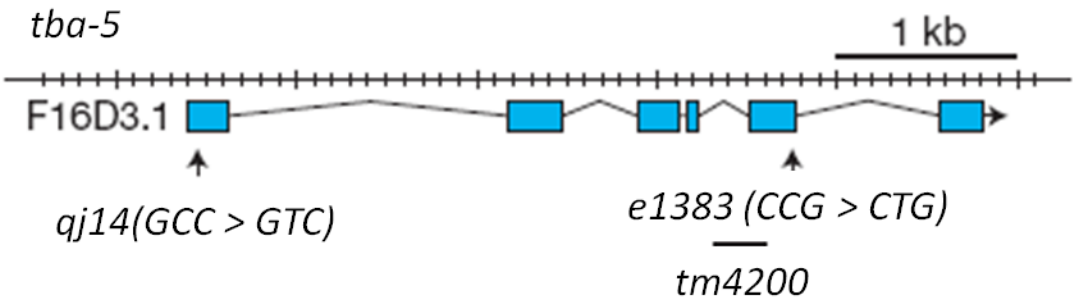


Figure 2.3. Alleles of tba-5

tba-5 has three alleles, *qj14, e1383* and *tm4200*. *qj14* features a C>T missense mutation in the first exon, resulting in a A19V amino acid change. *e1383* features a C>T missense mutation in the second last exon, resulting in a P360L amino acid change. *tm4200* is a null allele that features a 291bp deletion that spans part of an intron and part of the second last exon. Figure modified from Hao *et al.* (2011).

Since the *tba-5(tm4200)* null allele does not display any obvious ciliary defects, it is hypothesized that another alpha tubulin can compensate for the lack of TBA-5. Tubulins are a family of eukaryotic structural proteins that are the building blocks of microtubules. *C. elegans* have many isoforms of α- and β-tubulins that come together as heterodimers. There are nine and six α- and β-tubulins, respectively, encoded in the *C. elegans* genome (Gogonea *et al.*, 1999). Gene duplication followed by subfunctionalization typically in the expression domain is the main mechanism with which the tubulin paralogs were formed (Nielsen *et al.*, 2010); a paralog is a type of homolog that is separated by a gene duplication event. As a result, the high degree of similarity can be seen between the group of α- and β- tubulins (Figure 2.4). It has been found previously that TBA-1 and TBA-2 acts redundantly to function in pronuclear

migration and positioning of the first mitotic spindle (Phillips *et al.,* 2004). As a result, it is likely that there is another ciliary tubulin that acts redundantly with TBA-5. Hurd and colleagues identified candidate ciliary tubulins that are expressed in ciliated neurons by comparing four recent genomic analyses. Three of the studies identified genes expressed in ciliated neurons through serial analysis of gene expression (SAGE) (Blacque *et al.*, 2005), mRNA tagging followed by microarray analysis (Kunitomo *et al*., 2005), and microarray of genes down-regulated in *daf-19* mutants (Chen *et al.*, 2006). DAF-19 is currently the only *C. elegans* RFX-type transcription factor that regulates ciliary genes (Swoboda *et al.*, 2000) through the binding of the cis-regulatory element known as X-box motifs (Emery *et al.*, 1996). The fourth study used whole-organism microarray strategy to focus on genes that are expressed in the tail sensory rays, which contain sensory neurons (Portman and Emmons, 2004). Hurd and colleagues found that TBA-6, TBA-9 and TBB-4 are localized to the ciliary axonemes. Furthermore, of the nine α-tubulin genes, *tba-6* and *tba-9* ranked in the top 10th percentile in two and three out of the four studies, respectively. Using a transcriptional reporter, Hurd and colleagues found that *tba-9* showed expression in amphid neurons likely to be ADF, AFD, ASE, ASI, AWA and AWC. However, Hurd *et al*. only observed *tba-6* expression in the IL2 inner labial neurons and the HSN motor neurons. Thus, *tba-9* is more likely to be expressed in similar amphid neurons as *tba-5* than *tba-6*. Furthermore, the knockout allele of *tba-9*, *ok1858*, also results in grossly normal cilia similar to what was found for the *tm4200* allele; amphid and phasmid neurons of *tba-9(ok1858)* animals took up DiD, but dye filling of ASI amphid neurons was variable. Although *tba-9(ok1858)* animals were able to dye fill as normal, they did exhibit significant defects in responses such as mating and nose touch (Hurd *et al.*, 2010). As a result, it is hypothesized that *tba-9* can alone compensate for the absence of *tba-5* in regards to cilia formation and maintenance. To test this, it is expected that a double knockout mutant t*ba-5(tm4200)I;tba-9(ok1858)X* would result in defective cilia and therefore a Dyf phenotype.

	TBA-1b	TBA-2	MEC-12	TBA-4	TBA-5	TBA-6	TBA-7	TBA-8	TBA-9	TBB-1	TBB-2	TBB-4	BEN-1	MEC-7	TBB-6	TBG-1
TBA-1b	100.00	96.49	85.71	95.15	78.19	74.74	84.67	73.40	77.59	42.29	41.87	24.27	42.37	40.64	39.20	28.52
TBA-2	97.33	100.00	87.06	95.99	79.13	73.94	86.12	72.90	80.87	42.78	43.02	24.58	42.83	40.57	40.43	28.32
MEC-12	86.84	87.28	100.00	86.40	82.38	75.05	82.64	77.41	85.12	42.69	42.11	24.38	42.61	40.58	38.51	29.16
TBA-4	96.21	95.78	86.00	100.00	78.35	73.76	86.18	73.78	80.22	42.48	42.53	24.84	42.58	39.85	38.97	29.14
TBA-5	78.65	79.13	81.98	78.35	100.00	76.19	77.39	72.90	82.17	41.59	41.10	24.47	41.43	39.11	38.13	28.79
TBA-6	74.47	75.32	75.26	73.28	76.06	100.00	70.66	65.62	78.99	40.77	41.79	23.54	42.59	40.15	38.65	27.92
TBA-7	85.96	86.12	82.42	86.00	77.22	71.16	100.00	73.38	78.73	42.52	41.78	23.97	41.81	39.84	39.73	30.06
TBA-8	74.79	75.32	76.41	74.89	73.08	67.70	74.84	100.00	72.80	41.86	40.97	24.90	38.59	38.66	38.55	28.88
TBA-9	79.27	80.00	84.16	79.23	81.51	79.07	77.80	71.58	100.00	42.26	42.54	24.47	42.32	39.48	37.97	28.73
TBB-1	42.45	42.23	42.07	42.26	41.15	38.50	39.88	40.57	40.45	100.00	97.13	89.38	93.36	87.64	56.75	34.11
TBB-2	39.96	41.85	39.41	41.72	39.26	38.81	40.47	41.43	39.50	88.20	100.00	82.92	85.68	81.94	55.31	34.78
TBB-4	41.68	41.20	40.80	41.28	41.04	40.94	40.20	40.24	40.00	89.18	89.21	100.00	91.70	91.52	55.65	30.23
BEN-1	43.11	42.33	42.94	42.75	42.16	39.77	40.16	40.00	38.05	93.36	94.37	91.70	100.00	89.93	57.60	29.46
MEC-7	41.33	41.13	43.20	41.90	40.90	39.59	39.64	39.96	40.38	87.44	87.47	91.72	89.53	100.00	56.14	31.96
TBB-6	40.00	39.92	38.87	38.77	38.49	35.99	38.66	32.01	35.27	56.66	56.72	52.73	58.03	55.49	100.00	29.73
TBG-1	31.26	30.89	27.10	30.96	30.68	31.80	32.42	32.92	22.95	35.81	34.64	32.92	34.74	35.06	33.47	100.00

Figure 2.4. Peptide sequence similarity between different tubulins in C. elegans

Global Percentage Identity (PID) between predicted genBlastG (v138) gene, using *C. elegans* WS231 and query sequence. Figure kindly generated by Caleb Choo (Chen lab).

2.2. Materials and Methods

Strains and alleles. Growth and culture of *C. elegans* strains was carried out as described (Brenner, 1974). MT2179 *nDf25/unc-13(e1091)lin-11(n566)I* and RB1545 *tba-9(ok1858)X* were obtained from the *Caenorhabditis* Genetic Center (CGC). FX04200 *tba-5(tm4200)* was obtained from Mitani at the National BioResource Project (NBRP). *tba-5(qj14)* was obtained directly from Dr. Limin Hao (Hao *et al.,* 2011). BC9184 N2 hermaphrodites and BC9183 N2 males were obtained from Jun Wang. *dyf-10* and *dpy-5(e61)dyf-10* were obtained from Kyla Hingwing in the Hawkins lab.

Strain outcrossing protocol. The strains *tba-5(tm4200)* and *tba-9(ok1858)* were received without previous outcrossing. Furthermore, *tba-9(ok1858)* was generated by EMS mutagenesis by the OMRF knockout group and therefore may contain many other mutations. To rule out that other mutations in the genome might be affecting the dye filling results, the strains were outcrossed to WT N2 worms 4X in order to remove most of the background mutations. On a dot plate, two mutant hermaphrodites with eight BC9183 N2 males were plated together. The worms were incubated at 20°C for 3-4 days to give enough time for the worms to mate and generate L4 staged F1 progenies. If the cross was successful, meaning there are > 8 F1 males on the plate, ten L4 F1 hermaphrodites were plated individually. After incubation at 20°C for 1-2 days, the genotype of the F1 worms were checked by worm lysis and polymerase chain reaction (PCR). Only progeny of heterozygous mothers were kept. The F2 eggs were incubated for another 1-2 days to ensure they develop into L4 staged worms. Ten L4 F2 worms were plated individually from one heterozygous mother. After incubation at 20°C for another 3-4 days, the genotype of the F2 worms were checked by worm lysis and PCR. Progenies of homozygous F2 were kept. Finally, from one homozygous mother, another cross with BC9183 N3 males was set up. The process was repeated another three times until the strain has been outcrossed for a total of 4X.

Primer design. Primers were designed using the Primer3 (v.0.4.0) (Rozen and Skaletsky, 2000) and checked using Blast to make sure the primers can bind to the correct location uniquely. The following primers were designed to flank the *tm4200* allele (Figure 2.5A): *tm4200*-fwd TGCTTCACTTCGTTTCGATG and *tm4200*-rev AACCAGTTGGACACCAATCC. In *tba-5(tm4200)* worms, the *tm4200*-fwd and *tm4200*-rev primers (*tm4200* primer set) would give a DNA band of 374bp, whereas in N2 worms the PCR would give a DNA band of 665bp. The following primers were designed to flank the *ok1858* allele (Figure 2.5B): *ok1858*-fwd CATTTCCGGAGCCTCATAAA, *ok1858*-rev* GTATGTTCCGGTGCGAATCT and *ok1858*-rev CCGGTGGCAAGTATCAATTT. For *ok1858*, since the deletion is quite big, 1341bp, a forward, reverse and reverse-nested primer was designed. In *tba-9(ok1858)* worms, the *ok1858*-fwd, *ok1858*-rev* and *ok1858*-rev primers (*ok1858* primer set) give a band product of 432bp, whereas in N2 worms, the primers give a band product of 667bp. The larger PCR product in N2 worms would not be visible on the gel, since the product would be amplified relatively low compared to the shorter products that are easier to amplify.

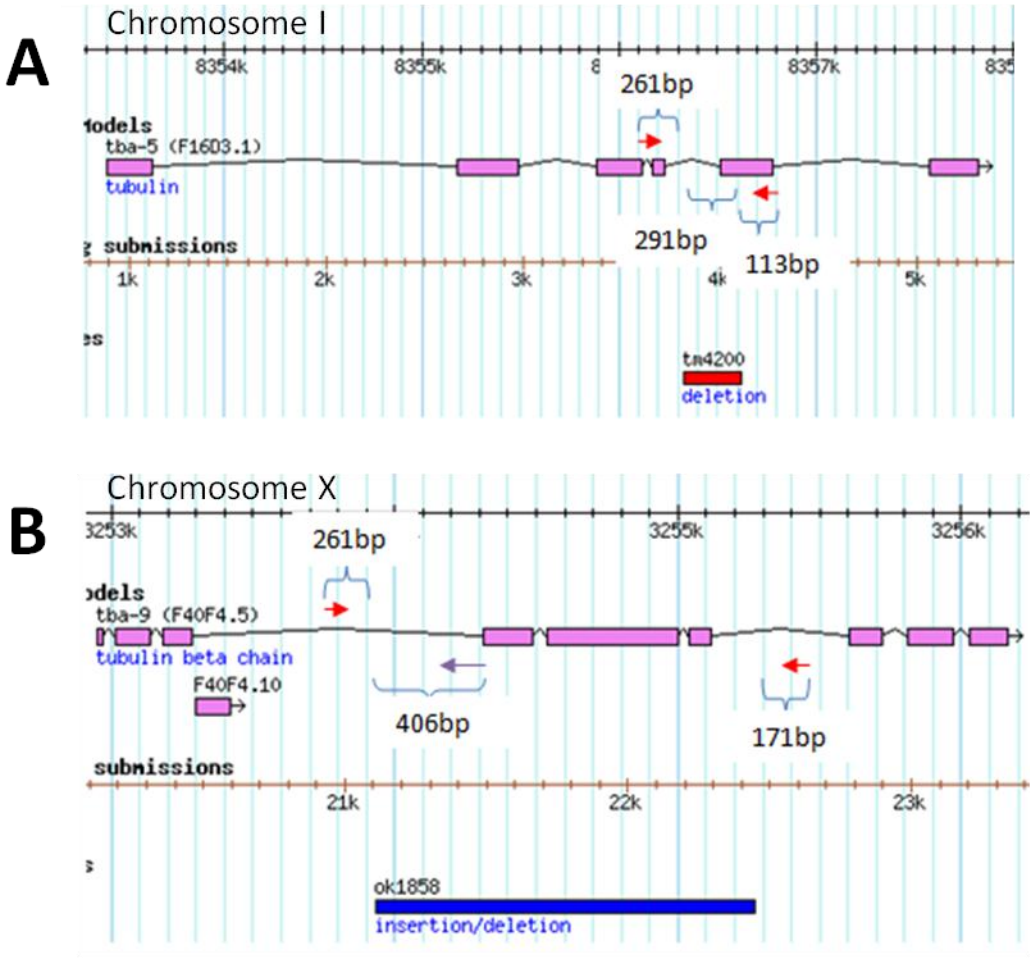


Figure 2.5. Primer Design to genotype tm4200 and ok1858 alleles

(A) *tm4200* allele features a 291bp deletion located on chromosome I that deletes part of an intron and exon. Red arrows indicated the approximate location of the forward and reverse primers that are designed to flank the *tm4200* allele. PCR genotyping of wild type N2 worms should give a 665bp band (261bp+291bp+113bp). PCR genotyping of *tba-5(tm4200)* worms should give a 374bp band (261bp+113bp). (B) *ok1858* allele features a 1341bp deletion and insertion of TT bases. Red arrows indicated the approximate location of the forward and reverse primers designed to flank the *ok1858* allele. The purple arrow indicated the approximate location of the nested reverse primer. Wild type N2 worms should give a 667bp band (261bp+406bp). The larger PCR product would not be visible on the gel since the product would be amplified at low quantities compared to the shorter product. *tba-9(ok1858)* worms should give a 432bp (261bp+171bp) band. Image modified from the *C. elegans* genome browser (http://www.wormbase.org).

Worm lysis and PCR genotyping protocol. The LB buffer consist of 10 mM Tris pH 8.4, 50 mM KCl, 2.5 mM $MgCl_2$, 0.45% Tween 20, and 0.5 mg/mL gelatin. 5 µL of 2 µg/µl proteinase K was added to 500 µL of the LB buffer and mixed. To extract DNA

from the cells, 1 worm was added to 5 μL of lysis buffer (LB) buffer with proteinase K. The mix was then incubated at -80°C for 10 min to facilitate the 'cracking' of the worm's body. Next the PCR tube is subjected to the worm 'lysis60' protocol in the PCR machine, which consists of incubation at 60°C for 60 min, 95°C for 15 min to deactivate proteinase K and finally storage at 4°C. The following PCR reaction mix with the *tm4200* primer set was used: 11.9 μL dH_2O, 2 μL TBA buffer, 2 μL $MgCl_2$, 2 μL 2.5 nM dNTP, 1 μL *tm4200*-fwd primer, 1 μL *tm4200*-rev primer and 0.1 μL Hutter Taq polymerase (Hutter lab) for a total of 20 μL per reaction. The following PCR reaction mix with the *ok1858* primer set was used: 10.9 μL dH_2O, 2 μL TBA buffer, 2 μL$MgCl_2$, 2 μL 2.5 nM dNTP, 1 μL *ok1858*-fwd primer, 1 μL *ok1858*-rev* primer, 1 μL *ok1858*-rev primer and 0.1 μL Hutter Taq polymerase for a total of 20 μL per reaction. The PCR reaction mix was mixed with the worm lysis and transferred to the PCR machine. The PCR reaction was subjected to the 'abshort' PCR protocol which consists of annealing at 65°C-67°C for 30 sec and elongation at 72°C for 1 min for 33 cycles. The PCR products are visualized on a 1% agarose gel mixed with EtBr and visualized using UV light.

Heat shock protocol. To generate males for crosses, three L4 hermaphrodites were plated per plate for three plates. The worms were incubated at 30°C for 5, 5.5 and 6 hours and store at 20°C for 3-4 days. F1 males were screened for on the plates and plated together with one or two F1 hermaphrodites from the heat shocked plates in order to generate more males from the strain. Overall, the strains *dyf-10(e1383)* and *tba-9(ok1858)* 4X outcrossed were easy to generate males with. However, only one male were found after heatshock with *tba-5(tm4200)* and *tba-5(qj14)*.

Complementation tests. The following crosses were dye filled: *dyf-10(e1383)* males crossed with *tba-5(qj14)* hermaphrodites, *dyf-10(e1383)* males crossed with *tba-5(tm4200)* 4X hermaphrodites and *tba-5(tm4200)* 4X male crossed with *tba-5(qj14)* hermaphrodites. Since only one *tba-5(tm4200) 4X* male was produced after the heat shock protocol, the *tba-5(tm4200) 4X* male was mated to many *tba-5(qj14)* hermaphrodites on separate plates. Following the cross, males and hermaphrodite mothers were removed, leaving only the F1 eggs. Once the F1 has reached L4 to adult stage, they were dye filled. Overall, each cross was replicated many times to generate

enough F1 for dye filling. Plates displaying > 8 F1 males, which indicate successful mating, were chosen for dye filling.

Dye filling assay protocol. Live animals were stained with a lipophilic carbocyanine dye DiO (3,3'-dioctadecyloxacarbocyanine percchlorate). Well-fed worms from a plate were transferred into a 5 mL microcentrifuge tube by washing the plate with 1 mL M9 buffer or by picking the worms and placing them into the microcentrifuge tube with M9 buffer. The worms were centrifuged at 2000 rpm for 1 min. After centrifugation, the supernatant was removed leaving the loose worm pellet and re-suspended in 0.5 mL M9 buffer. 1 μL of 2mg/mL DiO was added to the worm suspension, mixed, and the microcentrifuge tube was completely wrapped in aluminum foil. The mixture was incubated at room temperature, which is approximately 20°C, on a slow shaker for 30 min. After incubation, the worms were centrifuged at 2000 rpm for 1 min and washed once with 1 mL of M9 buffer. Washing consists of removing the supernatant after centrifugation, adding 1 mL of M9, mixing, and centrifugation. After washing and removing most of the supernatant, the worm pellet was transferred to a fresh plate. A yellow pipette tip was used to draw up the worm pellet and only 7 μL of worm pellet was drawn up at a time. This was to minimize the loss of worms through adhesion to the pipette tip. After double checking that the worms were transferred onto the plate, the plate was wrapped in aluminum foil and incubated at room temperature for 3-4 hours to allow time for the worms to recover and for the worms to expel the dye from their gut. After incubation, the worms were transferred into a 5 mL microcentrifuge tube by washing the plate with 1 mL M9 buffer or picking worms and placing them into the microcentrifuge tube with M9 buffer. Next, the worms were centrifuged at 2000 rpm for 1 min and the supernatant was removed. Finally 7 μL of worms were drawn up using a yellow pipette tip and deposited on to the 2% agarose pad on a slide. Sodium azide was added until the final [NaN_3] is 100mM. This is approximately 5 μL of 0.25M NaN_3. Before visualization on the GFP compound microscope, the number of worms on the slides was counted using the dissecting microscope. Finally, the number of worms that dye fill are counted and DIC and DiO pictures were taken using the GFP compound microscope. For each experiment, N2 was used as the positive control and *dyf-10(e1383)* was used as the negative control.

Generation of *tba-5(tm4200)I;tba-9(ok1858)X* double mutant. The strains *tba-5(tm4200) 4X* and t*ba-9(ok1858) 4X* were used to generate the double mutant. Figure 2.6 summarizes the construction of the double mutant. (1) On a dot plate, two *tba-5(tm4200) 4X* hermaphrodites and eight *tba-9(ok1858) 4X* males were plated. The worms were incubated for 3-4 days at 20°C to allow F1 progenies to develop into L4 staged worms. The cross is considered successful if there are > 6 F1 males. (2) From a successful cross, ten L4 hermaphrodites were plated individually and incubated for 1-2 days at 20°C. (3) For each plate, the F1 hermaphrodite mothers were lysed and genotyped by PCR. If the cross is successful, the genotype of the F1 worms should be *tba-5(tm4200)/+; tba-9(ok1858)/+*. It is expected that 100% of the F1 should be double homozygotes and it was observed that 88.9% (8 out of 9) worms are heterozygous for both *tba-5(tm4200)* and *tba-9(ok1858)* (Figure 2.7A). I was unable to determine whether worm #2 was heterozygous or homozygous for the ok1858 allele since there was not enough DNA for the PCR reaction. Furthermore, plate #10 was contaminated and therefore did not undergo PCR genotyping. Next, progeny of double heterozygous mothers was kept and incubated for 1-2 days at 20°C for the F2 eggs to develop into L4 staged worms. (4) Forty L4 staged F2 hermaphrodites were individually plated from one heterozygote mother (worm #3). The worms were incubated for 1-2 days at 20°C for the F2 hermaphrodites to lay eggs. (5) For each plate, the F2 hermaphrodite mothers were lysed and genotyped by PCR. Theoretically there is a 1/16 (6%) chance that the worms will have the genotype *tba-5(tm4200)/ tba-5(tm4200);tba-9(ok1858)/ tba-9(ok1858)*. It was observed that 10% (2 out of 20) of the worms are homozygous for both *tm4200* and *ok1858* (Figure 2.7B, worm #12 and 15). Next, progeny of double homozygous mothers were kept and incubated for 1-2 days at 20°C for the F3 eggs to develop into L4 staged animals. (6) F3 worms, which have a genotype of *tba-5(tm4200)/ tba-5(tm4200);tba-9(ok1858)/ tba-9(ok1858),* were dye filled with DiO and visualized with GFP compound microscopy. Furthermore, *tba-5(tm4200)/ tba-5(tm4200);tba-9(ok1858)/ tba-9(ok1858)* worms were genotyped again to confirm that they are indeed homozygous for both alleles (data not shown). It is noted that worms heterozygous for the *tm4200* allele display three bands instead of two bands (Figure 2.7A top panel). There appears to be an additional band that is very similar in size to the WT band. However, PCR of homozygous N2 and *tm4200* worms generate a clear WT and mutant band, respectively.

Since the band is not similar in size to the mutant band in heterozygotes, genotyping of the *tm4200* allele is still successful.

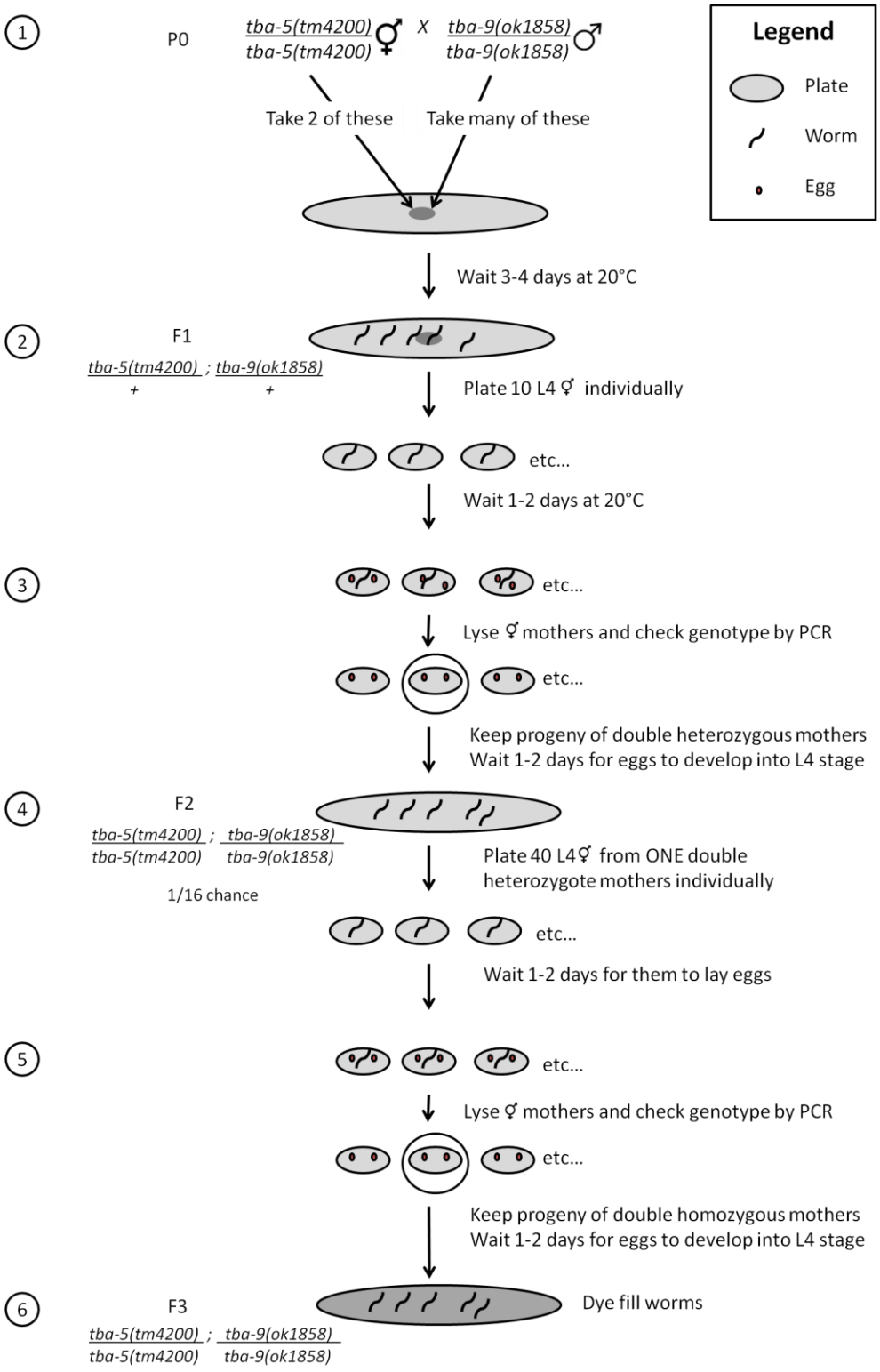
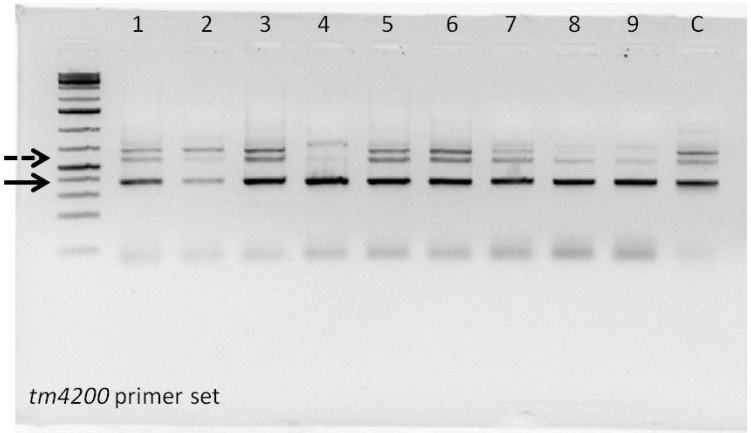
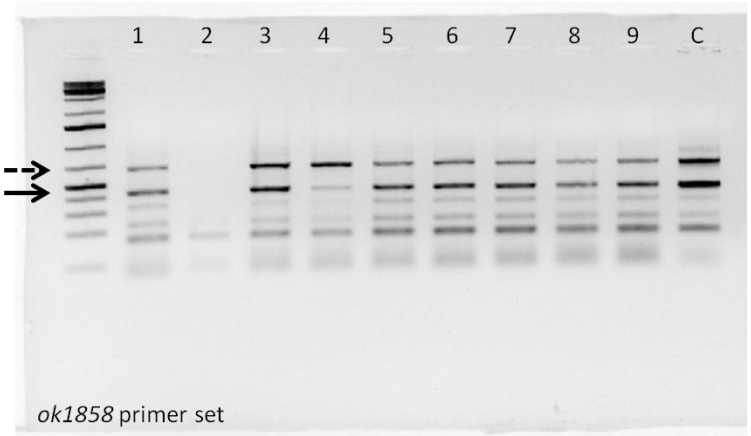


Figure 2.6. Method of Generating of tba-5(tm4200);tba-9(ok1858) double mutant

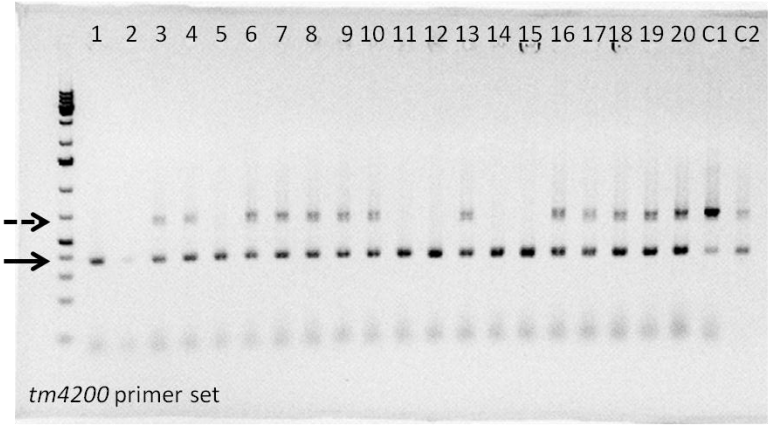
(1) Two *tba-5(tm4200)* hermaphrodites were plated with eight *tba-9(ok1858)* males. The worms were incubated for 3-4 days to give the worms enough time to mate and generate L4 staged F1. The cross is considered successful if there are > 6 F1 males. (2) Ten L4 hermaphrodites were plated individually and incubated for 1-2 days to give enough time for the hermaphrodites to lay eggs. (3) For each plate, F1 hermaphrodite mothers were lysed and genotyped by PCR. If the cross is successful, the genotype of the F1 worms should be *tba-5(tm4200)/+; tba-9(ok1858)/+*. Progeny of double heterozygous mothers were kept and incubated for another 1-2 days for the F2 eggs to develop into L4 stage worms. (4) Forty F2 L4 hermaphrodites were plated individually from one heterozygote mother. The worms were incubated at for 1-2 days. (5) For each plate, F2 hermaphrodite mothers were lysed and genotyped by PCR. There is 1/16 chance that the worms will be homozygous for *tba-5(tm4200);tba-9(ok1858)*. Progeny of double homozygous mothers were kept and incubated for 1-2 days for the F3 eggs to develop into L4 stage worms. (6) Finally, F3 worms were dye filled with DiO. The worms were kept at 20°C.

A: Genotyping of heterozygote mutants





B: Genotyping of homozygote mutants



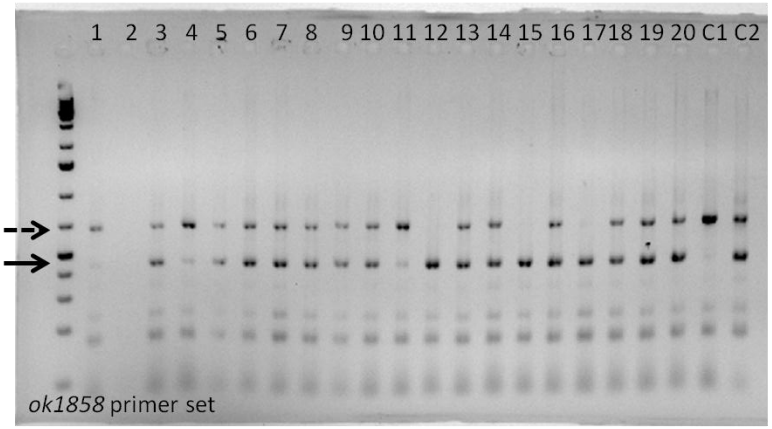


Figure 2.7. Generation of tba-5(tm4200)I;tba-9(ok1858)X mutants: PCR Genotyping

Method of genotyping for each PCR reaction: (1) lyse a single worm, (2) aliquot the lysate equally into two tubes, (3) run two PCR using the *tm4200* and *ok1858* primer set separately and (3) visualize the PCR reaction with agarose gel with UV light. (A) Genotyping of heterozygous F1 mothers in Figure 2.6 step 3. Top panel shows worms that have undergone PCR reaction with the *tm4200* primer set at a annealing temperature of 66°C. Bottom panel shows the same worms that have undergone PCR with *ok1858* primer set at a annealing temperature of 67°C. Lanes 1-9 represent worms #1-9, respectively. Lane C is a control showing *tm4200* and *ok1858* heterozygotes in the top and bottom panel, respectively. 8 out of 9 worms (#1, 3-9) are heterozygous for both *tm4200* and *ok1959*. (B) Genotyping of homozygous F2 mothers in Figure 2.6 step 5. Top panel shows worms that have undergone PCR with *tm4200* primer set at a annealing temp of 65°C. Bottom panel shows the same worms that have undergone PCR with *ok1858* primer set at annealing temp of 65°C. Lanes 1-20 represent worms #1-20, respectively. Lane C1 is a control showing a band for N2 worms. Lane C2 is another control showing bands for *tm4200* and *ok1858* heterozygotes in the top and bottom panel, respectively. 2 out of 20 worms (#12 and 15) are homozygous for both the *tm4200* and *ok1858* allele. All agarose gel was run at 100V for 34min. The dashed arrow indicates the band corresponding to the WT N2 PCR product. The solid arrow indicates the band corresponding to the mutant PCR product.

***Deficiency mapping: G*eneration of *dyf-10(e1383)/nDf25* worms for dye filling (Figure 2.8)**. For each dot plate, one *nDf25/unc-13(e1091)lin-11(n566)* hermaphrodite was crossed with four *dyf-10(e1383)* males for a total of fifteen dot plates. The worms were incubated for 3-4 for days at 20°C to allow F1 progeny to develop into L4 stages animals. From successful crosses (plates with > 8 F1 males), as many F1 hermaphrodites were plated individually as possible (total 337 plates). Plates where the F1 either were lost, found dead on the plate, displaying UncVul phenotype or developed into males were thrown away. After incubation for 1-2 days, the F1 mothers were separated from their eggs. The F1 mothers were incubated at 15°C to slow their growth and allow screening of their F2 progenies which were kept at 20°C. For the F2 progenies, plates with small brood size and absence of UncVul phenotype were kept and their corresponding F1 mothers were dye filled. A total of 59 F1 worms passed screening and were dye filled. UncVul phenotype is characterized by worms that are small, kinky, paralyzed and vulvaless. Often the UncVul phenotype result in “bags” of worms, where the progeny are trapped inside the hermaphrodite and eventually bursts out of their hermaphrodite mother.

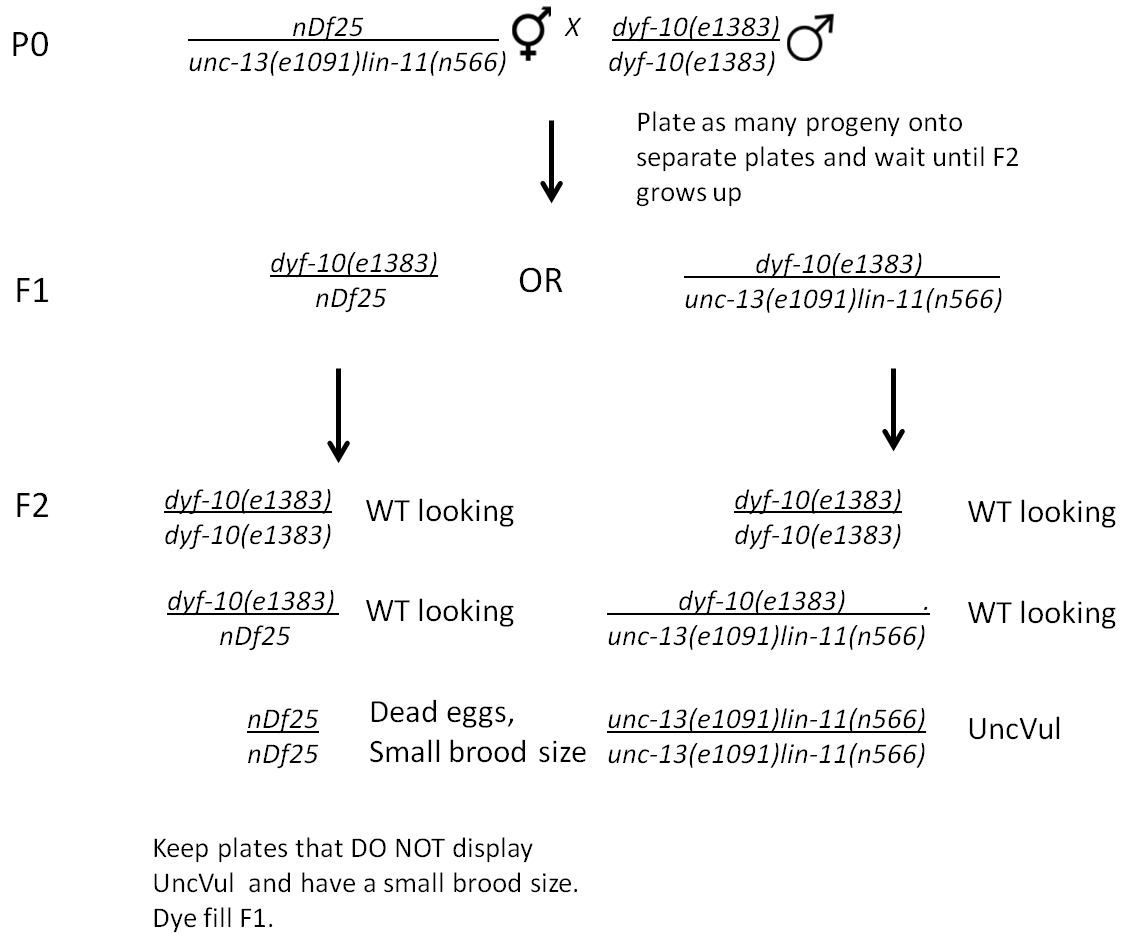


Figure 2.8. Method of generating dyf-10(e1383)/nDf25 worms for dye filling

For each dot plate, one *nDf25/unc-13(e1091)lin-11(n566)* hermaphrodite was crossed with four *dyf-10(e1383)* males for a total of fifteen dot plates. The worms were incubated at 20°C for 3-4 days to let the F1 progeny develop into L4 staged worms. From plates where the cross was successful (> 8 F1 males), as many F1 hermaphrodites were plated as possible (total 337 plates). Plates where F1 were either lost, found dead, gave rise to UncVul F1s or F1s that develop into males were thrown away. After 1-2 days, F1 mothers were separate from their eggs and stored at 15°C. F2 eggs were still incubated at 20°C. When the F2 progeny are adult staged, plates with small brood size and absence of UncVul phenotype were kept. Finally, F1 mothers from plates with small brood size and absence UncVul phenotype were dye filled. UncVul phenotype is characterized by worms that are small, kinky, paralyzed and vulvaless. Often UncVul result in "bags" of worms, where the progeny are trapped inside the hermaphrodite until they eventually bursts out.

Genotype of mutant strain. The mutant strain was kindly generated by Kyla Hingwing (Hawkins lab). The strain SP1709 from the Riddle lab, which contains *dyf-10(e1383)I,* was recombined with a *dpy-5(e61)I* containing strain. This resulted in a strain in which *dpy-5(e61)* is linked to and acts as a marker for *dyf-10(e1383)*. The recombined strain was then crossed to the strain SP1237 from the Riddle lab which contains *dyf-4(m158)V.* The final mutant strain contains *dpy-5(e61)I*, *dyf-10(e1383)I* and *dyf-4(m158)V* (Figure 2.9). For the purpose of the project, only *dpy-5(e61)* and *dyf-10(e1383)* will be focused on.

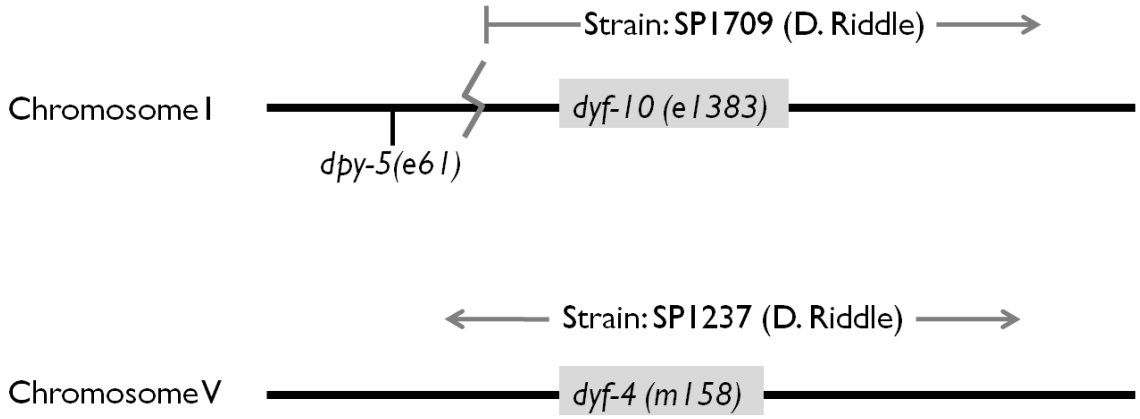


Figure 2.9. Mutant Strain Construction

Mutant strain was kindly generated by Kyla Hingwing (Hawkins lab). The strain SP1709 from the Riddle lab, which contains *dyf-10(e1383)I,* was recombined with a *dpy-5(e61)I* containing strain. This resulted in a strain in which *dpy-5(e61)* is linked to and acts as a marker for *dyf-10(e1383)*. The recombined strain was then crossed to the strain SP1237 from the Riddle lab which contains *dyf-4(m158)V.* The final mutant strain contains *dpy-5(e61)I*, *dyf-10(e1383)I* and *dyf-4(m158)V*. Note, diagram is not drawn to scale.

Whole genome sequencing and read alignment (Figure 2.10). The mutant strain was sequenced using the Illumina/Solexa paired end sequencing technology. The mean insert size or fragment length of the paired-end reads is 383 bp. Each read is 76 bp in length. The Hobert strain was also sequenced using the Illumina Solexa paired end sequencing technology, by Dr. O. Hobert at the Columbia University Medical Center, NY. Each reads is 35 base pairs in length. The Hobert sequencing reads were originally encoded in the "export" format; to convert the files into a STDFQ format, the 'fq_all2std.pl export2std' command from the MAQ software package was used. The mean insert size or fragment length of the paired-end reads is 227 bp. The average insert size for the strains was collected using Picard's 'CollectInsertSizeMetrics.jar' tool

(http://picard.sourceforge.net), version 1.40. Single end reads from the Horvitz *C. elegans* N2 Bristol strain was downloaded from The Genome Institute at Washington University. The sequencing reads were originally encoded in "srf" format. To convert the files into STDFQ format, the 'srf2fastq' command was used from the software io_lib (http://sourceforge.net/projects/staden/files/io_lib/), version 1.11.5. The sequencing reads from the mutant, Hobert and Horvitz strains were mapped separately to the WS224 Wormbase release of the N2 *C. elegans* reference genome using Novoalign version 2.07.13 (www.novocraft.com). The performances of Novoalign as well as other alignment algorithms were compared in a recent study using a Sequence Alignment evaluation suite (SEAL) that simulates short read sequencing runs and evaluates the output of each software. It was found that Novoalign along with Bowtie and BWA are the highly accurate even at high error rates, provided that a mapping quality threshold > 0 is applied; accuracy is defined as the number of correctly mapped reads over the sum of the number of correctly and incorrectly mapped reads. Novoalign is also highly accurate at varying InDel sizes and frequencies with a mapping quality threshold > 0, compared to other software (Rufallo *et al.,* 2011). Furthermore, Novoalign competes with MAQ on speed and is developed for Illumina sequencing platform. Novoalign also features paired-end gapped alignment, mapping qualities and adapter trimming functions. As a result, Novoalign was used in this project (www.novocraft.com). Default parameters in Novoalign was used except for the following parameters for the mutant and Hobert strains: -F STDFQ, which specifies the format of the read files, and -i PE 350,100, which specifies the approximate fragment length and standard deviation for reads in paired-end mode. Default parameters in Novoalign were used except for the following parameter for the Horvitz strain: -F STDFQ. After generating the BAM file from Novoalign, PCR duplicates were removed from the alignment using the 'samtools rmdup' command (Li *et al.,* 2009). The resulting BAM file was visualized on gbrowse. The WS224 FASTA file containing the genome sequence and the WS224 GFF3 file for GBrowse visualization were downloaded from the following sites: ftp://ftp.sanger.ac.uk/pub/wormbase/WS224/genomes/c_elegans/sequences/dna/c_elegans.WS224.dna.fa.gz and ftp://ftp.wormbase.org/pub/wormbase/genomes/c_elegans/genome_feature_tables/GFF3/c_elegans.WS224.gff3.gz. Sequencing of the Mutant strain generated approximately 83.2 million sequencing reads. After alignment, 84.02% of the mutant strain reads

mapped onto the WS224 *C. elegans* reference genome with an average read depth of 52 reads per position (Table 2.2). In the Hobert strain, 79.32% of the 78.3 million reads were mapped to the reference genome with an average read depth of 22 reads per position. Sequencing of the Horvitz N2 strain generated 130 million single-end reads. 68.70% of the reads were mapped to the reference genome with an average read depth of 31 reads per position.

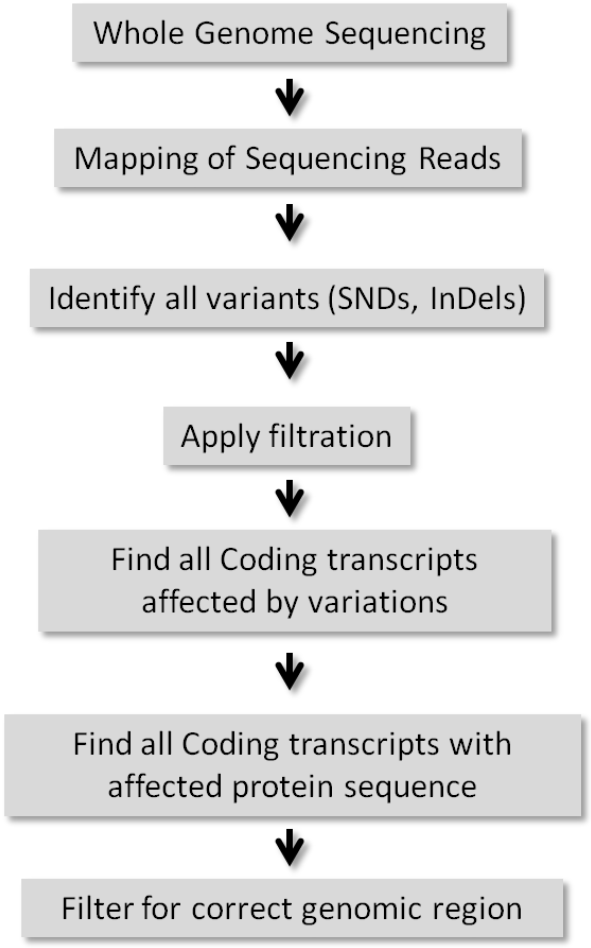


Figure 2.10. Variant Detection Pipeline for the Mutant Strain

Whole Genome Sequencing was completed on the mutant strain using the Illumin/Solexa platform. Novoalign (www.novocraft.com) was used to map the paired-end reads to the WS224 reference *C. elegans* genome. SNDs and small InDels were detected using VarScan (Koboldt, *et al.*, 2009) and filtered using the parameters: MinCov $\geq$ 9, VarFreq $\geq$ 30. Variant Analyzer (Chen lab, unpublished) was used to annotate coding transcripts affected by the filtered variations. Coding transcripts with affected protein sequence due to missense, nonsense, and frame shift mutations were identified and those that fall into the correct genomic region are considered as candidate genes for *dyf-10(e1383)*.

Table 2.2. Read Mapping Details for the Mutant, Hobert and Horvitz strains

Number of Reads			Category
Mutant Strain	**Hobert Strain**	**Horvitz Strain[a]**	
83164180	78275951	130277517	In total
0	0	0	QC failure
0	0	0	Duplicates
69873612 (84.02%)	62087932 (79.32%)	89500436 (68.70%)	Mapped
83164180	78275951	n/a	Paired in sequencing
41582198	39138012	n/a	Read1
41581982	39137939	n/a	Read2
68305491 (82.13%)	61614450 (78.71%)	n/a	Properly Paired
68400238	61686605	n/a	With itself and mate mapped
1473374 (1.77%)	401327	n/a	Singletons
35818	43542	n/a	With mate mapped to a different chromosome
35709	43197	n/a	With mate mapped to a different chromosome (mapQ>5)
52	22	31	Average read depth

Notes: Reads were aligned to WS224 reference *C. elegans* genome using Novoalign (www.novocraft.com). Information above was collected from the BAM alignments using the ‘’samtools flagstat’ command.

a The Horvitz strain consists of single reads, as a results, read pair information is not available for this alignment.

Variation detection (Figure 2.10). Small variations, including insertions and deletions (InDels), which are typically < 20bp in length, and single nucleotide differences (SNDs) were detected for the mutant, Hobert and Horvitz strains by first generating ‘pileup’ file using the ‘samtools pileup’ command, SAMtools version 0.1.7a. VarScan (version 2.2.3) was then used to detect SNDs and InDels using the ‘pileup2snp’ and ‘pileup2indel’ commands, respectively, by specifying 0 for the minimum coverage (MinCov) and 0 for minimum variant frequency (VarFreq), see discussion for explanation. Note, by default the minimum base quality (BaseQual) is set a 15 (Koboldt *et al.,* 2009). The variations are finally filtered for MinCov $\geq$ 9 and VarFreq $\geq$ 30% for the mutant strain, and MinCov $\geq$ 5 and VarFreq $\geq$ 30% for the Hobert and Horvitz strains

using PERL scripts (www.perl.org). The SND and InDel files are converted into GFF3 files using PERL scripts and visualized on generic genome browser (GBrowse). All PERL scripts were written by me.

Annotation of coding transcripts with variations and compilation of data (Figure 2.10). To extract a list of coding transcripts affected by the variations detected, first, a list of SNDs and small InDels were inputted into Variant Analyzer separately (Chen lab, unpublished). Second, the output was extracted for coding transcripts that have been affected by the following variations: missense, nonsense, synonymous, frame shifting and non-frame shifting variations. Third, using the Bio::DB::SeqFeature::Store bioperl module and accessing the WS224 *C. elegans* reference database, the variations are categorized into exon, intron, UTR and intergenic regions. Fourth, all variations from multiple strains, namely the mutant strain and the two WT strains (Hobert and Horvitz) were compiled and compared. Fifth, information extracted Variant Analyzer and Bio::DB:SeqFeature::Store bioperl module were annotated for each variation. Sixth, the list of variations were filtered for those that affect the protein sequence, are homozygous (VarFreq $\geq$ 75%), and does not appear in the wild type Hobert and Horvitz strains. Variations were considered non-WT if the positions of the variations did not match with any of the variations, that have MinCov $\geq$ 5 and VarFreq $\geq$ 30%, detected in the Hobert and/or Horvitz strains. Finally, the remaining variations were filtered for the correct genomic location.

Westgrid. For effective computational analysis, I aligned the sequencing data on the Western Canada Research Grid (WestGrid) which operates a high performance computing, collaboration and visualization infrastructure across Western Canada. In particular, I will be utilizing WestGrid's Bugaboo server, which consists of a network of computers where multiple tasks can run in parallel for faster computational analysis, and the Joffre server for data visualization (http://www.westgrid.ca).

2.3. Results

There are many variations detected in the alignment of the Hobert and Horvitz sequencing reads to the *C. elegans* reference genome (Table 2.3). It is expected that

there are much fewer variations detected in the Horvitz strain, since the Horvitz strain is much closer to the reference *C. elegans* genome than the Hobert strain that was derived from the CGC (personal communication). Furthermore, the Horvitz and Hobert strains are largely similar. 53.8% (1138 out of 2117) of SNDs, 95.2% (822 out of 863) of insertions and 84.5% (337 out of 399) of deletions in the Horvitz strain are shared with the Hobert strain (Table 2.3). Since *C. elegans* generate one mutation every three generations, the variations that are not shared between the two laboratory N2 strains, which are mostly SNDs, are likely due to genetic drift (Flibotte *et al.*, 2010). Overall, heterozygous and homozygous variations that were detected in the Hobert and Horvitz strains were compared to variations detected in the mutant strains. Mutations derived from EMS mutagenesis in the mutant strain should not be present in the two WT strains. Since the WT strains do not display the mutant phenotype of interest, variations that are shared with the WT strains are unlikely to be responsible of the mutant phenotype observed in the mutant strain.

Table 2.3. Summary of Variations Detected in the Hobert and Horvitz strains

Category	Hobert strain	Horvitz strain	Shared[b]
Total # SNDs[a]	5564	2117	1138
Total # small Insertions[a]	1077	863	822
Total # small Deletions[a]	543	399	337

a Variations have been filtered for MinCov $\geq$ 5 and VarFreq $\geq$ 30%

b Shared variations are based on start and end position information

There are many variations detected genome wide following alignment of the mutant strain to the *C. elegans* reference genome. After variant detection and filtration, 5436 SNDs and 1893 small InDels were detected by VarScan (Table 2.4). Of the total, 1930 SNDs and 1386 small InDels are homozygous, meaning they have a VarFreq $\geq$ 75%. After further filtration, 916 homozygous SNDs and 140 homozygous small InDels do not occur in the WT strains either as homozygous or heterozygous variations.

Table 2.4. Summary of Variations Detected in Mutant Strain

Total # SNDs[a]	5436		
Total # homozygous SNDs[b]	1930		
Total # homozygous and non-WT SNDs[c]	916		
Total # small InDels[a]	1893	(1221 insertions,	672 deletions)
Total # homozygous small InDels[b]	1386	(891 insertions,	495 deletions)
Total # homozygous and non-WT small InDels[c]	139	(58 insertions,	81 deletions)

a Variations have been filtered for MinCov $\geq$ 9 and VarFreq $\geq$ 30%

b Homozygous variations have VarFreq $\geq$ 75%

c Non-WT variations in the mutant strain do not occur in the Hobert and Horvitz strains. In other words, non-WT variations in the mutant strain do not have the same corresponding position as variations in the Hobert and Horvitz strains that have been filtered for MinCov $\geq$ 5 and VarFreq $\geq$ 30%.

Within the list of homozygous non-WT variations, the positive control, *dpy-5(e61*) marker, was found by the variation detection pipeline (Figure 2.11). The *e61* allele causes a recessive dumpy (Dpy) phenotype that is characterized by the general physical shortening of the worms (Thacker *et al.*, 2006). The C>A nonsense mutation that is responsible for the phenotype, was found with a VarFreq of 100% and at a depth of 19 reads, which is surprisingly much lower than the average read depth of 56 for the strain (Table 2.2). In addition, a C>T missense mutation upstream of the nonsense mutation was found with a VarFreq of 100% and at a depth of 18 reads. However, the missense mutation is not documented as part of the *e61* allele on WormBase, but it is present in other strains that contain the *dpy-5(e61)* allele. For example, Thacker and colleagues also sequenced the same base that coincides with the missense mutation (Thacker *et al.*, 2008). In addition, from personal communications with Dr. D. L. Baillie, it appears that the missense mutation is present in many more strains that carry the *dpy-5(e61)* allele; therefore the mutation may have been picked up by Sydney Brenner in 1969.

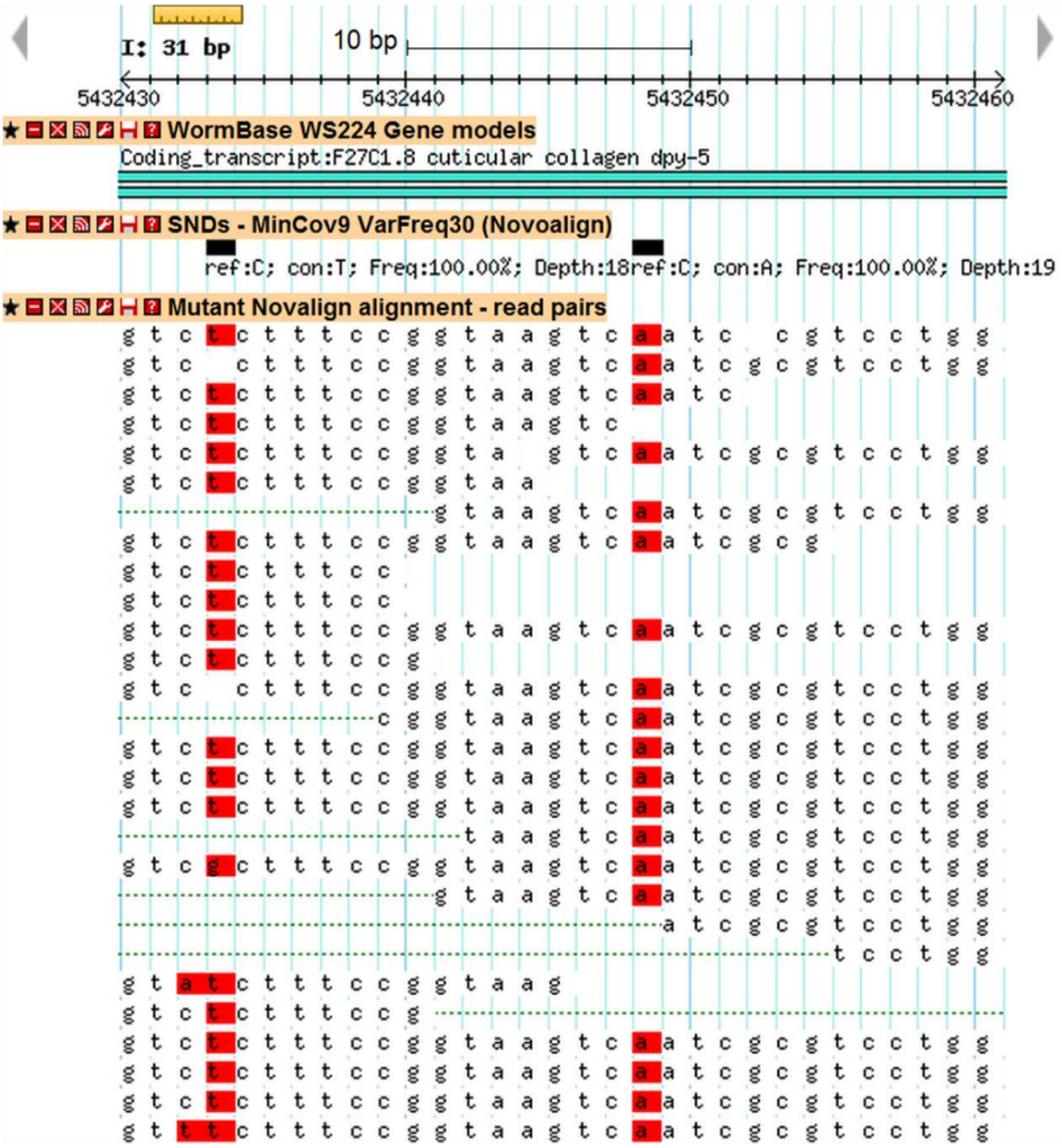


Figure 2.11. Positive Control: dpy-5(e61) detected by the variant detection pipeline

Region shown: I:5,432,430..5,432,460. *dpy-5(e61)* features a C>A nonsense mutation (right) at 100% variant frequency and a read depth of 19. In addition, the sequence alignment also reveals a C>T missense mutation (left) that is found in the sequence alignment at 100% variant frequency and read depth of 18. The mutant strain sequence reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Abbreviations: ref (reference base), con (consensus base), Freq (variant frequency) and Depth (read depth).

After filtering the homozygous non-WT variations for those that affect protein coding sequence that resides in the genomic area of interest, the only candidate gene for *dyf-10* is ZC308.1, or *gld-2* (Figure 2.12, Table 2.5). *gld-2* (germ-line development-2) features a missense mutation of a C>T change at 100% VarFreq and read depth of 42, which has not been characterized in previous literature. The variation itself is likely to be a true positive since (1) the C>T substitution is characteristic of EMS mutagenesis (Flibotte *et al.*, 2010), (2) the VarFreq at 100% clearly indicates the variation is homozygous, (3) the read coverage is very close to the average for the mutant strain, and (4) the mutation does not occur in both of the WT strains. *gld-2* codes for a catalytic subunit of the cytoplasmic poly(A) polymerase (cPAP). cPAP is responsible for maintaining or lengthening the poly(A) tails of a subset of mRNAs. It has also been found that GLD-2 enhances entry into the meiotic cell cycle (Suh *et al.*, 2006; Wang *et al.*, 2002). A Literature search of *gld-2* did not yield any dye filling assays of *gld-2* mutant worms. However, based on what is known of *gld-2*, none of *gld-2*'s function seems to be cilia related, or can explain the Dyf phenotype that is observed in *dyf-10(e1383)* worms. As a result, *gld-2* is not likely the identity of *dyf-10*.

It is possible that the *dyf-10(e1383)* was mis-mapped. The search for candidate genes was therefore expanded to the whole of chromosome I (Table 2.5). As expected from EMS mutagenesis, most of the substitutions are either C>T or G>A changes (Flibotte *et al.*, 2010). In addition, majority of the candidates have 100% VarFreq, read depth close to the average read depth for the mutant strain, and do not occur in the wild type Hobert and Horvitz strains. Again the table lists the two variations associated with the positive control marker *dpy-5(e61)* and the missense mutation associated with *gld-2*, which was the only candidate in the predicted region of interest.

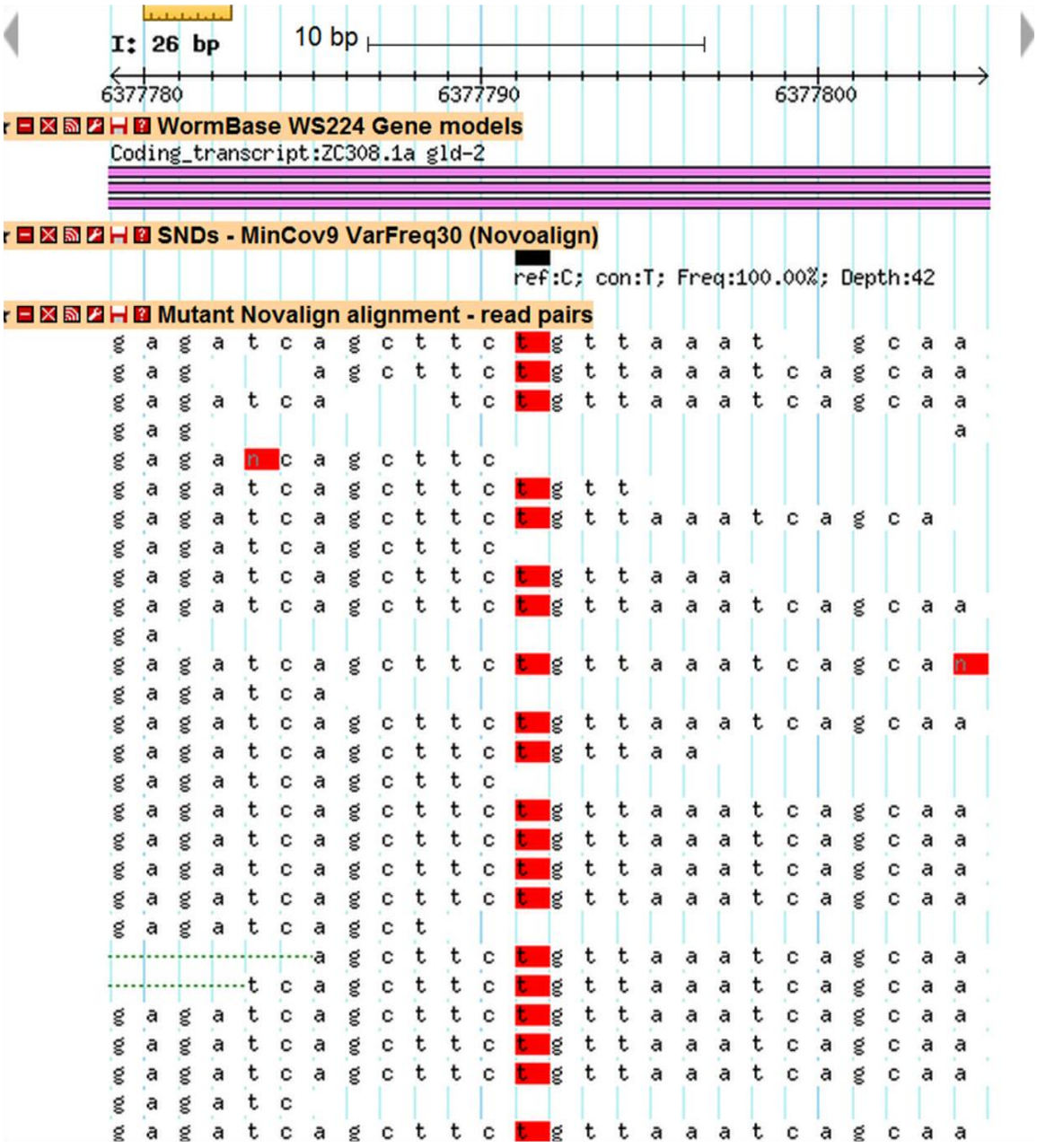


Figure 2.12. Only Candidate in Genomic Region of Interest: SND detected in gld-2 by Variant Detection Pipeline

Region displayed: I:6,377,779..6,377,804. The SND features a C>T substitution at 100% variant frequency and read depth of 42, and is not found in other wild type strains (Hobert and Horvitz). The missense mutation causes a P to L amino acid change. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Note: not all of the sequencing reads are shown in the Figure. Abbreviations: ref (reference base), con (consensus base), Freq (variant frequency) and Depth (read depth).

Table 2.5. List of dyf-10(e1383) candidates on Chromosome I

Pos	Change	VarFreq(%)	Depth	Trans	Gene	Type	Notes[a]
719554	G>A	97.25	109	Y18H1A.7	*n/a*	NON	HOM: neurofilament heavy polypeptide
2166094	G>A	100.00	31	W03F11.6	*afd-1*	RMIS	HOM: AFaDin (actin filament binding protein), GO: signal transduction
5344075	G>A	100.00	42	F55A12.7	*apm-1*	MIS	adaptin, Clathrin adaptor, EXP: ubiquitous but stronger in neurons
5432433	C>T	100.00	18	F27C1.8	*dpy-5*	MIS	PC
5432448	C>A	100.00	19	F27C1.8	*dpy-5*	NON	PC
6377791	C>T	100.00	42	ZC308.1	*gld-2*	MIS	catalytic subunit of PAP (poly(A) polymerase), GO: germline cell cycle switching and apoptosis
7588318	C>T	100.00	24	D2030.6	*prg-1*	RMIS	Argonaut/Piwi proteins, GO: embryonic development
8338997	C>T	100.00	55	C34B7.2	*n/a*	MIS	HOM: Polyphosphoinositide phosphatase, EXP: GABAergic motor neurons, head & tail neurons
8356761	C>T	100.00	40	F16D3.1	*tba-5*	MIS	alpha tubulin, GO: microtubule-based process
8659698	C>T	100.00	56	F39H2.1	*flp-22*	RMIS	FMRF-like peptide, GO: neuropeptide signaling pathway, EXP: head neurons
9286524	C>T	100.00	36	F14B4.3	*n/a*	MIS	HOM: DNA-directed RNA polymerase, GO: embryonic development
9365535	C>T	100.00	26	F16A11.3	*ppfr-1*	MIS	Protein Phosphatase 4 Regulatory subunit, GO: hatching
9642787	G>A	100.00	30	T05F1.8	*n/a*	MIS	HOM: mitochondrial phosphate carrier, GO: determination of adult life span

9891269	G>A	100.00	44	F52F12.1	*oct-1*	MIS	Organic cation transporter, GO: transmembrane transport, EXP: head and mechanosensory neurons
10270192	A>C	78.57	14	ZC247.1	*n/a*	MIS	GO: apoptosis, reproduction
13542278	C>T	100.00	94	Y87G2A.3	*atg-4.1*	MIS	HOM: cystein protease, EXP: Dorsal nerve cord, head neurons
14306866	A>T	95.00	20	F49B2.7	*n/a*	RMIS	HOM: FBXB-85, F-box associated domain

Note: All candidate genes listed were interrupted by homozygous, non-synonymous SNDs that do not occur in WT strains. Homozygous SNDs is defined as having a variant frequency $\geq$ 75%. Abbreviations: MIS (Missense), NON (Nonsense), RMIS (Radical Missense), PC (Positive Control), GO (Gene Ontology), HOM (Homologous), EXP (Expression pattern).

a Information collected from WormBase (http://www.wormbase.org)

From Table 2.5, the only obvious candidate that is closest to the region of interest is *tba-5*, which codes for an alpha tubulin that can be incorporated into the cilia axoneme. *tba-5* features a missense mutation of a C>T change which is characteristic of EMS mutagenesis. The missense mutation also has a VarFreq of 100%, read depth of 40, and does not occur in the Hobert and Horvitz WT strains (Figure 2.13). The missense mutation is approximately 900,000 bp from the right of *unc-13* which is located at the right breakpoint of the predicted genomic location.

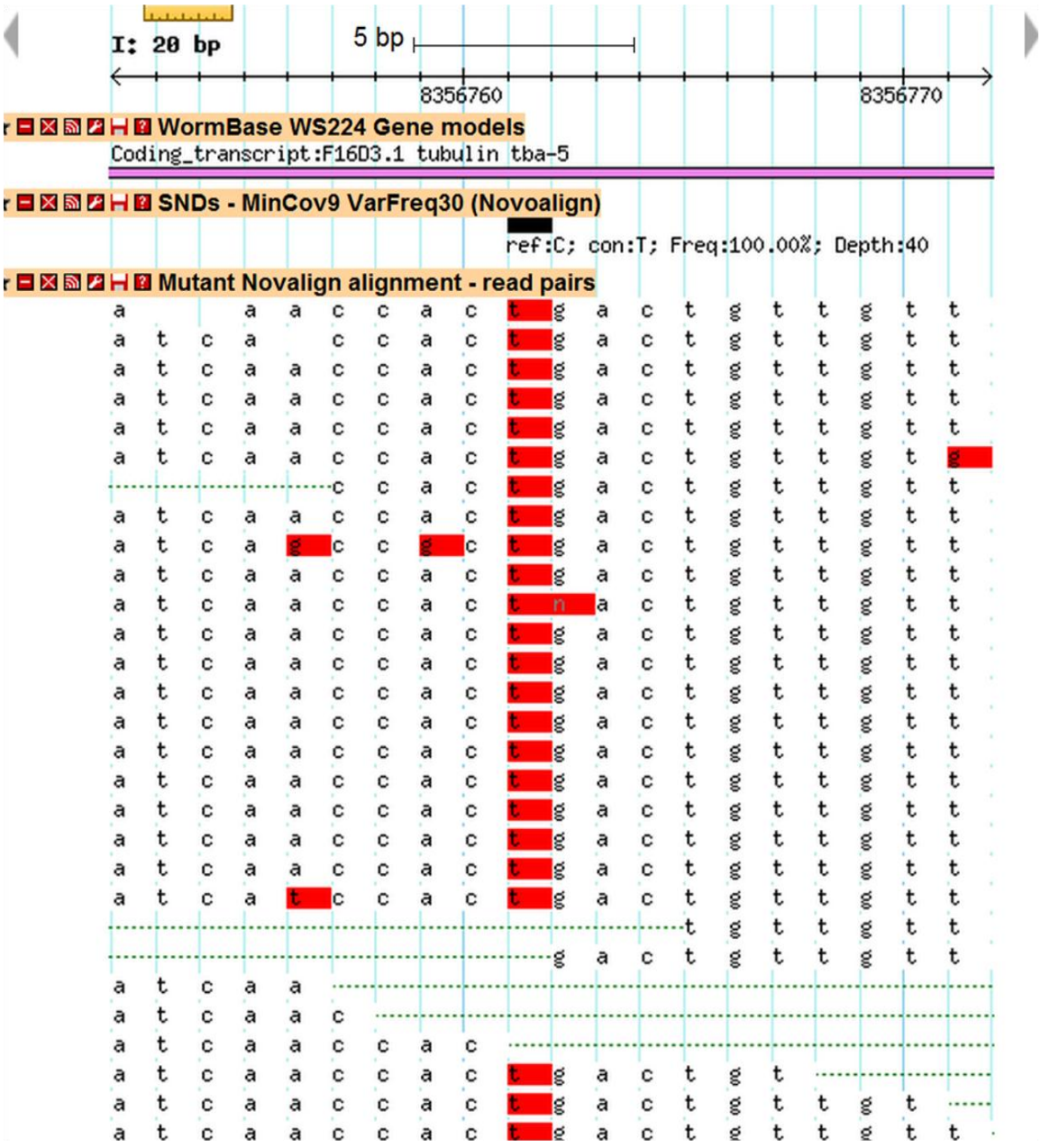


Figure 2.13. SND detected in tba-5 by the variant detection pipeline

Region shown: I:8,356,752..8,356,771. The SND features a C to T substitution at 100% variant frequency and read depth of 40. The SND is not found in the wild type strains (Hobert and Horvitz strains). The missense mutation causes an S to L amino acid change. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Note: not all the reads are displayed in the figure. Abbreviations: ref (reference base), con (consensus base), Freq (variant frequency) and Depth (read depth).

To provide support and evidence that *dyf-10* is *tba-5*, three-way complementation tests between *qj14, e1383* and *tm4200* were completed (Figure 2.14, Table 2.6 and Figure 2.15). First, the WT N2 positive control had 99.7% (N=318) of the worms dye filling with the majority of the worms dye filled in both the amphid and phasmid neurons, as expected. On the GFP compound microscope, the amphid and phasmid neurons can clearly be seen dye filling after DiO treatment in N2 worms (Figure 2.15A). Second, the *e1383* negative control was successful, 0.0% (N=334) of the worms dye filled. Third, the other *tba-5* alleles were dye filled to make sure the strains were behaving as expected. As was previously determined, *tm4200 4X* worms displayed a WT phenotype where 88.2% (N=34) were dye filling; this is comparable to the N2 positive control. All of the *tm4200 4X* worms dye filled in both the amphid and phasmid neurons. *qj14,* on the other hand, did not display a clear Dyf phenotype, although there was a clear decrease in the percentage of worms dye filling, 38.3% (N=141) of the worms dye filled, and most of them dye filled in the phasmid neurons only. Hao *et al.* previously determined that DiI staining of *qj14* worms was temperature sensitive. At 15°C, they observed 0% of the worms dye filling. At 20°C they observed less than 10% of the worms dye filling in the amphid and phasmid neurons. At 25°C they observed a dramatic increase in the percentage of *qj14* worms dye filling; 50% and 20% amphids and phasmids were dye filled, respectively (Hao *et al.*, 2011). Since the worms in this project was grown at room temperature, which is approximately 20°C, it is reasonable to observe 38.3% of the worms dye filling. Furthermore, the *qj14* worms only dye filled in a small portion of the sensory neurons, namely the two pairs of phasmid neurons. In strains that display a Dyf phenotype, when you compare the number worms that only dye fill in phasmid neurons to the number of worms that only dye fill in amphids, there are more worms in the former category (Table 2.6). This suggests that the phasmid neurons are more easily dye filled with DiO than the amphid neurons. Perhaps TBA-5 does not play a heavy role in cilia formation in the phasmids as compared to the amphids, thus mutations in TBA-5 does not affect the phasmid cilia formation as drastically. As a result, the number of worms dye filling in both amphid and phasmid neurons may be a better indicator of whether the strain displays a Dyf or WT phenotype. If we only consider the number of worms that dye fill in both the amphid and phasmid neurons, *qj14* clearly displays a Dyf phenotype. In the complementation tests, *dyf-10(e1383)/tba-5(qj14)* was observed to have 10.0% (N=201) of the worms dye filling, with the majority dye filling in the phasmid neurons

only. Overall, this is a large decrease of worms that dye filled when compared to the N2 positive control. Furthermore, only one worm was observed to dye filled in both the amphid and phasmid neurons. Hence, the strain *dyf-10(e1383)/tba-5(qj14)* displays a Dyf phenotype, which means that *dyf-10(e1383)* fails to complement *tba-5(qj14)* and are therefore alleles of the same gene; Hao and colleagues also reached the same conclusion (Hao *et al.,* 2011). *tm4200* and *qj14* are alleles of the same gene. *tba-5(tm4200)/tba-5(qj14)* mutants have 49.1% (N=112) of the worms dye filling. Only 12.5% of the worms dye filled in both the amphid and phasmid neurons, with the majority dye filling in only the phasmid neurons. Again, there is a clear difference in the number of worms dye filling in both the amphid and phasmid neurons when compared to the N2 positive control. As a result, *tba-5(tm4200)* also fails to complement *tba-5(qj14)*, which is expected since they are alleles of the same gene. *dyf-10(e1383)/tba-5(tm4200)* worms resulted in 15.6% (N=289) of the worms dye filling. Only 2.4% of the *dyf-10(e1383)/tba-5(tm4200)* worms dye filled in both the amphid and phasmid neurons. Again, the strain displays a Dyf phenotype where *dyf-10(e1383)* fails to complement *tba-5(tm4200)*, which is similar to *tm4200/qj14* mutants. This further proves that *dyf-10* is the same gene as *tba-5*.

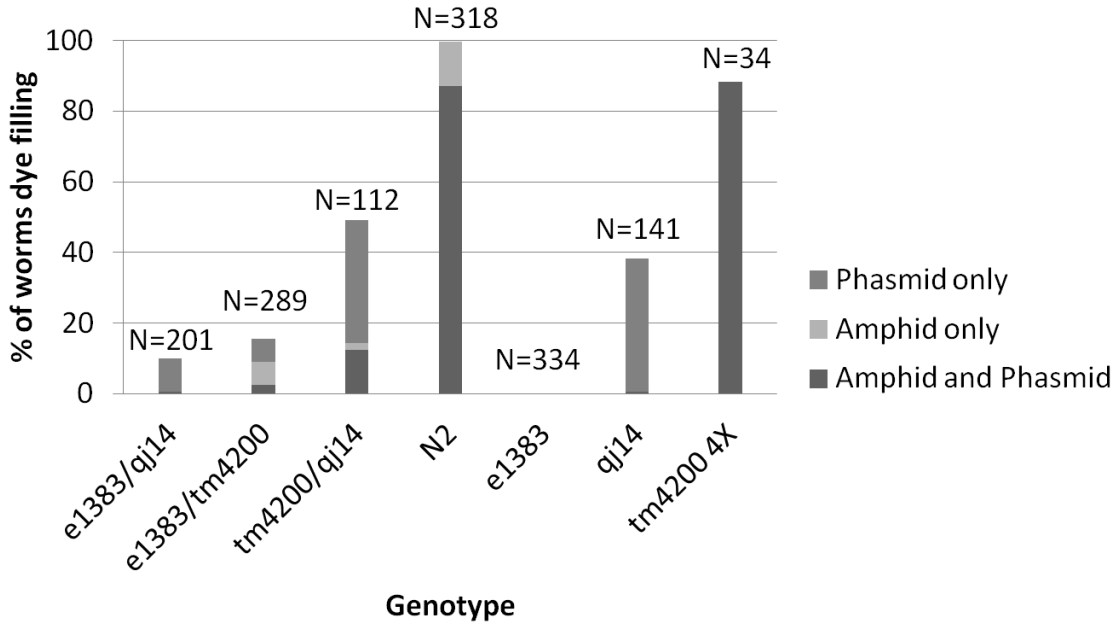


Figure 2.14. Percentage of Worms Dye Filling in Complementation Tests

Positive control, N2. Negative control, *dyf-10(e1383)*. *tm4200* worms have been outcrossed four times (4X). The percentage of worms that dye filled in both amphid and phasmid neurons, amphid neurons only and phasmid neurons only is displayed. Worms were dye filled using DiO and visualized using fluorescent compound microscopy. Abbreviations: N (Total number of worms)

Table 2.6. Complementation Tests: DiO Dye Filling Results

Genotype	Total[c]	Total Dye filling[d]		Amphid and Phasmid		Amphid Only (%)		Phasmid Only (%)		None (%)	
	N	N	%[e]	N	%[e]	N	%[e]	N	%[e]	N	%[e]
e1383[a] *qj14*	201	20	10.0	1	0.5	0	0.0	19	9.5	181	90.0
e1383[a] *tm4200*	289	45	15.6	7	2.4	19	6.6	19	6.6	244	84.4
tm4200[b] *qj14*	112	55	49.1	14	12.5	2	1.8	39	34.8	57	50.9
N2	318	317	99.7	277	87.1	40	12.6	0	0.0	1	0.3
e1383	334	0	0.0	0	0.0	0	0.0	0	0.0	334	100.0
qj14	141	54	38.3	1	0.7	0	0.0	53	37.6	87	61.7
tm4200 4X	34	30	88.2	30	88.2	0	0.0	0	0.0	4	11.8

Note: Positive control, N2. Negative control: *e1383*. Abbreviations: N (total number of worms), % (percentage of worms).

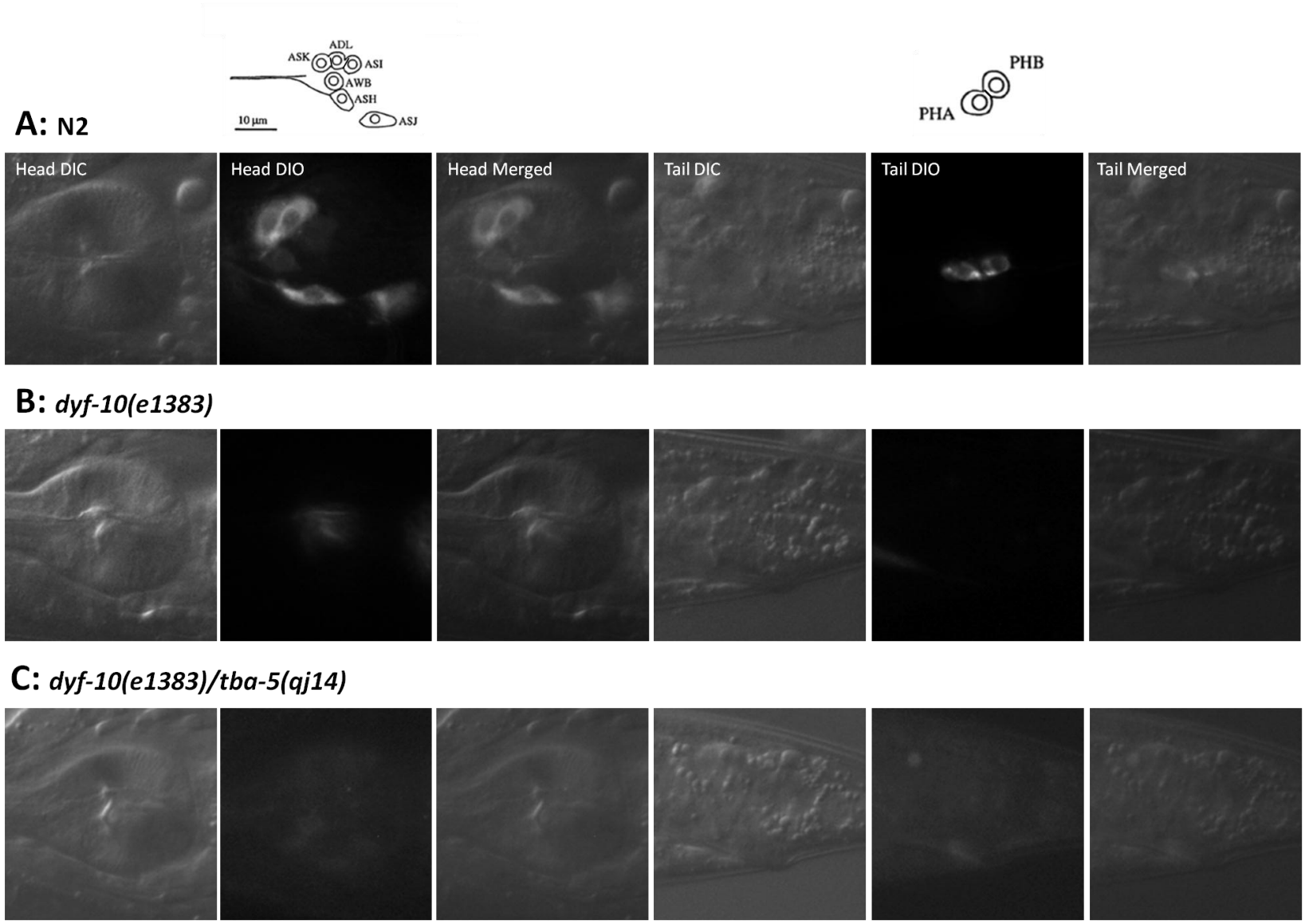
a Crosses completed using *dyf-10(e1383)* males

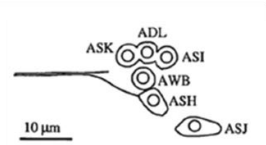
b Crosses completed using *tba-5(tm4200)* males

c Total number of worms in the assay.

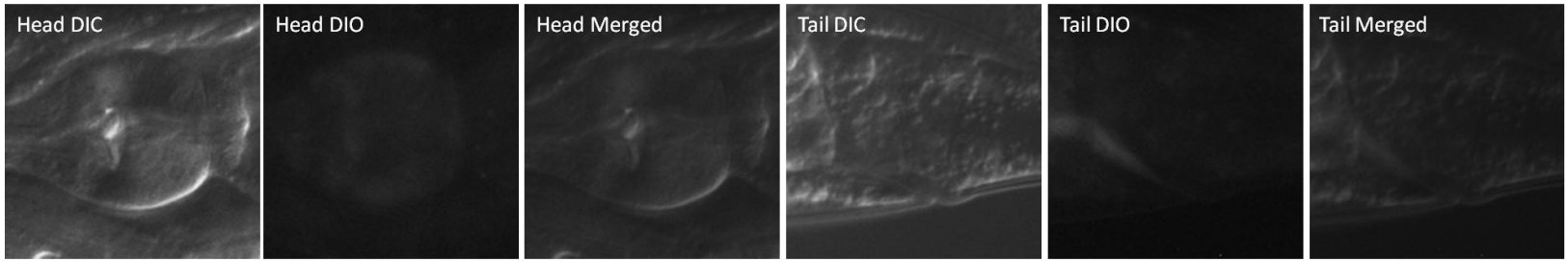
d Total number of worms that dye filled in amphid and phasmid together, amphid only and phasmid only.

e Percentage based on the total number of worms.

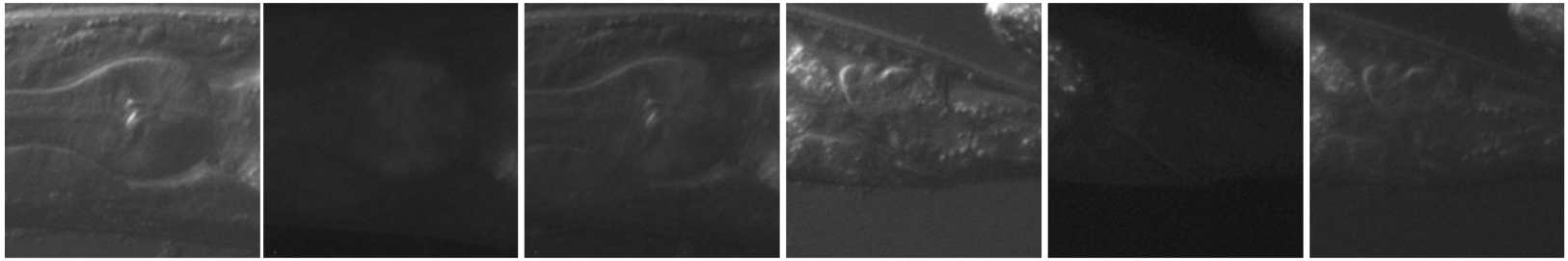

A: N2
ADL
ASK
ASI
AWB
ASH
ASJ
10 µm
PHB
PHA
Head DIC
Head DIO
Head Merged
Tail DIC
Tail DIO
Tail Merged
B: dyf-10(e1383)
C: dyf-10(e1383)/tba-5(qj14)



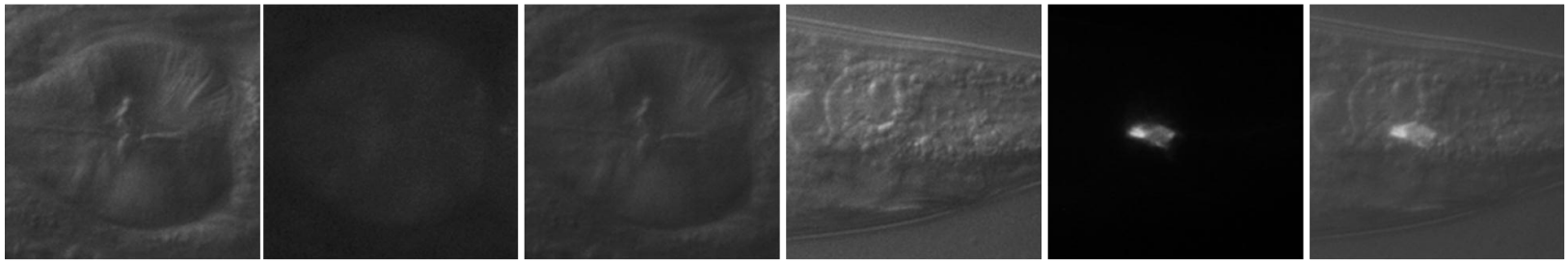
D: ***dyf-10(e1383)/tba-5(tm4200)***

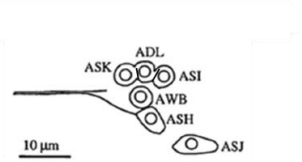


E: ***tba-5(tm4200)/tba-5(qj14)***

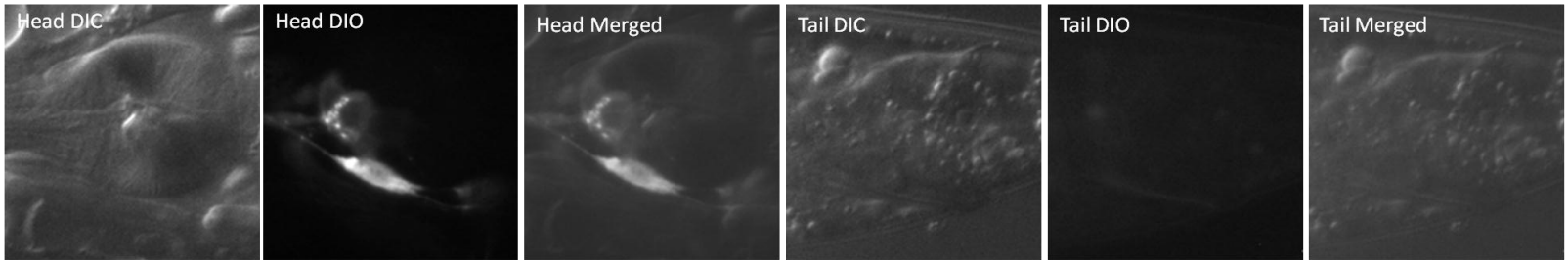


F: ***dyf-10(e1383)/tba-5(qj14)***

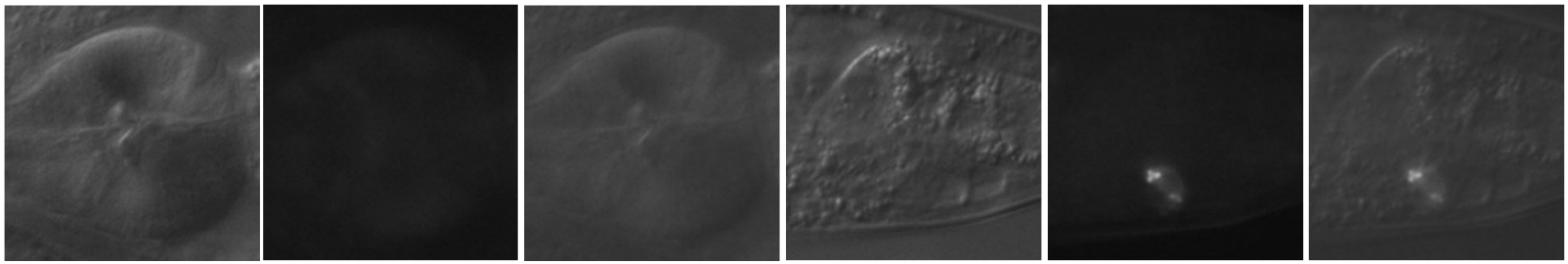




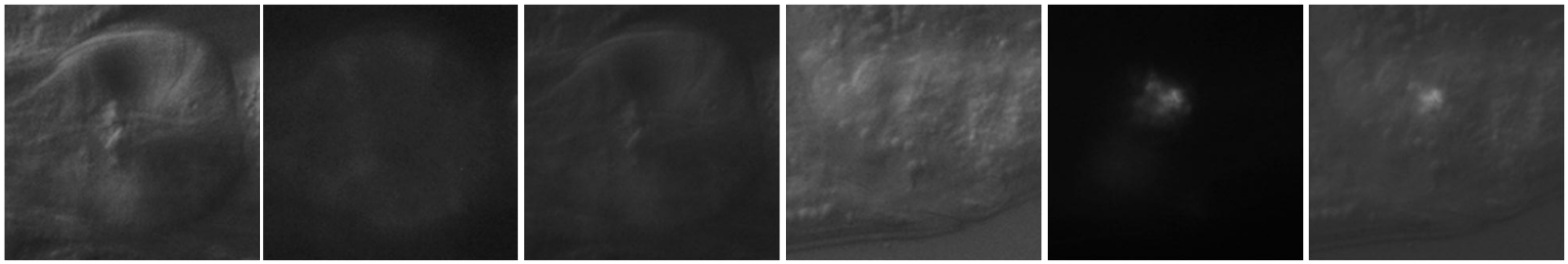
G: ***dyf-10(e1383)/tba-5(tm4200)***

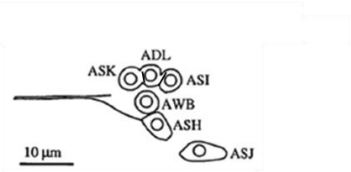


H: ***dyf-10(e1383)/tba-5(tm4200)***

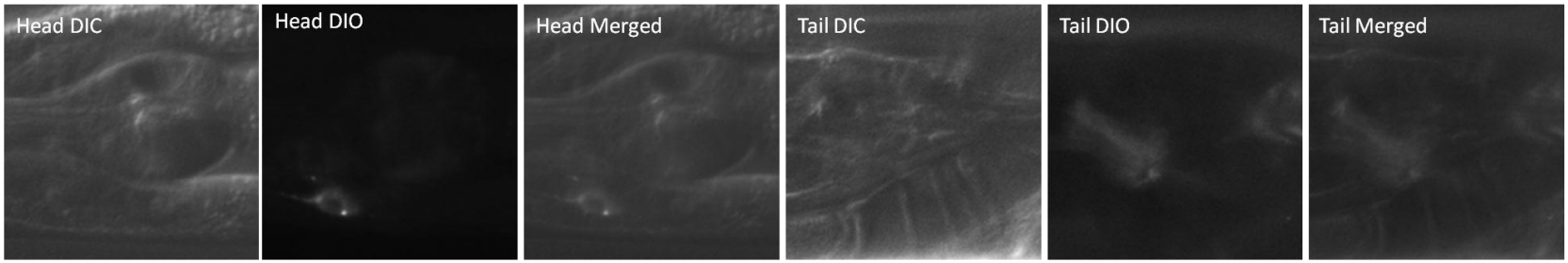


I: ***tba-5(tm4200)/tba-5(qj14)***





J: ***tba-5(tm4200)/tba-5(qj14)***



K: ***tba-5(tm4200)/tba-5(qj14)***

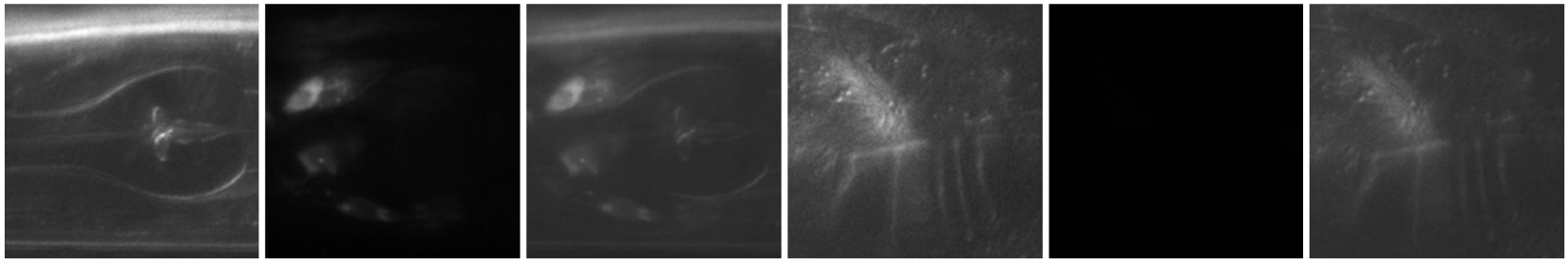
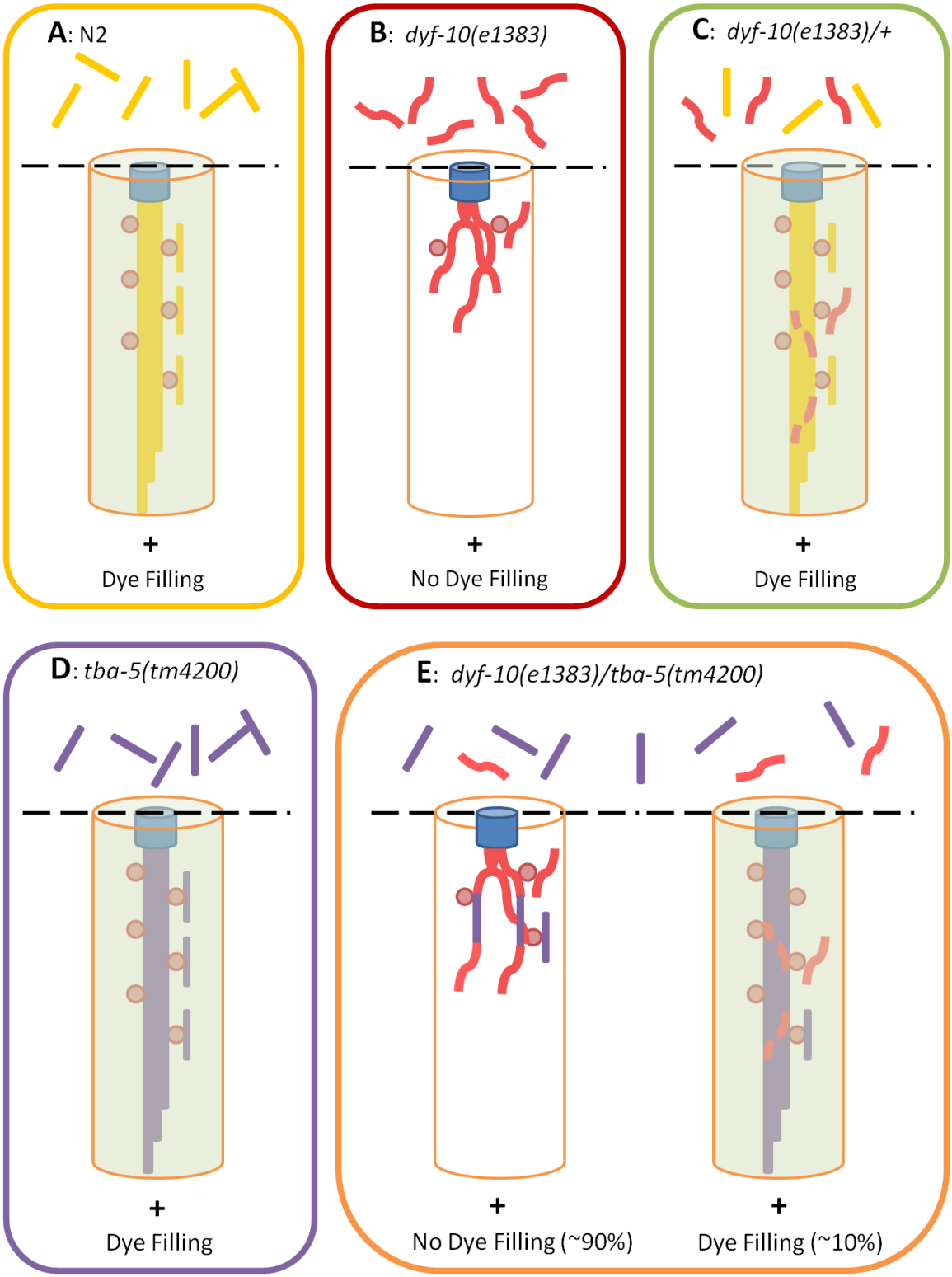


Figure 2.15. Complementation Tests: DiO Dye Filling Images of Amphid and Phasmid Neurons

(A) Positive Control, N2. (B) Negative control, *dyf-10(e1383)*. (C) *dyf-10(e1383)/tba-5(qj14)*, no dye filling. (D) *dyf-10(e1383)/tba-5(tm4200),* no dye filling. (E) *tba-5(tm4200)/tba-5(qj14)*, no dye filling. (F) *dyf-10(e1383)/tba-5(qj14)*, dye filling of phasmid neurons only. (G, H) *dyf-10(e1383)/tba-5(tm4200)*, dye filling of amphid and phasmid neurons, respectively. (I) *tba-5(tm4200)/tba-5(qj14)*, dye filling of phasmid neurons only. (J, K) *tba-5(tm4200)/tba-5(qj14)* males, dye filling in amphid and phasmid or amphid neurons only, respectively. DiO dye filling was visualized using fluorescent compound microscopy. Schematic diagram of amphid and phasmid neuronal bodies modified from Starich *et al.*,1995. Abbreviations: DIC (differential interference contrast), DiO (lipophillic dye).

Although *dyf-10(e1383)/tba-5(tm4200)* and *tba-5(tm4200)/tba-5(qj14)* worms display a Dyf phenotype, the percentage of worms dye filling in both the amphid and phasmid neurons is higher than *dyf-10(e1383)/tba-5(qj14)* mutants. *dyf-10(e1383)/tba-5(tm4200)* worms were created by crossing *dyf-10(e1383)* males with *tba-5(tm4200)* hermaphrodites. Hence, the dye filling may be explained by incomplete fertilization by the male, where some of the progeny is self fertilized and therefore has the genotype *tba-5(tm4200)* that can dye fill. However, this should not be the case for *tba-5(tm4200)/tba-5(qj14)* worms. *tba-5(tm4200)* males were crossed with *tba-5(qj14)* hermaphrodites, self fertilized progeny should have the genotype *tba-5(qj14)* that result in a Dyf phenotype. As further proof that the small percentage of dye filling in *dyf-10(e1383)/tba-5(tm4200)* and *tba-5(tm4200)/tba-5(qj14)* worms were not a result self fertilization, there were examples were male *tba-5(tm4200)/tba-5(qj14)* progeny that showed partial or full dye filling in their amphid neurons (Figure 2.15J and K). It may be argued that the temperature sensitivity of *qj14* mutants can explain the higher percentage of worms dye filling in *tba-5(tm4200)/tba-5(qj14)* worms, but *dyf-10(e1383)/tba-5(qj14)* mutants seem to only be affected minimally by the temperature sensitivity of the *qj14* allele. To explain this phenomenon, a simplistic model of TBA-5 in cilia structure was constructed. In N2 worms there are two copies of WT *tba-5*, resulting in the translation of WT TBA-5 that gets incorporated into the cilia by the IFT machinery and result in the formation of sensory cilia that dye fills (Figure 2.16A, Figure 2.14). In *dyf-10(e1383)*, both copies of *tba-5* are mutated (Figure 2.16B). The mutated TBA-5 gets incorporated into the cilia structure by the IFT machinery, resulting in defective cilia that does not dye fill (Figure 2.14). In *dyf-10(e1383)/+*, the worm produces one WT copy and one mutated copy of TBA-5 (Figure 2.16C). Assuming that the IFT machinery has a

higher binding affinity to the WT TBA-5 than the mutated TBA-5, the genotype produces normal-like cilia that can still dye fill (Figure 2.19). In *tba-5(tm4200)*, both copies of full length *tba-5* are missing (Figure 2.16D). It is hypothesized that tubulin paralogs can become incorporated into the cilia structure by IFT machinery to generate normal-like cilia that can dye fill (Figure 2.14). In *dyf-10(e1383)/tba-5(tm4200)* mutants, there is one mutated copy of *tba-5* and one missing copy of full length *tba-5* (Figure 2.16E). In this scenario, assuming the IFT machinery has a much higher binding affinity to the mutated TBA-5 than the TBA-5 paralogs, approximately 90% of the time, determined by the percentage of worms that dye filled in both the amphid and phasmid neurons and amphid neurons only, the cilia will results in a defective cilia that does not dye fill; the IFT machinery incorporated too many of the mutated TBA-5 compared to TBA-5 paralogs However, since the gene dosage of the mutated TBA-5 is halved approximately 10% of the time the IFT machinery will incorporate enough TBA-5 paralogs to result in a normal-like cilia that can dye fill (Figure 2.14). So this leads to the question of the identity of the TBA-5 paralog that can compensate for the absence of TBA-5.


A: N2
+
Dye Filling
B: dyf-10(e1383)
+
No Dye Filling
C: dyf-10(e1383)/+
+
Dye Filling
D: tba-5(tm4200)
+
Dye Filling
E: dyf-10(e1383)/tba-5(tm4200)
+
No Dye Filling (~90%)
+
Dye Filling (~10%)

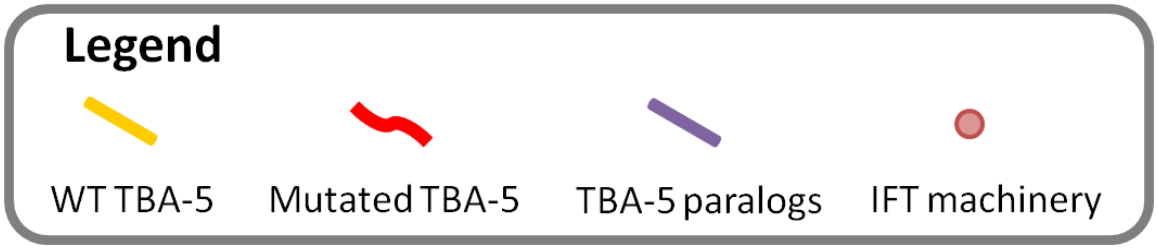


Figure 2.16. Model of TBA-5 in cilia structure

(A) N2 worms have two wild type (WT) copy of TBA-5, which results in normal cilia and dye filling. (B) *dyf-10(e1383)* worms have two mutated copy of TBA-5, which results in defective cilia and no dye filling. (C) *dyf-10(e1383)/+* worms have one mutated and one WT copy of TBA-5. IFT machinery has higher affinity to WT TBA-5 than mutated TBA-5, which results in normal cilia and dye filling. (D) *tba-5 (tm4200)* worms do not produce TBA-5. IFT machinery is able to incorporate TBA-5 paralogs into cilia structure, which results in normal-like cilia and dye filling. IFT machinery can bind to TBA-5 paralogs at a lower binding affinity than WT and mutated TBA-5. (E) *dyf-10(e1383)/tba-5(tm4200)* worms produce a lower dosage of mutated TBA-5 and no WT TBA-5. Lower dosage of mutated TBA-5 competes with a relatively higher level of TBA-5 paralogs for IFT binding. IFT machinery will have a higher binding affinity to mutated TBA-5 than TBA-5 paralogs. As a result, approximately 90% of cilia will incorporate mutated TBA-5 resulting in defective cilia and no dye filling, and approximately 10% of cilia will incorporate TBA-5 paralogs resulting in normal-like cilia and dye filling.

It is hypothesized that TBA-9 can alone compensate for the lack of TBA-5. To test this hypothesis a double knockout mutant, t*ba-5(tm4200)I;tba-9(ok1858)X,* was constructed and dye filled. Before the construction of the double mutant all *tba-5(tm4200)* and *tba-9(ok1858)* strains before and after 4X outcrossing into the N2 background were dye filled and were found to behave as expected. *tba-5(tm4200)* and *tba-9(ok1858)* worms before and after 4X outcrossing successfully dye fill (Figure 2.17, Table 2.7 and Figure 2.18). For *tm4200 0X* worms 98.8% (N=82) dye filled, for *tm4200 4X* worms 88.2% (N=34) dye filled, for *ok1858 0X* worms 100% (N=40) dye filled, and for *ok1858 4X* worms 100% (N=96) dye filled. Figure 2.18 panels D to G shows all the starting strains dye filling in amphid and phasmid neurons. The positive control, N2, resulted in 100.00% (N=317) of the worms dye filling and the negative control, *dyf-10(e1383)*, resulted in 0.4% (N=228) of the worms dye filling, as expected (Figure 2.17 and Table 2.7). Dye filling of the double knockout mutant, however, also resulted in a WT phenotype. 97.8% (N=178) of the worms dye filled with the majority dye filling in both the amphid and phasmid neurons (Figure 2.17, Table 2.7, Figure 2.18C). As a result, the hypothesis failed. Even without TBA-9 and TBA-5, there is cilia formation in the double mutants. Perhaps another α-tubulin or a combination of different α-tubulins

are responsible for the compensation of the absence of *tba-5*. It is also interesting to note that 28.7% of the *tba-5(tm4200);tba-9(ok1858)* mutants only dye filled in the amphid neurons. Perhaps mutating *tba-9* has an effect on the formation of the cilia in the phasmid neurons.

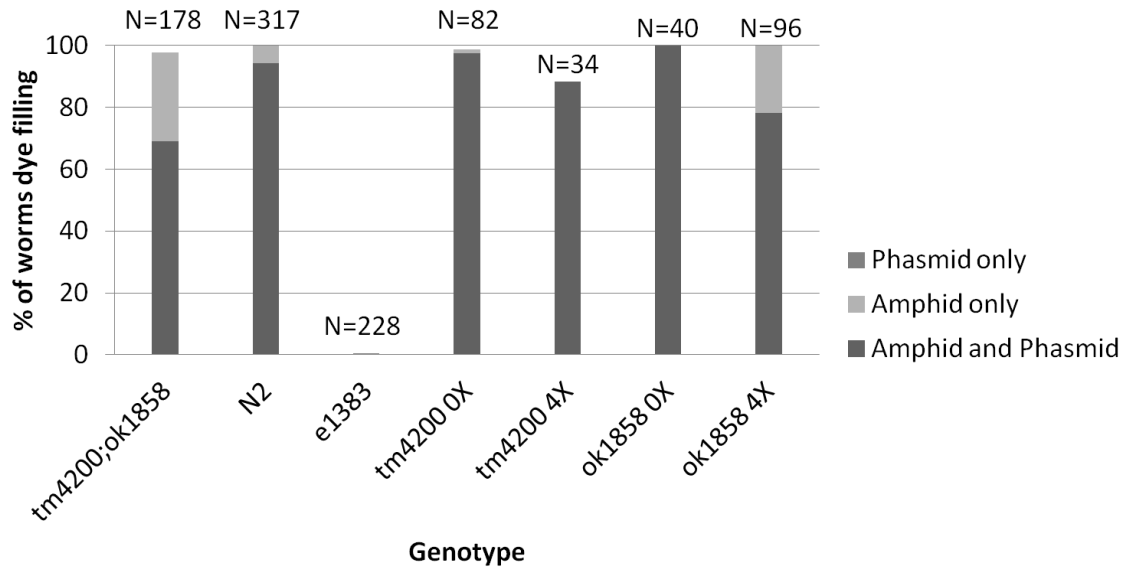


Figure 2.17. Percentage of Worms Dye Filling in Double Knockout Mutant

Positive control, N2. Negative control, *dyf-10(e1383)*. *tm4200;ok1858* strain was generated with *tm4200 4X* and *ok1858 4X* worms which have been outcrossed with N2 worms four times. Dye filling of double mutant shows dye filling comparable to N2 worms. Worms were dye filled using DiO and visualized using fluorescent compound microscopy.

Table 2.7. Double Knockout Mutant: DiO Dye Filling Results

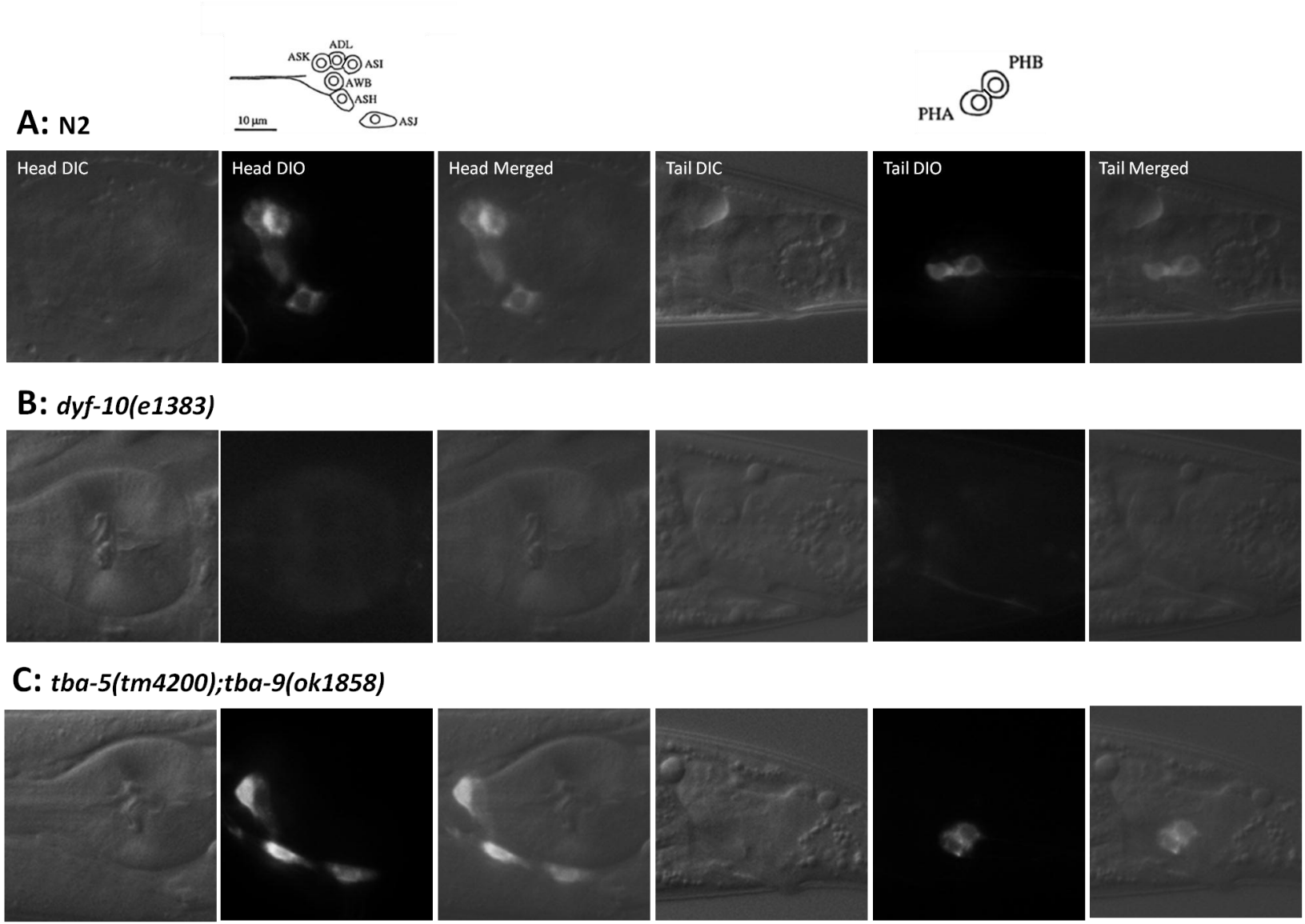
Genotype[c]	Total[a]	Total Dye filling[b]		Amphid and Phasmid		Amphid Only (%)		Phasmid Only (%)		None (%)	
	N	N	%	N	%	N	%	N	%	N	%
tm4200; ok1858	178	174	97.8	123	69.1	51	28.7	0	0.0	4	2.2
N2	317	317	100.0	299	94.3	18	5.7	0	0.0	0	0.0
e1383	228	1	0.4	0	0.0	0	0.0	1	0.4	227	99.6
tm4200 0X	82	81	98.8	80	97.6	1	1.2	0	0.0	1	1.2
tm4200 4X	34	30	88.2	30	88.2	0	0.0	0	0.0	4	11.8
ok1858 0X	40	40	100.0	40	100.0	0	0.0	0	0.0	0	0.0
ok1858 4X	96	96	100.0	75	78.1	21	21.9	0	0.0	0	0.0

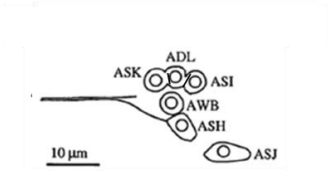
Note: Positive control, N2. Negative control: *e1383*. Abbreviations: N (total number of worms), % (percentage of worms).

a Total number of worms in the assay.

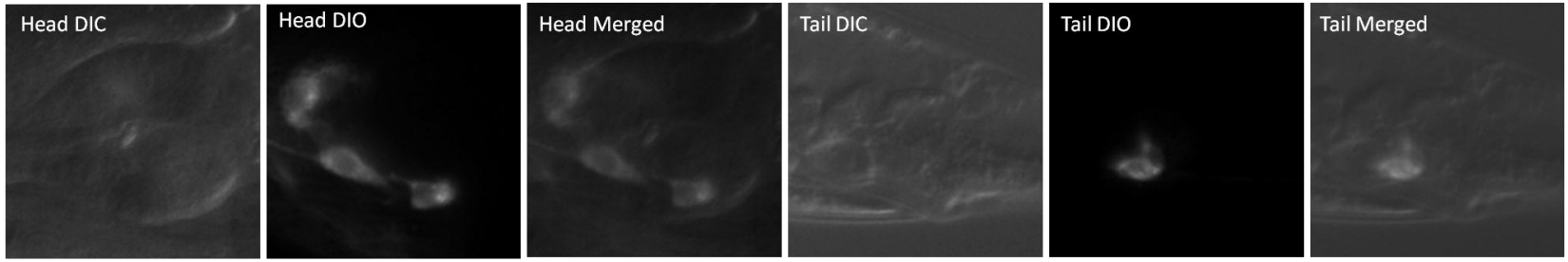
b Total number of worms that dye filled in amphid and phasmid together, amphid only and phasmid only

c The genotypes shown are homozygous

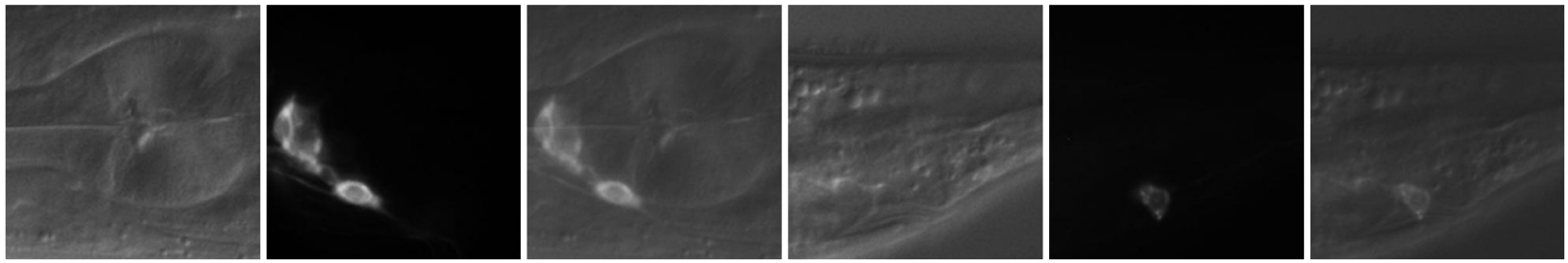

A: N2
ADL
ASK
ASI
AWB
ASH
ASJ
10 µm
PHB
PHA
Head DIC
Head DIO
Head Merged
Tail DIC
Tail DIO
Tail Merged
B: dyf-10(e1383)
C: tba-5(tm4200);tba-9(ok1858)

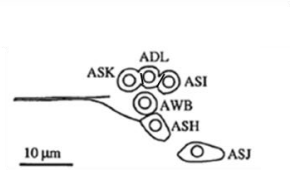


D: ***tba-5(tm4200) 0X outcrossed***

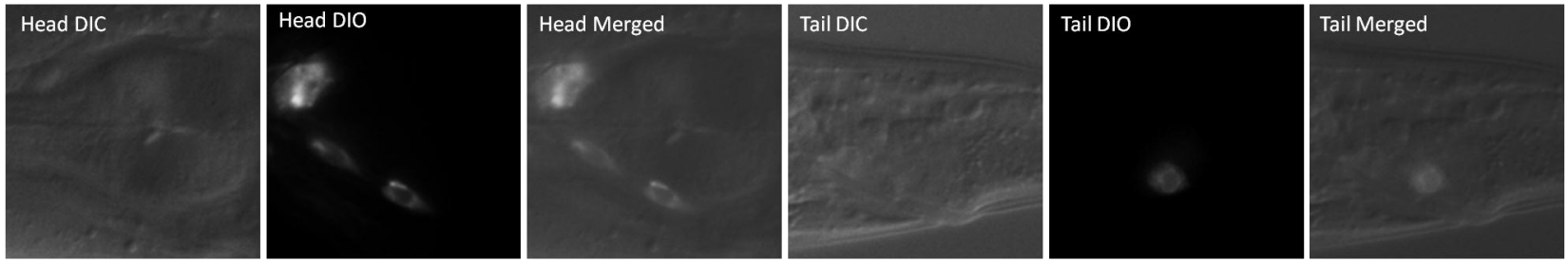


E: ***tba-5(tm4200) 4X outcrossed***





F: ***tba-9(ok1858) 0X outcrossed***



G: ***tba-9(ok1858) 4X outcrossed***

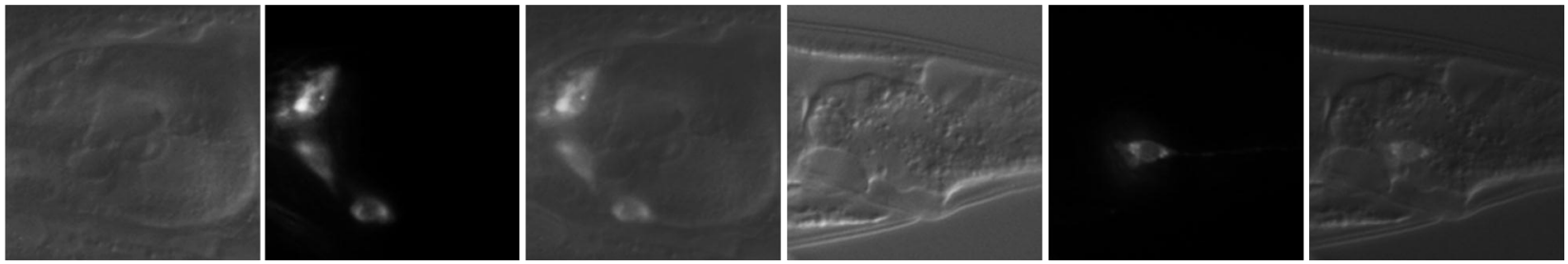


Figure 2.18. Double Knockout mutant: DiO Dye Filling Images of Amphid and Phasmid Neurons

(A) Positive Control, N2. (B) Negative control, *dyf-10(e1383)*. (C) *tba-5(tm4200);tba-9(ok1858)*, dye filling. (D) *tba-5(tm4200) 0X*, dye filling. (E) t*ba-5(tm4200) 4X*, dye filling. (F) *tba-9(ok1858) 0X, dye filling*. (G) *tba-9(ok1858) 4X, dye filling*. DiO dye filling was visualized using fluorescent compound microscopy. Schematic diagram of amphid and phasmid neuronal bodies modified from Starich *et al.*,1995. Abbreviations: DIC (differential interference contrast), DiO (lipophillic dye).

Hao and colleagues suggested that *dyf-10(e1383)* is a recessive gain of function mutant (Hao *et al.,* 2011). To test this hypothesis, the following strains were dye filled: *dyf-10(e1383)*, *dyf-10(e1383)/+, dyf-10(e1383)/tba-5(tm4200)* and *dyf-10(e1383)/nDf25* (Figure 2.19, Table 2.8, Figure 2.20). As expected of *dyf-10(e1383)* worms that have two mutated copies of *tba*-5, 0.2% (N=643) of the worms dye filled. The N2 positive control was also successful, 99.9% (N=684) of the worms dye filled, with 91.4% of the worms dye filling in both the amphid and phasmid neurons. Dye filling of *dyf-10(e1383)/+* where there is one mutated copy of *tba-5* and one WT copy of *tba-5* resulted in a WT phenotype with 100% (N=88) of worms dye filling in both the amphid and phasmid neurons. This data clearly indicates that *dyf-10(e1383)* is a recessive mutant where one WT copy of *tba-5* in the genome is sufficient for normal cilia formation. Dye filling of *e1383/tm4200* and *e1383/nDf25* mutants, however, seems to suggest that the *tm4200* still produces partially functional TBA-5 products. *dyf-10(e1383)/tba-5(tm4200)* worms resulted in 15.6% (N=289) of the worms dye filling, whereas *dyf-10(e1383)/nDf25* worms resulted in 0% (N=66) of the worms dye filling. *tm4200* is an inframe deletion, it is hypothesized that there is a slightly functional truncated copy of *tba-5* in *tm4200* mutants. Since *e1383/tm4200* mutants do not display the strong WT phenotype as the *e1383/+* mutants, the truncated form of *tba-5* in the presence of a mutated copy of TBA-5 must not be fully but slightly functional. Furthermore; since *tm4200 4X* worms resulted in 88.2% of the worms dye filling in the amphid and phasmid neurons (Figure 2.14 and Table 2.6) and *e1383/tm4200* worms resulted in 15.6% of the worms dye filling, it is hypothesized that the *e1383* mutation further dampens the function of the truncated *tba-5.* Finally, if we only consider *nDf25* as the null allele of *tba-5*, the data seems to suggest that *e1383* is a loss-of-function allele: *e1383* and *e1383/nDf25* worms share the same dye filling result.

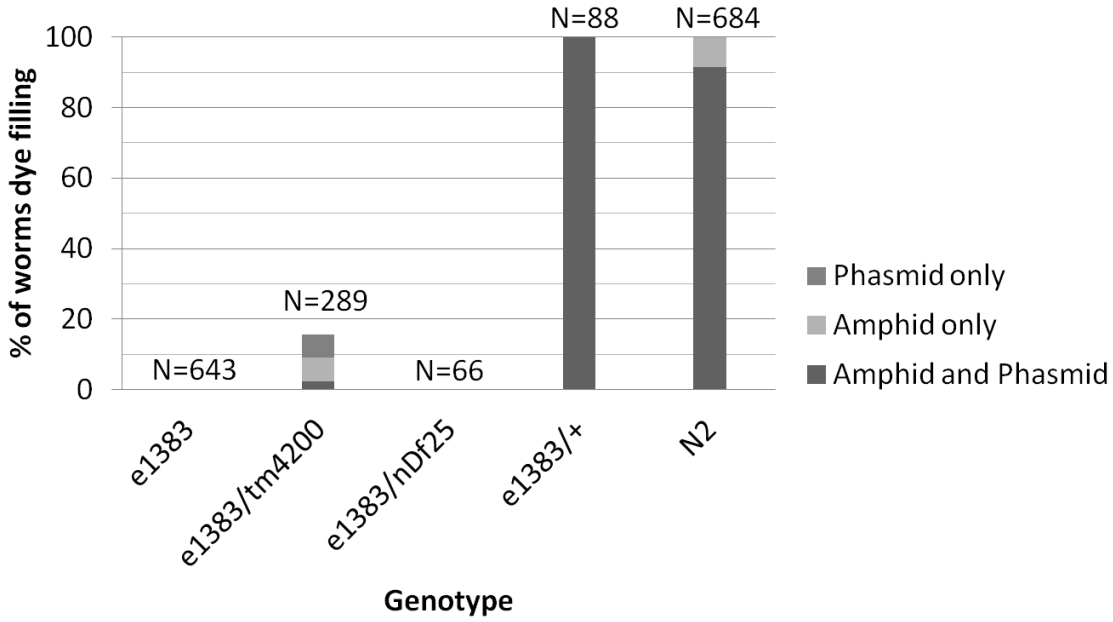


Figure 2.19. Recessive Gain-of-function Mutant: Percentage of Worms dye filling

Positive control, N2. Negative control, *dyf-10(e1383)*. *dpy-5(e61) dyf-10(e1383)* worms were crossed to N2 males to generate *e1383/+* worms. *e1383, e1383/tm4200* and *e1383/nDf25* worms show little to no worms dye filling. *e1383/+* worms show dye filling comparable to N2 worms. Worms were dye filled using DiO and visualized using fluorescent compound microscopy.

Table 2.8. Recessive Gain-of-function Mutant: DiO Dye Filling Results

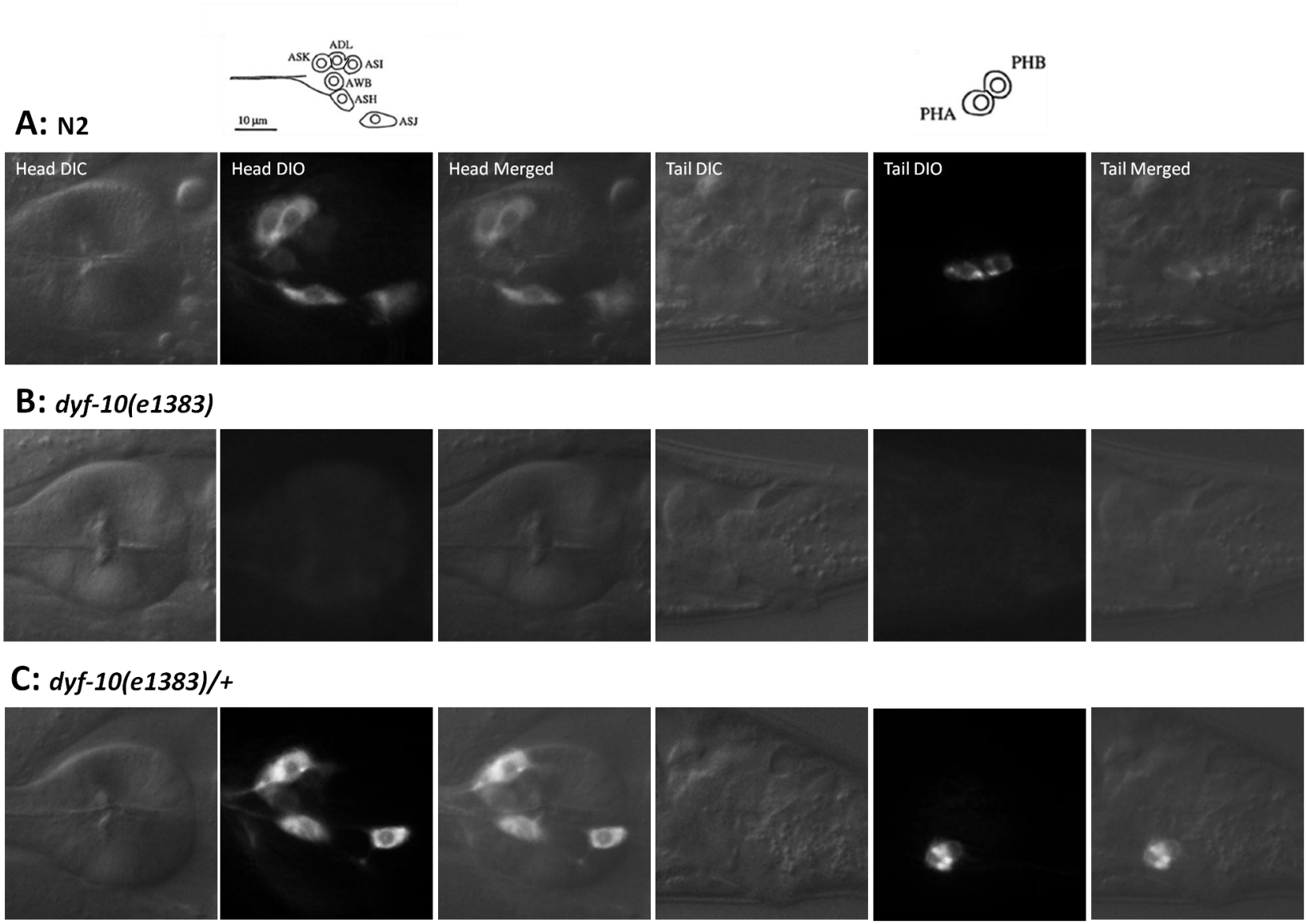
Genotype[c]	Total[a]	Total Dye filling[b]		Amphid and Phasmid		Amphid Only (%)		Phasmid Only (%)		None (%)	
	N	N	%	N	%	N	%	N	%	N	%
e1383	643	1	0.2	0	0.0	0	0.0	1	0.2	642	99.8
e1383 tm4200	289	45	15.6	7	2.4	19	6.6	19	6.6	244	84.4
e1383 nDf25	66	0	0.0	0	0.0	0	0.0	0	0.0	66	100.0
e1383/+[c]	88	88	100.0	88	100.0	0	0.0	0	0.0	0	0.0
N2	684	683	99.9	625	91.4	58	8.5	0	0.0	1	0.1

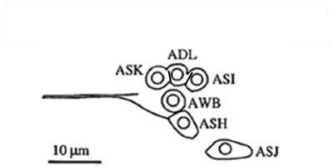
Note: Positive control, N2. Negative control: *e1383*. Abbreviations: N (total number of worms), % (percentage of worms).

a Total number of worms in the assay.

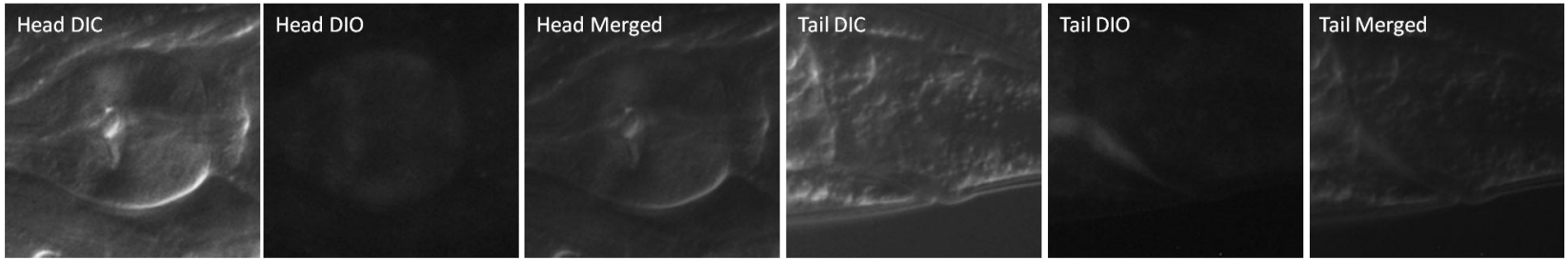
b Total number of worms that dye filled in amphid and phasmid together, amphid only and phasmid only

c dpy-5(e61) dyf-10(e1383) worms were crossed to N2 males to generate *e1383/+*. Full genotype is *e61 e1383/+ +*.


A: N2
ADL
ASK
ASI
AWB
ASH
ASJ
10 µm
PHB
PHA
Head DIC
Head DIO
Head Merged
Tail DIC
Tail DIO
Tail Merged
B: dyf-10(e1383)
C: dyf-10(e1383)/+



D: ***dyf-10(e1383)/tba-5(tm4200)***



E: ***dyf-10(e1383)/nDf25***

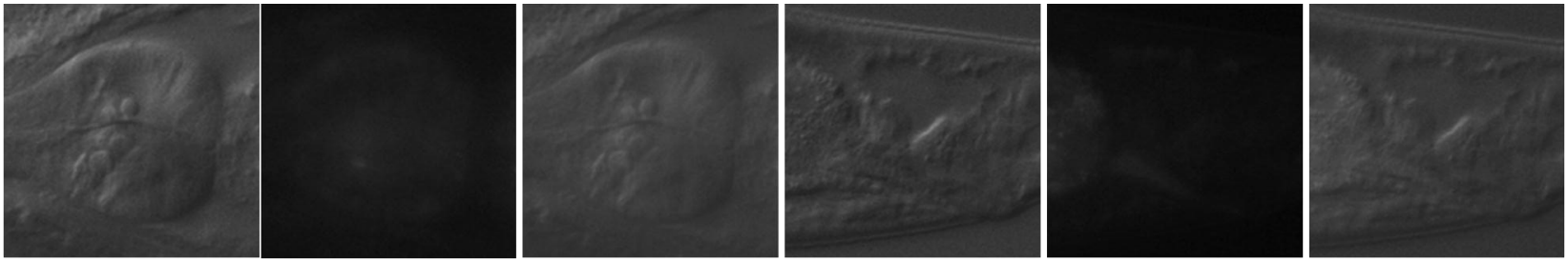


Figure 2.20. Recessive Gain-of-function: DiO Dye Filling Images of Amphid and Phasmid Neurons

(A) Positive Control, N2. (B) Negative control, dyf-10(e1383). (C) *dyf-10(e1383)/*+, dye filling. (D) *dyf-10(e1383)/tba-5(tm42000*, no dye filling. (E) *dyf-10(e1383)/nDf25*, no dye filling. DiO dye filling was visualized using fluorescent compound microscopy. Schematic diagram of amphid and neuronal bodies modified from Starich *et al*.,1995. Abbreviations: DIC (differential interference contrast), DiO (lipophillic dye).

2.4. Conclusions and Future Work

The identity of *dyf-10* is *tba-5*. The variation detection pipeline successfully identified the marker *dpy-5(e61)* as well as the homozygous and non-WT missense mutation that corresponds to *dyf-10(e1383)* in *tba-5* (Figure 2.13). Complementation tests confirmed that *dyf-10* and *tba-5* are of the same gene. Crosses of the three alleles, *e1383, qj14* and *tm4200*, with each other resulted in a Dyf phenotype, meaning all three alleles did not complement each other (Figure 2.14 and Table 2.6). Furthermore, *qj14/e1383* crosses had the same dye filling result as Hao and colleagues' work (Hao *et al.*, 2011). However, *dyf-10(e1383)* is not located in the predicted region of interest as was determined by three-point factor mapping (Starich *et al.*, 1995). *dyf-10(e1383)* has been mis-mapped since *tba-5* is located approximately 900,000 bp to the right of the predicted region of interest. To further confirm that *dyf-10* is *tba-5*, future work may include injecting a construct expressing the full length copy of *tba-5* into *dyf-10(e1383)* and/or *tba-5(qj14)* worms. It is expected that the WT copy of TBA-5 should be able to rescue the Dyf-phenotype in the two mutated strains. If the *tba-5* construct does not rescue, possible reasons may include the dosage of *tba-5* that is injected, whether all the necessary promoter elements are included in the construct and whether *tba-5* is able to travel into ciliated cells from the injection point.

Dye filling assays indicate that *e1383* is a recessive loss-of-function mutation and *tm4200* is not a null mutation. *e1383/tm4200* worms, where the *tm4200* allele consist of a small in-frame deletion, resulted in partial dye filling. On the other hand, *e1383/nDf25* worms, where the nDf25 deficiency that deletes the whole *tba-5* transcript, displayed no dye filling. The finding suggests that *tm4200* is not a null allele of *tba-5* compared to *nDf25*. If *tm4200* was a null allele, we would expect to see the same dye filling result for *e1383/tm4200* and *e1383/nDf25*. *e1383* in the presence of a WT copy of *tba-5* is

sufficient for normal cilia formation, which suggests that *e1383* is a recessive allele. Furthermore, the absence of dye filling in both *e1383* and *e1383/nDf25* worms suggests that the *e1383* mutation is indeed a loss-of-function mutation.

The dye filling results led to the model that the IFT machinery has variable binding affinities to different tubulins which can affect the formation of the cilia. Since the *tm4200* is likely not a null allele, in *tba-5(tm4200)* mutants it is likely the truncated TBA-5 in addition to TBA-5 paralogs that are incorporated into the cilia structure to result in a grossly normal cilia that dye fills. It is hypothesized that the binding of IFT to tubulin from highest to lowest affinity are: WT TBA-5, mutated TBA-5, TBA-5 paralogs and truncated TBA-5. *e1383* mutation occur in a highly conserved residue, so it is reasonable to assume that the allele affect the contact between the tubulin and the IFT machinery in such a way that it makes the binding of the IFT machinery to the TBA-5 less effective. Disruption of the transport of tubulin components in the cilia axoneme will eventually result in defective cilia. If the dosage of mutated TBA-5 has been halved, the chances of the IFT machinery incorporating other tubulin isoforms into the cilia and/or, in the case of e1383/tm4200 mutants, incorporating slightly functional truncated copies of TBA-5, increases, which can result in functioning cilia. However, since the percentage of worms dye filling in these cases are still very low, it is reasonable to assume that the IFT machinery has a much lower binding affinity to these tubulin paralogs and truncated version of TBA-5 than WT or mutated TBA-5. Future work can test the binding affinity of the IFT machinery; mutant TBA-5 can be over expressed at different dosage in N2 worms and then dye filled. It is expected that a higher dosage of the mutated TBA-5 will have a better chance at competing with WT TBA-5 for IFT binding and therefore result in a less percentage of the worms dye filling when compared to N2 worms.

Dye filling was observed in the knockdown of TBA-9 in *tm4200* worms (Figure 2.17). Now considering that the *tm4200* allele is not a null allele and assuming TBA-9 is the paralog that can compensate for the loss of TBA-5, we can hypothesize that the truncated TBA-5 is sufficient by itself to generate normal-like cilia that can dye fill. However, assuming that TBA-9 is not the correct paralog that can compensate for the loss of TBA-4, we can also hypothesize that the functional cilia was formed by the truncated TBA-5 and supplemented by another tubulin paralog. To determine which paralogs can compensate for the loss of TBA-5, we can generate a *ok1858;nDf25*

double mutant, but the phenotype will be lethal due to the large deficiency. As a result, we first have to generate a null allele of *tba-5* that won't be lethal when homozygous, and then generate double knockout mutants. Future work on finding the TBA-5 paralog(s) would be to test all of the alpha tubulins by injecting a construct carrying the WT alpha tubulin being tested, or a combination of the alpha tubulins, into *dyf-10(e1383)* and/or *tba-5(qj14)* worms. It is expected that the over expression of the tubulin(s) responsible for the compensation effect should be able to rescue the Dyf phenotype in the mutant worms.

Many ciliary genes expressed in ciliated neurons identified thus far are regulated by the RFX transcription factor DAF-19 (Swoboda *et al.*, 2000). DAF-19 binds to an X-box motif (Emery *et al*., 1996) usually within a 250bp promoter region (Chu *et al.,* 2012). Future work may include predicting potential X-box motifs in the promoter region of *tba-5* and testing whether the X-box motif is functional. If *tba-5* does not have an X-box motif, the expression of *tba-5* is likely not dependent on DAF-19. Conversely, if *tba-5* does have an X-box motif, it is likely that *tba-5* expression is dependent on DAF-19. Preliminary results have identified several X-box motif candidates approximately 1 kbp upstream from the transcriptional start site of *tba-5*. The X-box motifs were predicted by the program Conservation-aided transcription factor binding site finder (COTRASIF), using the Hidden Markov model (HMM)-based method of finding transcription factor binding sites (Tokovenko *et al.,* 2009). A set of validated X-box motifs were inputted to train the program to predict X-box motifs genome wide.

2.5. Discussion

VarFreq is often miscalculated for InDels when using VarScan. Often, I find that there are many InDels detected by VarScan with VarFreq > 100%, VarFreq = ∞ or a VarFreq that does not match with what is shown in the read alignment on GBrowse. Information on SeqAnswers (www.seqanswers.com) by the author acknowledges the parsing error and is fixed in a later version of VarScan, version 2.2.8. Since an earlier version of VarScan was used in the analysis, to bypass the parsing error, the VarFreq in the VarScan SND and InDel files were re-calculated. First, a file detailing the read count for each base using the 'readcount' command in VarScan was generated. Second, the

variant frequency for all the variations was re-calculated using the correct q15 (BaseQual 15) read count for each base; q15 specifies that each base that is counted has a quality > 15. Third, all the predictions were re-filtered using Perl scripts with the following criteria: MinCov $\geq$ 9 and VarFreq $\geq$ 30%. Fourth, the resulting SND and InDel files were converted into GFF3 files using Perl scripts. Fifth, a check was made to make sure the filtered variation files are unique, meaning there is only one line for each variation. Often VarScan may predict two different variations in the same position. Take for example a heterozygous SND, in which 35% of the reads support a consensus base of C, 30% of the reads support a consensus base of G, and 35% of the reads support the reference base of T. Varscan will output two separate lines detailing both of the variant bases. To avoid confusion in the downstream analysis, variation with the lower VarFreq were removed. In addition, in order for the deletions to display correctly on gbrowse, the Varscan deletion positions were shifted by 1 bp.

qj14 is a less severe mutation of TBA-5 when compared to *e1383*. A19V involves a change of a hydrophobic amino acid into another hydrophobic amino acid in *qj14* mutants. On the other hand, P360L involves a change of a proline into a hyprophobic amino acid in *e1383* mutants. Prolines are important for providing rigidity in a protein structure. Structurally, P360L will have a larger effect on the protein structure of TBA-5 compared to A19V. The temperature sensitivity of *qj14* compared to the temperature insensitivity of *e1383* (Hao *et al.,* 2011) lends support that *e1383* has a more damaging effect on the protein structure of TBA-5. Dye filling experiments also support the notion that *qj14* is a less destructive mutation than *e1383*. More worms dye filled in both amphid and phasmid neurons in *tm4200/qj14* mutants than *tm4200/e1383* mutants (Figure 2.14, Table 2.6). In strains where the dosage of the mutant allele is halved, the chances of proper cilia formation is higher. This is especially true for *qj14* mutants. Since *qj14* affects protofilament-protofilament interactions (Hao *et al.*, 2011), the IFT machinery is able to transport and incorporate the mutated tubulin into the ciliary axoneme. Furthermore, with the abundance of many different tubulin paralogs, the lack of protofilament-protofilament interactions can easily be compensated by other fully functional tubulin isotypes.

Since *tba-5(tm4200)* worms were able to dye fill, it is possible that there is an extra WT copy of *tba-5* in the *tba-5(tm4200)* worms. To test this, cDNA of *tba-5(tm4200)*

worms were generated and PCR of the *tba-5* cDNA with several sets of primers yielded non-specific bands that were confirmed by Sanger re-sequencing to be non-*tba-5* sequence (data not shown, Sanger re-sequencing completed by Dr. Maja Tarailo-Graovac). Primers consisting of SL1 paired with a reverse primer at the end of the cDNA sequence, and a shorter primer pair that closely flanks the tm4200 allele, produced non-specific bands in two cDNA libraries; the cDNA libraries were generated using oligodT or the *tba-5* gene specific primer. If the experiment had been successful, a band corresponding to *tba-5* with the tm4200 deletion would have been detected. It is likely that PCR of the cDNA library failed, because the concentration of *tba-5(tm4200)* mRNA may be low and thereby did not get reverse transcribed into cDNA, and cannot be detected by PCR. However, based on other evidence the *tba-5(tm4200)* strain does not seem to contain a WT copy of *tba-5* elsewhere in the genome. First, during outcrossing of the *tba-5(tm4200)* 0X strain with N2 background, genotyping clearly showed homozygous WT, homozygous tm4200 allele bands or heterozygous bands as expected. If there was a WT copy of *tba-5,* only heterozygous bands would be observed in all steps of the *tm4200* allele genotyping. Second, if we assume that there was a WT copy of *tba-5* on another chromosome in the *tba-5(tm4200)* strain, we would expect more than 15% of the worms dye filling in *e1383/tm4200* worms (Figure 2.14). Since *e1383* and *qj14* are recessive alleles and the dosage of the mutated TBA-5 is halved, the presence of a WT TBA-5 from a WT copy of *tba-5* elsewhere in the genome should be sufficient for cilia formation and result in dye filling. As a result, it is safe to conclude that there are no other WT copies of *tba-5* in the *tba-5(tm4200)* strain.

nDf25 is a large deletion that spans genes *lin-10* at I:8107823 to *unc-29* at I:8901657(Ferguson and Horvitz, 1985). Based on Wormbase, nDf25 does not appear to overlap with hDf8. Within the nDf25 region, genes related to sensory cilia that are deleted in the deficiency include: *tba-5* and *daf-8*. *daf-8* encodes an R-Smad protein that represses dauer development. Park and colleagues determined that *daf-8* is expressed in a subset of head and tail neurons and was specifically abundant in the amphid neurons ASI and ADL (Park *et al.,* 2010). However, it is not known how or if *daf-8* affects the formation of the cilia.

3. Bioinformatic Analysis of *dsh-2(or302)* Suppressors

3.1. Background

Asymmetric cell division is important for neuronal diversity. The mechanism refers to the division of a polar mother cell that ultimately results in the differential fate of two sister cells, fate 'A' and fate 'B', which may differ in size, shape, or in other morphological or biochemical features (Horvitz and Herskowitz, 1992). The importance is highlighted in the fact that all 302 neurons in a *Caenorhabditis elegans* hermaphrodite are generated by asymmetric cell division through distinct lineages (cited in Hawkins *et al.*, 2005). For example in the PHB lineage, ABpl/rapppap, the HSN/PHB neuroblast undergoes asymmetric cell division to produce an anterior cell that undergoes apoptosis and a posterior cell called the HSN/PHB precursor. The HSN/PHB precursor then undergoes asymmetric cell division to produce the HSN motor neuron and the PHB phasmid neuron (reviewed in Hawkins *et al.,* 2005) (Figure 3.1A).

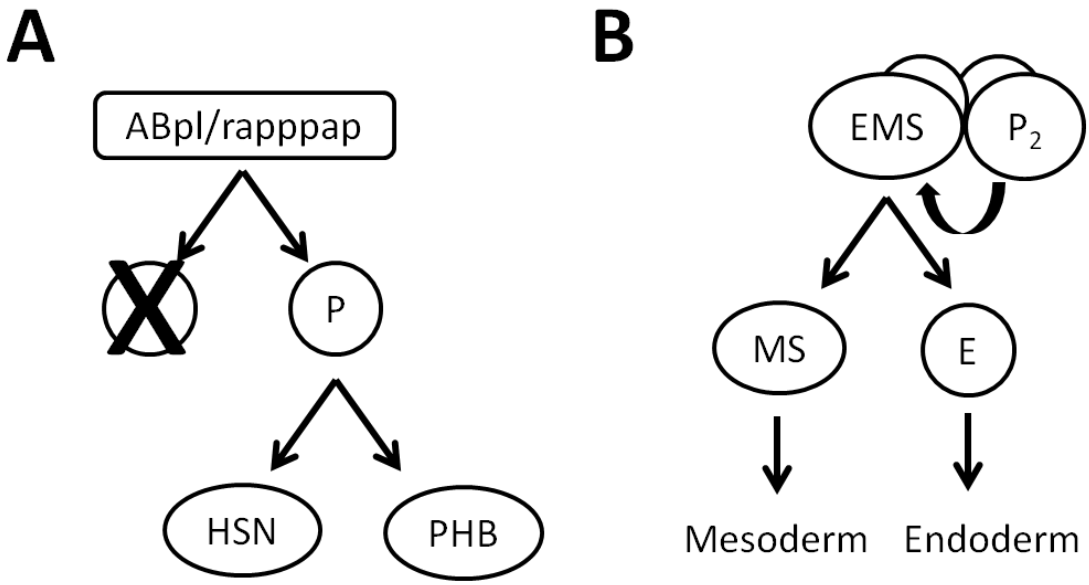


Figure 3.1. Examples of Asymmetric Cell division in C. elegans

(1) ABpl/rapppap, the HSN/PHB neuroblast undergoes asymmetric cell division to produce an anterior cell that undergoes apoptosis and a posterior cell (P) that is the HSN/PHB precursor. The HSN/PHB precursor then undergoes asymmetric cell division to produce the HSN motor neuron and the PHB phasmid neuron (reviewed in Hawkins *et al.,* 2005). (2) At the four-cell stage, the posterior most cell (P_2) induces its anterior sister cell (EMS) to divide into an anterior cell (MS) and a posterior cell (E). MS eventually gives rise to the mesoderm, while E will give rise to the endoderm in the animal (Bei *et al.,* 2002).

Wnt signalling controls many asymmetric cell divisions in *C. elegans*. The use of Wnt ligands (wingless in *Drosophila*) for cell signaling is evolutionarily conserved amongst metazoans and is used extensively during animal development. Work on *Drosophila* and vertebrates have shown that there is a 'canonical' or Wnt/β-catenin pathway and a 'non-canonical' pathway; *C. elegans* utilizes both pathways. In *C. elegans* it has been found that processes utilizing the β–catenin homolog BAR-1 generally uses the canonical Wnt pathway, whereas processes involving the β–catenin homolog WRM-1 use the non-canonical Wnt pathway. For the canonical pathway in the absence of a signal, the transcription factor β–catenin is targeted for degradation by the destruction complex composed of CK1α, GSK3β, APC and Axin. In the presence of a signal (Figure 3.2), Porcupine (Porc) protein promotes the secretion of the Wnt ligands. Next, the binding of the Wnt ligand to the Frizzled receptor and co-receptor LRP on the target cell's surface leads to the activation of Dishevelled (Dsh) and the stabilization of β–catenin. β–catenin can then interact with TCF/LEF proteins in the nucleus to activate gene transcription. Non-canonical Wnt signalling utilizes many of the same signal transduction components. However, the β–catenin homolog WRM-1 and parallel input from the mitogen-activated kinase (MAPK) pathway, consisting of kinases MOM-4/Tak1, LIT-1/Nlk and the binding protein TAP-1/Tab1 that interacts with WRM-1, is needed to down-regulate the TCF/LEF protein POP-1. This is in contrast to the up-regulation of POP-1 by β–catenin homolog BAR-1 in the canonical Wnt signalling pathway. Generally, it is the non-canonical Wnt signalling that is responsible for regulating asymmetric cell division. In *C. elegans*, known genes encoding Wnt ligands includes *lin-44, egl-20, mom-2, cwn-1* and *cwn-2*. Known genes encoding the Frizzled family of Wnt receptors includes *lin-17, mom-5, mig-1* and *cfz-2.* Known genes encoding the Disheveled proteins include *mig-5, dsh-1* and *dsh-2*. Finally, known *C. elegans* genes homologous to Porc, CK1α, GSK3β, Axin and TCF include *mom-1, lin-19, gsk-3, pry-1* and *pop-1*, respectively. To date there are no genes encoding a clear LRP homolog identified in *C. elegans* (reviewed in Eisenmann, 2005; reviewed in Korswagen, 2002).

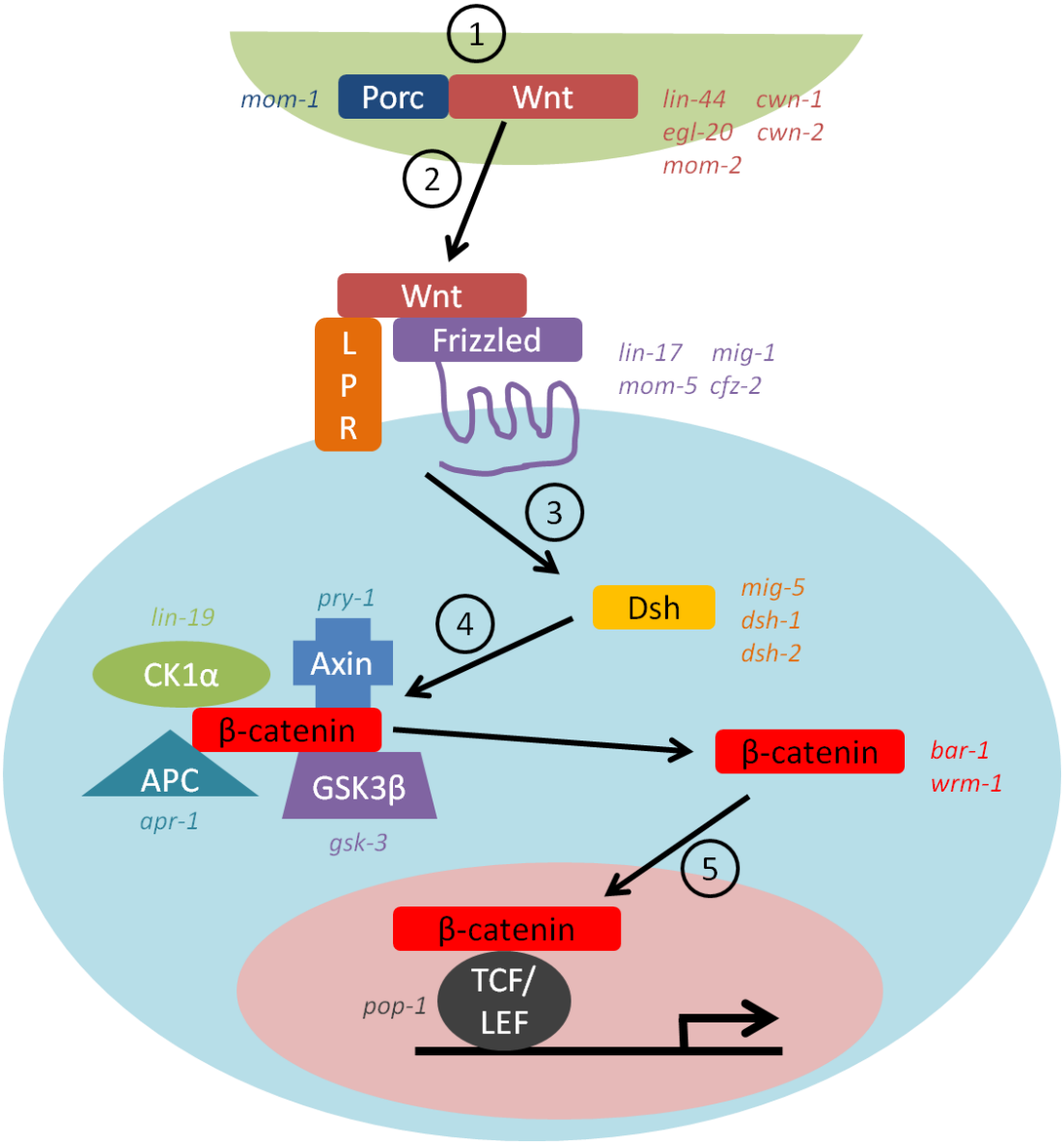


Figure 3.2. Diagram of the activation of the canonical Wnt Signaling pathway in C. elegans

When the canonical Wnt signaling pathway is activated by a signal, (1) the Porc protein helps aid the secretion of Wnt. (2) Wnt then binds onto the Frizzled transmembrane receptor and the co-receptor LPR on the cell surface to (3) activate Dsh in the cytoplasm. (4) Dsh then prevents the transcription factor β–catenin from being degraded by the destruction complex, composed of CK1α, GSK3β, APC and Axin. (5) The stabilized β-catenin finally migrates into the nucleus to bind to TCF/LEF to activate transcription. Known *C. elegans* homolog for Porc is the gene *mom-1*. Known homologs for Wnt are: *lin-44, egl-20, mom-2, cwn-1* and *cwn-2*. Known homologs for Frizzled are: *lin-17, mom-5, mig-1* and *cfz-2*. Known homologs for Dsh are: *mig-5, dsh-1* and *dsh-2*. Known homolog for CK1α, Axin, APC and GSK3β are *lin-19, pry-1, apr-1* and *gsk-3*, respectively. Known homologs of β-catenin are *bar-1 and wrm-1*. Known homolog of TCF/LEF is *pop-1*. Diagram adapted from http://www.stanford.edu/group/nusselab/cgi-bin/wnt/pathway_diagram [*access July 17, 2012*].

DSH-2 belongs to the Dishevelled protein family that regulates many asymmetric cell division through the Wnt signalling pathway. *dsh-2* and *mig-5* have been implicated in EMS polarization for endoderm induction using the non-canonical Wnt signalling pathway. At the four-cell stage the posterior most cell (P_2) induces its anterior sister cell (EMS) to divide into the anterior (MS) and posterior (E) cell. The anterior cell (MS) and posterior (E) eventually gives rise to the mesoderm and endoderm, respectively (Figure 3.1B) (reviewed in Bei *et al.*, 2002). Bei and colleagues propose that Wnt signalling functions in parallel with Src signaling to specify E cell fate. They found that removing *src-1* with both *dsh-2* and *mig-5* resulted in 100% of the embryos that lack endoderm. However, removing only *dsh-2* and/or *mig-5* did not result in endoderm defects (Bei *et al*., 2002). In another study, Hawkins and colleagues recently defined a role for DSH-2 in asymmetric cell division in the lineage that generates the phasmid neuron PHA. Experiments with *dsh-2* RNAi resulted in a high frequency of embryonic lethality. Hawkins and colleagues also isolated and experimented on a null allele, *dsh-2(or302). dsh-2(or302)* consists of a large frame shifting deletion that starts in exon 2 and extends into exon 6. Animals with the *dsh-2(or302)* allele are homozygous viable, maternal-effect lethal. In this phenotype, animals heterozygous for *dsh-2(or302)* are healthy. F1 animals homozygous for *dsh-2(or302)* lack both copies of *dsh-2* (zygotic), but still contain maternally provided DSH-2. As such, they are viable and healthy but partly sterile due to defects in gonadal morphogenesis. F2 animals, on the other hand, died as embryos or young larvae since they lacked both maternal and zygotic *dsh-2*. Importantly, an extra PHA neuron is often found in L1 escapers where there is neither functional maternal or zygotic DSH-2 in the animal. As a result, Hawkins and colleagues found that the loss of maternal and zygotic *dsh-2* resulted in symmetric cell division of the ABpl/rpppa neuroblast, and therefore the duplication of the PHA neuron (Hawkins *et al.,* 2005).

To identify genes functioning with DSH-2 in asymmetric neuroblast division, an ethyl methanesulfonate (EMS) genetic suppressor screen was performed by the Hawkins lab. The screen identified dominant suppressors S*up245*, *Sup327* and *Sup305* that suppress the defects found in *dsh-2(or302)*. Mapping by the Hawkins lab has placed *Sup245* and *Sup327* on chromosome I and *Sup305* on chromosome IV. As a result, all three suppressors are categorized as extragenic, where the suppression of the

mutant phenotype is due to a change elsewhere in the genome. Suppression can be caused by altered gene dosage where the lack of DSH-2 can be compensated by up-regulation or de-regulation of an alternative protein or pathway. Suppression can also affect signal transduction by modifying the strength of the signal. In this case, the loss of DSH-2 can be suppressed by antagonistic mutations acting downstream in the signal transduction pathway (reviewed in Hodgkin, 2005). To determine the mechanistic nature of the suppression and therefore identify genes that function with DSH-2, the goal of the project is to determine the location of *Sup245, Sup327* and *Sup305*. First, putative mutations will be identified using next generation sequencing (NGS) for each of the strains carrying the suppressors. Second, candidate genes that are affected by the putative mutations will be identified. Since EMS mutagenesis primarily produces GC to AT transitions (Flibotte *et al.*, 2010), genes that are affected by single nucleotide differences (SNDs) would be ideal suppressor candidates.

3.2. Materials and Methods

Genotype of suppressor strains that are sequenced. The strain Sup327 contains the suppressor variation *Sup327*, which has been mapped onto the left arm of chromosome I, and the *dsh-2(or302)* knockout allele on chromosome II (Figure 3.3A). The strain Sup245 contains the suppressor variation *Sup245*, which has been mapped onto the right arm of chromosome I, and the *dsh-2(or302)* knockout allele on chromosome II (Figure 3.3B). The strain A01396 contains the markers *lin-17(n671)* and *unc-54(e190)* on chromosome I. Presently, it is not known whether variations for either or both *Sup327* and *Sup245* are present in the strain on chromosome I. A01396 also contains *dsh-2(or302)* balanced by *mIn1[mIs14 dpy-10(e128)]* (Figure 3.3C). *mIn1* is a inversion of a large central portion of chromosome II extending from *lin-31* to *rol-1*. *mIn1* contains *mIs14,* an insertion of a transgene that gives the animals a semi-dominant pharyngeal GFP phenotype, as well as the *dpy-10(e128)* allele (Edgley and Riddle, 2001). Finally, the strain A00842 contains the suppressor variation *Sup305*, which has been mapped on chromosome IV, with *dsh-2(or302)* on chromosome II (Figure 3.3D). All the suppressor strains were generated and sequenced by the Hawkins lab.

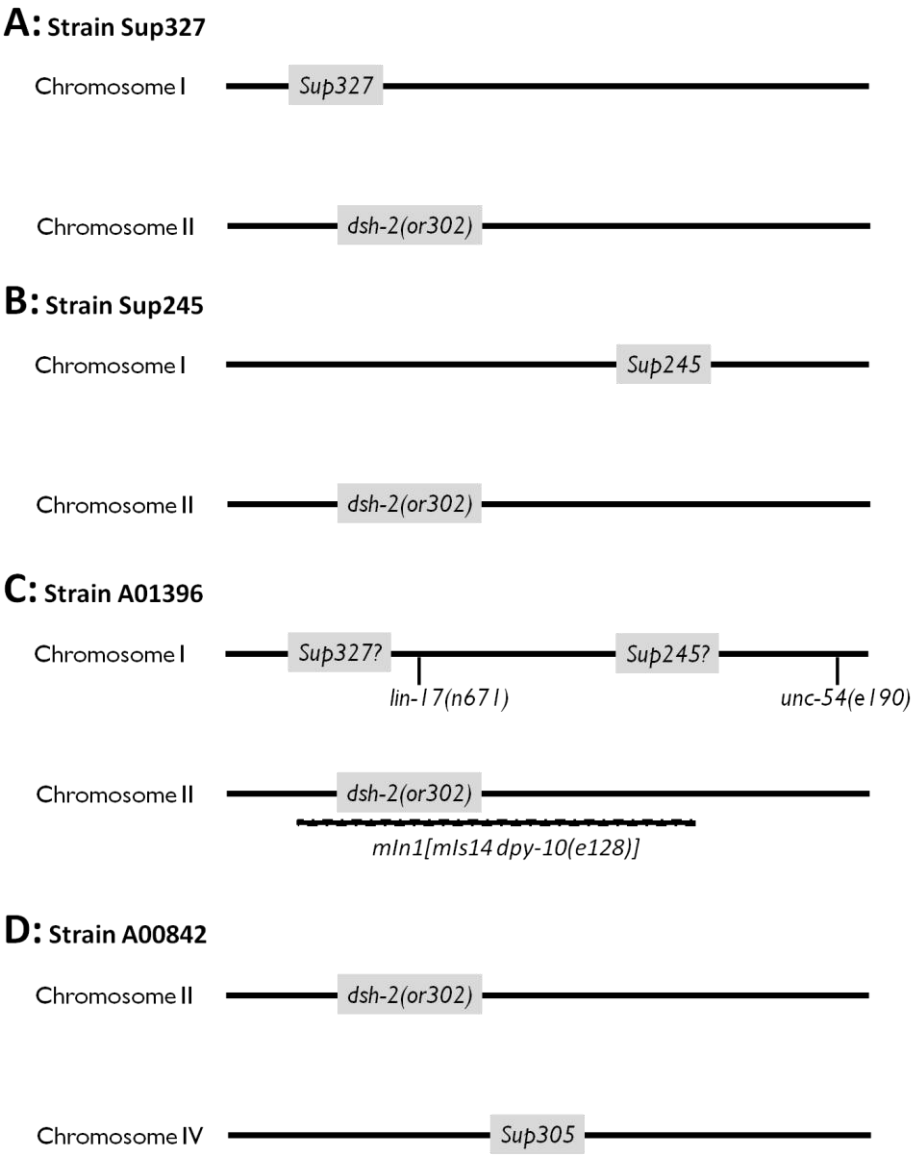


Figure 3.3. Overview of Genotype for Suppressor Strains: Sup327, Sup245, A01396 and A00842

(A) *Sup327*, the suppressor gene of interest is mapped to the left arm of chromosome I. Knockout *dsh-2(or302)* allele, featuring a 1076bp deletion, is located on chromosome II. (B) *Sup245*, the suppressor gene of interest is mapped to the right arm of chromosome I. Knockout *dsh-2(or302)* allele, featuring a 1076bp deletion, is located on chromosome II.. (C) The strain may or may not carry the suppressors *Sup327* and *Sup245*, which are both mapped to chromosome I. Knockout *dsh-2(or302)* allele, featuring a 1076bp deletion, is located on chromosome II and is balanced by the rearrangement *mIn1[mIs14 dpy-10(e128)]*. *mIn1* features a large inversion that covers most of chromosome II, extending from *lin-31* to *rol-1*. The markers *lin-17(n671)* and *unc-54(e190)* are located on chromosome I, featuring a nonsense mutation and a large 401bp deletion, respectively. (D) *Sup305,* the suppressor of interest is mapped to chromosome IV. Knockout *dsh-2(or302)* allele is located on chromosome II. Note, diagram is not drawn to scale.

Whole genome sequencing and read alignment. All strains have been sequenced using the Illumina/Solexa paired-end sequencing platform (Figure 3.4). Strains Sup327 and Sup245 have a mean insert size or fragment length of 163 and 164, respectively, and a read length of 75 bp. Strains A01396 and A00842 have a mean insert size or fragment length of 350 and 345, respectively, and a read length of 76 bp. The average insert size for each strain was collected using Picard's 'CollectInsertSizeMetrics.jar' tool (http://picard.sourceforge.net), version 1.40. Since read information for strains Sup327 and Sup245 were encoded in BAM alignment file, read information in standard FASTQ (STDFQ) format was first extracted using 'bam2fastq' software version 1.1.0 from the Genomic Services Lab at HudsonAlpha (http://www.hudsonalpha.org/gsl/software/bam2fastq.php) using default parameters. Read information for strains A01396 and A00842 were encoded in 'qseq' files. A PERL script (www.perl.org) was used to convert the 'qseq' file into a STDFQ format in preparation for read alignment. All four strains were aligned to the WormBase WS224 *C. elegans* reference genome using Novoalign version 2.07.13 (www.novocraft.com) (Figure 3.4). Default parameters were used for strains Sup327 and Sup245 except for the following: -F STDFQ, which specifies the format of the read files, and -I PE 250,100, which specifies the approximate fragment length and standard deviation for reads in paired-end mode. Default parameters were used for strains A01396 and A00842 except for the following: -F STDFQ and -I PE 350,100. After generating the BAM file from Novoalign, PCR duplicates were removed from the alignment using the 'samtools rmdup' command, version 0.1.12a (Li *et al.,* 2009). The resulting BAM file was visualized with Generic Genome Browser (GBrowse), version 2.39. Sequencing of the Sup327 strain generated 112 million reads, of which 58.84% were mapped to the WS224 *C. elegans* reference genome (Table 3.1). The average read depth in Sup327 is 50 reads per base. Sequencing of the Sup245 strain generated a much higher number of reads, around 236 million. 82.94% of the reads were mapped to the reference genome, generating an average read depth of 146 reads per base. Sequencing of the A01396 strain yielded only 75 million reads, of which 75.5% of them were mapped to the reference genome. This resulted in an average read depth of 46 reads per base. Finally, the sequencing of the A00842 strain generated 66 million reads. 84.78% of the reads were mapped and the average read depth is 47 reads per base.

Table 3.1. Read Mapping Details for Suppressor Strains: Sup327, Sup245, A01396 and A00842

Number of Reads				Category
Sup327	**Sup245**	**A01396**	**A00842**	
112804888	236259093	74935364	66099513	In total
0	0	0	0	QC failure
0	0	0	0	Duplicates
66374968 (58.84%)	195952163 (82.94%)	56576999 (75.50%)	56041025 (84.78%)	Mapped
112804888	236259093	74935364	66099513	Paired in sequencing
56402616	118130602	37467804	33049815	Read1
56402272	118128491	37467560	33049698	Read2
65418613 (57.99%)	193080205 (81.72%)	53166814 (70.95%)	54056068 (81.78%)	Properly Paired
65564842	193352763	54723524	54744307	With itself and mate mapped
810126 (0.72%)	2599400 (1.10%)	1853475 (2.47%)	1296718 (1.96%)	Singletons
37084	54126	1226328	502266	With mate mapped to a different chromosome
36584	53836	1225822	502116	With mate mapped to a different chromosome (mapQ>5)
50	146	46	47	Average read depth

Notes: Reads were aligned to WS224 reference *C. elegans* genome using Novoalign (www.novocraft.com). Information above was collected from the BAM alignments using the ‘’samtools flagstat’ command.

Variation detection and filtration (Figure 3.4). Small variations, including InDels and SNDs were detected by first generating either a ‘pileup’ file or ‘mpileup’ file from SAMtools version 0.1.7a and version 0.1.12a, respectively. Strains, Sup327 and Sup245, were analyzed from the ‘mpileup’ file using the ‘-B’ option. The ‘-B’ parameter disables the calculation of base alignment quality (BAQ), which is the Phred-scaled probability of a read base being misaligned. Strains A01396 and A00842 were originally analyzed from the extended ‘pileup’ file that was generated by ‘samtools pileup -c’ command. VarScan , version 2.2.3, was then used to generate a list of SNDs and small InDels using the ‘pileup2snp’ and ‘pileup2indel’ commands, respectively, by specifying 0 for the minimum coverage (MinCov) and 0 for minimum variant frequency (VarFreq) (see discussion section in chapter 2). In addition, by default the minimum base quality

(BaseQual) is set a 15 (Koboldt *et al.,* 2009). The variations are finally filtered for the MinCov $\geq$ 5 and Varfreq $\geq$ 30% using PERL scripts. Pindel v0.2.0 was used to detect large deletions. Specifically, two of the positive controls are large deletions that cannot be detected by VarScan. Pindel can detect deletions by aligning read pairs where one end is mapped and the other mate is not mapped. For these cases, it is possible that the read pair that is unmapped spans a large deletion event. By splitting the unmapped read into two and re-mapping both fragments back to the reference genome, Pindel is able to compute the exact breakpoint of large deletions (Ye *et al.,* 2009). Since the alignment was generated by Novoalign, first, SAMtools was used to convert the BAM file into a SAM file. Next, the 'sam2pindel.cpp' script was used to generate a Pindel text input file. Second, the Pindel text input file was inputted into Pindel to generate predictions for large deletions, short insertions, long insertions, inversions and tandem duplications. Third, the large deletions file was converted into a GFF3 file using a Perl script; only large Pindel deletions were analyzed and were kept unfiltered. All variations were uploaded onto GBrowse for visualization. All PERL scripts used for filtering the data was written by me.

Annotation of coding transcripts with variations and compilation of data (Figure 3.4). To extract a list of coding transcripts affected by the variations detected, first, a list of SNDs and small InDels were inputted into Variant Analyzer separately (Chen lab, unpublished). Second the output was extracted for coding transcripts that have been affected by the following variations: missense, nonsense, synonymous, frame shifting and non-frame shifting variations. Third, using the Bio::DB::SeqFeature::Store bioperl module and accessing the WS224 *C. elegans* reference database, the variations are categorized into exon, intron, UTR, intergenic and repeat regions (as defined by Repeat Masker). Fourth all variations from multiple strains, namely Sup327, Sup245, A01396, A00842, and the two WT strains (Hobert and Horvitz) were compiled and compared. Fifth, information extracted Variant Analyzer and Bio::DB:SeqFeature::Store bioperl module were annotated for each variation. Sixth, the list of variations were filtered for those that affect the protein sequence, are homozygous in at least one of the strains (VarFreq $\geq$ 75%) does not appear in the wild type (WT) Hobert and Horvitz strains, and do not overlap with repeat regions. Finally, the remaining variations were filtered for the correct genomic location.

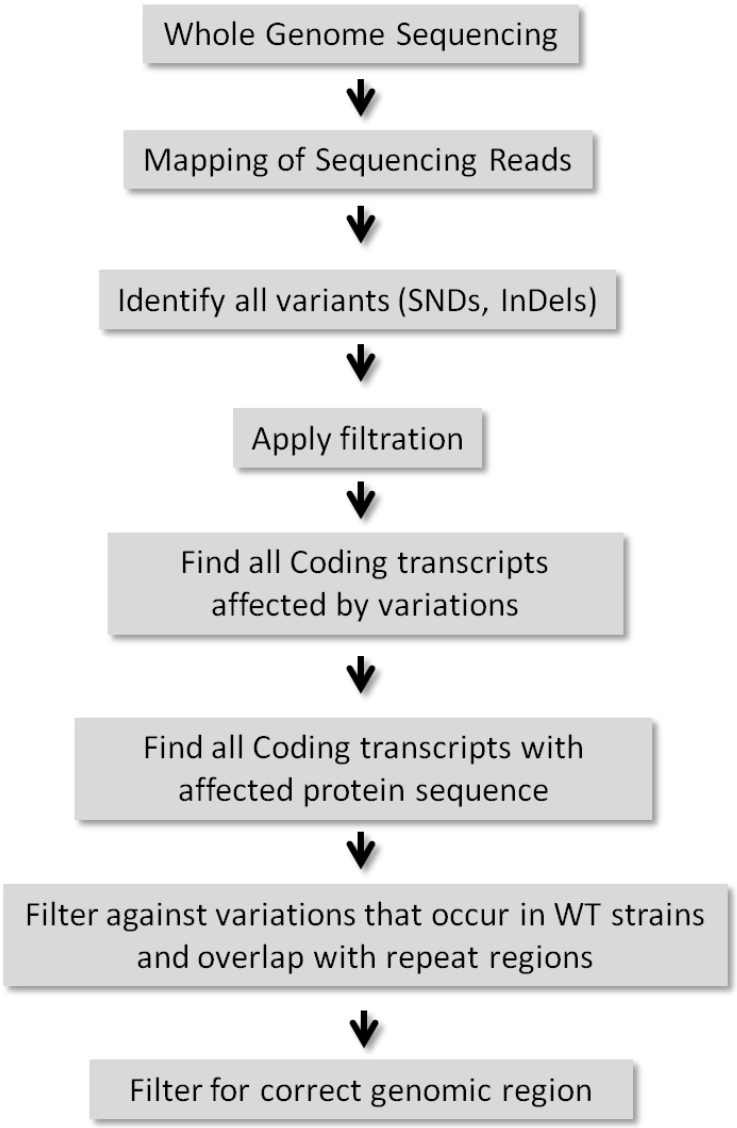


Figure 3.4. Variant Detection Pipeline for all Suppressor Strains

Whole Genome Sequencing was completed on the Suppressor strains using the Illumina/Solexa platform. Novoalign [www.novocraft.com] was used to map the paired-end reads to the WS224 reference *C. elegans* genome. SNDs and small InDels were detected using VarScan (Koboldt *et al.*, 2009) and filtered using the parameters: MinCov $\geq$ 5, VarFreq $\geq$ 30. Large deletions were detected using Pindel (Ye *et al*., 2010). Variant Analyzer (Chen lab, unpublished) was used to annotate coding transcripts affected by the filtered variations. Coding transcripts with affected protein sequence due to missense, nonsense, and frame shift mutations were identified. All variations that are also found in the two wild type strains (Hobert and Horvitz) and/or correspond with repeat regions and are not homozygous (VarFreq $\geq$ 75%) on at least one strain are filtered out. Finally, variations that fall into the correct genomic region, namely chromosome I and IV, are considered as candidate genes for *Sup327, Sup245* and *Sup305.*

3.3. Results

All positive controls in the four suppressor strains are detected in the variant detection pipeline. The *dsh-2(or302)* allele was detected in all four strains as expected. The *dsh-2(or302)* allele can be visualized by a sharp decrease of reads in the region where the deletion allele occurs (Figure 3.5). Strain Sup327, Sup245 and A00842 all show a homozygous deletion of 1076bp with clear breakpoints that are at the same location for all three strains. The deletion matches the *or302* allele description reported by Hawkins and colleagues (Hawkins *et al.*, 2005). A01396 shows a heterozygous deletion (Figure 3.5 and 3.6), which is expected as *dsh-2(or302)* is balanced by the inversion *mIn[mIs14 dpy-10(e128)]* in the strain. In Figure 3.5, strain A01396 shows a sharp decrease of reads at the deletion breakpoint with a smaller read coverage along the length of the deletion. In the zoomed in version of the *or302* deletion allele for strain A01396 (Figure 3.6), it is clear that a portion of the reads spans the deletion; these are the reads that show an insert size larger than the average calculated for the whole genome, it is these reads that supports the deletion. There are also a portion of the reads that map within the deletion. These reads map to the *mIn[mIs14 dpy-10(e128)]* genetic balancer that covers the deletion. The allele *unc-54(e190)* was also detected in the A01396 strain (Figure 3.7). *unc-54 (e190)* features a 401bp deletion that was detected by Pindel and is only found in the strain A01396, as expected. In addition, *lin-17(n671)* was also detected in the strain A01396 (Figure 3.8). The allele *n671* consist of a C>T nonsense mutation and this is supported by the read alignment since the nonsense mutation was detected at 100% VarFreq and at a read depth of 48. The mutation is only detected in the strain A01396 as expected.

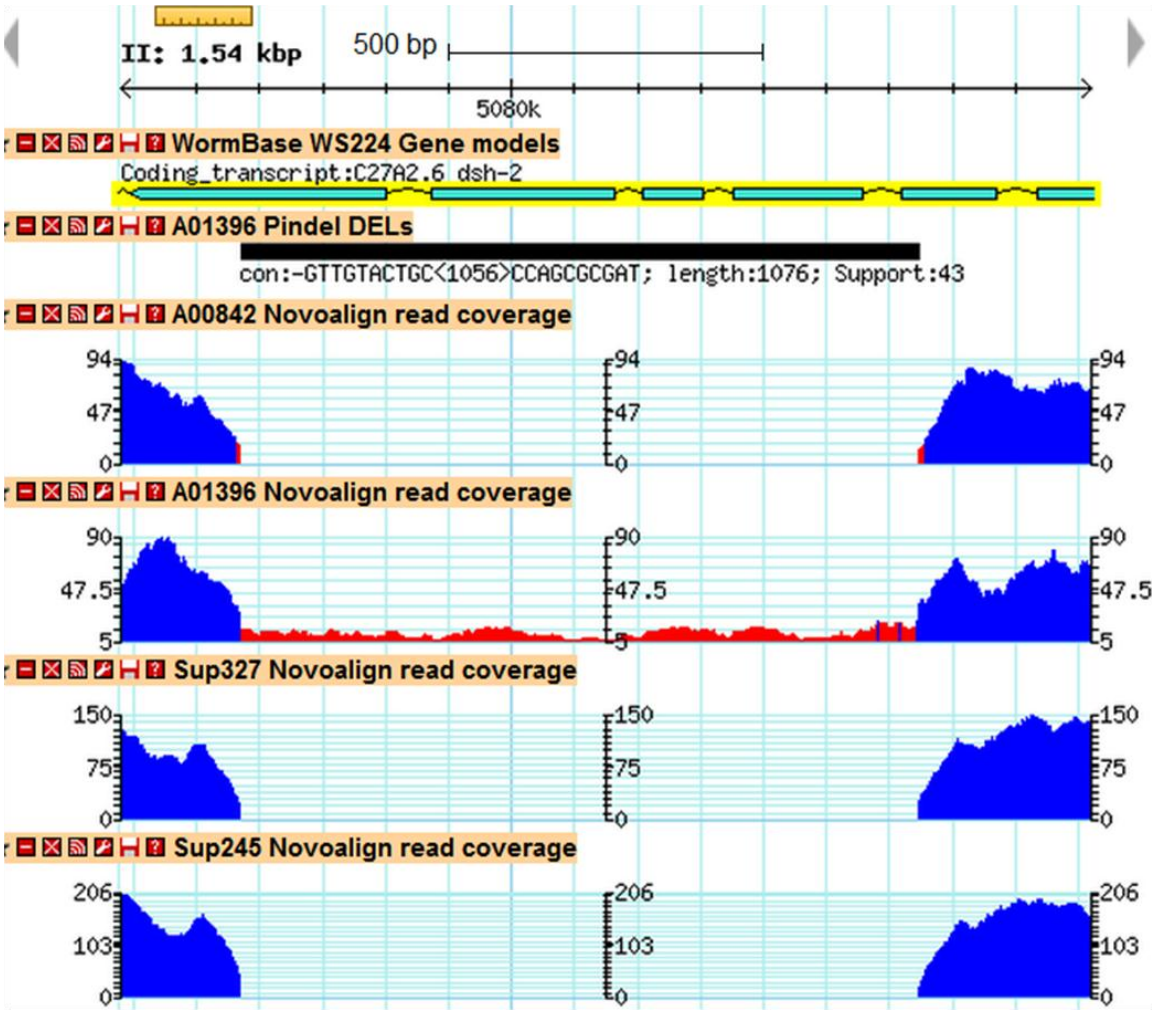


Figure 3.5. Positive Control: dsh-2(or302) Detected by Variant Detection Pipeline for all Suppressor Strains

Region shown: II:5,079,380..5,080,919. *dsh-2(or302)II* features a large deletion that starts 458bp downstream of the start codon and finishes in exon 6. A deletion of 1076bp, corresponding to the *dsh-2(or302)/mIn1[dpy-10(e128) mls14]* in strain A01396 and *dsh-2(or302)* in strains Sup327, Sup245 and A00842 was detected by Pindel at the correct location. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Abbreviations: con (consensus of deleted sequence, only the beginning and end 10bp are shown), length (length of deletion) and Support (number of split reads that support the deletion).

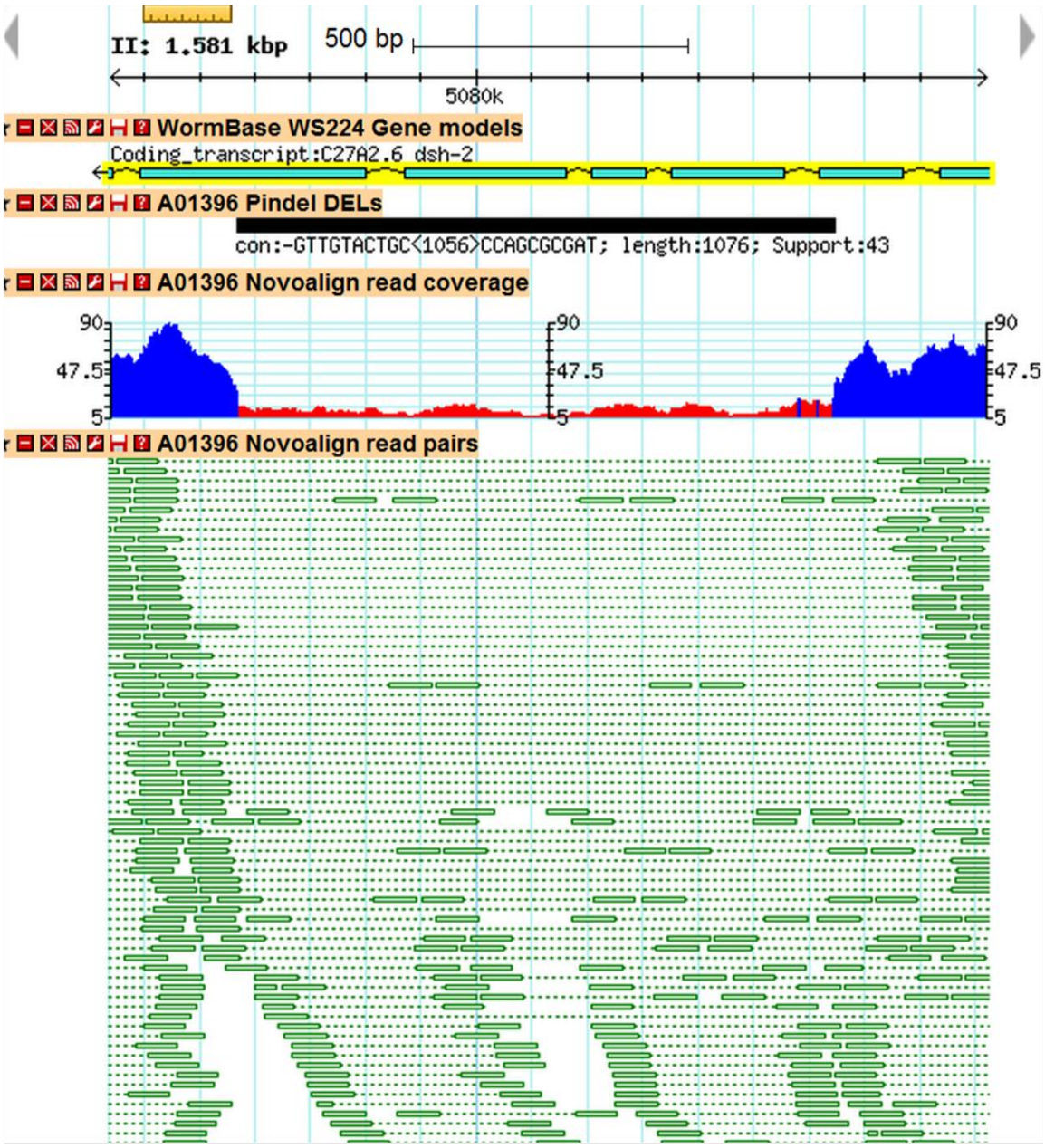


Figure 3.6. Positive Control: dsh-2(or302)/mIn[dpy-10(e128) mIs14]II Detected by Variant Detection Pipeline in strain A01396

Region shown: II:5,079,340..5,080,920. *dsh-2(or302)II* features a large deletion that starts 458bp downstream of the start codon and finishes in exon 6. A heterozygous deletion of 1076bp, corresponding to the *dsh-2(or302)/mIn1[dpy-10(e128) mls14]* was detected by Pindel at the correct location in A01396. A portion of the sequencing reads, characterized by large insert size support the *or302* deletion. A portion of the sequencing reads that map within the deletion support the balancer *mIn1[dpy-10(e128) mls14]*. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Abbreviations: con (consensus of deleted sequence, only the beginning and end 10bp are shown), length (length of deletion),Support (number of split reads that support the deletion), DEL (deletion). Note, not all of the sequencing reads are shown in the figure.

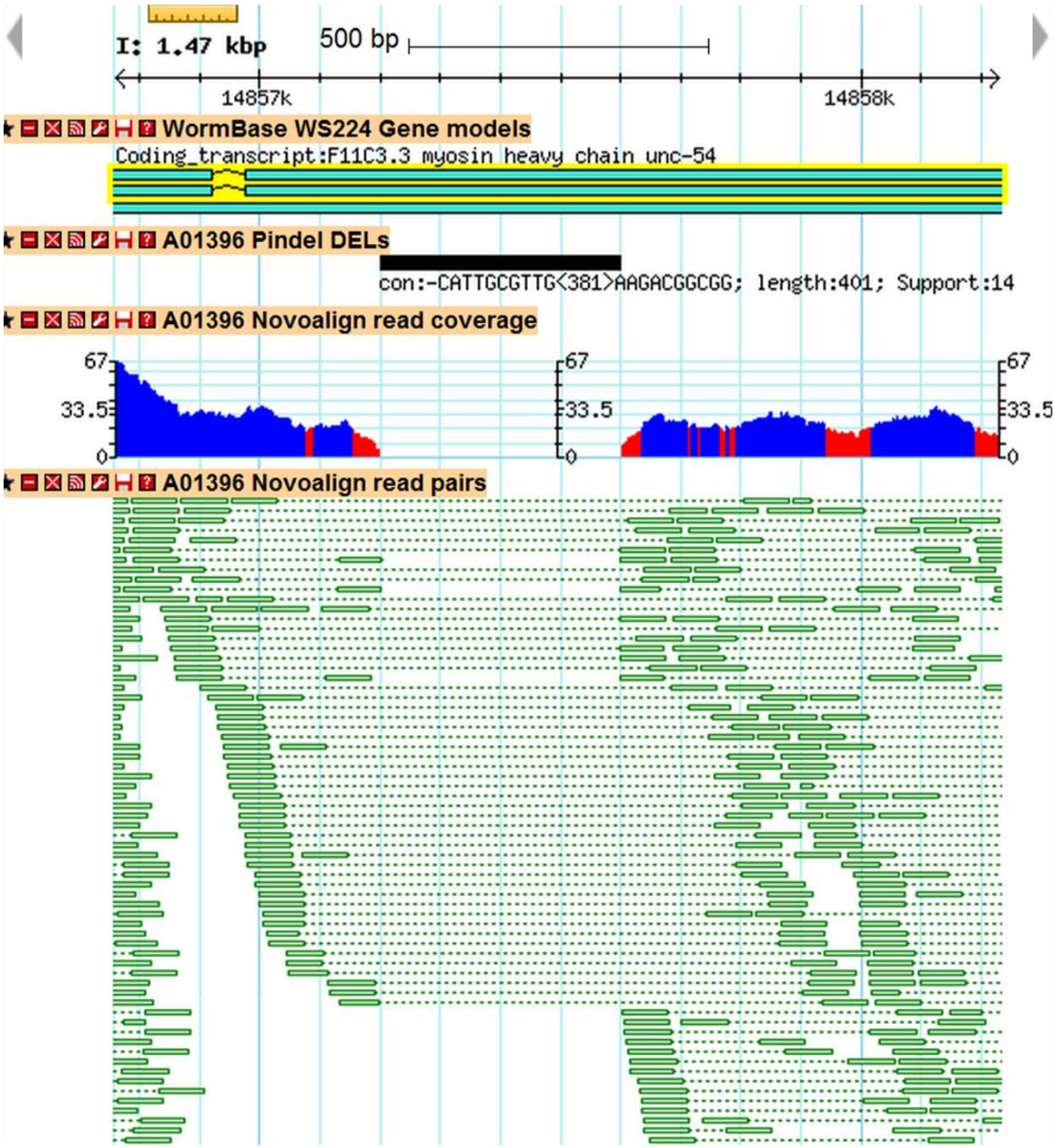


Figure 3.7. Positive Control: unc-54(e190)I Detected by Variant Detection Pipeline in strain A01396

Region shown: I:14,856,760..14,858,229. *unc-54(e190)I* features a 401bp deletion. A homozygous deletion with the correct deleted sequence was detected by Pindel and support by 14 split reads. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Abbreviations: con (consensus of deleted sequence, only the beginning and end 10bp are shown), length (length of deletion), Support (number of split reads that support the deletion), DEL (deletion). Note, not all of the sequencing reads are shown in the figure.

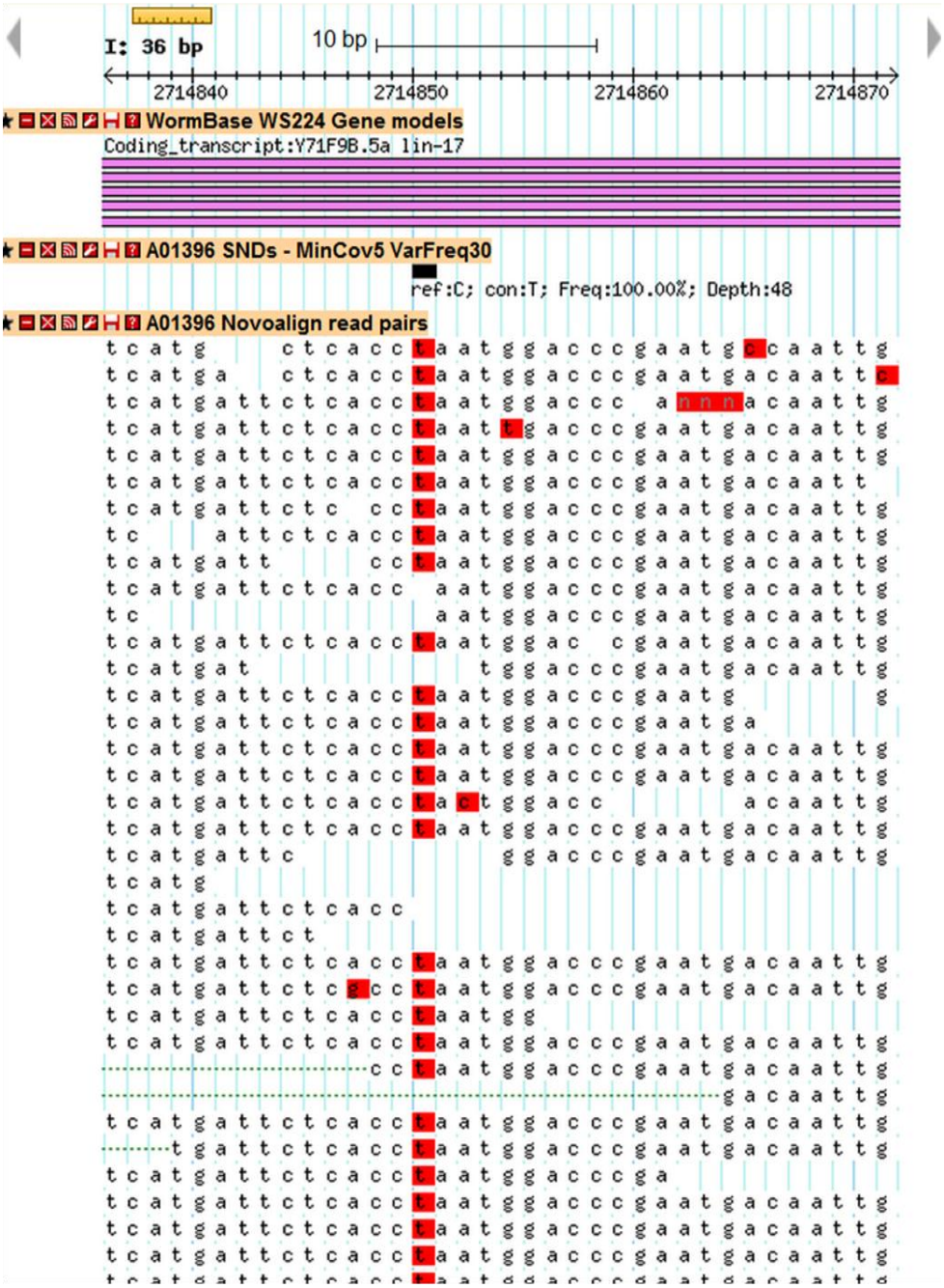


Figure 3.8. Positive Control: lin-17(n671)I Detected by Variant Detection Pipeline in strain A01396

Regions shown: I:2,714,836..2,714,871. *lin-17(n671)I* features a C to T substitution at I:2714850. The homozygous SND that corresponds to *n671* was detected by VarScan at 100% variant frequency and read depth of 48. Sequencing reads were mapped by Novoalign (www.novocraft.com) to WS224 reference *C. elegans* genome and displayed on Gbrowse. Abbreviations: ref (reference base), con (consensus base), Freq (variant frequency), Depth (read depth). Note, not all of the sequencing reads are shown in the figure.

Majority of the homozygous SND variations are validated to be true positives (Table 3.2). Sanger re-sequencing validation and PCR digest validation was done by Kyla Hingwing (Hawkins lab) and mostly on SNDs in strains A00842 and Sup245. 97% (34 out of 35) of the homozygous SNDs, defined as having VarFreq $\geq$ 75%, were confirmed to be true positives. I:2046288 is the only homozygous SND that was confirmed to be a false positive; the SND has a read depth of 114, which is twice as high as the average read depth in A00842. This is an indicator that the SND occurs in a low complexity region. Even though the variations that occur in repeat regions are filtered, there are still examples of variations with very high coverage. Out of the two heterozygous SNDs that were tested (I:1644810 and I:1926286), both of them are verified as false positives. Since the SNDs that were validated to be true positive are all at 100% VarFreq and SNDs that were not validated were lower than 40% VarFreq, most of the true positives likely have a VarFreq much higher than 40% and probably closer to 100% VarFreq.

Majority of the homozygous deletions found in the strain A00842 are validated by Sanger re-sequencing to be true positives (Table 3.2). 75% (3 out 4) of the homozygous deletions, with VarFreq $\geq$ 75%, are validated to be true positives. The deletion at I:1580681 was also detected in strains Sup327, Sup245, A01396, A00842 and Hobert WT. After re-sequencing, the deletion was validated to be a true positive in strains A01396, A00842 and Hobert WT; the deletion was not tested in strain Sup327 and Sup245. The deletion at I:1873302, was confirmed to be a false positive even though the deletion have a VarFreq of 100%. The low read depth (11 reads) and the length of the deletion (12 bp), likely indicates that the deletion occurs in a region that is hard to sequence and therefore is a sequencing error. 100% (1 out of 1) heterozygous deletions tested were validated to be a true positive. The deletion at I:2515851 with a VarFreq of 65.79% was confirmed to be a false positive. Overall, this indicates that deletions detected with VarFreq $\geq$ 75% are likely to be true positives.

Table 3.2. Validated SNDs and small InDels

Chrom	Position	Ref	Consensus	Type	VarFreq (%)	Depth	Confirmed? (Strain tested)
I	991025	G	A	SND	100.00	18	yes (A00842)
I	1484483	-	-A	DEL	75.00	76	yes (A00842)
I	1577273	T	A	SND	100 .00	62	yes (A00842)
I	1580681[a]	-	-ATCCG	DEL	100.00	23	yes (A00842, A01396, Hobert)
I	1644810	A	G	SND	33.33	12	no (A00842)
I	1873302	-	-AAGAA ATTTTTT	DEL	100.00	11	no (A00842)
I	1926286	C	T	SND	40 .00	15	no (A00842)
I	1978049	A	G	SND	100 .00	40	yes (A00842)
I	1984892	G	C	SND	100 .00	28	yes (A00842)
I	2046288	G	A	SND	99.12	114	no (A00842)
I	2151403	G	A	SND	100 .00	32	yes (A00842)
I	2178445	C	T	SND	100 .00	54	yes (A00842)
I	2515851	-	-A	DEL	65.79	38	no (A00842)
I	2687856	C	T	SND	100 .00	18	yes (A00842)
I	3110819	A	G	SND	100 .00	30	yes (A00842)
I	3446896	-	-C	DEL	79.17	48	yes (A00842)
I	3842946	T	A	SND	100 .00	42	yes (A00842)
I	3842947	G	A	SND	100 .00	42	yes (A00842)
I	4164042	G	T	SND	100 .00	29	yes (A00842)
I	8874623[b]	G	T	SND	100 .00	29	yes (A00842)
I	11978602	G	A	SND	100.00	168	yes (Sup245)
I	12109226[c]	G	T	SND	100.00	146	yes (Sup245)
I	12451877	G	A	SND	100.00	145	yes (Sup245)
I	13010426	G	A	SND	100.00	121	yes (Sup245)
I	14937124	C	A	SND	99.20	125	yes (Sup245)
III	13182074[d]	A	G	SND	100.00	33	yes (A00842)
IV	1794828	G	A	SND	100.00	44	yes (A00842)
IV	3876987	G	C	SND	100.00	32	yes (A00842)
IV	7112288	C	G	SND	100.00	14	yes (A00842)
IV	7482513	G	A	SND	100.00	29	yes (A00842)

IV	7558891	G	A	SND	100.00	37	yes (A00842)
IV	7634026	G	A	SND	100.00	24	yes (A00842)
IV	7772648	G	A	SND	100.00	30	yes (A00842)
IV	7896966	C	A	SND	100.00	24	yes (A00842)
IV	8481723	G	A	SND	100.00	61	yes (A00842)
IV	10202816	C	T	SND	100.00	30	yes (A00842)
IV	11930743	G	A	SND	100.00	31	yes (A00842)
IV	12493304	G	A	SND	100.00	57	yes (A00842)
IV	12921589	G	A	SND	100.00	44	yes (A00842)
V	4535526	G	A	SND	100.00	27	yes (A00842)
V	1688659[e]	A	G	SND	100.00	20	yes (A00842)
V	6609962	G	A	SND	100.00	41	yes (A00842)
V	14535102	G	A	SND	100.00	26	yes (A00842)

Note: 'Chrom' refers to chromosome. 'Ref' refers to reference base affected. 'SND' refers to single nucleotide difference. 'DEL' refers to deletion. 'Depth' refers to the total read depth at the position indicated.

a Deletion at I: 1580681 predicted in strains Sup327, Sup245, A01396, A00842 and Hobert WT. Deletion confirmed to be a true positive in strains, A01396, A00842 and Hobert WT. Deletion was not tested in strains Sup327 and Sup245.

b SND at I: 8874623 predicted in strains Sup245 and A00842. SND confirmed to be a true positive in strain A00842. SND was not tested in strain Sup245.

c SND at I:12109226 predicted in strains Sup327, Sup245 and A01396. SND confirmed to be true positive in strain Sup245. SND was not tested in strains Sup327 and A01396.

d SND at III:13182074 predicted in strains Sup327, Sup245, A01396 and A00842. SND confirmed to be a true positive in strain A00842. SND was not tested in strains Sup327, Sup245 and A01396.

e SND at V:1688659 predicted in strains Sup327, Sup245 and A00842. SND confirmed to be a true positive in strain A00842. SND was not tested in strains Sup327 and Sup245.

Candidates for *Sup327* are listed in Table 3.3. *Sup327* specific variations occur in strains Sup327 and may be present in strains A01396 and/or Sup245, provided that *Sup327* and *Sup245* are allelic. The variations in the candidates are not predicted to be in the other strains, namely A00842, Hobert and Horvitz. The candidates listed also do not occur in repeat regions and do have an effect on the protein coding sequence of a coding transcript. In addition, all candidates listed are affected by a SND. The three candidates listed are: *col-63*, W05H12.1 and W04A8.6. W05H12.1 seems to be very promising as (1) the variation is a G>A substitution which is characteristic of EMS mutagenesis (Flibotte *et al.*, 2010), (2) the variation has 100% VarFreq, and (3) the read depth is close to the average read depth for the Sup32 strain. In addition, according to WormBase, W05H12.1 is homologous to chromatin assembly factor. Previously, Nakano and colleagues found that the chromatin assembly factor-1 (CAF-1) is required for neuronal asymmetry, specifically in the MI-e3D motor neuron. CAF-1 protein complex is an evolutionarily conserved histone chaperone that mediates nucleosome formation. CAF-1 deposits histone H3-H4 proteins onto replicating DNA and mutations in one of the *C. elegans* histone H3 genes has been found to eliminate MI asymmetry (Nakano *et al.,* 2011). *col-63* is also a good candidate since (1) the variation is a C>T change which is characteristic of EMS mutagenesis, (2) the variation has 100% VarFreq, and (3) the read depth is also close to the average read depth. According to Gene Ontology (GO), *col-63* is a structural constituent of the cuticle. Seam cells are lateral hypodermal cells that function by secreting cuticle and mediating the elongation of the embryo by generating a contractile force (Altun and Hall, 2009). It has been found that seam cell contacts are important for epidermal differentiation. In particular, expression of *nhr-25* in seam cells is important for seam cell elongation after asymmetric divisions. Defects in seam cell elongation affect the subsequent fate of the seam-cell anterior daughters (Silhankova *et al*., 2005). Since *col-63* is a structural constituent of cuticle, which is likely secreted by the seam cell, perhaps *col-63* may play a role in epidermal differentiation. The least likely candidate for *Sup327* is W04A8.6, since the missense mutation does not consist of a C>T or G>A change. W04A8.6 is homologous to a structural component of the synaptonemal complex. The synaptonemal complex is a structure that binds and keeps the homologous chromosomes together during the pachytene stage of the cell cycle, and the complex remains intact until later in meiotic prophase (Colaiacovo *et al.*, 2003). However, there are no studies that suggest that the

synaptonemal complex is associated with asymmetric neuroblast division. Currently, none of the three candidate variations have undergone validation by PCR digest or Sanger re-sequencing.

Table 3.3. Summary of Candidates for *Sup327* on Chromosome I in strain Sup327

Pos	Change	VarFreq(%)	Depth	Trans	Gene	Type	Notes[b]
8243822[a]	C>T	100.00	49	ZK265.2	*col-63*	MIS	Collagen, GO: structural constituent of cuticle, integral to membrane
13423110	G>A	100.00	20	W05H12.1	*n/a*	MIS	HOM: chromatin assembly factor, Huntingtin-associated protein, GO: integral to membrane, EXP: dorsal/ventral nerve cord, head/tail neurons
13855278	G>A	100.00	17	W04A8.6	*n/a*	MIS	HOM: transverse filament protein of synaptonemal complex

Note: All candidate genes listed were interrupted by homozygous SNDs that do not occur in WT strains (Hobert and Horvitz N2 strain) and do not overlap with repeat regions, as defined by Repeat Masker in WS224 reference *C. elegans* GFF3 file. Homozygous SNDs is defined as having a variant frequency $\geq$ 75%. Variations are all present in Sup327 strain and maybe present in the A01396 and/or Sup245, but do not occur in the A00842 strain. Abbreviations: MIS (Missense), GO (Gene Ontology), HOM (Homologous) and EXP (Expression pattern).

a I: 8243822 variation also occur in A01396 strain: C>T change, 100%VarFreq, Cov 14.

b Information collected from WormBase (http://www.wormbase.org)

Candidates for *Sup245* are listed in Table 3.4. The candidates listed are present in the strain Sup245 and may be present in A01396 and/or Sup327, provided that Sup245 and Sup327 are allelic. The candidates are not predicted to be in the strains A00842, Hobert and Horvitz. The majority of the candidate variations feature a G>A substitution, which is characteristic of EMS mutagenesis. It is interesting that there is a missense mutation of T>C found in the gene *pop-1* in the strain Sup245. *pop-1* is a TCF/LEF protein that activates gene transcription following Wnt signalling. However, the mutation in *pop-1* may not be *Sup245* as *Sup245* has been mapped to the right of chromosome I, from candidate C35E7.4 to candidate *kin-1*(personal communication from Kyla Hingwing). Furthermore, the substitution found in *pop-1* is not characteristic of

EMS mutagenesis. *kin-1* maybe an ideal candidate since KIN-1 is a conserved serine/threonine kinase that phosphorylates histones and is involved in cell polarity, microtubule stability or cell cycle regulation. It has been found previously that kin1p in yeast accumulates asymmetrically at the cell cortex and affects cell shape (La Carbona *et al*., 2004). In addition, *par-1* (partitioning defective) in *C. elegans* is a serine/threonine kinase that helps regulate cell polarity and asymmetric cell division by ensuring polar distribution of cell fate determinants. Loss of PAR-1, or any of the PAR proteins, led to defects in asymmetric cell division and embryonic lethality (Spilker *et al.*, 2009). *kin-1* features a C>A substitution that have been validated (Table 3.2) in strain Sup245, but the substitution is not typical of EMS mutagenesis. *kin-1* also has a VarFreq of 99.2% and a read depth of 125 which is close to the average read depth for strain Sup245. The variation in candidate C54C8.4 has also been confirmed to be a true variation in strain Sup245. The variation features a G>A substitution that is characteristic of EMS mutagenesis. In addition, the variation has a VarFreq of 100% and a read depth of 145 that is also close to the average read depth for strain Sup245. C54C8.4 is predicted to be located on the membrane. There may be a chance that C54C8.4 is involved in a signal transduction cascade.

Table 3.4. Summary of Candidates for *Sup245* on Chromosome I in strain Sup245

Pos	Change	VarFreq (%)	Depth	Trans	Gene	Type	Notes[b]
2825128	T>C	78.30	894	W10C8.2	*pop-1*	MIS	HMG box-containing protein, TCF/LEF family of Transcription factor, GO: Wnt receptor signaling pathway, apoptosis, mesodermal cell fate determination
2924334	G>A	81.44	87	Y71F9AL.18	*pme-1*	MIS	Poly(ADP-ribose) polymerase, GO: cellular response to DNA damage stimulus
6445889	G>A	100.00	132	T23H2.3	*n/a*	MIS	GO: ATP binding, DNA binding, helicase activity, HOM: transcription termination factor 2

7499261	G>A	99.28	138	C26C6.1	*pbrm-1*	MIS	Interacts with components of EGF.RAS signaling pathway, HOM: chromatin remodeling, transcriptional regulation, GO: growth, embryonic development
8329541	G>A	100.00	126	F52B5.6	*rpl-25.2*	RMIS	Large ribosomal subunit, GO: apoptosis, collagen/cutculin-based cuticle development, embryonic development
8728826	G>A	100.00	145	C36B1.3	*rpb-3*	MIS	RNA Polymerase II, GO: embryonic development
9140042	G>A	99.40	168	ZK858.6	*n/a*	RMIS	HOM: protein with role in cellular adhesion, filamentous growth and endosome-to-vacuole sorting, GO: integral to membrane, EXP: intestine
9548609	G>A	100.00	108	C36F7.2	*n/a*	MIS	HOM: myotrophin homolog, ankyrin repeat, GO: protein binding
9861275	G>A	100.00	160	K10C3.4	*n/a*	MIS	GO: integral to membrane, EXP: vulva, rectal epithelium, neuron, pharyngeal cell, seam cell
10826002	G>A	99.33	150	C35E7.5	*n/a*	MIS	HOM: Huntingtin-associated protein, protein involved in nonsense-mediated mRNA decay
10833785	G>A	100.00	145	C35E7.4	*n/a*	MIS	EXP: head/tail neuron, vulval muscle, intestine
12451877	G>A	100.00	145	C54C8.4	*n/a*	MIS	GO: integral to membrane
13859444[a]	G>A	100.00	59	W04A8.6	*n/a*	MIS	HOM: transverse filament protein of synaptonemal complex
14490673	A>T	100.00	138	Y105E8A.20	*n/a*	MIS	HOM: Methionyl-tRNA synthetase, mitochondrial, GO: embryonic development, methionyl-tRNA aminoacylation, reproduction

14937124	C>A	99.20	125	ZK909.2	*kin-1*	MIS	Serine/threonine protein kinase, phosphorylate histone H2B, GO: oviposition, morphogenesis of epithelium, larval development, EXP: ventral cord neuron, intestine, nervous system, excretory cell

Note: All candidate genes listed were interrupted by homozygous SNDs that do not occur in WT strains, such as the Hobert and Horvitz N2 strain, and do not overlap with repeat regions, as defined by RepeatMasker in WS224 reference *C. elegans* GFF3 file. Homozygous SNDs is defined as having a variant frequency $\geq$ 75%. Variations are all present in Sup245 strain and maybe present in the A01396 and/or Sup327 strain, but do not occur in the A00842 strain. Abbreviations: MIS (Missense), RMIS (Radical Missense), GO (Gene Ontology), HOM (Homologous) and EXP (Expression pattern).

a I: 13859444 variation also occur in A01396 strain: G>A change, 100%VarFreq, Cov 8.

b Information collected from WormBase (http://www.wormbase.org)

Candidates for *Sup305* are listed in Table 3.5. The candidate variations listed does not occur in other strains such as Sup327, Sup245, A01396, Hobert and Horvitz. There are 33 candidates on the whole of chromosome IV. Most of the candidates in the list seem to consist of G>A substitutions, which is expected. Through personal communications with Kyla Hingwing, she has narrowed down the list of candidates using SNP mapping to two genes, (1) *plp-1* and (2) *hcf-1*. Both variations have been validated to be true variations in strain A00842. The missense variation in *plp-1* consist of a G>A substitution, that is expected of EMS mutagenesis. The VarFreq is at 100% but the read depth of 24 is approximately half of the average read depth in A00842. *plp-1* codes for a PUR alpha transcriptional activator that is important for embryonic development. PLP-1 has already been found to interact with END-1 and POP-1 in Wnt/MAP kinase signaling paths for endoderm differentiation (Witze *et al.*, 2009). *hcf-1* also features a G>A missense mutation at 100% VarFreq and read depth of 30. *hcf-1* is a transcriptional regulator that functions in cell cycle progression and mitotic histone modification. Herpes simplex virus (HSV) host cell factor-1 (HCF1) has been found to be a component in multiple chromatin modulating complexes (Peng et al., 2010). Although chromatin factors have not yet been found to function in asymmetric cell division, it is likely that

chromatin factors are involved in the process (Cui and Han, 2007). Both genes are excellent candidates for Sup305.

Table 3.5. Summary of Candidates for *Sup305* on Chromosome IV in strain A00842

Pos	Change	VarFreq (%)	Depth	Trans	Gene	Type	Notes[a]
1818697	C>T	100.00	14	Y38C1BA.3	col-109	RMIS	Cuticular collagen
3045137	C>T	100.00	28	Y67D8C.5	eel-1	MIS	Modifier of transcription factor efl-1/E2F, involved in embryonic patterning, HOM: Hect E3 ubiquitin ligase, GO: asymmetric protein localization involved in cell fate determination, EXP: head/tail neuron, cholinergic/GABAergic neuron, intestine, pharynx
3082642	C>T	97.14	35	Y67D8C.9	n/a	MIS	HOM: Aminopeptidase, GO: proteolysis
3876987	G>C	100.00	32	F37C4.6	n/a	MIS	HOM: Pyridine nucleotide-disulfide oxidoreductase domain-containing protein, GO: oxidation reduction
5471042	G>A	100.00	29	T11F8.3	rme-2	MIS	LDL receptor, required for yolk uptake during oogenesis, EXP: developing oocytes, localizes near cell surface
6791314	G>A	100.00	33	C17H12.14	vha-8	MIS	vacuolar proton-translocating ATPase, GO: embryonic development, molting cycle, necrotic cell death, EXP: hypodermis, reproductive system, excretory cell, vulval muscle, head, intestine
7102104	C>A	100v	46	ZC477.5	n/a	SPL	HOM: Ribonuclease Zc3h12a-like
7341738	G>A	100.00	37	Y40C5A.3	n/a	MIS	HOM: cell surface glycoprotein, GO: embryonic development

7482513	G>A	100.00	29	F55G1.4	rod-1	MIS	HOM: Kinetochore-associated protein, GO: embryonic development
7634026	G>A	100.00	24	F45E4.2	plp-1	MIS	protein that contains 3 PUR repeats, HOM: transcriptional activator, GO: embryonic development, DNA binding, EXP: intestine, body wall muscle, nervous system, ventral nerve cord, head/tail neurons, nuclei of blastomeres
7772648	G>A	100.00	30	C46A5.9	hcf-1	MIS	associates with histone modification enzymes, plays role in cell cycle progression and mitotic histone modification, determines adult lifespan. HOM: human host cell factor, transcriptional regulator, GO: cell cycle progression
7896966	C>A	100.00	24	C55F2.2	ilys-4	MIS	Invertebrate Lysozyme, GO: reproduction, lysozyme activity
8430396	C>T	100v	27	T26A8.4	n/a	MIS	HOM: protein that regulates transcription and RNA degradation, GO: embryonic development, hermaphrodite genitalia development, larval development
8481723	G>A	100.00	61	C28C12.5	spp-8	RMIS	Saposin-like protein family
8836921	C>A	100.00	46	T28C6.7	n/a	RMIS	HOM: myosin-like protein associated with nuclear envelope, EXP: intestine, nervous system, pharynx, hypodermis, body wall musculature
9590942	G>A	100.00	24	F27C8.5	n/a	NON	HOM: BTBVPOZ-like, GO: protein binding
9865251	+GC	100.00	29	K07F5.15	n/a	INS	HOM: Protein kish-B, GO: integral to membrane

10071628	C>A	100.00	24	W01B6.3	n/a	MIS	HOM: Major facilitator superfamily domain-containing protein, GO: transmembrane transport
10125800	C>T	100.00	59	K08F4.1	n/a	SPL	HOM: Chromosome transmission fidelity protein, GO: cellular response to DNA damage stimulus
10202816	C>T	100.00	30	K04D7.5	gon-4	RMIS	Nuclear protein required for gonadogenesis in both sexes, proper germline and vulval development, EXP: somatic gonadal cells
10814004	C>T	100.00	31	R13.4	miz-1	MIS	MIZ-type zinc finger putative transcription factor, GO: larval development, EXP: pharynx, pharyngeal-intestinal valve, anal depressor muscle, reproductive system, nervous system
11202960	T>A	92.86	14	R07H5.11	n/a	RMIS	GO: positive regulation of growth rate
11202960	+CTTTCAC	92.86	14	R07H5.11	n/a	INS	GO: positive regulation of growth rate
11225687	G>A	100.00	16	C29F4.1	col-125	MIS	HOM: cuticle collagen, GO: body morphogenesis, lipid storage, larval development, receptor-mediated endocytosis
11371000	T>C	100.00	18	T01G1.3	sec-31	MIS	HOM: component of Sec13p-Sec31p complex of COPII vesicle coat, vesicle transport
11412645	G>T	100.00	39	F22B3.4	n/a	MIS	HOM: glucosamine-fructose 6-phosphate aminotransferase, GO: apoptosis, carbohydrate metabolic process, cell division, embryonic development, larval development

12102555	G>A	100.00	26	F11A10.4	mon-2	MIS	Arf-GEF-like protein, has been identified as suppressor of mutationsi in ipla-1, HOM: predicted to function in endosome to Golgi retrograde transport
12493304	G>A	100.00	57	T04A11.6	him-6	MIS	RecQ-like ATP-dependent DNA helicase, required for normal levels of recombination during meiosis, GO: chromosome organization/segregation, DNA repair/replication
12921589	G>A	100.00	44	C32H11.4	n/a	MIS	HOM: epoxide hydrolases, CUB-like domain
14203565	G>A	100.00	61	F02H6.3	n/a	RMIS	n/a
15924539	G>A	100.00	55	Y105C5B.8	n/a	MIS	GO: integral to membrane
16775907	G>A	100.00	15	Y43D4A.5	n/a	SPL	HOM: cell surface glycoprotein

Note: All candidate genes listed were interrupted by homozygous SNDs that do not occur in WT strains, such as the Hobert and Horvitz N2 strain, and do not overlap with repeat regions, as defined by RepeatMasker in WS224 reference *C. elegans* GFF3 file. Homozygous SNDs is defined as having a variant frequency $\geq$ 75%. Variations are all present in A00842 strain, but do not occur in the Sup327, Sup245, and A01396 strains. Abbreviations: MIS (Missense), SPL (Splice Junction), NON (Nonsense), RMIS (Radical Missense), INS (Insertion), GO (Gene Ontology), HOM (Homologous) and EXP (Expression pattern).

a Information collected from WormBase (http://www.wormbase.org)

3.4. Conclusion and Future Work

The variation detection pipeline used in this analysis is sensitive enough to detect all of the positive control variations. The presence of *dsh-2(or302)* was detected in all four suppressor strains using Pindel. The presence of *unc-54(e190)* was detected in strain A01396 by Pindel. Finally, the presence of *lin-17(n671)* was detected in strain A01396 by VarScan.

A high percentage of the variations that were tested are validated to be true positives using Sanger re-sequencing and PCR digest. 97% of the homozygous SNDs and 75% of the homozygous deletions that were tested are true positives. To optimize

the variation detection pipeline, future work can include determining the sensitivity and specificity of the variation detection pipeline. By validating a random set of SNDs and InDels, we can optimize the filtration parameters to maximize the number of true positives and minimize the number of false negatives and true negatives.

Candidate suppressors detected through the variation pipeline include genes that are known to act in Wnt signaling, act with Wnt signaling substrates, or are homologous to proteins that are known to be involved in asymmetric cell division. In the list of *Sup327* candidates, W05H12.1 and *col-63* are good potential candidates. W05H12.1 is homologous to chromatin assembly factors that are known to be involved in asymmetric cell division. *col-63* is a component of cuticle and involved with seam cells that are important for epidermal differentiation. In the list of *Sup245* candidates, *kin-1* and C54C8.4 are good potential candidates. *kin-1* is a conserved serine/threonine kinase that may be involved in asymmetric cell division. Although not much is known about C54C8.4, since the transcript is predicted to be on the membrane, it may be involved in signal transduction cascade, such as Wnt signaling. In the list of *Sup305* candidates, *plp-1* and *hcf-1* are good potential candidates. PLP-1 has been previously found to interact with Wnt/MAPK signaling components such as END-1 and POP-1 in endoderm differentiation. Furthermore, HCF-1 is a component of chromatin modulating complexes that are likely to be involved in asymmetric cell division.

3.5. Discussion

Novoalign by default, suppresses most reads with multiple alignment through the '-r None' parameter, however, there are still a number of reads that are falsely misaligned. Yu and colleagues have tested Novoalign alignment accuracy using CpG island simulation data, which contain a lot of repetitive regions. They found that alignment accuracy is substantially improved when multiple alignments are suppressed, however, there are still some false alignments which is also seen in other aligners (Yu *et al.*, 2012). This explains how after browsing through the read alignment on Gbrowse, there are still instances of regions in the genome with abnormally high read coverage. These regions contain many reads that map to multiple locations in the genome which were not filtered out.

BAQ calculation was disabled in 'mpileup' in order to reduce the number of false negative variations. BAQ is the Phred-scaled probability of a read base being misaligned and enabling BAQ calculation can dramatically increase specificity, which is recommended by the author (personal communication). However, since the focus is on homozygous variations that affect protein coding transcripts, only a small subset of the variations will be considered. As a result, it is important that the number of false negative variations is minimized, so as not to lose any potential gene candidates for the suppressors. In addition, Spencer Myrtle has previously determined that the number of SNDs detected using 'mpileup –B' and 'pileup' is similar (data not shown).

Strain Sup327 had a very low percentage of reads, 58.84%, that mapped to the reference genome when compared to other strains. The reason for the low number of reads mapped could be due to the quality of the read data. One clue is the number of reads generated for Sup327 and Sup245. Both strains were sent for Illumina/Solexa sequencing at the same time. The lower number of reads generated for Sup327 compared to Sup245 may indicate sequencing difficulties for Sup327. Yu and colleagues also found that for low quality data set and without data quality improvements such as read trimming, Novoalign aligned less than 50% of the reads. With read trimming, however, the percentage of reads mapped increases to approximately 70%. The trend is similar in the other aligners tested, but the differences are more pronounced using Novoalign. It may be worthwhile to trim off low quality ends of reads before alignment since data quality improvements have larger effect on Novoalign (Yu *et al.*, 2012)

Many filtration thresholds and parameters were used to filter the variations detected from all four strains. The filtration threshold for MinCov is set at 5, which is very lenient, considering the average read coverage of the four strains ranges from 46 to 146 reads per base (Table 3.1). The lenient MinCov is set to minimize false negatives (FN). VarFreq is first set to 30% in order to detect heterozygous variations since homozygous variations in certain strains may be manifested as heterozygous variations in other strains. Theoretically, heterozygous variations should occur at a VarFreq of 50%, but to account for chances of sequencing errors, the threshold for VarFreq was lowered to 30%. After compilation of all the variations that occurred in all four strains, as well as the characteristics of each variations, suppressor candidates were chosen based

on whether they occur in protein coding genes, homozygous in at least one strain, do not occur in the wild type strains, and do not occur in repeat regions. Since the goal of the project is to identify genes functioning with DSH-2, ideal suppressors would have variations that affect another protein coding gene. It is expected that the suppressors are homozygous, as a result, VarFreq was further set at 75%. However, to account for heterozygosity differences between strains, the variation only needs to be homozygous in at least one of the four strains. VarFreq threshold for homozygous variations have previously been used by Cord and colleagues (Drogemuller *et al.*, 2010). It is also expected that the suppressors do not occur in the Hobert and Horvitz wild type strains. As a result, variations that are shared with the two wild type strains are filtered out. Finally, in order to filter out variations that are detected in high coverage repetitive regions, variations that occur in repeat regions, as defined by Repeat Masker were filtered out. Repeat regions are regions that are occupied by short exact tandem and inverted repetitive elements, and/or long repetitive elements. Repetitive regions often consist of reads that have low mapping qualities and therefore are often areas of read misalignments. As such, variations detected in such regions are not reliable.

Since the suppressors are dominant, there is a slight chance that the suppressor may be manifested as a heterozygous variation. However, due to the hermaphroditic nature of *C. elegans*, it is likely that the suppressors are homozygous in the strains that are sequenced. If none of the homozygous variations turns out to be the suppressors, it will be worthwhile to look at heterozygous variations and variations that do not affect protein coding sequence.

4. Bioinformatic Analysis of *C. briggsae* genome

4.1. Background

Caenorhabditis briggsae is the most extensively studied sister species of *C. elegans.* Although the two species diverged from a common ancestor roughly 100 million years ago, both species share similarities in morphology, behaviour and hermaphroditic life style. Furthermore, the similarity in morphology suggests that the two species share conserved genes and pathways (Gupta *et al.,* 2007; Stein *et al*., 2003; Coghlan and Wolfe, 2002). After the *C. briggsae* genome was sequenced in 2003 (Stein *et al.,* 2003), following the sequencing of *C. elegans* genome in 1998 (*C. elegans* Sequencing Consortium, 1998), it was found that there were also extensive similarities among the two genomes. The two species had similar number of chromosomes, protein coding and non-protein coding genes and similar genome size (Gupta *et al*., 2007).

The *C. briggsae* reference genome was generated mostly by paired-end whole-genome shotgun (WGS) sequencing reads in combination with a physical map produced by a high-throughput fingerprinting scheme. The Phusion assembler was used to align the WGS Sanger sequencing reads. Phusion assembled the Sanger reads into contigs base on overlap information, and then further into super-contigs based on read pair information. The sequence assembly was then integrated with the physical map. Finally, the previously finished 12MB of clone-based sequence was also integrated with the whole genome assembly to produce the draft assembly (Mullikin and Ning, 2003; Stein *et al.*, 2003)

The *C. briggsae* genome sequencing project also gave valuable insights towards the understanding of *C. elegans* gene structures. The annotation of *C. briggsae* genome helped the discovery of almost 1300 new genes in *C. elegans* (Stein *et al.,* 2003). Furthermore, with the possibility of many more potential discoveries, for example identification of cis-regulatory elements and evolution of gene families, the *C. briggsae*

genome and annotations was immediately established as an excellent platform for comparative genomics (Gupta and Sternberg, 2003). As a result, high quality genome annotations in both *C. briggsae* and *C. elegans* are important for many downstream analyses.

However, annotation of *C. briggsae* genome lagged behind the annotations for *C. elegans* genome. The *C. briggsae* gene set was generated entirely by computation gene finding (Gupta *et al.,* 2007). The *C. elegans* gene set, on the other hand, was generated by computation gene finding (Spieth and Lawson, 2006), followed by extensive manual curation using evidence from expression sequence tags (ESTs) (Kohara, 1996; Shin *et al.,* 2008), open reading frame sequence tags (OSTs) (Reboul *et al.,* 2003; Lamesch *et al.,* 2004; Wei *et al.,* 2005), serial analysis of gene expression (SAGE) tags (Ruzanov *et al.,* 2007; Nesbitt *et al*., 2010; Ruzanov and Riddle, 2010), RNA-seq results (Hillier *et al*., 2009; Allen *et al.,* 2011), and translational expression evidence (Shim and Paik, 2010). In addition, a larger number of labs are studying *C. elegans* than *C. briggsae,* which provides more data for accurately curating *C. elegans* gene models. While the *C. elegans* gene set has been improving and updated in WormBase every three weeks, the *C. briggsae* gene set had remained unchanged for three years following the publication of the *C. briggsae* genome analysis (Gupta *et al.,* 2007).

As a result, Uyar *et al.* set to validate and further improve the *C. briggsae* gene annotations through RNA-seq analysis. The transcriptome data consists of 42 bp paired-end reads sequenced from L1 and mixed staged worms; the strain that was sequenced, AF16, is the same strain used to generate the reference *C. briggsae* genome (reviewed in Stein *et al.*, 2003). First, Uyar *et al.* applied a homology-based gene predictor, genBlastG (She *et al.*, 2011), to computationally annotate the *C. briggsae* genome. Using *C. elegans* gene models as query, the predictor helped improve the *C. briggsae* gene models; the models with the higher protein sequence level similarities (PID) were kept. Next, with the RNA Sequencing (RNA-seq) data, which provides details of exon-intron boundaries when mapped to the *C. briggsae* genome, the data can validate and refined the improved gene model set. RNA-seq utilizes the massively parallel sequencing of complementary DNAs (cDNAs) that are generated from the RNA of interest. The reads that are generated can be mapped to a reference

genome in order to construct a transcriptome map (Nagalakshmi *et al*., 2010). The combined approach of homology gene prediction and RNA-seq based improvements have resulted in the first validated *C. briggsae* gene set of 23,159 genes, of which 7347 genes (33.9% of all genes with introns) have all of their introns confirmed. However, many introns predicted from the RNA-seq data cannot be used to revise existing introns, because their integration would cause frame shifts. Uyar *et al.* observed 636 of such Illumina introns (un-incorporated Solexa introns) that are supported by two or more independent reads. The un-incorporated Solexa introns cannot be incorporated into the *C. briggsae* genome due to frame shifts that would be produced in the gene model (Figure 4.1). Uyar and colleagues also noticed that there are a lot of frame shifting differences, such as insertions and deletions (InDels) that were detected following the alignment of the transcriptomic reads to the reference *C. briggsae* genome (Figure 4.1) (Uyar *et al.*, 2012).

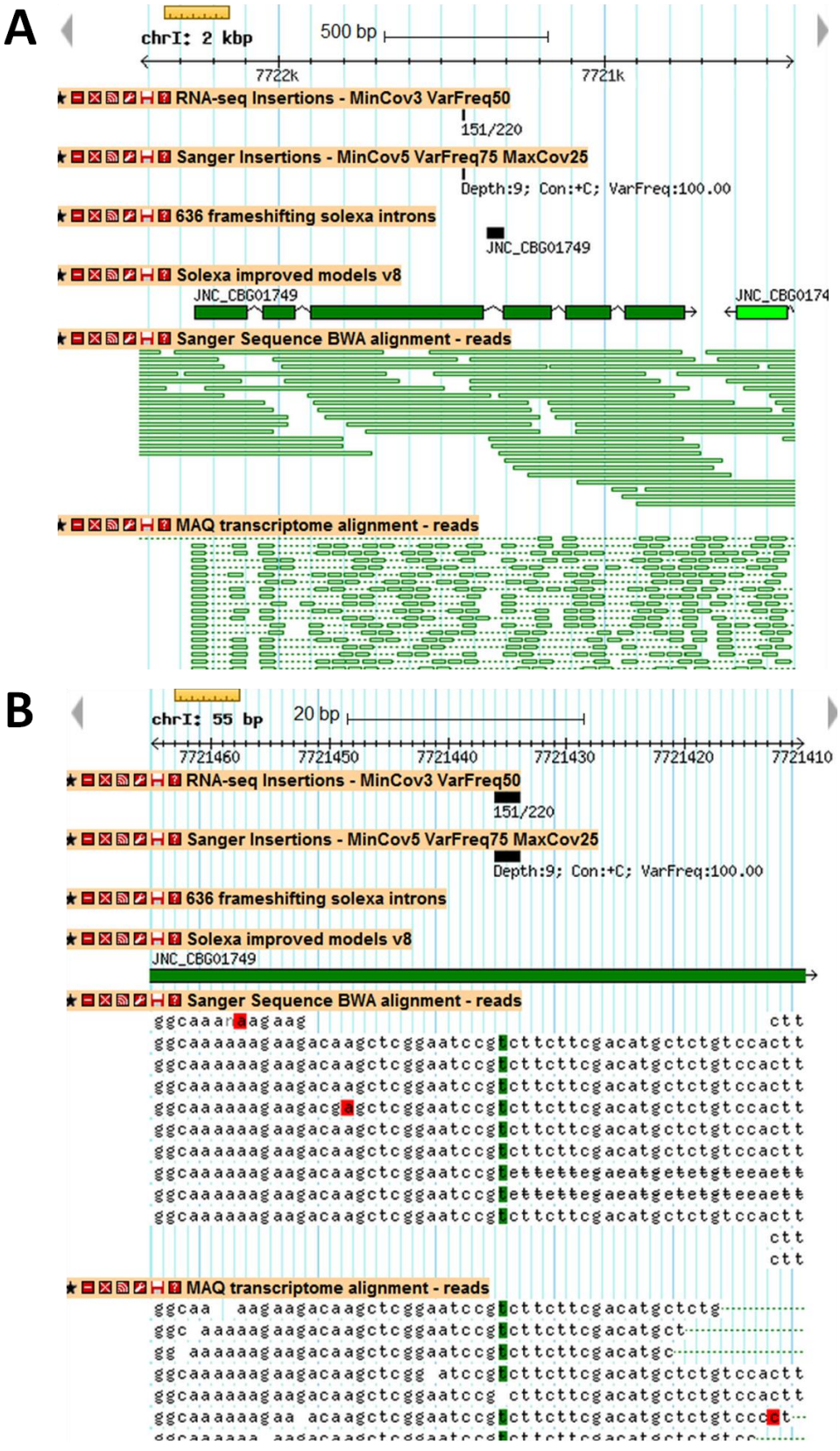

A
chrI: 2 kbp
500 bp
7722k
7721k
RNA-seq Insertions - MinCov3 VarFreq50
151/220
Sanger Insertions - MinCov5 VarFreq75 MaxCov25
Depth:9; Con:+C; VarFreq:100.00
636 frameshifting solexa introns
JNC_CBG01749
Solexa improved models v8
JNC_CBG01749
JNC_CBG0174
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads
B
chrI: 55 bp
20 bp
7721460
7721450
7721440
7721430
7721420
7721410
RNA-seq Insertions - MinCov3 VarFreq50
151/220
Sanger Insertions - MinCov5 VarFreq75 MaxCov25
Depth:9; Con:+C; VarFreq:100.00
636 frameshifting solexa introns
Solexa improved models v8
JNC_CBG01749
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads

Figure 4.1. Frame shifting un-incorporated Solexa intron in close proximity to a frame shifting exonic insertion

(A) Region displayed: chrI:7,720,422..7,722,421. The figure shows an example of an insertion that is found in the transcriptomic alignment and genomic alignment. The transcriptomic alignment has 151 out of 220 reads supporting the insertions. The genomic alignment has 9 reads supporting the insertion of a C. The insertion that are shared between the genomic and transcriptomic alignment is affecting the exon that is directly upstream of an un-incorporated solexa intron. (B) Region displayed: chrI:7,721,410..7,721,464. The figure shows a zoomed in view of the transcriptomic and genomic alignments in (A). The location of the insertion is highlighted in green in both the genomic and transcriptomic alignments. This panel highlights the agreement of the reads in the genomic and transcriptomic alignment. The 'Solexa improved models v8' track displays the gene model set improved by genBlastG v135 and the 42bp RNA-seq. Abbreviations: Depth (read depth), Con (Consensus variation) and VarFreq (variant frequency). Note, not all of the reads in the transcriptome alignment are shown. In addition, the region displayed has been 'flipped' in GBrowse.

If the transcriptome sequencing reads were generated from the same strain used in the *C. briggsae* genome assembly, why are there so many frame shifting differences detected? The first hypothesis is that there are genome assembly errors. The second hypothesis is that the strain used to generate RNA-seq data is genetically different from the strain used in the assembly of the *C. briggsae* genome. The strains can be genetically different due to genetic drift or spontaneous accumulation of mutations throughout generation to generations (Flibotte *et al.*, 2010). To test this hypothesis, WGS Sanger sequencing reads used in the *C. briggsae* genome assembly are re-aligned to the same reference *C. briggsae* genome as was done with the transcriptomic sequencing reads. Variations are detected from both the RNA-seq alignment and the Sanger sequence alignment and compared (Figure 4.2). If the differences between the two strains are due to genetic drift, we would expect to see very little overlap between the InDels detected in the transcriptomic and the genomic data. On the other hand, if there is great overlap between the InDels detected in the transcriptomic and genomic data that would be an indication that the reference genome still contains many assembly errors that would have to be corrected (Figure 4.1). As such, gene models that are affected the assembly errors would have to be modified and improved yet again.

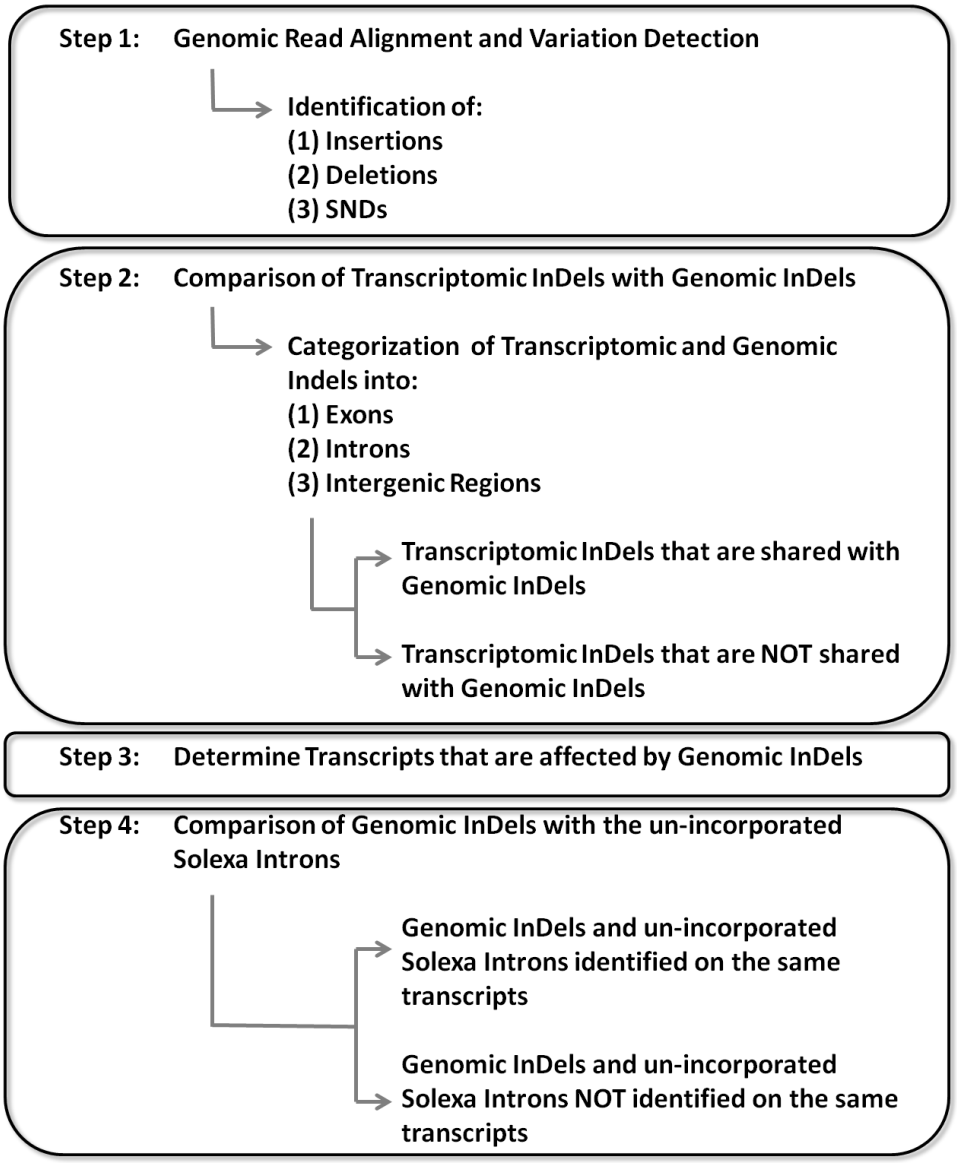


Figure 4.2. C. briggsae Analysis Flowchart

In step 1, genomic Sanger sequencing reads (Stein *et al.*, 2003) used in *C. briggsae* genome sequencing project were aligned by BWA (Li and Durbin 2009) to the WS215 reference *C. briggsae* genome. SNDs from genomic sequence alignment were filtered using the parameters: MinCov $\geq$ 5, VarFreq $\geq$ 75%, MaxCov $\leq$ 25 and SNPQ $\geq$ 20. Small InDels were filtered using the parameters: MinCov $\geq$ 5, VarFreq $\geq$ 75% and MaxCov $\leq$ 25. In addition, deletions of 'N' ambiguous base(s) were filtered out from the deletion set. In step 2, genomic InDels and previously filtered transcriptomic InDels are categorized into exons, introns and intergenic regions. Furthermore, the genomic and transcriptomic InDels are compared to each other, in order to determine (1) whether the two strains used in the two data set are identical and (2) if there are genome assembly errors in the reference genome. In step 3, the number of transcripts that are affected by the genomic InDels are determine. Finally, in step 4, to determine whether genome assembly errors are responsible for the un-incorporation of the Solexa introns, the transcripts that they affect are compared together.

Finally, what is preventing the incorporation of the 636 un-incorporated Solexa introns? The Solexa introns were predicted from the RNA-seq data, yet their incorporation would cause frame shifting changes to the gene models. The hypothesis is that genome assembly errors are preventing the incorporation of many Solexa introns. To test this hypothesis, transcripts affected by InDels detected in from the genomic alignment will be compared to transcripts affected by the un-incorporated Solexa introns.

4.2. Materials and Methods

Sequencing Data. Data for the analysis consist of the WGS Sanger sequencing reads used in the *C. briggsae* genome assembly (Stein *et al.*, 2003). The Sanger sequencing reads were generated at The Wellcome Trust Sanger Institute and the Genome Sequencing Center, Washington University School of Medicine in St. Louis. The Sanger sequencing reads and their corresponding quality files was downloaded from the following site: ftp://ftp.ncbi.nih.gov/pub/TraceDB/caenorhabditis_briggsae. For both data sets, the data consists of FASTA read files with their corresponding qualities file. To prepare the read data for read alignment, the FASTA and quality files were converted into FASTQ format using the 'qualfa2fq.pl' script from Burrows-Wheeler Alignment (BWA) (Li and Durbin, 2009). Analysis of the sequencing data is summarized in Figure 4.2. cDNA library for RNA-seq was prepared using Trizol extraction and the Superscript II reverse transcriptase kit (Invitrogen, SKU# 18080-085) by the Baillie lab. The primer used to initiate reverse transcription was a modified oligo d(T) primer (5′-CCAGACACTATGCTCATACGACGCAGT(16) VN-3′) (Invitrogen). The protocol accompanying the kit was followed, and the samples were treated with Ribonuclease H (Invitrogen, SKU# 18021-014). Finally, DNA sequencing was performed on the Illumina cluster station and 1G analyzer (Uyar *et al.*, 2012).

Read Alignment. The RNA-seq (transcriptome) data, consisting of 42 bp paired-end reads, was aligned to WS215 *C. briggsae* reference genome. Uyar and colleagues aligned the transcriptome data to the WS215 reference *C. briggsae* genome using the program Mapping and Assembly with Qualities (MAQ) (http://maq.sourceforge.net) (Uyar *et al.,* 2012). The Sanger sequencing reads were also aligned to the reference *C. briggsae* genome (WormBase release WS215), but

using the 'BWA-SW' algorithm with default parameters (BWA version 0.5.7) (Step 1 in Figure 4.2); 'BWA-SW' was designed for aligning long reads (Li and Durbin, 2010). The advantage of using BWA is that it enables gapped alignment for single-end reads with similar alignment accuracy to MAQ. In addition, BWA outputs the alignment in Sequence Alignment/Map (SAM) format which is readily used by many programs in the downstream analysis. The resulting SAM alignment was then converted into the Binary Alignment/MAP (BAM) format using 'samtools import' command from SAMtools, versions 0.1.7a (Li *et al.*, 2009) and visualized with Generic Genome Browser (Gbrowse), version 2.39. The total number of reads that were used by BWA-SW in the Sanger sequence alignment is 2258364, which resulted in an average read depth of 13 reads per base (Table 4.1).

Table 4.1. Genomic Sanger Sequence Alignment: Read Mapping Details

Number of Reads	Category
2258364	In total
0	QC failure
0	Duplicates
2258364 (100.00%)	Mapped
13	Average read depth

Notes: Sanger reads were aligned to WS215 reference *C. briggsae* genome using BWA, 'BWA-SW' algorithm (Li and Durbin, 2010).

Variation Detection and Filtration in the genomic and transcriptomic alignments. All small variations present in the genomic Sanger sequence alignment were extracted using PERL scripts (www.perl.org) from the extended 'pileup' file that was generated by 'samtools pileup -c' (Step 1 in Figure 4.2). The '-c' parameter additionally outputs the consensus base, consensus quality, SNP quality (SNPQ) and maximal mapping quality of each base. InDels from the sequence alignment were then filtered using the following criteria and in the following order: minimum read coverage (MinCov) $\geq$ 5, minimum variant frequency (VarFreq) $\geq$ 75% and maximum read coverage (MaxCov) $\leq$ 25. Deletion of 'N' ambiguous nucleotide(s) was additionally filtered from the deletion set. Single nucleotide differences (SNDs) were filtered using the following criteria: MinCov $\geq$ 5, VarFreq $\geq$ 75%, MaxCov $\leq$ 25 and SNPQ $\geq$ 20. All variations were filtered using PERL scripts that was generated by myself. InDels were

generated from the transcriptomic alignment by Bora Uyar and Jeff Chu. The InDels were previously filtered using the following criteria: MinCov $\geq$ 3 and VarFreq $\geq$ 50%. All PERL scripts were written by me.

Analysis (Step 2-4 of Figure 4.2). Categorization of InDels, using PERL scripts, are based on the gene model set improved by genBlastG v135 and the 42 bp RNA-seq Solexa reads (Uyar *et al.*, 2012). In this analysis, the start and end position of the insertion refers to the left and right base flanking the insertion, respectively. Regarding InDels that occur at exon-intron or exon-intergenic boundaries, as long as either the start or the end position overlaps with an exon, the InDel is categorized as an exonic InDel. InDels from the genomic and transcriptome data were compared to each other based on chromosome, start and end positions only. Next, transcripts from the gene model set improved by genBlastG and RNA-seq data that are affected by genomic InDels were extracted. Finally, transcripts affected by the genomic InDels were compared to the transcripts that were affected by the 636 un-incorporated Solexa introns using PERL scripts. Again, all PERL scripts were written by me.

4.3. Results

Detection of many variations in the genomic alignment suggests there are many errors in the *C. briggsae* reference genome. Table 4.2 lists the number of variations that were detected from the alignment of the genomic Sanger sequencing reads to the WS215 *C. briggsae* reference genome. Overall, there are 7388 insertions, 421 deletions and 2995 SNDs detected after filtration in the genomic alignment. Since the WGS reads that were originally used in the *C. briggsae* genome assembly were used in this analysis, the variations detected could indicate a number of errors, such as alignment or assembly errors.

Table 4.2. Summary of Variations Detected in C. briggsae genomic alignment

Insertions	Deletions	SNDs
7388[a]	421[b]	2995[c]

a Filtration parameters: MinCov ≥ 5, VarFreq > 75% and MaxCov ≤25

b Filtration parameters: MinCov ≥ 5, VarFreq > 75%, MaxCov ≤25 and removal of deletions of 'N' nucleotides

c Filtration parameters: MinCov ≥ 5, VarFreq > 75%, MaxCov ≤25 and SNPQ≥20. There are 507 exonic SNDs, 1082 intronic SNDs and 1406 intergenic SNDs

The strains used for RNA-seq and WGS are very similar if not the same strain, as determined by the comparison of the InDels found in both alignments There are 895 filtered insertions and 575 filtered deletions detected in the transcriptome alignment (Table 4.3). After a comparison between the transcriptomic and genomic InDels, it is found that the majority of the transcriptome insertions are also detected in the genomic alignment. Overall, 71% (638 out of 895) of the transcriptome insertions and specifically 73% (271 out of 369) of the exonic transcriptome insertions are found in both alignments. Additionally, 66% (269 out of 406) of the intronic transcriptome insertions and 82% (98 out of 120) of the intergenic transcriptome insertions are found in the genomic alignment. Intergenic insertions that are found in both the transcriptomic and genomic alignments can be explained by 5' and 3' untranslated regions (UTRs) that are present in cDNAs and therefore also sequenced with the exons (Mangone *et al.*, 2010). Intronic insertions that are found in both the transcriptomic and genomic alignments may indicate the presence of additional exons in the transcript affected or an additional unpredicted transcript that overlaps with known transcript in the same region. On the other hand, there are very few transcriptome deletions that are shared with the genomic deletions. Overall, 4% (21 out of 575) of the transcriptome deletions are found in the genomic alignment. This lack of agreement can be explained by partial read misalignment due to the nature of the alignment tool. Since MAQ (http://maq.sourceforge.net) is unable to perform gapped alignments, reads that span

more than one exon will have the larger portion of the read mapped to the intended exon while the smaller portion of the read misaligned to the adjacent intron (Figure 4.3). In such cases, there are often many InDels and mismatches located on the fragment of the read that is misaligned, especially in the introns, intergenic regions and near the exon-intron boundary. Upon further investigation of all transcriptomic InDels that are not found in the genomic alignment, majority of the InDels are detected in introns, intergenic regions, and exon-intron boundaries (Table 4.4). 88% (226 out of 257) of the transcriptome insertions, which are not detected in the genomic alignment, occur in introns, intergenic regions and exon-intron boundaries. 98% (543 out of 554) of the transcriptome deletions, which are not detected in the genomic alignment, occur in introns, intergenic regions and exon-intron boundaries. In other words, the majority of the transcriptome InDels that are not present in the genomic alignment occur outside of exons, indicative of read misalignment by MAQ. Exonic InDels that are only found in the transcriptomic alignment may indicate novel differences between the AF16 strain used in RNA-seq and WGS and/or the exonic InDels may indicate heterozygosity in the region. However, these differences are minimal; only 12% (31 out of 257) of the insertions and 2% (11 out of 554) of the deletions, which are found in the transcriptomic alignment and not the genomic alignment, are located in exonic regions. Overall, since the majority of transcriptomic InDels are found in the genomic alignment and the majority of the transcriptomic InDels that are not found in the genomic alignment are due to alignment errors, the AF16 strains used in both the RNA-seq and WGS sequencing are highly similar if not the same strain.

Table 4.3. Comparison of transcriptomic InDels with Genomic InDels

Category[a]	Transcriptome[b]	Genome[c]	Shared[d]
# of exonic variations	369 insertions	899 insertions	271 insertions
	30 deletions	48 deletions	10 deletions
# of Intronic variations	406 insertions	3056 insertions	269 insertions
	330 deletions	159 deletions	7 deletions
# of intergenic variations	120 insertions	3433 insertions	98 insertions
	215 deletions	214 deletions	4 deletions
Total # variations	895 insertions	7388 insertions	638 insertions
	575 deletions	421 deletions	21 deletions

a Definition of exons and introns based on the Uyar and colleagues' gene model set improved by genBlastG v135 and 42bp transcriptomic Solexa reads (Uyar *et al.*, 2012)

b Filtration parameters: MinCov $\geq$ 3 and VarFreq $\geq$ 50%. Detection and filtration completed by Bora Uyar and Jeff Chu

c Filtration parameters: MinCov $\geq$ 5, VarFreq > 75%, MaxCov $\leq$25 and removal of deletions of 'N' nucleotides

d Shared InDels have identical chromosome, start and end coordinates only.

Table 4.4. Transcriptomic InDels that are not shared with Genomic InDels are detected due to read mis-alignment

Category	Insertions[a]	Deletions[a]
# of exonic InDels	31[b]	11[c]
# of intronic InDels	137	232
# of intergenic InDels	22	211
# of InDels in exon-intron boundary	67	100
Total # InDels	257	554

a Transcriptomic InDels that are not shared with genomic InDels. Filtration parameters: MinCov $\geq$ 3 and VarFreq $\geq$ 50%. Detection and filtration completed by Bora Uyar and Jeff Chu.

b 21 exonic insertions have VarFreq < 75%. 10 exonic insertions have VarFreq $\geq$ 75%.

c 5 exonic deletions have VarFreq < 75%. 6 exonic deletions have VarFreq $\geq$ 75%.

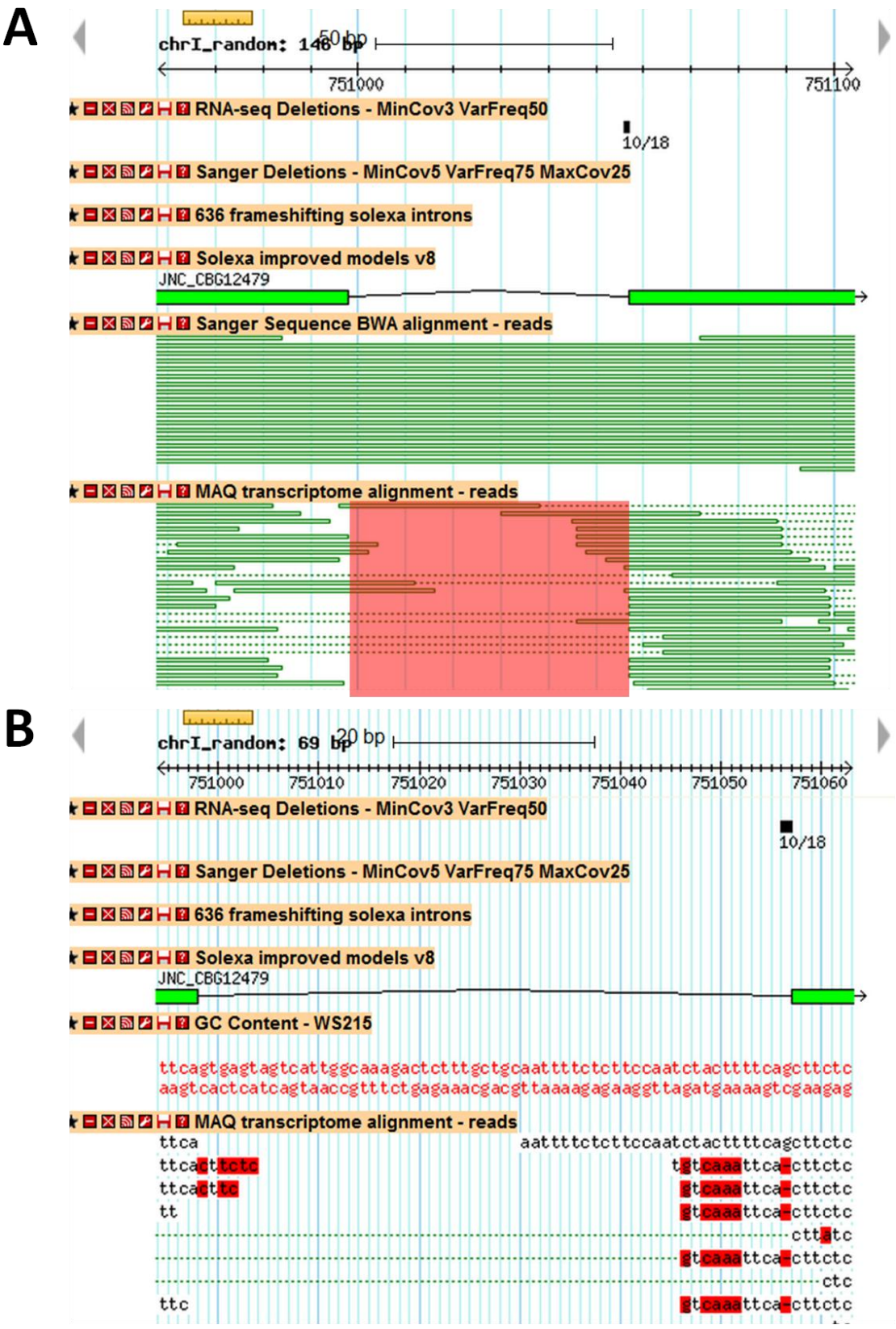

A
chrI_random: 148 bp
50 bp
751000
751100
RNA-seq Deletions - MinCov3 VarFreq50
10/18
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
Solexa improved models v8
JNC_CBG12479
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads
B
chrI_random: 69 bp
20 bp
751000
751010
751020
751030
751040
751050
751060
RNA-seq Deletions - MinCov3 VarFreq50
10/18
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
Solexa improved models v8
JNC_CBG12479
GC Content - WS215
ttcagtgagtagtcattggcaaagactctttgctgcaattttctcttccaatctacttttcagcttctc
aagtcactcatcagtaaccgtttctgagaaacgacgttaaaagagaaggttagatgaaaagtcgaagag
MAQ transcriptome alignment - reads
ttca
aattttctcttccaatctacttttcagcttctc
ttcacttctc
tgtcaaattca-cttctc
ttcacttc
gtcaaattca-cttctc
tt
gtcaaattca-cttctc
cttgtc
gtcaaattca-cttctc
ctc
ttc
gtcaaattca-cttctc

Figure 4.3. Exonic transcriptome InDels that are not shared with Sanger exonic InDels are due to partial read misalignment (alignment error) at exon-intron boundaries

(A) Region shown: chrI_random:750,958..751,103. The figure shows the misalignment of reads (red) at the exon-intron boundary. The insertion that is detected in the RNA-seq transcriptome data but not in the genomic alignment is due to misalignment of reads. (B) Region shown: chrI_random:750,994..751,062. The zoomed in figure shows fragments of the reads that are mis-aligning to the intron. The sequence highlighted in red (mismatches) matches perfectly to the adjacent exon. 'RNA-seq Deletions – MinCov3 VarFreq50' track displays filtered deletions detected from transcriptome alignment. The deletion displayed has 10 out 18 reads supporting the deletion. 'Sanger Deletions – MinCov5 VarFreq75 MaxCov25' track displays filtered deletions detected from genomic alignment. Note, not all the reads in the MAQ transcriptome alignment are shown in the figures.

InDels that are detected represent genome assembly errors that can have a major effect on gene model predictions. By re-aligning the Sanger reads that was used in the *C. briggsae* genome sequencing project, identifying the presence of InDels and confirming the existence of these InDels through a different data set (RNA-seq), it is very likely that these InDels represent genome assembly errors. In other words, transcriptomic InDels that are not due to alignment errors support the InDels detected in the genomic alignment. However, since coverage of RNA-seq reads depends on the expression level of the transcripts, not all transcripts are expressed or have high read coverage, hence, not all transcripts have RNA-seq read coverage and not all InDels will have been detected in the transcriptomic alignment; this is indicated by the large discrepancy in the number of exonic Indels between the transcriptomic and genomic alignments (Table 4.3). Thus, genomic InDels will represent more of a complete list of genome assembly errors, which can have a major effect on gene model predictions. There are 3348 transcripts that are affected by genomic InDels, or assembly errors (Table 4.5). Specifically, 897 transcripts are affected by frame shifting exonic InDels. These transcripts may need to be updated to incorporate the genome assembly errors.

Table 4.5. Transcripts affected by genomic InDels and/or un-incorporated Solexa introns

Category	# of Transcripts[a]
# of Transcripts affected by genomic InDels	3348 (897 affected by exonic InDels)
# of Transcripts affected by the un-incorporated Solexa introns	567
# of Transcripts affected by both genomic InDels and the un-incorporated Solexa introns	257

a Transcripts are from the gene model set improved by genBlastG v135 and 42bp transcriptomic Solexa reads (Uyar *et al.*, 2012) . Since gene models have not been annotated for UTRs, the numbers listed may under-represent the true number of transcripts affected.

If frame shifting genome assembly errors affect many transcripts, it is likely that they also affect the incorporation of the 636 un-incorporated Solexa introns. After comparing the genome assembly errors with the un-incorporated Solexa introns, there are 282 un-incorporated Solexa introns that affect transcripts whose genomic read alignment also supports an assembly error (Figure 4.1, Table 4.6). As a result, the frame shifting assembly errors may have a large effect on the incorporation of many Solexa introns. Particularly, the assembly errors may play a role in the un-incorporation of some of the 636 Solexa introns that were determined by Uyar and colleagues (Uyar *et al.*, 2012). If the assembly errors have been corrected in the reference genome, many of the 282 Solexa introns may have been successfully incorporated in to the gene models.

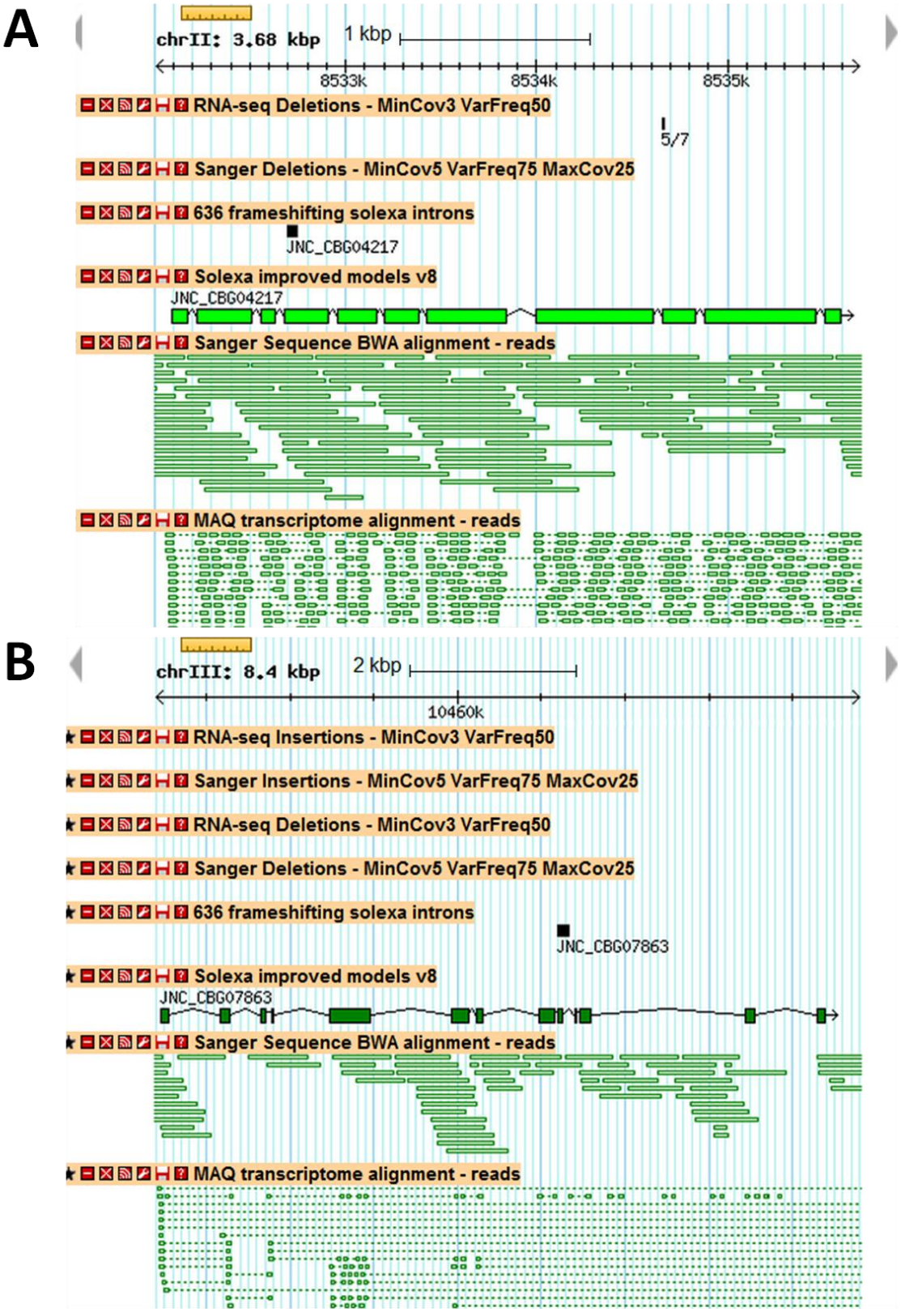
Table 4.6. Comparison between the un-incorporated Solexa introns with genomic InDels

Category	# of InDels or Solexa introns
# of exonic and intronic Genomic InDels	4162 InDels
# of un-incorporated Solexa introns	636 introns
# of un-incorporated Solexa intron that occur in the same transcripts as the genomic InDels[a]	282 introns
# of genomic InDels that occur in the same transcripts as the un-incorporated Solexa introns[a]	336 InDels (161 exonic, 175 intronic)

a Transcripts are from the gene model set improved by genBlastG v135 and 42bp transcriptomic Solexa reads (Uyar *et al.*, 2012) . Since gene models have not been annotated for UTRs, the numbers listed may be an under-representation of the true numbers.

Majority of the un-incorporated Solexa Introns cannot be explained by genome assembly errors. Thus, most of the un-incorporated Solexa Introns are likely to represent different isoforms of the same transcripts or may be falsely predicted introns. To further explore the reason why there are still so many Solexa introns that are un-incorporated, 10 of the un-incorporated Solexa introns were randomly picked and examined on GBrowse. In 2 out of 10 cases, there is an un-incorporated Solexa intron corresponding to a transcriptomic InDel affecting the same transcript, but not corresponding genomic InDel; there appears to be sufficient genomic read coverage, hence the lack of the genomic InDels is not due to low read coverage (Figure 4.4A). In 1 out 10 cases there was low genomic read coverage in some areas of the transcripts (Figure 4.4B). Although there was sufficient read coverage in the transcriptomic alignment, no InDel was detected. For a transcript to be predicted in a region of low genomic read coverage in the first place, it is possible that the transcripts were predicted from the previously finished 12MB of clone based sequence that was incorporated into the genome assembly afterwards. In 2 out of 10 cases had transcripts were affected by two un-incorporated Solexa introns. Perhaps the two frame shifting Solexa intron, when incorporated simultaneously into the gene model, can compensate each other and result in a non-frame shifting transcript (Figure 4.4C). Finally, in 5 out of 10 cases, there appears to be sufficient genomic and transcriptomic read coverage, but no InDel was detected in either alignment (Figure 4.4.D). As a result, majority of the un-incorporated Solexa introns that cannot be explained by assembly errors either represent different isoforms of the same transcripts, or are false positive Solexa introns

.


A
chrII: 3.68 kbp
1 kbp
8533k
8534k
8535k
RNA-seq Deletions - MinCov3 VarFreq50
5/7
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
JNC_CBG04217
Solexa improved models v8
JNC_CBG04217
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads
B
chrIII: 8.4 kbp
2 kbp
10460k
RNA-seq Insertions - MinCov3 VarFreq50
Sanger Insertions - MinCov5 VarFreq75 MaxCov25
RNA-seq Deletions - MinCov3 VarFreq50
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
JNC_CBG07863
Solexa improved models v8
JNC_CBG07863
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads

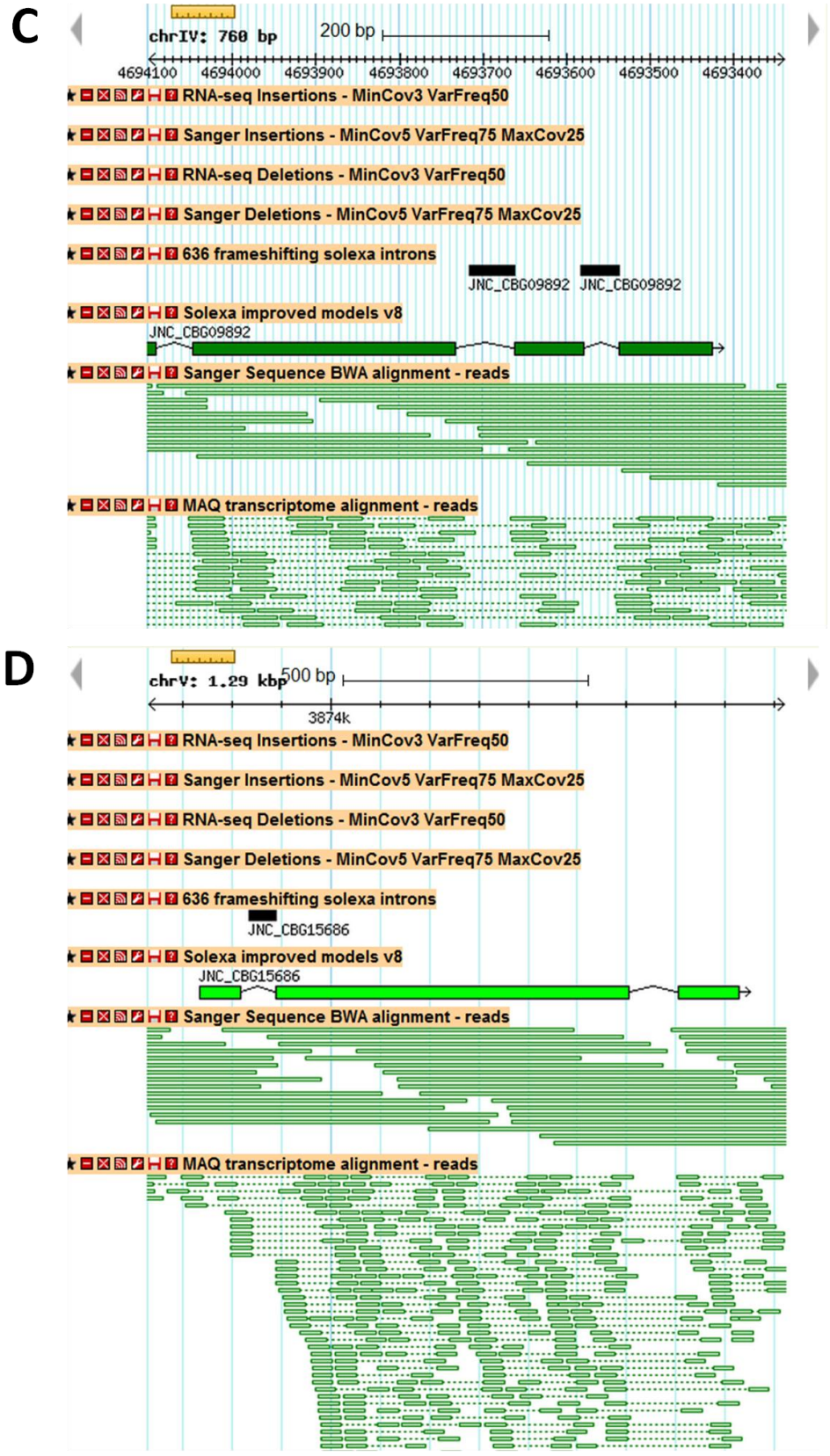

C
chrIV: 760 bp
200 bp
4694100
4694000
4693900
4693800
4693700
4693600
4693500
4693400
RNA-seq Insertions - MinCov3 VarFreq50
Sanger Insertions - MinCov5 VarFreq75 MaxCov25
RNA-seq Deletions - MinCov3 VarFreq50
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
JNC_CBG09892
JNC_CBG09892
Solexa improved models v8
JNC_CBG09892
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads
D
chrV: 1.29 kbp
500 bp
3874k
RNA-seq Insertions - MinCov3 VarFreq50
Sanger Insertions - MinCov5 VarFreq75 MaxCov25
RNA-seq Deletions - MinCov3 VarFreq50
Sanger Deletions - MinCov5 VarFreq75 MaxCov25
636 frameshifting solexa introns
JNC_CBG15686
Solexa improved models v8
JNC_CBG15686
Sanger Sequence BWA alignment - reads
MAQ transcriptome alignment - reads

Figure 4.4. Un-incorporated Solexa introns that do not correspond to 'Shared InDels'

(A) Region shown: chrII:8,532,010..8,535,689. InDel is detected in transcriptomic alignment but not in the genomic alignment. (B) Region shown: chrIII:10,455,200..10,463,599. Low genomic read coverage in some parts of the transcripts maybe responsible for the lack of InDels called in the transcript. (C) Region shown: chrIV:4,693,340..4,694,099. Two un-incorporated Solexa introns in the same transcript. (D) Region shown: chrV:3,873,630..3,874,919. Sufficient genomic and transcriptomic read coverage, but there are no InDel detected in the same transcript as the un-incorporated Solexa intron. Note, the genomic region in (B) and (C) is 'flipped' on GBrowse.

4.4. Conclusions and Future Work

The *C. briggsae* genome contains a number of genome assembly errors, namely InDels that is detected in both the re-alignment of the Sanger reads to the reference genome and the alignment of the transcriptome reads to the reference genome. These frame shifting genome assembly errors can have an effect in gene model predictions, particularly in the incorporation of solexa introns; 282 un-incorporated Solexa introns are associated with an assembly error (Table 4.6). The next step in this analysis would be to experimentally validate the genome assembly errors and to fix the genome assembly errors in the *C. briggsae* genome sequence. In addition, this analysis did not focus on SNDS. There may be many SND genome assembly errors. Future work can include detecting SNDs from the transcriptome alignment and comparing the SNDs between the genomic and transcriptomic alignment.

The strain used in the RNA-seq analysis and the strain used to assembly the reference *C. briggsae* genome are very similar. Essentially, the majority of the transcriptomic InDels are found in the genomic alignment. The high number of exonic genomic InDels that are not found in the transcriptomic alignment can be explained by low or non-existent read coverage in the transcriptome alignment (Table 4.3). Furthermore, transcriptome InDels that were not found in the genomic alignment can be explained by alignment errors (Table 4.4). To filter out InDels in the transcriptomic alignment due to alignment errors, future work can include re-aligning the transcriptome reads using alignment tools that are dedicated to RNA-seq data, such as TopHat, which is a fast splice junction mapper (Trapnell *et al.*, 2009). I expect that with high transcriptome read coverage and alignment with RNA-seq dedicated tools, such as

TopHat, that there would be a higher percentage of genomic InDels that are shared with transcriptomic InDels.

Although genome assembly errors can explain the un-incorporation of many Solexa introns, it does not provide the whole story. There are situations where multiple solexa introns in the same transcript can affect each other's incorporation. To determine whether this is true, multiple frame shifting entities have to be incorporated into the gene model simultaneously, to be able to determine whether the multiple frame shifting entities can compensate each other and result in a non-frame shifting transcript. Furthermore, many of the un-incorporated Solexa introns that are not associated with frame shifting mutations have normal genomic and transcriptomic read coverage. First, the Solexa introns may be false introns that are predicted. Again, using an RNA-seq dedicated aligner may help determine which Solexa introns are incorrect. Second, the un-incorporated Solexa introns may represent alternative isoforms of the transcript. As a result, future work can include identifying all the possible alternative isoforms for all transcripts, based on RNA-seq data.

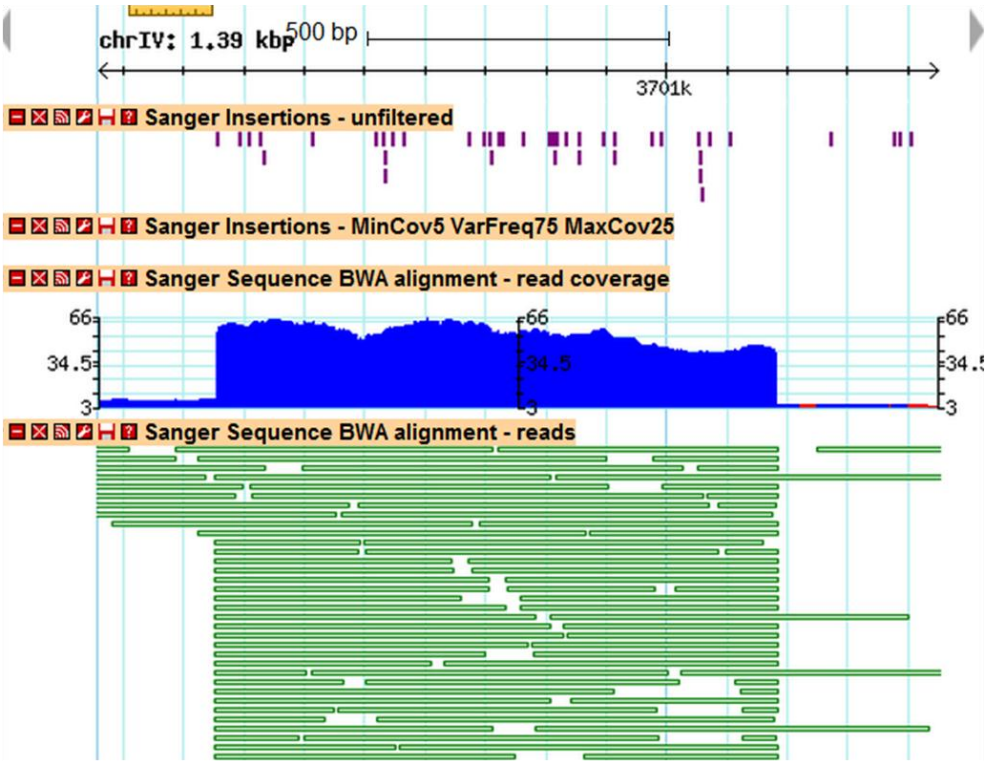
4.5. Discussion

The alignment statistics of the genomic alignment are very similar to the ones generated by the Phusion assembler using the same Sanger sequencing reads. In the Phusion assembly, the overall sequence coverage is estimated at 11-fold. In addition, Phusion utilized 2,085,214 reads for the assembly (Mullikin and Ning, 2003). In the BWA alignment, the average read depth is 13 reads per base and the total number of reads utilized by BWA is 2,258,364. The average read depth between the Phusion and BWA alignment are in agreement with each other.

Many filtration thresholds were used to filter the variations detected from the genomic alignment. The filtration threshold for MinCov is set at 5, which is relatively stringent considering the average read coverage of the genomic alignment is 13. The stringent MinCov was set to minimize false positives (FP) due to genome sequencing errors. VarFreq is set at 75% since true genome assembly errors should manifest as homozygous variations. VarFreq threshold for homozygous variations have previously

been used by Cord and colleagues (Drogemuller *et al.*, 2010). In order to filter out variations that are detected in high coverage regions (Figure 4.2A) a threshold for MaxCov was determined. First, the percentage of InDels and SNDs that are filtered out at MaxCov 20, 30, 40 and 50 were plotted (Figure 4.5). For both insertions and deletions, there is a big difference in the percentage of variations that are removed between thresholds MaxCov $\geq$ 20 and MaxCov $\geq$ 30 (Figure 4.6). By randomly picking 20 variations with read coverage between 20 to 25 inclusive (Set A) and 20 variations with a read coverage of 26 to 30 inclusive (Set B) using a PERL script, the MaxCov $\leq$ 25 was set. In Set A, 15 out 20 variations did not occur in regions of high read coverage. In Set B, 9 out of 20 variations did not occur in regions of high read coverage. Since there is a higher percentage of variations that did not occur in high read coverage regions in Set A, a threshold of 25 was set for MaxCov. Regions of high read coverage are often areas of read misalignment; they occur due to the mapping of reads to low complexity regions and transposon elements (Emmons *et al.*, 1983). Reads that map to repetitive regions often map to multiple locations. For reads that map to multiple locations, BWA-SW will output multiple alignments (Figure 4.2B) which will be shown in the genome browser. Also, since no post processing of the reads were applied, for example to filter for reads that map to unique locations only, repetitive regions are characterized by abnormally high read coverage. In regards to variations detected in repetitive regions, there are situations where regions of the genome are imperfectly repeated; in order words, there may be slight differences between the repeated sequences. Sequencing read corresponding to the repeated sequences will be slightly different, manifesting as variations and aligned to multiple locations. As a result, these variations are unreliable, as we cannot tell apart which repeated region the variations came from. Continuing with the filtration thresholds, SNDs were additionally filtered for SNPQ. SNPQ is a Phred-scaled probability that the consensus is identical to the reference; the lower the probability that the consensus is identical to the reference, the higher the SNPQ value. Jia and colleagues determined that the quality scores of variants called by ‘pileup’ and ‘mpileup’ were quite similar and suggested a cutoff value of 20 for ‘mpileup’ (Jia *et al.*, 2012). Since quality scores for ‘pileup’ and ‘mpileup’ are very similar, the SNPQ threshold was also set at 20 for the ‘pileup’ data. Inspection of SNDs with SNPQ =0, uncovered SNDs that were called at bases where the reference is an ‘N’ ambiguous nucleotide. Hence, SNDs with SNPQ=0 are unreliable as we cannot

tell if there is indeed a variation if reference base is unknown. These cases are filtered out with the current SNPQ threshold.



B

gnl|ti|36381785 16 chrIV 3700252 195 110S556M43S * 0 0
TTTTTTGATATTTTTCGTGATTTTTATTTTTTTTTTCCCTGATTTTTTATACCTTTTTATCCTTTTGGACTTCTGAAATTTCGAATTTTTAGCTGA
ATCTCGAATTTTTAAATAAAATACATAGAATTTTAGATGAATTGGTAAAAATGAAAAAAATAAAGAAAAAAGATTTGAAAAATATTTATT
CTAATATTGAAATGATTTAATTTTTAAGGTAAAATTGAAATTTTATTTTTATAATATTTTATACTTCTTTCAGAACTAGACTCATTAACTTTA
TTTTTTGCTTCTTGTCCCTTGTTGTCCGGACTTGTCTCATTTAGTTTTATATTTAATTCCATTAAACCTCTATTTAATGTTTTTACTGATACATC
ATGTAATTTACACCATTTCGATTTCATAATTGGAAATTTTCTGTTATCTTCTAAATAAAAATGATGCTATAATTTAAAAATGATGCTATAAT
TTTCTTATCGAGCATAATTCTTTCAGAACTAGACTCGTCAAGTTTATTTGTTAATTTAAAGGAGAAAAATTTATTCTATTAAATTAACAAAA
TGGAAATTCAGCTCTTCAAACAACTAAATATTAATATCGATAAAGCCGAATTAAACGCTTATTTAGATAATGATTTGACGAAAGAAACTG
TTATAAGACATCTAGATGGGATCCTCTAGAGTCAGACCTGCGAGTGCATGCAANGCAATGC ..79:55../<;FQ<9598><9>..3<=BQQQ
Q>:530/6<<AA<>--05/--3:42283,-
3:6:57:AAA;:947:<>>@???BAA=88?<<93;396CFKLTTTTTTTTYYYYYKKKKKKKKKKOOYKFFFFDD=>ADDITOFFFFFHYYOKIINNIINN
TYYYYYFIFFFIYTTNNIFFDHDIIINNIFFFIIIFFIIITTTTTTYTTTTTTTTTTNTLLTTTTNNNNTTYYYYYTTTNNNLJJOOLLLLLLNYYYYYYYYYYYYY
NNNNNNNNNNNTNTYYLJLLLLGILNNNNLYYLLLLEEJYYOLJNNNNNTNNNNNNTTTTTTNNNNNNTTYTTYYYTLLLLLNNNNNLTTNNNT
NLIILLNTTTTTTTTYTTTTTTYYTTTTTNTTTTTTTTTTTMMYYOTNNLLOIILLLLLLLLLLTLNNNNNYTTTTTTTTTTYYYYTTYYYYYTTTTLMJJJ
NTTLLTTLLTNNLYYTTLLLLLLLLLTTTTTTTTTTTNLLLLLNTTNNNNLLDDDDDLYYYYYYYYYYYNNLLLLLLNNNNNNNIIIIIITTGGGGHHY
YYTTTTTTYYYYYYTTNHHHHHHINDDDDDDIIIIIIIFIINNNNNNNNNIFFFHHHIIIIYTNNDDDDDBBBBADYYYYYTODDDDDHHHHDDICCCC
AA;;;;;>I>777::=60+'''''..(''''))+239>52%!%%%!!!! AS:i:556 XS:i:149 XF:i:3 XE:i:11 XN:i:0 NM:i:0 MD:Z:556

gnl|ti|36381785 16 chrIV 3776715 170 109M600S * 0 0
TTTTTTGATATTTTTCGTGATTTTTATTTTTTTTTTCCCTGATTTTTTATACCTTTTTATCCTTTTGGACTTCTGAAATTTCGAATTTTTAGCTGA
ATCTCGAATTTTTAAATAAAATACATAGAATTTTAGATGAATTGGTAAAAATGAAAAAAATAAAGAAAAAAGATTTGAAAAATATTTATT
CTAATATTGAAATGATTTAATTTTTAAGGTAAAATTGAAATTTTATTTTTATAATATTTTATACTTCTTTCAGAACTAGACTCATTAACTTTA
TTTTTTGCTTCTTGTCCCTTGTTGTCCGGACTTGTCTCATTTAGTTTTATATTTAATTCCATTAAACCTCTATTTAATGTTTTTACTGATACATC
ATGTAATTTACACCATTTCGATTTCATAATTGGAAATTTTCTGTTATCTTCTAAATAAAAATGATGCTATAATTTAAAAATGATGCTATAAT
TTTCTTATCGAGCATAATTCTTTCAGAACTAGACTCGTCAAGTTTATTTGTTAATTTAAAGGAGAAAAATTTATTCTATTAAATTAACAAAA
TGGAAATTCAGCTCTTCAAACAACTAAATATTAATATCGATAAAGCCGAATTAAACGCTTATTTAGATAATGATTTGACGAAAGAAACTG
TTATAAGACATCTAGATGGGATCCTCTAGAGTCAGACCTGCGAGTGCATGCAANGCAATGC ..79:55../<;FQ<9598><9>..3<=BQQQ
Q>:530/6<<AA<>--05/--3:42283,-
3:6:57:AAA;:947:<>>@???BAA=88?<<93;396CFKLTTTTTTTTYYYYYKKKKKKKKKKOOYKFFFFDD=>ADDITOFFFFFHYYOKIINNIINN
TYYYYYFIFFFIYTTNNIFFDHDIIINNIFFFIIIFFIIITTTTTTYTTTTTTTTTTNTLLTTTTNNNNTTYYYYYTTTNNNLJJOOLLLLLLNYYYYYYYYYYYYY
NNNNNNNNNNNTNTYYLJLLLLGILNNNNLYYLLLLEEJYYOLJNNNNNTNNNNNNTTTTTTNNNNNNTTYTTYYYTLLLLLNNNNNLTTNNNT
NLIILLNTTTTTTTTYTTTTTTYYTTTTTNTTTTTTTTTTTMMYYOTNNLLOIILLLLLLLLLLTLNNNNNYTTTTTTTTTTYYYYTTYYYYYTTTTLMJJJ
NTTLLTTLLTNNLYYTTLLLLLLLLLTTTTTTTTTTTNLLLLLNTTNNNNLLDDDDDLYYYYYYYYYYYNNLLLLLLNNNNNNNIIIIIITTGGGGHHY
YYTTTTTTYYYYYYTTNHHHHHHINDDDDDDIIIIIIIFIINNNNNNNNNIFFFHHHIIIIYTNNDDDDDBBBBADYYYYYTODDDDDHHHHDDICCCC
AA;;;;;>I>777::=60+'''''..(''''))+239>52%!%%%!!!! AS:i:109 XS:i:0 XF:i:3 XE:i:2 XN:i:0 NM:i:0 MD:Z:109

Figure 4.5. Short Region of high read coverage indicative of read misalignments in Genomic Sanger sequence alignment

(A) Region shown: chrIV:3,700,060..3,701,449. The region shows an area with abnormally high read coverage, reaching 66 reads per base. The high read coverage area is characterized by distinct read break points and are typically covered by shorter reads. There are also a lot of unfiltered insertions that are found in the high coverage region. After using MaxCov threshold of 25, as well as the other filtration parameters, the insertions located in the high coverage region are filtered out, as indicated by the lack of insertions detected in the 'Sanger Insertions – MinCov5 VarFreq75 MaxCov25' track. (B) Example of a read mapped to a high coverage region, shown in SAM format. The read shown starts at the left breakpoint of the high coverage region. BWA-SW outputs two alignments for the read. One of the alignments is mapped to the region shown in (A), at chrIV:3,700,252 and alignment maps the read to chrIV:3776715170, which is further downsteam. Both alignments of the reads features extensive soft clipping. In the top alignment in the CIGAR column,110S556M43S indicates that there is 110bp and 43bp clipped from the left and right side of the read, respectively. In the bottom alignment in the CIGAR column, 109M600S indicates that there is 600bp of the read clipped from the right end. This is indicative of a bad read alignment.

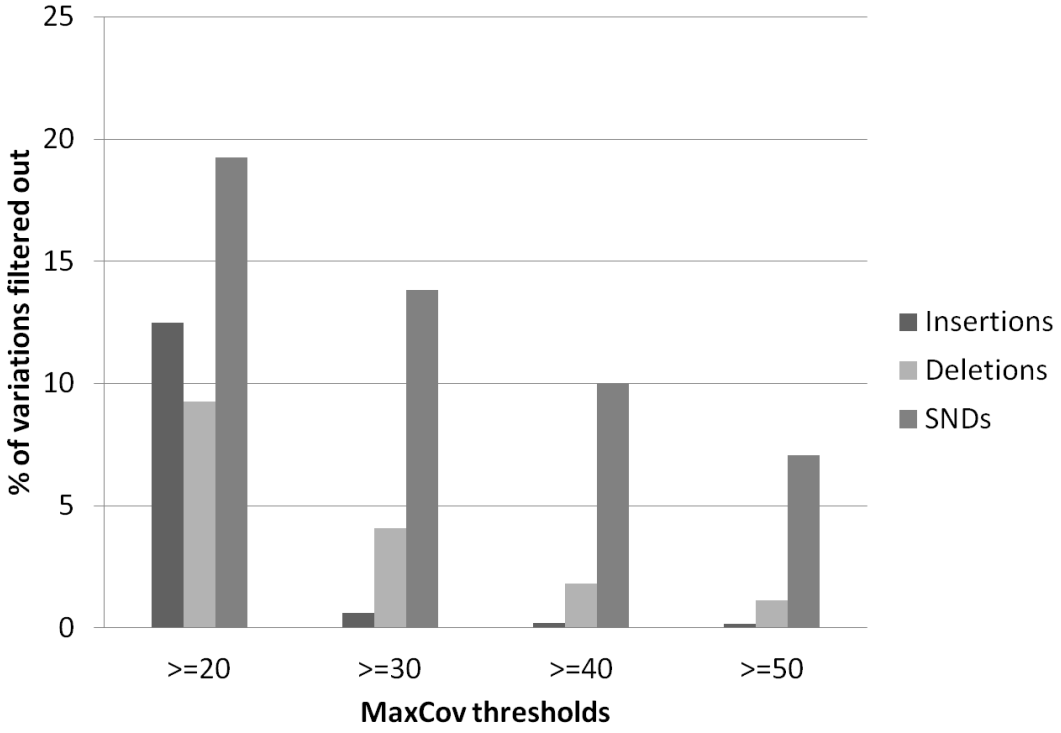


Figure 4.6. Percentage of variations filtered at different MaxCov thresholds detected through Sanger sequence alignment

To determine the optimal threshold for MaxCov in the genomic Sanger sequence alignment, the percentage (%) of variations that were filtered out after applying the MaxCov threshold was plotted. The variations have previously been filtered for MinCov $\geq$ 5 and VarFreq $\geq$ 75%. MaxCov 30, 40 and 50 appears to have little effect on the percentage of InDels filtered out, compared to MaxCov 20. There also seems to be a gradual percentage of SNDs filtered out at each threshold.

Finally, there are noticeably more insertions than deletion detected in the Sanger sequence alignment (Table 4.2); there are approximately 17X more insertions than deletions. After examining the distribution of variations throughout the chromosomes, I find that the percentages of the total InDels and SNDs in each chromosome are very similar (Figure 4.8), with the exception of chrUn which shows a much higher percentage of SNDs than InDels. Furthermore, the percentages of total variants throughout all the non-random chromosomes are around the 10-15% region, meaning a similar number of InDels and SNDs were detected in all the non-random chromosomes. It is also expected that there are fewer variations detected in the random chromosomes, as the random chromosomes consist of a much smaller sequence. Since the distributions of the variants throughout the chromosome are vastly similar, it does not explain why there are a lot more insertions than deletions detected. One possible explanation is that the original pipeline used in the *C. briggsae* genome assembly is bias towards insertions, for example, it is unable to incorporate insertions effectively into the draft genome assembly.

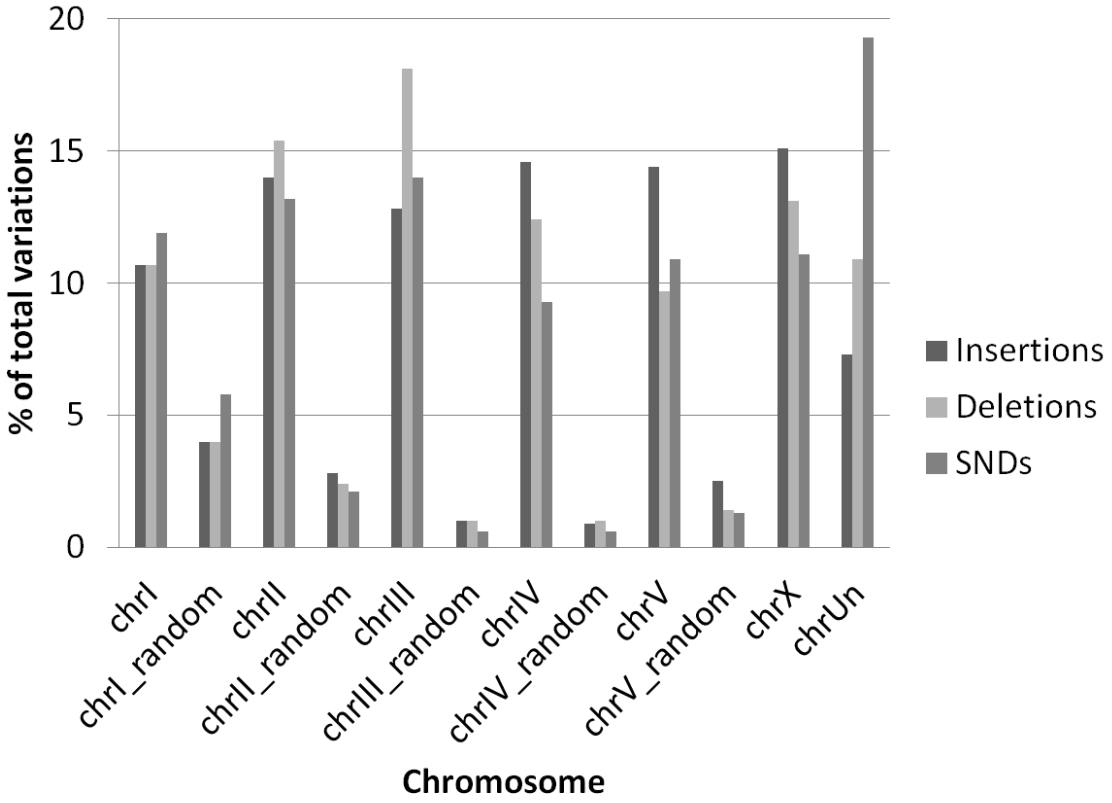


Figure 4.7 Distribution of Sanger variations throughout the chromosomes

The number of variations, represented at the percentage (%) of the total number of variations for each type, is plotted according to chromosome. 10-15% of the InDels and SNDs are distributed amongst the non-random chromosomes. Fewer variations are detected in the random chromosome due to smaller sequence size compared to the non-random chromosomes. In chrUn, there appears to be much more SNDs than compared to the other non-random chromosomes.

5. General Conclusion and Discussion

Variation detection for gene identification using next generation sequencing (NGS) data are often plagued by sequencing errors, alignment errors and variations due to strain differences. For example in Illumina platforms, the 3' ends of sequencing reads are often low in quality due to phasing artifacts (Minoche *et al.*, 2011; Yu *et al.*, 2012). To circumvent the high frequency of sequencing errors that is inherent in sequencing platforms, several measures are taken. One obvious measure is to sequence the genomes at a high read coverage. Sequencing errors distributed throughout the genome do not make as big of an impact in variation calling at high read coverage than at low read coverage. In a high read coverage genome, sequencing errors would represent a smaller percentage of the bases at each position. In projects 1 and 2, the mutant and suppressor strains were sequenced with an average read depth ranging from 46 to 146 reads per base. The read coverage is much higher than what was used by Sarin and colleagues to successfully clone *lys-12.* Another measure is to set a threshold for the variant frequency (VarFreq), minimum coverage (MinCov) and SNP quality (SNPQ). Variants with high VarFreq, and for single nucleotide differences (SNDs), high SNPQ are less likely to be sequencing errors. In lower read coverage regions, sequencing errors will represent a larger percentage of the reads. To avoid detecting SNPs in regions that have low read coverage, a MinCov threshold was set in the variant detection pipeline. To circumvent alignment errors, longer reads are used since they are less likely to be mis-mapped. The mutant and suppressors strains all have reads that are either 75 or 76 bp in length. In comparison, Sarin *et al.* worked with 35 bp reads. Reads that are mis-mapped are usually mapped to multiple locations, for example using BWA. To avoid these low complexity regions, a maximum read coverage (MaxCov) threshold was set. Finally, to avoid variations due to strain differences, Sarin et al. filtered out variations that are also found in the starter strain and N2 strain (Sarin *et al.*, 2008). In projects 1 and 2, variations that are found in the wild type Hobert and Horvitz strains were filtered. Starter strains for the mutant and suppressor strains were not sequenced.

Apart from avoiding false positives, the variation detection pipeline also has to be sensitive enough to detect all the known true positives. In projects 1 and 2, all the known positive controls were detected by VarScan and Pindel in the proper strains, at the correct location and with a high VarFreq. Since all the positive controls were detected, the filtration thresholds are sensitive enough to identify variations that correspond to the genes of interest. Indeed, a missense mutation in *tba-5* detected by the variation detection pipeline was found to be *dyf-10(e1383).* Furthermore, validation of a set of homozygous SNDs and InDels in the suppressor strains resulted in the validation of 97% of the SNDs and 75% of the deletions. Another verification that the filtration thresholds are optimal, is that the majority of the SNDs detected in the mutant and suppressor strains consist of a GC to AT transition, which is characteristic of the mutagen, ethyl methanesulfonate (EMS), that was used on the strains (Flibotte *et al.*, 2010).

NGS uncovered that *dyf-10(e1383)* was previously mis-mapped. *tba-5* is not located in the region of interest. In fact, *tba-5* is approximately 900,000 bp away from the right break point, *unc-13*. It is no wonder that *dyf-10(e1383)*, which was discovered in 1995 (Starich *et al.*, 1995), was un-cloned for more than 10 years. Furthermore, this is evidence that minimal mapping to a chromosome is sufficient for gene cloning (Sarin *et al*., 2008). With NGS, it is easy to identify potential candidates in the whole chromosome simultaneously, making the cloning process much faster and less labour intensive.

Gene cloning is highly dependent on the quality of the reference sequence for reference-dependent read alignment methods. Errors in the reference sequence can result in a higher rate of false positive and false negatives variations. In addition, the errors may affect gene annotations. In the *C. briggsae* genome, there were many genome assembly errors detected. Differences found from the re-alignment of the reads originally used in the *C. brigssae* genome assembly were supported by differences detected from the RNA-seq of the same strain. Utilizing different types of NGS data gathered from the same strain is an efficient way to identify the genome assembly errors.

References

Ahringer, J. (2006) Reverse genetics. *WormBook, ed*. The *C. elegans* Research community, WormBook, doi/10.1895/wormbook.1.47.1,http:www.wormbook.org.

Allen, MA. *et al.* (2011) A global analysis of *C. elegans* trans-splicing. *Genome Res.* **21(2)**:255-64.

Altun, ZF. and Hall, DH. (2009) Epithelial system, seam cells. In WormAtlas. doi:10.3908/wormatlas.1.14.

Badano, JL. *et al.* (2006) The Ciliopathies: An Emerging Class of Human Genetic Disorders. *Annu Rev Genomics Hum Genet*. **7**:125-48.

Bargmann, CI. (2006) Chemosensation in *C. elegan. WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1123.1, http://www.wormbook.org.

Bei, Y. *et al.* (2002) SRC-1 and Wnt Signaling Act Together to Specify Endoderm and to Control Cleavage orientation in Early *C. elegans* Embryos. *Dev Cell.* **3(1)**:113-25.

Blacque, OE. *et al.* (2005) Functional Genomics of the Cilium, a Sensory Organelle. *Curr Biol.* **15(10)**:935-41.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics.* **77(1):**71-94.

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science.* **282(5396)**:2012-8.

Chen, N. *et al.* (2006) Identification of ciliary and ciliopathy genes in *Caenorhabditis elegans* through comparative genomics. *Genome biol.* **7(12)**:R126.

Chu, JSC. *et al.* (2012) Fine tuning of RFX/DAF-19-regulated target gene expression through binding to multiple sites in *Caenorhabditis elegans*. *Nucleic Acids Res.* **40(1)**:53-64.

Coghlan, A. and Wolfe, KH. (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res*. **12(6):**857-67.

Colaiacovo, MP. *et al.* (2003) Synaptonemal complex assembly in *C. elegans* is dispensable for loading strand-exchange proteins but critical for proper completion of recombination. *Dev Cell.* **5(3)**:463-74.

Davis, MW. *et al.* (2005) Rapid single nucleotide polymorphism mapping in *C. elegans. BMC Genomics*. **6:**118.

Davis, MW. and Hammarlund, M. (2006) Single-Nucleotide Polymorphism Mapping. In K. Strange (Ed.). *Methods in Molecular Biology, vol.351: C. Elegans: Methods and Applications (pp85-86). Totowa, NJ: Human Press Inc.*

Dolan, PC. and Denver, DR. (2008) TileQC: A system for tile-based quality control of Solexa data. *BMC Bioinformatics*. **9**:250.

Drogemuller, C. *et al.* (2010) Identification of the Bovine Arachnomelia Mutation by Massively Parallel Sequencing Implicates Sulfite Oxidase (SUOX) in Bone Development. *PLoS Genet*. **6(8)**:e1001079.

Cui, M. and Han, M. (2007) Roles of chromatin factors in *C. elegans* development. *WormBook,* ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.139.1, http://www.wormbook.org.

Edgley, ML. and Riddle, DL. (2001) LG II balancer chromosomes in *Caenorhabditis elegans*: mT1(II;III) and the mIn1 set of dominantly and recessively marked inversions. *Mol Genet Genomics*. **266(3)**:385-95.

Eisenmann, DM. (2005) Wnt signaling. *WormBook, ed*. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.7.1, http://www.wormbook.org.

Emery, P. *et al.* (1996) RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom. *Nucleic Acids Res.* **24(5)**:803-7.

Emmons, SW. *et al.* (1983) Evidence for a Transposon in *Caenorhabditis elegans*. *Cell.* **32(1)**:55-65.

Fay, D. (2006) Genetic mapping and manipulation: Chapter 6-Mapping with deficiencies and duplications, *WormBook, ed*. The *C. elegans* Research community, WormBook, doi/10.1895/wormbook.1.95.2, http;//www.wormbook.org.

Fay, D. and Bender, A. (2006) Genetic mapping and manipulation: Chapter 4-SNPs: Introduction and two-point mapping. *WormBook, ed*. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.93.1, http://www.wormbook.org.

Fay, D. and Bender, A. (2008) SNPs: Introduction and two-point mapping. *WormBook, ed.* The *C. elegans* research community, Wormbook, doi/10.1895/wormbook.1.93.2, http://www.wormbook.org.

Ferguson, EL. and Horvitz, HR. (1985) Identification and characterization of 22 genes that affect the vulval cell lineages of the nematode *Caenorhabditis elegans*. *Genetics.* **110(1)**:17-72.

Flibotte, S. *et al.* (2010) Whole-Genome Profiling of Mutagenesis in *Caenorhabditis elegans. Genetics*. **185(2):**431-41.

Forsythe, E. and Beales, PL. (2012) Bardet-Biedl syndrome. *Eur J Hum Genet.* **doi: 10.1038/ejhg.2012.115**

Gogonea, CB. *et al.* (1999) Computational prediction of the three-dimensional structures for the *Caenorhabditis elegans* tubulin family. *J Mol Graph Model.* **17(2):**90-100.

Gupta, BP. *et al.* (2007) Genomics and biology of the nematode *Caenorhabditis briggsae. WormBook, ed*. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.136.1, http://www.wormbook.org.

Gupta, BP. and Sternberg, PW. (2003) The draft genome sequence of the nematode *Caenorhabditis briggsae*, a companion to *C. elegans*. *Genome Biol*. **4(12)**:238.

Hao, L. *et al.* (2011) Intraflagellar transport delivers tubulin isotypes to sensory cilium middle and distal segments. *Nat Cell Biol.* **13(7)**:790-8.

Hawkins, NC*. et al*. (2005) MOM-5 Frizzled regulates the distribution of DSH-2 to control *C. elegans* asymmetric neuroblast division. *Dev Biol.* **284(1)**:246-59.

Hedgecock, EM. *et al.* (1985) Axonal guidance mutants of *Caenorhabditis elegans* identified by filling sensory neurons with fluorescein dyes. *Dev Biol.* **111(1):**158-170.

Hillier, LW. *et al.* (2009) Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans. Genome Res*. **19(4)**:657-66.

Hobert, O. (2010) The Impact of Whole Genome Sequencing on Model System Genetics: Get Ready for the Ride. *Genetics*. **184(2**):317-9.

Hodgkin, J. (1999) 5.2 Standard 2-factor mapping. In IA. Hope (Ed.). *C. elegans: A practical approach (pp263). Oxford: Oxford University Press*.

Hodgkin, J. (2005) Genetic suppression. *WormBook, ed*. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.59.1, http://www.wormbook.org.

Hong, GF. (1982) Sequencing of large double-stranded DNA using the dideoxy sequencing technique. *Biosci Rep*. **2(11):**907-12.

Hopp, K. *et al.* (2011) B9D1 is revealed as a novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing and deletion analysis. *Hum Mol Genet.* **20(13)**:2524-34.

Horvitz, HR. and Herskowitz, I. (1992) Mechanisms of Asymmetric Cell Division: Two Bs or Not Two Bs, That Is the Question. *Cell*. **68(2):**237-55.

Huang, L. *et al.* (2011) TMEM237 is mutated in individuals with Joubert syndrome related disorders and expands the role of the TMEM family at the ciliary transition zone. *Am J Hum Genet.* **89(6)**:713-30.

Hurd, DD. *et al.* (2010) Specific α- and β-Tubulin Isotypes Optimize the Functions of Sensory Cilia in *Caenorhabditis elegans. Genetics.* **185(3):**883-896.

Inglis, PN. (2007) The sensory cilia of *Caenorhabditis elegans. WormBook, ed*. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.126.2, http://www.wormbook.org.

Inglis, PN. *et al*. (2009) Functional Genomics of Intraflagellar Transport-Associated Proteins in *C. elegans*. *Methods Cell Biol*. **93:**267-304.

Jia P, *et al*. (2012) Consensus Rules in Variant Detection from Next-generation Sequencing Data. *PLoS One.* **7(6)**:e38470.

Jorgensen, EM. and Mango, SE. (2002) The art and design of genetic screens: *Caenorhabditis elegans*. *Nat Rev Genet*. **3(5):**356-69.

Kato, K. (2009) Impact of next generation DNA sequencers. *Int J Clin Exp Med*. **2(2)**:193-202.

Koboldt, DC. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. **25(17)**:2283-5.

Kohara, Y. (1996) Large scale analysis of *C. elegans* cDNA. *Tanpakushitsu kakusan koso.* **41(5)**:715-20.

Korswagen, HC. (2002) Canonical and non-canonical Wnt signaling pathways in *Caenorhabditis elegans*: variations on a common signaling theme. *Bioessays*. **24(9)**:801-10.

Kunitomo, H. *et al.* (2005) Identification of ciliated sensory neuron-expressed genes in *Caenorhabditis elegans* using targeted pull-down of poly(A) tails. *Genome Biol.* **6(2)**:R17.

La Carbona, S., *et al.* (2004) The protein kinase *kin1* is required for cellular symmetry in fission yeast. *Biol Cell.* **96(2):**169-79.

Lamesch, P. *et al.* (2004) *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14(10B)**:2064-9.

Levy, S. *et al.* (2007) The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* **5(10)**:e254.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. **25(14)**:1754-60.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*. **26(5)**:589-95.

Li, H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25(16)**:2078-9.

Luo, C. *et al.* (2012) Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS One*. **7(2):**e30087.

Mangone, M. *et al.* (2010) The Landscape of *C. elegans* 3'UTRs. *Science.* **329(5990):**432-5.

Medvedev, P. *et al.* (2009) Computational methods for discovering structural variation with next-generations sequencing. *Nat Methods.* **6:**S13-S20.

Metzker, ML. (2010) Sequencing technologies – the next generation. *Nat Rev Genet.* **11(1)**:31-46.

Minoche, AE. *et al.* (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*. **12:**R112.

Morozova, O. *et al.* (2009) Applications of new Sequencing Technologies for Transcriptome Analysis. *Annu Rev Genomics Hum Genet.* **10**:135-151.

Mullikin, JC. and Ning, Z. (2003) The phusion assembler. *Genome Res*. **13(1)**:81-90.

Nagalakshmi, U. *et al.* (2010) RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol.* Chapter 4:Unit 4.11.1-13.

Nakano, S. *et al.* (2011) Replication-coupled chromatin assembly generates a neuronal bilateral asymmetry in *C. elegans*. *Cell.* **147(7)**:1525-36.

Niedringhaus, TP. *et al.* (2011) Landscape of Next-Generation Sequencing Technologies. *Anal Chem*. **83(12)**:4327-41.

Nielsen, MG. *et al.* (2010) Tubulin evolution in insects: gene duplication and subfunctionalization provide specialized isoforms in a functionally constrained gene family. *BMC Evol Biol.* **10:**113.

Nesbitt, MJ. *et al.* (2010) Identifying novel genes in *C. elegans* using SAGE tags. *BMC Mol Biol*. **11(1)**:96.

O'Neil, M.J. et al. (2006) The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals Fourteenth Edition. In MJ. O'Neil (Ed.). *Whitehouse Station, NJ: Merck & Co., Inc.*

Pan, J. *et al.* (2005) Cilium-generated signaling and cilia-related disorders. *Lab Invest*. **85(4)**:452-63.

Park, D. *et al.* (2010) Antagonistic Smad transcription factors control the dauer/non-dauer switch in *C. elegans*. *Development*. **137(3)**:477-85.

Peng, H. *et al.* (2010) Transcriptional coactivator HCF-1 couples the histone chaperone Asf1b to HSV-1 DNA replication components. *Proc Natl Acad Sci U S A*. **107(6)**:2461-6.

Phillips, JB. *et al.* (2004) Roles for Two Partially Redundant α-Tubulins During Mitosis in Early *Caenorhabditis elegans* Embryos. *Cell Motil cytoskeleton.* **58(2):**112-26.

Portman, DS. and Emmons, SW. (2004) Identification of *C. elegans* sensory ray genes using whole-genome expression profiling. *Dev Biol*. **270(2):**499:512.

Reboul, J. *et al.* (2003) *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet.* **34(1)**:35-41.

Rozen, S. and Skaletsky, HJ. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S., Misener, S. (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.

Ruffalo, M. et al. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*. **27(20):**2790-6.

Ruzanov, P. *et al.* (2007) Discovery of novel alternatively spliced *C. elegans* transcripts by computational analysis of SAGE data. *BMC Genomics*. **8**:447.

Ruzanov, P. and Riddle, DL. (2010) Deep SAGE analysis of the *Caenorhabditis elegans* transcriptome. *Nucleic Acids Res*. **38(10)**:3252-62.

Sarin, S. *et al.* (2008) *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat Methods.* **5(10):**865-7.

She, R. *et al.* (2011) genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics*. **27(15)**:2141-3.

Shim, YH. and Paik, YK. (2010) *Caenorhabditis elegans* proteomics comes of age. *Proteomics.* **10(4)**:846-57.

Shin, H. *et al.* (2008) Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biol.* **6:**30.

Silhankova, M. *et al*. (2005) Nuclear receptor NHR-25 is required for cell-shape dynamics during epidermal differentiation in *Caenorhabditis elegans*. *J Cell Sci*. **118**:223-32.

Spieth, J. and Lawson, D. (2006) Overview of gene structure. *WormBook, ed*. The *C. elegans* Research community, WormBook, WormBook, doi/10.1895/wormbook.1.65.1, http://www.wormbook.org.

Spilker, AC. *et al.* (2009) MAP Kinase Signaling Antagonizes PAR-1 Function During Polarization of the Early *Caenorhabditis elegans* Embryo. *Genetics.* **183(3)**:965-77.

Starich, TA. *et al.* (1995) Mutations Affecting the Chemosensory Neurons of *Caenorhabditis elegans*. *Genetics*. **139(1)**:171-188.

Stein, LD. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1(2)**:E45.

Suh, N. *et al.* (2006) The GLD-2 poly(A) polymerase activates gld-1 mRNA in the *Caenorhabditis elegans* germ line. *Proc Natl Acad Sci U S A.* **103(41):**15108-12.

Swoboda, P. *et al.* (2000) The RFX-type Transcription Factor DAF-19 Regulates Sensory Neuron Cilium Formation in *C. elegans*. *Mol Cell.* **5(3):**411-21.

Thacker, C. *et al*. (2006) *Caenorhabditis elegans dpy-5* is a cuticle procollagen processed by a proprotein convertase. *Cell Mol Life Sci.* **63(10):**1193-204.

Tokovenko, B. *et al.* (2009) COTRASIF: conservation-aided transcription-factor-binding site finder. *Nucleic Acids Res*. **37(7)**:e49.

Trapnell, C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. **25(9)**:1105-11.

Uyar, B. *et al.* (2012) RNA-seq analysis of *C. briggsae* transcriptome. *Genome Res*. doi:10.1101/gr.134601.111.

Wang, L. *et al*. (2002) A regulatory cytoplasmic poly(A) polymerase in *Caenorhabditis elegans*. *Nature*. **419(6904)**:312-6.

Wei, C. *et al*. (2005) Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res*. **15(4):**577-82.

White, JG. *et al*. (1986) The structure of the nervous system of *Caenorhabditis elegans*. *Philos Trans R Soc Lond B Biol Sci.* **314(1165)**:1-340.

Witze, ES. *et al*. (2009) *C. elegans* pur alpha, an activator of end-1, synergizes with the Wnt pathway to specify endoderm. *Dev Biol.* **327(1)**:12-23.

Ye, K. *et al*. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. **25(21)**:2865-71.

Yu, X. *et al*. (2012) How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Min.* **5(1)**:6.