UNIVERSITY OF CALIFORNIA

Santa Barbara

A novel forward genetic pipeline for the identification of genes required for transdifferentiation based on computational analysis of whole-genome sequences of large numbers of mutants

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Molecular, Cellular, and Developmental Biology

by

Tsung-Han Yeh

Committee in charge:

Professor Joel Rothman, Chair

Professor Denise J. Montell

Professor Julie H. Simpson

Professor William Smith

December 2022

The dissertation of Tsung-Han Yeh is approved.

Denise J. Montell

Julie H. Simpson

William Smith

Joel H. Rothman, Committee Chair

December 2022

A novel forward genetic pipeline for the identification of genes required for transdifferentiation based on computational analysis of whole-genome sequences of large numbers of mutants

By

Tsung-Han Yeh

ACKNOWLEDGEMENTS

I want to thank Sihe Tan, Martin Koch, and Gianna Giovannoni for their dedication to this project. Thank you to my advisor, Joel Rothman, for the years of opportunities and mentorship, and my committee members, Denise Montell, Julie Simpson, and William Smith, for their advice and wisdom. I’m grateful to all members of the Rothman lab for the camaraderie, feedback, and friendship. Thank Dr. Pan-Young Jeong, for starting this fantastic project and providing data for me to analyze.

I want to thank Dr. Jenn-Yah Yu for his mentorship in my master's study and his support in encouraging me to study abroad. I want to thank my loyal feline friend Jerry for his company on endless nights. I am thankful to my mother, Shu-Fen Li, and my two sisters, Yi-Chun and Tze-Chun, for their unconditional love and encouragement.

VITA OF TSUNG-HAN YEH

December 2022

EDUCATION

Ph.D. in Molecular, Cellular, and Developmental Biology, December 2022, University of California, Santa Barbara, Santa Barbara, CA
Master in Life Sciences and Genome Sciences, Jun 2013, National Yang-Ming University, Taiwan
Bachelor of Life Sciences, Jun 2011, Fu Jen Catholic University, Taiwan

PROFESSIONAL EMPLOYMENT

University of California, Santa Barbara
Teaching Assistant
MCDB 111: Human physiology, Summer, 2022
MCDB 101B: Molecular Genetics II, Spring 2019, 2020, 2021, 2022
MCDB 101B: Molecular Genetics II, Winter 2021, 2022
MCDB 101A: Molecular Genetics I, Fall 2017, 2019, 2021
MCDB 111: Human physiology, Fall 2020
MCDB 101B: Molecular Genetics II, Summer 2020
MCDB 103L: Molecular Cell Biology Lab, Winter 2018, 2019, 2020
MCDB 103: Molecular Cell Biology, Fall 2018
MCDB 104L: Recombination DNA Lab, Spring 2018
MCDB 1BL: Intro Biology Lab II, Winter 2017

National Taiwan University Hospital
Research Assistant (Aug 2015- Jul 2016)
Established the organoids culture protocol for human adipose-derived stem cells and demonstrated its potential in endothelial differentiation.

PUBLICATIONS

Huang Y.C., Chen K.H., Chen Y.Y., Tsao L.H., **Yeh T.H.**, Chen Y.C., Wu P.Y., Wang T.W., Yu J.Y. (2021) "βPS-Integrin acts downstream of Innexin 2 in modulating stretched cell morphogenesis in the Drosophila ovary." G3 (Bethesda), 11(9):jkab215.

Hsu T.H., Yang C.Y., **Yeh T.H.**, Huang Y.C., Wang T.W., Yu J.Y. (2017) "The Hippo pathway acts downstream of the Hedgehog signaling to regulate follicle stem cell maintenance in the Drosophila ovary."

Yeh T.H., Huang S.Y., Lan W.Y., Liaw G.J., Yu J.Y. (2015) "Modulation of cell morphogenesis by Tousled-like kinase in the Drosophila follicle cell." *Developmental Dynamics,* 244(7):852-65

Lin T.H. *, **Yeh T.H. ***, Wang T.W., Yu J.Y. (2014) "The Hippo pathway controls border cell migration through distinct mechanisms in outer border cells and polar cells of the Drosophila ovary." *Genetics*, 193(3):1087-99 (* contribution equally)

AWARDS AND FELLOWSHIPS

Higher Education Emergency Relief Fund III, UCSB, 2022
Jane Altman Memorial Fellowship, UCSB, 2018

FIELDS OF STUDY

Major Field: Developmental Biology

Investigating cell migration and stem cell maintenance in the Drosophila ovary system with Dr. Jenn-Yah Yu, National Yang-Ming University, Taiwan
Investigating the effect of oncogene $P15^{PAF}$ on cancer metastasis in the cell culture system with Dr. Jin-Mei Lai, Fu Jen Catholic University, Taiwan

ABSTRACT

A novel forward genetic pipeline for the identification of genes required for transdifferentiation based on computational analysis of whole-genome sequences of large numbers of mutants

By

Tsung-Han Yeh

How cell fate is decided and maintained has been an ongoing question for decades. Early embryogenesis studies showed that embryonic stem cells are pluripotent and can differentiate into every cell type in vivo or in vitro by introducing the corresponding factors. Later, the cellular reprogramming discovered by Yamanaka opened a new area of generating induced pluripotent stem cell (iPSC) from a terminal differentiated post-mitotic cell. In this study, we aimed to develop a new forward genetic screening pipeline to illuminate the genetic requirements underlying developmental plasticity and to investigate how post-mitotic differentiated cells in an intact animal can be reprogrammed and remodeled into new cell types in the process of transdifferentiation (Td). In C. elegans, forced ubiquitous expression of the ELT-7 GATA-type transcription factor, which functions in intestinal differentiation, is capable of converting differentiated, post-mitotic cells of two organs, the pharynx and uterus, into cells with gene expression patterns and ultrastructural characteristics of normal intestine cells. The developmental arrest phenotype was also observed in ELT-7-mediated Td animals, plausibly due to the conversion of pharyngeal cells into intestine-like cells. Reverse genetic approaches have been applied to knock down or knock out hundreds of candidate genes to uncover the molecular mechanism of ELT-7-mediated Td; however, no positive result was found. This result suggests that none of the candidate genes was involved in Td, or the down-regulation of candidate genes approach was not a proper experimental design for identifying Td. For example, genes required for intestinal Td may also be crucial for normal animal development; thus, a regular gene knock-down approach is unsuitable. To address this problem and widen the search for candidate genes in Td, a large-scale forward genetic screening using the auxin-inducible degradation system to ubiquitously express ELT-7 was performed by selecting suppressors of development arrest. We hypothesize that genes required for ELT-7-mediated Td will more likely be selected in the forward genetic screening and show many unique mutations. Mutations in those Td-required genes would also have a high incidence of disrupting their protein function by affecting the conserved residues or the active domain. A total of 660 mutant lines were isolated that escaped developmental arrest after ELT-7 overexpression. Mutant lines were pooled into different group sizes for whole-genome sequencing and a hundred thousand SNPs were predicted and evaluated with the three analyses: 1) The mutation density in a gene, 2) The percentage of mutations in conserved amino acids, and 3) The distribution of mutations in a gene. This analysis identified the positive experimental control, *elt-7* transgene, and among the top candidates identified, *cdk-12* was confirmed as a causally associated gene required for ELT-7-mediated Td through complementation tests and rescue experiments. This novel pooled mutant sequencing strategy was highly efficient since it did not require procedures to pre-screen mutations that affect the system or clean up the background mutations before fine mapping. It also applies to non-viable phenotypes or genetic manipulation unavailable in the animal. Finally, pooled mutant sequencing was cost-effective, with 10x less expense compared to single animal sequencing. This method was ideal for organisms with large brood sizes, such as bacteria, yeast, worms, and flies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One

Introduction

Modern Developmental Landscape

In 1957, C.H. Waddington proposed his concept of how gene regulatory networks regulate development without insight into the understanding of DNA and the function of genes. Later, this concept became known as Waddington's epigenetic landscape theory (Goldberg et al., 2007). In Waddington's epigenetic landscape, life starts with a single pluripotent cell which has the potential to differentiate into every cell type. This cell will then divide into two daughter cells, two daughter cells into four cells, and so on, ultimately resulting in the construction of a complete organism. Throughout this process, daughter cells of the pluripotent cell gradually lose their pluripotency and cell fate ultimately becomes generally restricted to that of a single type. C.H. Waddington proposed that the epigenetic tension, which is the differentiation status and gene regulatory networks of individual cells, would gradually increase and is required for proper cell differentiation.

Several discoveries have shown that the notion of Waddington's epigenetic landscape is generally correct. One example is the multipotent-to-commitment transition, which occurs during *C. elegans* embryogenesis, most or all blastomeres can be transformed into a particular cell type by ectopically expressing the transcription factors that normally specify these identities, suggesting they remain multipotent; however, the differentiation potential of the blastomeres is lost when embryogenesis progresses to ~100 cells (4E stage) and they fully commit to their appropriate cell fate at ~200 cells (8E stage) (Spickard et al., 2018). Similar phenomena are observed in mammalian embryogenesis. In mice, pluripotent primitive ectoderm/epiblast (PE) and embryonic stem cells (ESCs) derived from inner cell mass can both contribute to three germ layer formation *in vitro* and are capable of forming a chimera when injected into host blastocysts; however, only cultured ESCs have the capability to undergo unlimited proliferation and retention of pluripotency, revealing fundamental differences in plasticity between PE and ESCs (Keller, 2005).

Waddington's epigenetic landscape has been a milestone in the understanding of developmental biology and epigenetics, but the hypothesis was focused on an idea of a linear and irreversible development progression, which is that cells will eventually differentiate into a terminal fixed cell type; however, recent studies have shown that the developmental process is not always one-way. For example, a fully differentiated somatic cell can be reprogrammed to adopt the fate of its ancestor, i.e., induced pluripotent stem cell (iPSC) (Takahashi and Yamanaka, 2006). Further, the fate of a fully differentiated somatic cell can be directly transformed into that of another fully differentiated cell type with or without de-differentiation of its progenitor status, i.e., Td (Graf, 2011). These findings not only open a window for potential cell and gene therapies in humans but also demonstrate that cell fate can be more plastic than originally thought.

Reprogramming (Induced Pluripotent Stem Cells, iPSCs)

ESCs were first discovered in early mouse embryos (Evans and Kaufman, 1981). Later, this study led to the isolation of ESCs by culturing inner cell mass from early human embryos with condition medium *in vitro* (Thomson et al., 1998). Although both mouse ESCs and human ESCs displayed unlimited pluripotency and proliferation; however, gene regulation networks and signaling pathways that control differentiation were significantly different in these two cell types; therefore, human ESCs have an irreplaceable role in studying human embryogenesis (Romito and Cobellis, 2016). However, conducting research on human ESCs have been challenging due to the limited resource to obtain ESCs and the ethical issues. K. Takahashi and

S. Yamanaka first demonstrated that pluripotent stem cells could be induced from somatic fibroblasts under ES cell culture conditions with the introduction of four transcription factors, Octamer-binding transcription factor 3/4 (Oct3/4), Sex determining region Y (SRY)-box 2 (Sox2), Krüppel-like factor 4 (Klf4), and cellular-Myelocytomatosis (c-Myc) (OSKM), which provided an alternative way to study early human development, cellular reprogramming, and providing the potential on cellular therapy (Al Abbar et al., 2020; Takahashi and Yamanaka, 2006).

The general mechanisms that lock down differentiated cell fates include DNA methylation, histone modification, and chromosome remodeling (Miller and Grant, 2013). These chromosome modifications usually suppress pluripotency and proliferation genes and open chromatin regions that allow differentiation gene expression. OSKM-mediated reprogramming has shown the opposite effect on these chromosome modifications and the outcome through direct or indirect down-regulation of differentiation genes and up-regulation of pluripotency genes. Transcription analysis revealed a high correlation between gene expression profiles in iPSC and ESC; specifically, three transcription factors, Oct4, Sox2, and Nanog, play a critical role in pluripotency and maintenance of ESC identity (Narayan et al., 2017). Oct4 and Sox2 function as transcription activators, forming an interconnect autoregulatory circuitry with other essential regulators to express the essential genes required for reprogramming and pluripotency, such as Nanog. C-Myc has been reported to bind to methylated H3K4me2 and H3K4me3 in a large region of the somatic genome, allowing the Oct4 and Sox2 to access genes necessary for reprogramming (Martinato et al., 2008). The transcription factor, Klf4, also showed direct protein interaction with Oct4 and Sox2 that results in the activation of pluripotency genes (Wei et al., 2009).

The detailed reprogramming mechanism remains largely unknown due to the complex autoregulatory circuitry and gene interactions. Reprogramming efficiency is different between cell types as well. For instance, 1% of foreskin keratinocytes can be reprogrammed into iPSCs within 2-3 weeks of culture, but less than 0.01% of foreskin fibroblasts can be reprogrammed and require 1 month of culture (Egusa et al., 2012). The low accessibility of reprogramming in foreskin fibroblasts may be due to the higher epigenetic tension since a fibroblast is a terminally differentiated fate, whereas keratinocytes remain progenitors. Genes required for reprogramming also vary from species to species. For instance, LIF and Bmp4 are required for mESCs maintenance, but LIF is dispensable for hESCs, and the presence of BMP4 in the culture media leads to trophoblast differentiation in hESCs (Hirai et al., 2011). These findings again highlight that cellular reprogramming is not only different between species but also largely depends on the cellular context, i.e., epigenetic status, between different cell types.

Transdifferentiation (Td) and Metaplasia

Unlike reprogramming, which is a transformation from unipotent to pluripotent ancestor cells, Td transforms one differentiated cell type to another by de-differentiating to the progenitor cells or through direct transformation without rewiring to their previous developmental status. Td can happen both naturally and experimentally. Natural Td has been observed in lens regeneration, in which iris cells undergo partial de-differentiation to their progenitors followed by differentiation into lens cells (Tsonis and Del Rio-Tsonis, 2004). Td has been reported as part of the natural development of *C. elegans* as well. The post-mitotic epithelial cell Y cell is part of the rectum, but later on, it migrates forward and transdifferentiates into a motor neuron named PDA (Jarriault et al., 2008). Natural Td in humans has been discovered and considered a disease, including Barrett's esophagus which is Td of the stratified squamous to columnar epithelium, and liver fibrosis which is Td of hepatic stellate cells to myofibroblasts (Thowfeequ et al., 2007). Td rarely happens naturally, probably because the epigenetic tension builds up during differentiation to prevent cell type transformation; therefore, most examples of Td are seen under experimental conditions *in vitro* and *in vivo*. For instance, B cells can transdifferentiate into macrophages by introducing CEBPα and CEBPβ which suppress B cell-specific gene, Pax5, and activate macrophage-specific genes. Hepatocytes can transdifferentiate into pancreatic cells by overexpressing PDX1, which suppresses the hepatic transcription factor CEBPβ (Jopling et al., 2011). Both cases involved de-differentiation since the downregulation or deletion of Pax5 results in B cell de-differentiation. Downregulation of CEBPβ also cause reduced mature hepatocyte gene expression, which is required for PDX1-mediated Td (Jopling *et al.*, 2011).

Similar to the epigenetic tension in reprogramming, generally only a certain cell type can be transdifferentiated into another, presumably due to the certain epigenetic modifications that may physically restrict the binding of ectopic expressing transcription factor to the target sites or a positive feedback loop is involved in maintaining a terminal cell fate, thus preventing random fluctuation of gene expression.

ELT-7 mediated intestinal Td in *C. elegans*

The entire intestine in *C. elegans* arises from a single founder cell, the E blastomere, in the 8-cell embryo. Specification of the endoderm begins with the maternally-derived bZIP-like transcription factor, SKN-1, in the EMS blastomere of the 4-cell embryo. The EMS blastomere then asymmetric divides into EMS and E cells caused by the Wnt signaling pathway from the

P2 cell, leading to a low level of endoderm suppressor POP-1 in the nucleus and, at the same time, converting POP-1 into an endoderm gene activator in the E cell. With the role of transcription factor POP-1 in the E cell, SKN-1 directly activates the endoderm-specifying GATA-type transcription factors END-1 and END-3 to specify the endoderm cell fate of the E cell. The expression of END-1 and END-3 *3* starts at the 1E stage and lasts until the 8E stage; meanwhile, END-1 and END-3 activate ELT-2 and ELT-7, single zinc-finger GATA-type transcription factors, to activate thousands of intestine genes regulating terminal intestinal differentiation (McGhee, March 27, 2007). The signaling pathways involved in endoderm specification and differentiation have shown redundancy. For instance, knockdown or knockout of *skn-1* or *pop-1* alone is not sufficient for the complete elimination of the endoderm development (Torres Cleuren et al., 2019). Single mutation of *end-1* and *elt-7* also showed no detectable phenotype of gut development defect (Ewe et al., 2022). These multiple redundancies ensure the proper development of the intestine and may be an evolution outcome that supports the quick development cycle of *Caenorhabditis*.

Early studies suggest that Td cannot happen in the post-mitotic embryo stages; however, Riddle *et al*. showed that ubiquitously overexpressing END-3, ELT-2, or ELT-7 could trigger intestinal Td in the differentiated pharynx and somatic gonad (uterus) with varying efficiency (Riddle et al., 2016). *C. elegans* undergoing Td also show development arrest possible due to the starvation caused by the transformation of the pharynx into the intestine and thus loss of the muscle pumping ability, or Td happens in unidentified cells that are essential for animal development.

The detailed mechanism of ELT-7-mediated Td remains largely unknown. Two major questions about this process are: (1) why does Td happen only in the pharynx and uterus but not in other tissue? And (2) what are the molecular mechanisms that allow this Td to occur? Ubiquitously overexpressing ELT-7 causes extensive expression of ELT-2 in the animal, implying that the activation of ELT-2/ELT-7 positive feedback loop is not restricted to the pharynx and uterus. However, tissues other than the pharynx and uterus failed to express the late intestine differentiation gene IFB-2, suggesting intestinal Td required other factors. For example, (1) pioneer factors may be required before or during the Td process to establish the proper cellular contexts. Indeed, knockdown *pha-4*, the pioneer factor required for pharynx differentiation, abolishes pharynx-to-intestine Td due to the lack of pharyngeal cells (Riddle *et al.*, 2016). This finding suggested that ELT-7-mediated Td largely depended on the cellular context. In this case, only the fully differentiated pharynx is susceptible to such Td. (2) Studies have shown that histone modification is one of the crucial factors for successful cellular reprogramming and Td. Spickard's study demonstrated that reduced uterus-to-intestine Td can be found when knockdown of *epc-1*, one of the highly expressed genes in the intestinal Td process and predicted function in the regulation of transcription and histone acetyltransferase activity. However, the role of EPC-1 in ELT-7-mediated Td remained unclear since slow development and gonad development defects were found when *epc-1* was down-regulated; therefore, the reduced intestinal Td in the uterus may be a result of the lack of the uterus cellar context. (3) Hypothetically, the protein degradation system may be an important factor in Td as well since the cellular structure and context are required dramatically changing during Td. Proteins from the previous cell fate needed to be quickly degraded and recycled for new cell fate. Spickard showed that knockdown *smo-1*, another highly expressed gene in the intestine Td process and functions as a ubiquitin-like modifier involved in SUMO pathway, reduced uterus-to-intestine Td. However, similar to the *epc-1* knockdown, down-regulation of *smo-1* also causes significant gonad development defects, making it difficult to separate the requirement of protein degradation and the cellular context of the uterus in ELT-7-mediated Td. (4) Studies have also shown that immune response activation is required in Td, in the process called (Meng et al., 2017). Indeed, Spickard's work revealed that many immune response genes, specifically the intracellular pathogen response, are highly upregulated in intestinal Td based on transcriptomic analysis. However, knockdown of genes involved in intracellular pathogen response or innate immune response did not affect ELT-7-mediated Td. It is possible that the transflammation is regulated by multiple immune response pathways in this case; therefore, down-regulate one gene or pathway at a time is not sufficient to block Td. It is also possible that the immune response is a secondary effect that reflects the dramatic cellular changes in the animal and is not directly involved in ELT-7-mediated Td.

Conclusions and Perspectives

Reprogramming of iPSCs has shown great potential in clinical cell and gene therapy; however, how to induce proper differentiation and ensure the safety of using ultimate proliferation stem cells remains a major challenge. Td directly transforms a terminal differentiated somatic cell into another, allowing a more direct differentiation route and shortening the *in vitro* incubation timing, providing another option for clinical usage. In addition, transdifferentiated cells *in vivo* have shown an increased probability of developing into cancer cells, showing the urgency of studying Td for better health quality. Understanding the fundamental mechanism of Td not only satisfies our curiosity about cellular plasticity but can also provide insight into carcinogenesis and possible treatments.

In this study, we will first focus on developing a novel bioinformatic analysis pipeline to facilitate discovering the underlying mechanism of ELT-7-mediated Td. The analysis pipeline will then apply to a large-scale forward genetic screening to identify the potential causal gene in the intestinal Td.

Chapter Two

Developing a novel genetic and bioinformatic analysis pipeline based on pooled mutant whole genome sequencing approach

Summary

Forward genetic screening of randomly mutagenized model organisms has been a reliable method in many studies for discovering unknown gene functions and mechanisms. One of the biggest challenges of this strategy is identifying the relevant causal genes mutated in the identified mutants. Several approaches have been proven effective, including classical genetic linkage mapping, single-nucleotide polymorphism (SNP) mapping, and whole genome sequencing (WGS); however, all these methods require laborious crossing schemes. In this study, we developed a novel mapping strategy that relied on whole genome sequencing of pooled genomic DNA obtained from large numbers of mutants and processed the sequencing data in statistical and bioinformatics studies. Three analyses were applied to find the candidate genes: (1) The normalized frequency of unique mutation in the gene; (2) The percentage of mutation on conserved residue that potentially affects protein function; and (3) The distribution of mutations within a gene to evaluate whether a particular region, such as functional domain, was being selected during the mutagenesis screening. The experimental positive control was identified and confirmed by these three analyses. Among the top candidate genes, *cdk-12*, which has a major function in transcription regulation, was confirmed its requirement for ELT-7-mediated TD. Different from most of the existing mapping methods, which required various genetic manipulations in the animal to identify causal genes, the pooled mutant sequencing method utilized the power of big data bioinformatic analysis to predict the potential candidate genes without any genetic crossing. The pooled mutant sequencing method is also fast and cost-efficient compared to other WGS-based mapping techniques. It has the potential to reveal multiple candidate genes at once and discover the potential gene regulation network. Finally,

this method can be adapted to model organisms with large offspring sizes, such as bacteria, yeast, worms, or flies, showcasing its broad application potential.

Introduction

Strategies for discovering gene function.

Two major approaches to linking gene identity with function are loss-of-function and gain-of-function experiments. The first method studied gene function by observing the changes in individuals at the phenotypic level when the gene is down-regulated or mutated and speculating the natural function of its. It is the most common method to study gene function because the experiments were performed in a more natural environment; however, this method cannot answer the sufficiency of a gene nor the effect when the gene is misexpressed in different cells or at different development times. To address these issues, the gain-of-function was introduced. The gain-of-function experiment includes artificially increased gene expression by introducing ectopic transgene in cells that normally don't express such the gene or developmental timing that the gene is usually turned off and mutations that increase gene expression or alter gene function. Gain-of-function experiments are often considered artificial; however, it has revealed important discoveries in biology, for instance, the proto-oncogene *Ras* in the tumor development (Pylayeva-Gupta et al., 2011), the *Hox* genes in *Drosophila* segment development (Mark et al., 1997), the cellular reprogramming (Takahashi and Yamanaka, 2006), and ELT-7 mediated Td in *C. elegans* (Riddle et al., 2016).

Forward genetic methods.

Forward genetics provides an unbiased, objective approach to identifying unknown gene functions. In contrast to reverse genetics, which utilizes the known sequence identity of a gene to customize sequence-specific modification by homologous recombination or CRISP-Cas9 techniques, forward genetics involves randomly mutagenizing the whole genome and selecting or screening individual mutants with phenotypes that are the interest of the project. The process does not require prior knowledge of gene function and so it is an ideal method to identify unknown gene functions. Since the phenotype of interest will only be selected when the causal gene is mutated in a way that mutes or change its original function; therefore, it is crucial to have a large number of mutagenized progenies to create mutations that cover the whole genome. Two parameters have to be considered in such screening: the mutation rate of the gene and the nature of the experiment animals, ideally the animal with a short life cycle and a large brood size. For example, if a gene mutates in low frequency, then a large number of animals will be needed for the screening to ensure every gene has the chance to be mutated; vice versa, the number of animals subjected to the screening can be reduced if the mutation rate is high. Therefore, increasing the mutation rate and finding the ideal model organism for the screening is important.

Naturally occurring mutations exist with very low mutation rates, from 10^{-3} to 10^{-5} per base per generation in RNA viruses and 10^{-8} to 10^{-9} per site per generation in most animals; therefore, it is challenging to rely on naturally occurring mutations for experiments (Domingo et al., 2021; Kumar and Subramanian, 2002; Lynch, 2010). Artificially triggered mutation events are therefore used to increase the rate of mutations. For example, point mutations can be triggered by using chemical mutagens ethyl methanesulfonate (EMS), which reacts with guanine in DNA, causing replacement of thymine for cytosine during DNA replication with a mutation rate of $5x10^{-4}$ per gene. Therefore, in theory, every gene would be mutated at least once in every 2000 EMS-treated animals ($5x10^{-4}$ x 2000 = 1).

Animal brood size is another critical factor needed to be taken into consideration since not every mutation would give the phenotype of interest; therefore, a large number, from thousands to millions, of mutagenized animals are required for the phenotypic screening to allow mutants to be isolated. Organisms with short life spans and many offspring are ideal candidates for such a framework; bacteria, yeast, worms, and fruit fly have been shown to be successful in this approach. For example, with the short life cycle (3 days) and large brood size (~300 progenies/ adult) of *C. elegans*, depending on the screening or selection procedures, it is possible to analyze up to millions of animals for genetic screening in just two weeks, making *C. elegans* a powerful model organism for such approach (Hodgkin and Barnes, 1991).

Identifying causal genes through Next Generation Sequencing.

Next generation sequencing, specifically whole genome sequencing, has become a reliable tool for identifying mutations from forward genetic studies, combined with traditional gene cloning methods such as preliminary genetic mapping, complementation tests, and rescue experiments. Assuming the animal carries a homozygous recessive mutation, the single nucleotide variant (SNV) of the causal mutation would have an allele frequency, which is how common an allele is in a population or the genome, of a nucleotide that is different from that of the reference genome equal to 1. However, the individual mutant carries not only the causal SNV but also numerous SNVs throughout the genome since mutagenesis is random. Several methods have been applied to solve such a problem. For example, backcrossing the phenotypic individuals to the parental or reference strain is the most straightforward and effective method. However, in most cases, a handful of background SNVs remain even after 10 rounds of backcrosses. Therefore, substantial downstream analysis is needed to confirm the causal gene (Fay, 2013). Other techniques that apply the power of whole-genome sequencing to identify the causal gene have shown success, such as SNP mapping and Genome-wide association studies. Still, they either required intensive outcrossing procedures or a large amount of sequence information from individuals that can become problematic in animals challenging for genetic crossing or that lack genetic diversity or genomic information between genetically distinct lines.

In this study, we have developed a novel method to identify potentially causal genes by analyzing pooled sequencing of 660 mutants computationally to assess the distribution of mutations across the entire pooled set, allowing us to narrow in the relevant sequence variants. To identify the causal mutations, we assess the density of mutations within a gene, the fraction of mutations that affect conserved residues, and the mutation distribution within a gene. Mutations in the *elt-7* transgene, the experimental positive control, are utilized for evaluating and refining the computational analyses. The sequencing data from the Million Mutation Project (MMP) is used as the experimental control since no selection was imposed on the animal other than the viability. Finally, a ranked candidate gene list is composed by comparing the performance of the three analyses in the AID mutant lines and the MMP data. A phenotypic causal gene required in the process of ELT-7-mediated Td is confirmed.

Results

Overexpressing ELT-7 results in pharynx-to-intestine Td and development arrest in the auxin-inducible degron (AID) system

The heat-shock (HS) system provided a rapid and efficient way to overexpress ELT-7. We found that *C. elegans* undergoes development arrest after a minimum of 2 mins of HS. This arrest reached 100% at 5 min of HS (Fig 2.1). However, the HS-*elt-7* system had several disadvantages in this project: 1) The system can be easily disrupted when the HS response genes are mutated. For example, *hsf-1*, a major transcription responsible for the HS response, was identified in our piolet selection for suppressor of ELT-7 mediated Td in the HS system (data not shown). 2) The ELT-7 is overexpressed through an integrated extrachromosomal array with hundreds of copies of the transgene; thus, the highly expressed ectopic ELT-7 maybe mask genes involved in Td in a subtle way. 3) The selection is made when the conditional switch is operating during the selective condition. To address these questions, we decided to overexpress ELT-7 using the AID system with single copy insertion.

The AID system is a post-translational regulation system that includes three major components: 1) TIR1, an E3 ubiquitin ligase, 2) Degron, a 45 amino acid that can be recognized by TIR1, and 3) Auxin, indole-3-acetic acid (IAA), that is required for TIR1-mediated protein degradation. When IAA is presented in the system, TIR1 will bind to the degron sequence and add ubiquitin to the target, leading to the degradation of the target protein (Zhang et al., 2015). JR3904, a transgenic strain carrying ubiquitously expressed Degron::ELT-7, TIR1, and a Td reporter IFB-2::CFP was constructed for this study. Like the HS *elt-7* system, JR3904 showed 100% development arrest at L1/L2 stages when embryos grew in the IAA-absent environment (Fig 2.2A). Interestingly, not all the arrest animals showed Td; only 68% (19/28) of them expressed the Td marker in the anterior bulb of the pharynx (Fig 2.2B, C). Td-positive arrest animals have slightly greater body lengths than Td-negative arrest animals, implying that development arrest may be independent of pharyngeal Td. In addition, the process of Td may require energy; thus, only animals that can intake bacteria prior to the development arrest could have enough energy for Td. Furthermore, animals with only one copy of *elt-7* transgene have greater body length than homozygous two copies of *elt-7* transgene, suggesting an ELT-7 dosage-dependent Td event (Fig 2.3A). Surprisingly, some of the one-copy *elt-7* transgene animals propagated with a minimal brood size with a Td pharynx on Day 6, possibly due to the incomplete or delayed Td (Fig 2.3B).

Although intestinal Td and development arrest was found in both the HS system and the AID system, the expression pattern of the Td marker, IFB-2::CFP, was different. The HS-*elt-7* system could transform the whole pharynx and uterus into intestine-like tissue; however, Td was only found in the anterior bulb of the pharynx and not the uterus in the AID-*elt-7* system. This difference is likely due to the *elt-7* transgene copy numbers difference. For example, hundreds of copies in the HS system vs. two in the AID system. In addition, studies have shown that Td is highly dependent on the correct cellular context. In the AID-*elt-7* animals, development arrest happened before the uterus was developed; therefore, lacking a uterus may result in the lack of Td in the uterus. In conclusion, Td in the AID-*elt-7* system was more subtle and sensitive than in the HS-*elt-7* system. It can be an ideal system for identifying subtle changing of Td in genetic screening.

660 development arrest resistance lines were isolated in the EMS-based forward genetic screening.

The AID-*elt-7* transgene animal, JR3904, was utilized for a large-scale forward genetic screening for discovering the mechanism underlining ELT-7-mediated Td. 58 EMS-mutagenized JR3904 founder animals ensuring broad independence were propagated on the IAA plate, and a total of 12.5 million F2 progenies were transferred to the NGM plate for selection of development arrest resistance phenotype. 660 viable mutant lines were established, frozen, and isolated with their genomic DNA for pooled sequencing. Mutant lines were divided into six groups based on the number of isolated lines per founder ratio to minimize the potential founder effect (Fig 2.4). For instance, group 1 contained the lowest ratio, 18 mutant lines from 10 founders, suggesting that every line was selected independently, and group 6 had the highest ratio, 223 mutant lines from 7 funders, implying that some of the progenies maybe share a similar or identical genetic mutation. Each group was then sequenced and mapped to the reference strain N2. The sequencing depth of every group ranged from 176 to 213 and had 98.8% mapping coverage for most groups, indicating that the sequencing data is of good quality. However, the sequencing depth per strain for every group showed huge variants since they were pooled with different mutant lines. For example, group 1 had about 10x sequencing depth per line, whereas group 6 had less than 1x (Table 2.1). Although the ideal sequencing depth is recommended with a minimal 10x to ensure accuracy, we decided to include all data (groups 1 to 6) regardless of the low sequencing depth for downstream analysis for complete sequence information. Variant calling was performed on all mutant lines using CRISP software, which identifies rare variants in multiple pooled samples based on the contingency table and evaluates the probability along with sequencing errors.

Four mutant lines, #2, #13, #22, and #24, were randomly picked to measure their development and Td to verify the development arrest resistance phenotype (Fig 2.5). All of them showed a normal developmental speed but various Td levels. For instance, #2 had Td in the pharynx at 86% (6/7) and in the somatic gonad at 29% (2/7); #13 had Td in the pharynx at 71% (5/7); #22 and #24 had no sign of Td at all. This finding suggested that mutation in each line affected Td differently. It may be due to the nature of the mutation, affected in different genes, or depending on where the pathway was affected. Overall, the screening was successful, and suppressors of development arrest caused by ELT-7 were selected.

elt-7 transgene was heavily mutagenized and selected in the screening.

A positive selection in forward genetic screening can result from the gene of interest being mutated or due to the disruption of the screening system. For example, the HS response can be blocked when *hsf-1*, a transcription factor required for HS response, is down-regulated by mutations or RNAi knockdown (Bar-Ziv et al., 2020; McMillan et al., 1998). We hypothesized that mutations on the *elt-7* transgene construct would also have a catastrophic effect on Td in the AID-*elt-7* expression system through decreased transgene expression. The *elt-7* transgene construct in mutant lines was examined by PCR sequencing to verify the hypothesis (Table 2.2). 36% (21/57) of lines had defects on the transgene, including nonsense and missense mutation on the *elt-7* transgene coding sequence (19%, 11/57) and deletion in the whole or part of the construct (17%, 10/57). This result proved that our hypothesis was correct. The escape of development arrest phenotype can be selected when mutations in the *elt-7* transgene construct disrupt the ectopic gene expression. The finding also suggested that an estimated one-third of the isolated lines were positive results due to the failure of the system.

Challenges on variants calling for pooled sequencing data

99,729 unique SNVs were detected using CRISP variants calling on all mutant lines (Fig 2.6). A 150-base pair sliding window was applied to visualize the density of SNVs on chromosomes. 714 windows were found to have equal or more than 20 SNVs (Fig 2.7B). The 714 windows were primarily found in the intergenic region where non-coding RNAs were located. Studies have shown that non-coding RNAs are essential in cellular and genetic regulations. Unfortunately, a closer examination revealed that most SNVs were in the repeated/duplicated sequences where the sequence reads have a high incidence of being misaligned to the reference. Indeed, SNVs in the 714 windows were mostly found in the variable nucleotides in the repeated motif, implying the false-positive result of SNVs due to the misalignment of the sequence in the intergenic region (Fig 2.7C).

The accuracy of variant calling is another factor that needs to be considered since most groups' sequencing depth was low. The *elt-7* transgene mutations confirmed by PCR sequencing were compared to the predicted SNVs. A total of 8 mutations were confirmed in the *elt-7* transgene, and the CRISP program successfully predicted 50% (4/8) of them. This prediction accuracy may result from the low sequencing depth of each group or the allele frequency of SNVs being too low to be distinguished from sequencing errors. Although the program failed to capture all the confirmed mutations, a total of 23 *elt-7* transgene mutations were predicted and 26% (6/23) of them were nonsense, which is considered one of the highest impacts of all the mutations (Table 2.3). This result supported the hypothesis that the AID-*elt-7* system can be easily interrupted by *elt-7* transgene mutations, and the predicted mutations confirmed it despite the average accuracy.

Developing mutation analysis pipeline based on the *elt-7* transgene mutations.

Unlike single mutant WGS, which usually existed ~200 SNVs in one animal before backcrossing, the pooled mutant sequencing provided 99,729 unique predicted SNVs from 660 mutant lines. With this tremendous amount of information, it is crucial to have a known positive control gene to validate the bioinformatic analysis pipeline. Therefore, we took advantage of the *elt-7* transgene mutations as the positive control. The result was compared with the mutations in endogenous *elt-7* from the Million Mutation Project (MMP) EMS-mutagenized data, which performed random mutagenesis without imposing any selection other than viability and isolated 748 lines as our experimental comparison (Thompson et al., 2013).

The first analysis strategy hypothesized that genes required for the resistance of development arrest would have a greater chance of being selected and showed a high density/frequency of unique mutations in that gene. The analysis counted the number of unique mutations on the protein-coding sequence (CDS) normalized by length. Indeed, when comparing *ttn-1*, one of the largest genes in *C. elegans* (45,483 bp CDS), with the *elt-7* transgene (597 bp CDS), 409 unique SNVs were predicted in *ttn-1* and 23 were in the *elt-7* transgene; however, after normalization, *ttn-1* downranked to 9 SNVs/kb and *elt-7* transgene unranked to 39 SNVs/kb (Fig 2.8, Fig 2.9). This method emphasized the fundamental strategy of the pooled mutant sequencing strategy and effetely identified *elt-7* transgene as the top candidate gene.

The second analysis strategy emphasized the impact of missense mutations on protein function (Fig 2.10). During evolution, protein residues responsible for their function tend to be highly conserved. We examine the percentage of mutations that happened on the conserved residues using The Sorting Intolerant From Tolerant (SIFT) program, which predicts whether an amino acid substitution affects protein function based on sequence homology and physical properties of amino acids (Ng and Henikoff, 2003). 94% (15/16) predicted missense mutations in the *elt-7* transgene were evaluated as deleterious, and 0% (0/2) of the endogenous *elt-7* mutations in the MMP were predicted as deleterious (Fig 2.10A). This result strengthened the role of *elt-7* transgene as a positive control and emphasized the importance of conserved residues in regulating protein function. With this analysis, the impact of various missense mutations can be evaluated and weighted accordingly, providing valuable insight into recognizing potential candidate genes.

Similar to the second analysis where conserved resides was essential for its function, the function domain(s) within a protein was also important, and mutations can be destructive in the active domain despite the conserved residues. Therefore, the Kolmogorov–Smirnov test (KS test) was imposed on examining the possibility that mutations impaired a function domain. The KS test calculated the most significant distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution and provided a statistical p-value. Furthermore, the KS test was used to calculate whether the distribution of mutations in a protein differed from the uniform distribution, suggesting that mutations clustered in a particular region or domain in a linear 2D structure. Not surprisingly, *elt-7* transgene mutation in the AID-*elt7* system was significantly different from the uniform distribution ($p < 0.001$), and the endogenous *elt-7* mutations in the MMP showed uniform distribution ($p = 0.45$) despite the lack of statistical power due to the small sample size (Fig 2.10B, C, E).

Finally, ELT-7 mutations were visualized in a 3D protein structure. Consistent with the KS test result, ELT-7 transgene mutations from the AID-*elt-7* system were primarily located in the DNA-binding domain (DBD). In contrast, endogenous ELT-7 mutations from the MMP were scattered outside DBD (Fig 2.10D). This result again validated the previous analysis and showed the power of visualizing mutation in a 3D structure. This analysis method can be most beneficial for proteins with a scattered function domain or the domain that only exists in a 3D structure, such as channel proteins with the multi-pass transmembrane feature.

The three bioinformatic analyses evaluated candidate genes.

The three bioinformatic analysis pipeline effectively recognizes *elt-7* transgene as the top candidate gene with the highest mutation density, high deleterious ratio, and significantly low p-value of the KS test. The same analysis was then applied to all the coding genes and filtered genes with greater mutation density, greater deleterious ratio, and lower p-value of the KS test in the AID-*elt-7* screening compared to the MMP. As a result, the top six candidate genes ranked by their mutation density were *elt-7* transgene, *D1086.9*, *ifb-2, fust-1, D1086.9,* and *cdk-12* (Table 2.4).

Not surprisingly, the *elt-7* transgene was the first candidate gene since it was a positive experimental control. The second gene, *D1086.9*, showed slightly better performance in all three analyses. However, this gene was a nematode-specific gene only present in *Caenorhabditis* and *Pristionchus*; due to the interest in understanding the general effect of ELT-7-mediated Td, the gene was not selected as our top priority. The third gene, *ifb-2*, was identified due to the hundreds of copies of *ifb-2p::IFB-2::CFP* transgene in the selection strain, JR3904. Therefore, predicted mutations on the endogenous *ifb-2* would likely result from the mutations on the *ifb-2* transgene. The fourth gene, *fust-1*, had an excellent performance on mutation density and the KS test. Despite a relatively low deleterious ratio, none of the five mutations in the MMP were deleterious; therefore, *fust-1* maybe was one of the causal genes in Td and needed more investigation. The fifth gene on the list, *K09B11.4*, similar to *D1086.9*, was a nematode-specific gene and lacked proper comparisons in the MMP; therefore, it was not in our best interest. The sixth gene, *cdk-12*, outperformed all analyses, having a 67% of deleterious ratio and relatively low KS test p-value; thus, we considered that *cdk-12* had a high incidence to be the actual causal gene.

The selection of candidate genes for further investigation considered all three analyses and a subjective point of view. *D1086.9* and *K09B11.4* were still possibly involved in ELT-7-mediated Td by regulating *C. elegans*-specific cellular physiology, and *fust-1* had a very promising probability of being a causal gene just like *cdk-12* as well. Unfortunately, the number of mutations in a gene dramatically decreased after *cdk-12*.

Discussion

Identifying the causality in genetics with efficiency and accuracy remained one of the biggest challenges in forward genetic screening. Through pooled mutant sequencing, the candidate genes can be evaluated with three aspects: the mutation density reveals how often a gene was mutated, the deleterious ratio indicates the severity of mutations affecting protein function, and the mutation distribution within a gene recognizes if a certain region or domain was targeted in the selection. Almost a hundred thousand predicted SNPs were evaluated with the three analyses, and the pipeline successfully identified the *elt-7* transgene, the *ifb-2* transgene, and at least two potential causal genes, *fust-1* and *cdk-12*. Furthermore, the pooled mutant sequencing identified candidate genes through pure bioinformatic analysis; therefore, the process did not require any genetic manipulation in the animal. In conclusion, we demonstrated a novel method for identifying candidate genes in a forward genetic screening. This method could be applied to any organism with a large brood size to support the initial screening and downstream bioinformatic analysis.

The positive control in the experimental design

Since the ELT-7-mediated Td was governed by ectopic expression of the *elt-7* transgene, it was not surprising that mutations in the *elt-7* transgene, specifically the *elt-7* transgene CDS, would disrupt the system and result in positive selection. PCR sequencing of the *elt-7* transgene CDS discovered 19% (11/57) of the lines contained 4 nonsense and 4 missense mutations. The variants calling prediction also indicated a potential total of 23 mutations on *elt-7* transgene CDS with 6 nonsense and 17 missense mutations; furthermore, 2 additional nonsense mutations were predicted in the degron sequence resulting in a non-functional truncated transgene protein. Mutations in the *elt-7* transgene not only served as a positive control for the experimental design and were an indicator for the successful prediction of the analysis methods with the highest mutation density among all genes, a high deleterious ratio, and a significantly low p-value of KS test.

Positive selection due to the system failure was usually considered a flaw in experimental design and required an initial cleaned up before any fine mapping; however, the pooled mutant sequencing approach included all the SNVs for bioinformatic analysis, and so the factors contributed to the system failure can be recognized and served as a positive control. The background SNVs will be down-ranked after the analysis pipeline. The pooled mutant sequencing approach provided an efficient pipeline to uncover candidate genes without prior genetic manipulations.

The false-positive result in the experimental design

The top third candidate gene on the list was *ifb-2*, a downstream differentiation gene to produce gut-specific intermediate filament. Knockout *ifb-2* in *C. elegans* prevents endotube formation and mildly affects the morphology of the microvillar, showing slow development, reduced brood size, altered survival, and increased sensitivity to microbial toxins. Although *ifb-2* was directly involved in intestine maturation, the mutations in the *ifb-2* were considered a false positive result due to: 1) The strain subjected to genetic screening, JR3904, carried an integrated multi-copy extrachromosomal array of *ifb-2p::IFB-2::CFP* transgene; therefore, there was a significant probability of mutagenizing the *ifb-2* transgene, and the mutations were mapped to the endogenous *ifb-2*. 2) Mutations in the endogenous *ifb-2* will be automatically rescued by the *ifb-2* transgene in the screening strain JR3904 and, therefore, would never be selected. This result again highlights the power of pooled mutant sequencing method that can identify multiple genes or factors that may contribute to a positive selection, depside their nature being an artifact or actual causality.

Advantages of pooled mutant sequencing strategy compared to single mutant sequencing.

The pooled mutant sequencing strategy was highly efficient because it did not require procedures to pre-screen mutations affecting the system or clean up the background mutations before fine mapping. Pooled mutant sequencing strategy also allows the application of non-viable phenotype or genetic manipulation unavailable in the animal. The whole analytic pipeline is done computationally through big data bioinformatic analysis. The pooled mutant sequencing provides additional prediction power that traditional single mutant sequencing cannot. For instance, novel functional domains can be recognized by analyzing the location of mutations in a gene. Genetic interaction or gene ontology analysis can identify gene regulatory networks or pathways on candidate genes. The function of non-coding sequences, such as non-coding RNAs or promoters, can be identified with an improved sequencing approach. Pooled mutant sequencing was also a cost-effective strategy. Sequencing technology is improving rapidly with a dramatically reduced cost. For example, sequencing a single C. elegans with 30x sequencing depth costs ~ 100 USD today; however, sequencing all 660 isolates will still cost ~66,000 USD. With the pooled sequencing method, only ~7,700 USD was spent on sequencing the 660 mutant lines with the tradeoff of low sequencing depth.

In conclusion, pooled mutant sequencing provided a novel way of performing and analyzing forward genetic screening data. Unlike most published methods of finding the candidate genes, the pooled approach did not require genetic crossing and had a high tolerance for potential system failure. The pooled approach was also highly cost-efficient; valuable genetic information was obtained in the study. The pooled method can be applied to organisms with large brood sizes, such as bacteria, yeast, worms, and flies, showing its broad potential application.

Limitation of pooled mutant sequencing strategy and potential improvements

The fundamental limitation of pooled mutant sequencing strategy was on the technical level, precisely the sequencing parameters. C. elegans genome contains tandem repeats and duplication sequences within the intergenic regions, so misalignment occasionally happened in those regions when sequencing was done with the short-read method. This problem can be solved by using soft or hard-masked reference genomes that ignore the intergenic regions or call variants with high allele frequency when sequencing is done in a single animal. Since the pooled sequencing method includes every SNV in consideration regardless of their allele frequency, the false-positive hit will be recognized even when a small fraction of misalignment was presented; therefore, the information in those repeated regions had to be discarded. The practical way to solve this problem was to use long-read sequencing methods such as Oxford

Nanopore sequencing and PacBio single-molecule real-time (SMRT) sequencing. However, both methods had higher sequencing error rates than paired-end illumine (10% vs. 0.1%), which can be another problem in acquiring accurate variants calling (Li et al., 2022; Workman et al., 2019). Obtaining accurate information in the intergenic regions remained challenging unless the sequencing method or mapping algorism improved.

This study demonstrated that much valuable genetic information could be identified even when the data was sequenced with a shallow sequencing depth (<1x). However, much of the information needed to be investigated more. Indeed, in group 6, only 81% of the genome was covered by sequence reads. The ideal practical depth to call variants was a minimum of 10x; therefore, the best practice would be to increase the depth to a minimum of 5x and more to obtain more accurate information (Jiang et al., 2019; Koboldt, 2020). Improving the sequencing depth will provide more accurate variants prediction, and new information may emerge and change the current ranking of the candidate genes, allowing a better decision of the downstream analysis.

Three bioinformatic analyses introduced in this study each showed their value for finding candidate genes, but none can be used alone. For example, mutation density analysis highlights *elt-7* transgene and downranks *ttn-1*; however, *ttn-1* also has significant mutation distribution bias in the AID and MMP database. *ttn-1* was finally ruled out based on the deleterious ratio, which was lower in AID compared to the MMP database. Some improvements could be made to the current and future analyses. For example, the mutation distribution analysis (KS test) was performed based on the liner 2D position, which can be problematic if some multiple functional domains or domains only exist in a 3D structure. AlphaFold could be adapted to this strategy and used for calculating the distance of SNVs in a

3D structure. It is also possible to adapt machine learning to facilitate the discovery of the causal genes if there is more information, such as different ways of analyzing the data with confirmed causal genes as the learning subject for the machine.

Material and Methods

C. elegans incubation and genetics

The strains used in this study were: JR3899 (*eft-3p::TIR1::mRuby::unc-54 3'UTR; kcIs6 [ifb-2p::IFB-2::CFP] (IV))*), JR3904 (*unc-119 (-) wIS167[eft-3p::degron::ELT-7::EMGFP::unc-54 3'UTR]*), JR3904 (*ieSi57[eft-3p::TIR1::mRuby::unc-54 3'UTR + Cbr-unc-119(+)] II; wIS167[eft-3p::degron::ELT-7::EMGFP::unc-54 3'UTR] kcIs6 [ifb-2p::IFB-2::CFP] (IV)*), JR3642 *(wIs125[hsp-16-2::ELT-7 hsp-16–41::ELT-7]; rrIs01[elt-2p::lacZ::GFP + unc-119(+); kcIs6 [ifb-2p::IFB-2::CFP] (IV)],* EG8081 (*oxTi177 [ttTi5605 + NeoR(+) + unc-18(+)]*). Strains carrying the AID-*elt-7* transgene were grown on plate supplements with 500 μM IAA, except the AID-*elt-7* mutagenesis screening isolated lines. All the experiments were performed at 20 °C and kept the stocks at room temperature (20 °C to 23 °C).

C. elegans growth rate measurement

Development of N2 and JR3904 was measured in a liquid culture environment followed by standard protocol. In short, eggs were harvested from gravid hermaphrodites by the bleaching technique and cultured in the M9 buffer overnight to have synchronized L1s. L1s were then transferred to the S medium supplement with pelleted OP50 to grow with or without IAA. Animals were collected for imaging and measurement from Day 0 to Day 5. The growth rate of AID-*elt-7* mutagenesis isolates lines followed a similar procedure but culture on regular NGM plates.

Imaging

Animals were immobilized using 10 mM Levamisole and mounted on 4% agarose pads. All the images were captured using Nikon Eclipse Ti-E inverted microscope and Nikon NIS-Elements AR v4.13.05. The same exposure and threshold setting was used within each experiment to maintain consistency. Adobe Photoshop CC 2019 v20.0.7 was used for the figure display.

Building transgene construct and transgenic animal

The *elt-7* transgene vector, pCC1FOS (*ttTi5605::unc-119p::UNC-119::unc-119 3'UTR::eft-3p::degron::ELT-7::EMGFP::unc-54 3'UTR:: ttTi5605)*, from *pLZ29* was cloned by using Gibson assembly and injected into EG8081. *Elt-7* transgene animals, JR3903, were maintained on *elt-7* RNAi plate. AID-*elt-7* transgenic stain, JR3904, was generated by crossing JR3903 with JR3899 and maintained on the IAA plate.

Forward genetic screening

Synchronized L4 JR3904 was treated with EMS. 58 founder JR3904 were grown and propagated on the IAA plates for one generation. Generation 2 (F2) was then synchronized from gravid F1 hermaphrodites by the bleaching technique and placed on NGM plates for selecting the development arrest resistance phenotype. Development arrest escapers were singled and propagated as independent lines.

PCR sequencing

DNA samples were prepared by single worm PCR protocol. In short, a single or several animals were lysed in worm lysis buffer (100 mM Tris pH 8.3 (Fisher), 500 mM KCl (Fisher), 20 mM MgCl2 (Fisher), 1mg/mL proteinase K (GoldBio)) and incubated at 65 °C for 1 hour followed by 85 °C for 15 mins. Primers were synthesized from Integrated DNA Technologies IDT. PCR was performed with PCRBIO HiFi Polymerase (PCR Biosystems Inc.) and purified and sequenced by UC Berkeley DNA Sequencing Facility with magnetic SPRI bead technology and Sanger sequencing dGTP protocol. Sequences were visualized by using the ApE program.

WGS mapping and variants calling

Samples were sent to The McDonnell Genome Institute at Washington University for library construction and sequencing on HiSeqX. Reads were assembled and mapped to the reference genome assembly Wbcel235 by bwa tools with the command: *bwa mem -t 8 -C -p -reffile -bamfile*. Duplicated reads were marked by samblaster tools with the command: *samblaster -I stdin -o stdout–-excludeDups–-addMateTags -d -discordantfile -s splitterfile–-maxSplitCount 2–-maxUnmappedBases 50–-minIndelSize 50–-minNonOverlap 20*. Variants calling was done by CRISP tool with the command: *./CRISP -bams file_bam_paths -ref reference.fasta -VCF variantcalls.VCF -poolsize poolsize -bed targets.bed > variantcalls.log*

Bioinformatic analysis

All the analyses were done by R and shell with customized scripts. The number of variants within a 150-bp window was done by using slider package v0.1.0. In short, a 150-bp sliding window was moving on the chromosomes from the beginning to the end, and the number of variants presented in the window was counted. Findpeaks function from the pracma package v1.9.9 was used to identify the most concentrated variants region within a 150-bp window with a minimal 150-bp distance, and regions with equal or more than 20 variants were selected. Tandem repeat in the 150-bp window was identified using the Tandem Repeats Finder (TRF) program, which reported the repeated sequence motif and number of occurrences. The effect of the variants at the protein level was evaluated with Variant Effect Predictor (VEP) webtool. Variants on the protein-coding genes were then selected and normalized by their longest CDS if alternative splicing existed. The deleterious ratio was calculated by SIFT4G program. The mutation distribution was calculated using the ks.test function in r with uniform distribution. Information, including gene ID, name, feature, and position, was obtained from Wormbase. MMP data was obtained from Wormbase and the VC20xxx strains were selected for this study.

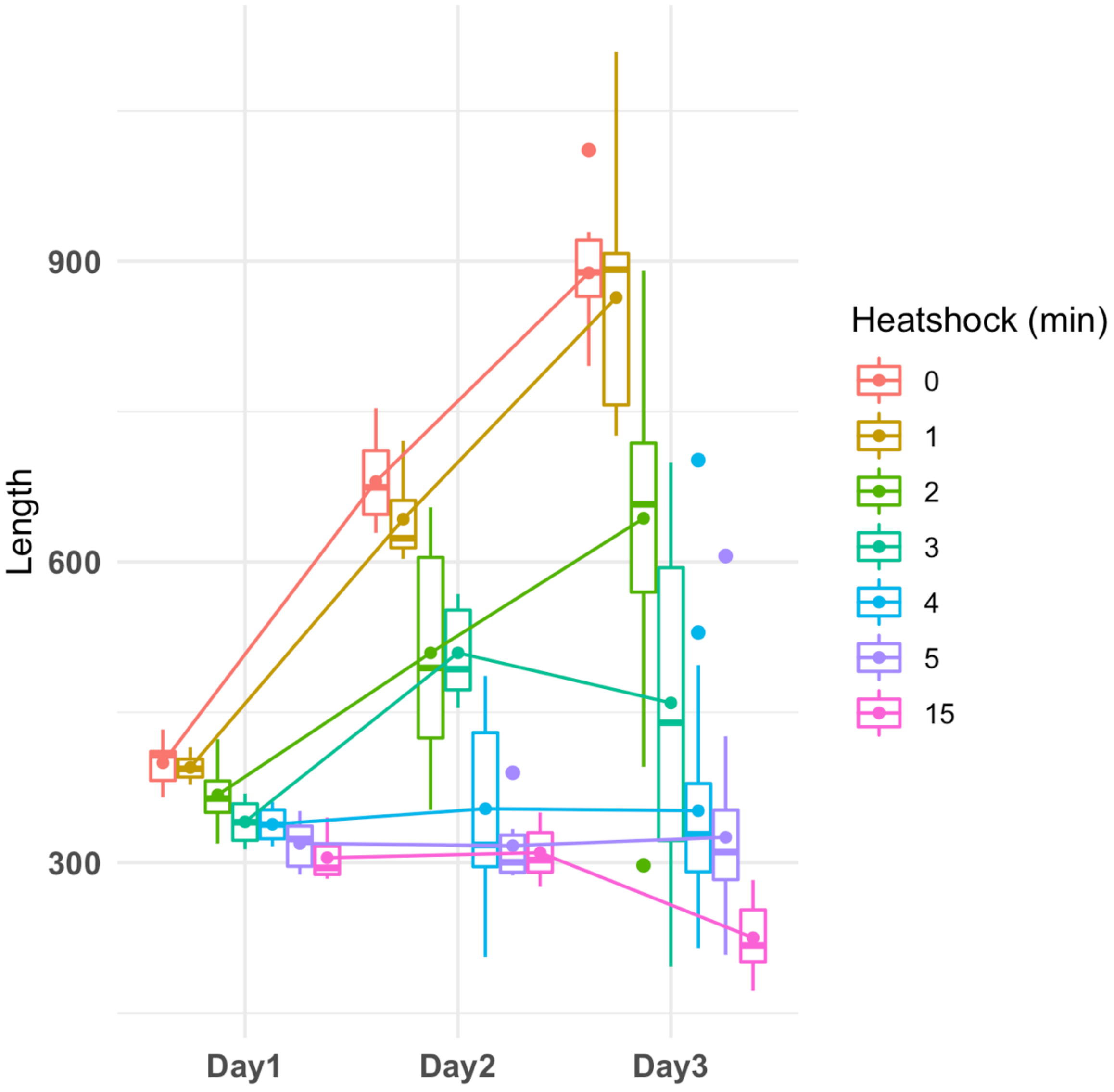


Fig 2.1. A quick and sensitive response to the ELT-7 overexpression in the HS system. Synchronized JR3642 L1 animals were fed on OP50 for 3 hours before HS with a thermocycler at various times. The development of heat-shocked animals was measured by their body length from Day 1 to Day 3. Animals showed development arrest after a minimal 3 mins HS and complete arrest with 5 mins HS. Each group on Day 1 contains 4-7 animals, 6-10 animals on Day 2, and 13-35 animals on Day 3.

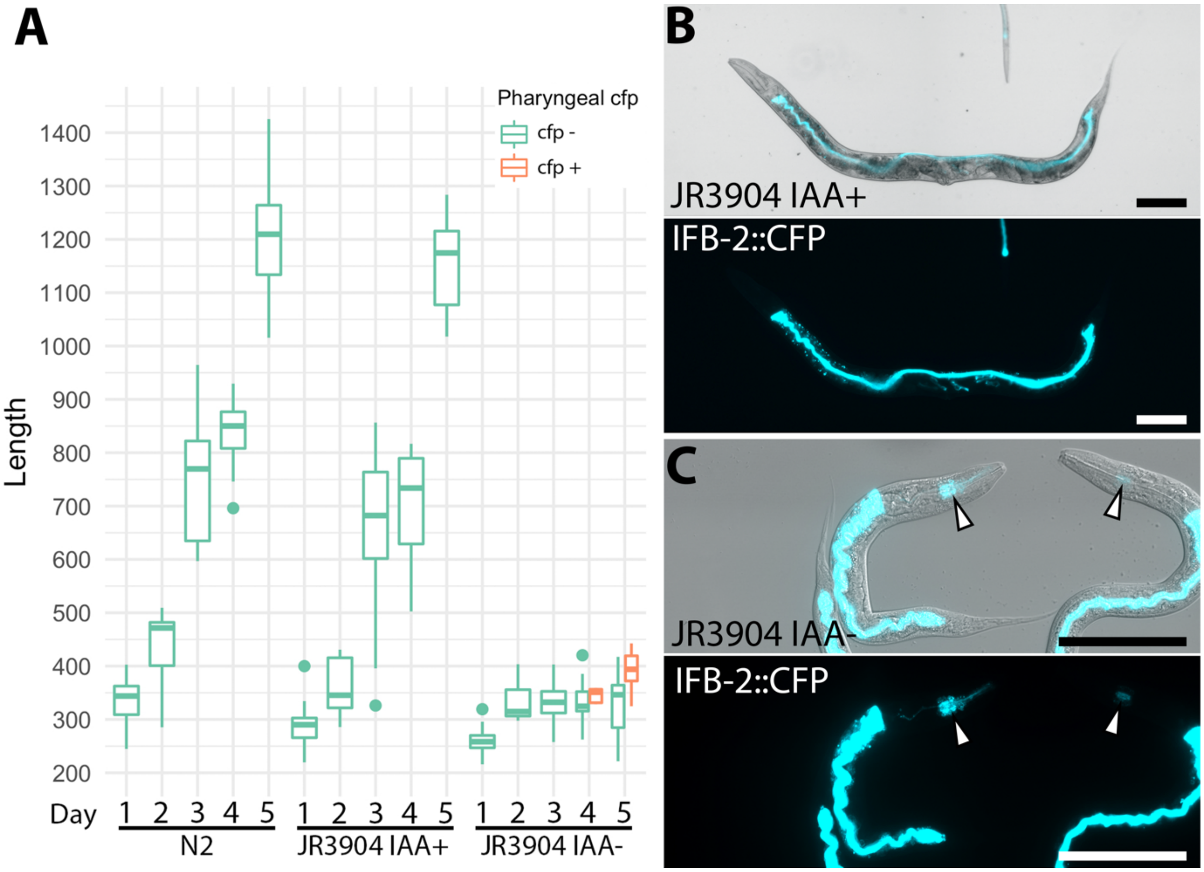


Fig 2.2. Overexpressing ELT-7 caused pharyngeal Td and development arrest in the AID-*elt-7* system.

(A) Synchronized L1 animals were cultured in liquid culture conditions with the presence or absence of IAA. Body length was measured from Day 1 to Day 5 as an indicator of animal growth. JR3904, the AID-*elt-7* strain, grew into adults and propagated normally when IAA was present in the system but failed to develop when IAA was absent. 29.4% (5/17) of JR3904 grew in the absence of IAA showing ectopic intestinal marker, IFB-2::CFP, in the pharynx on Day 4 and 67.9% (19/28) on Day 5. All groups had 6-42 animals. (B) JR3904 showed normal development when IAA was present in the system. (C) Ectopic intestinal marker was found in the anterior bulb of the pharynx, metaCorpus, when IAA was absent in the system (arrowheads). Scale bar = 100 µm.

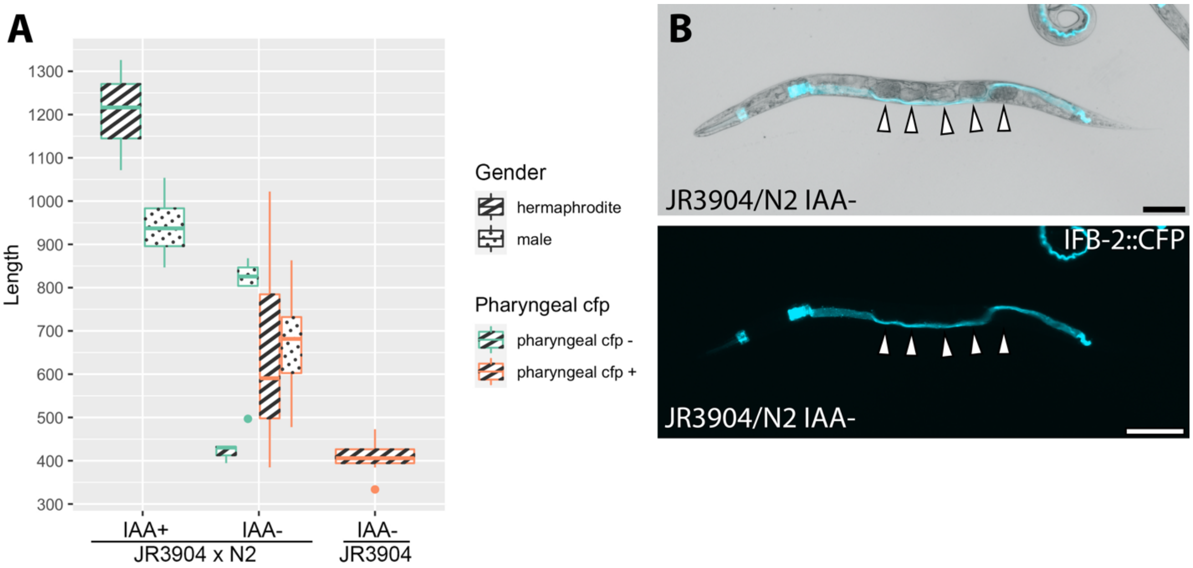


Fig 2.3. ELT-7-mediated development arrest showed a dosage-dependent effect in the AID-*elt-7* system.

N2 hermaphrodites were crossed to JR4458 males and laying eggs in a liquid culture environment overnight (Day 0). P0 adults were removed on Day 1 and the body length of F1 progenies was measured on Day 5. (A) Heterozygous F1 growing without IAA showed various levels of development arrest based on their pharynx-to-intestine Td condition and gender. Heterozygous F1 hermaphrodites arrested at ~L1/2 stage (370-480 µm) when the Td marker failed to express in the pharynx and arrested at ~L3 stage (640 µm) when the pharynx Td into the intestine. F1 males mostly grew into adults when Td failed to happen in the pharynx and arrested ~L3 stage similar to hermaphrodites when pharynx Td. All groups had 7-26 animals. (B) Occasionally F1 heterozygous hermaphrodites were observed with Td pharynx and had several progenies (arrowhead). Scale bar = 100 µm.

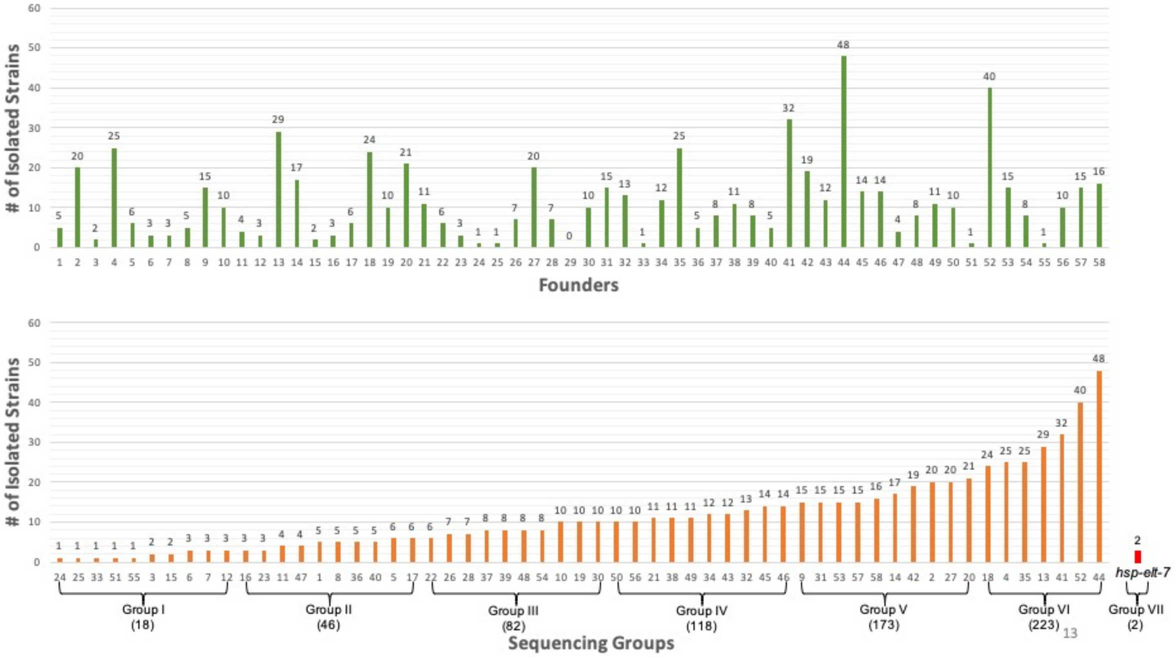


Fig 2.4. 660 Td-resistance strains were isolated by EMS-based mutagenesis screening. (A) A total of 58 founder animals were subjected to EMS mutagenesis and their F2 progenies were selected based on the resistance phenotype of development arrest caused by overexpression of ELT-7 transgene. (B) 660 isolated lines were separated into 6 groups based on the ratio of isolated lines per founder to avoid potential founder's effect. Group I contained the lowest ratio, 18 lines from 10 founders, and group XI had 223 lines in 7 founders. Additionally, 2 Td resistance lines were isolated in the HS-*elt-7* system and pooled for sequencing as group XII.

	G1		G2		G3		G4		G5		G6	
	Total Reads	**Depth**	**Total Reads**	**Depth**	**Total Reads**	**Depth**	**Total Reads**	**Depth**	**Total Reads**	**Depth**	**Total Reads**	**Depth**
AcrossGenome	1.82E+10	181.87	1.74E+10	173.19	1.86E+10	185.95	2.24E+10	222.95	1.80E+10	179.66	2.07E+10	206.38
chI	2.77E+09	183.83	2.63E+09	174.63	2.83E+09	187.61	3.38E+09	224.57	2.74E+09	181.6	3.09E+09	204.94
chII	2.80E+09	183.48	2.80E+09	183.48	2.85E+09	186.76	3.43E+09	224.18	2.76E+09	180.43	3.14E+09	205.61
chIII	2.49E+09	180.5	2.49E+09	180.5	2.56E+09	185.66	3.07E+09	222.71	2.47E+09	178.97	2.79E+09	202.64
chIV	3.09E+09	176.64	3.09E+09	176.64	3.19E+09	182.48	3.82E+09	218.55	3.09E+09	176.5	3.59E+09	205.24
chV	3.78E+09	180.76	3.78E+09	180.76	3.86E+09	184.71	4.63E+09	221.35	3.73E+09	178.07	4.30E+09	205.48
chX	3.30E+09	186.34	3.30E+09	186.34	3.35E+09	188.94	4.02E+09	226.91	3.24E+09	182.91	3.78E+09	213.34
Percentage of Coverage	98.77%		98.78%		98.78%		98.80%		98.77%		81.22%	
Coverage per strain	10.1		3.76		2.27		1.89		1.04		0.93	

Table 2.1. Sequencing depth and genomic coverage of groups 1 to 6 in the AID-*elt-7* mutagenesis screening.

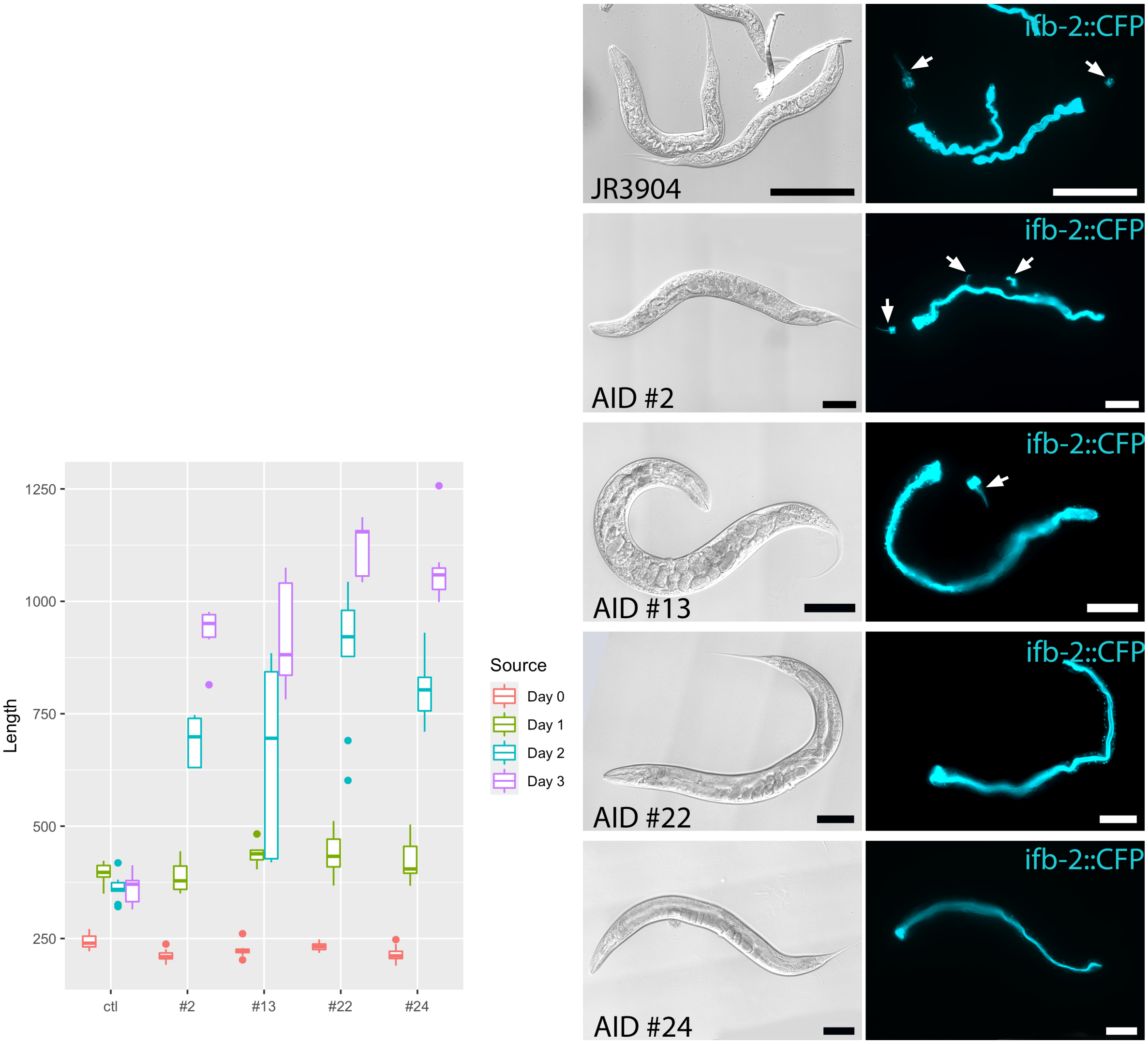


Fig 2.5. Development arrest resistance strains showed various levels of Td in the pharynx and uterus.

Synchronized L1 animals were placed on NGM plates on Day 0, and the body length was measured for the next 3 days. (A) JR3904 was arrested at ~L2 stage on the NGM plate, and the AID-*elt-7* mutant lines, AID #2, #13, #22, and #22 grew normally. (B) Td was examined on Day 3. JR3904, AID #2, and AID #13 had various pharynx Td from 100%, 85.7%, and 71.4%, respectively (arrow). AID #13 further showed uterus Td (71%) on Day3 (arrow). All groups had 6-8 animals. Scale bar = 100 µm.

Group	Funder	Strain	eft-3p+degron+ELT-7	GFP+3'UTR
1	3	35	Arg178Cys(532C>T)	Correct
1	3	36	Correct	Correct
1	6	71	Correct	Correct
1	6	72	Correct	Correct
1	6	74	Correct	Gly287Glu (860G>A)
1	7	77	Correct	Correct
1	7	79	Correct	Correct
1	7	80	Correct	Correct
1	12	129	Correct	Correct
1	12	130	Correct	Correct
1	12	132	Correct	Correct
1	15	191	Ala199>Val (497C>T)	Correct
1	15	192	Correct	Correct
1	24	314	Correct	Correct
1	25	316	Correct	Correct
1	33	421	Gln70>Stop (208 C>T)	Correct
1	51	734	No PCR Product	Deletion
1	55	812	Correct	Gly267Arg (799G>A)
2	1	1	NO PCR product	Correct
2	1	2	Correct	Correct
2	1	3	Correct	Correct
2	1	4	Gly160Arg (478G>A)	Correct
2	1	6	Correct	Correct
5	2	10	Correct	Correct
5	2	11	Correct	Correct
5	2	13	Correct	Correct
5	2	14	Gln194Stop (580C>T)	Correct
5	2	16	NO PCR product	NO PCR product
5	2	17	Cys167Tyr (500G>A)	Correct
5	2	18	Arg190Stop (568C>T)	Correct
5	2	19	Val176Met (526G>A)	Correct
5	2	20	Correct	Correct
5	2	21	Correct	Correct
5	2	22	Correct	Gly314Arg (939G>A)
5	2	24	Correct	Correct
5	2	25	No PCR product	NO PCR product
5	2	26	Correct	Correct
5	2	29	Correct	Gly314Arg (939G>A)
5	2	30	Correct	Gly287Arg (859G>A)
5	2	31	NO PCR product	NO PCR product
5	2	32	Correct	Correct
5	2	33	Correct	Correct
5	2	34	NO PCR product	NO PCR product
5	9	92	Ala199>Val (497C>T)	Correct
6	4	38	Correct	Deletion

6	4	39	Gly160>Glu(479G>A)	Correct
6	4	40	Correct	Correct
6	4	42	Correct	Correct
6	4	43	Correct	Deletion
6	4	44	Gln115Stop (343C>T)	Correct
6	4	45	Gln115Stop (343C>T)	Correct
6	4	46	Correct	Gly279Glu (839G>A)
6	4	47	Cys164Stop (491C>A)	Correct
6	4	48	Correct	Gly262Gln (845G>A)
6	4	49	Correct	Gly262Gln (845G>A)
6	4	50	Correct	Gly262Gln (845G>A)

Table 2.2. Confirmed mutations on *elt-7* transgene construct by PCR sequencing.
Two primer sets, eft-3p+degron+ELT-7 and GFP+3'UTR, were used for PCR and sequencing the *elt-7* transgene construct to confirm their integrity in the AID-*elt-7* mutant lines. The mutation position was labeled based on the start codon of each protein, i.e., Arg178Cys(532C>T) indicated a C to T transition on *elt-7* CDS position 532, Gly287Glu (860G>A) indicated a G to A transition on *emgfp* CDS position 860.

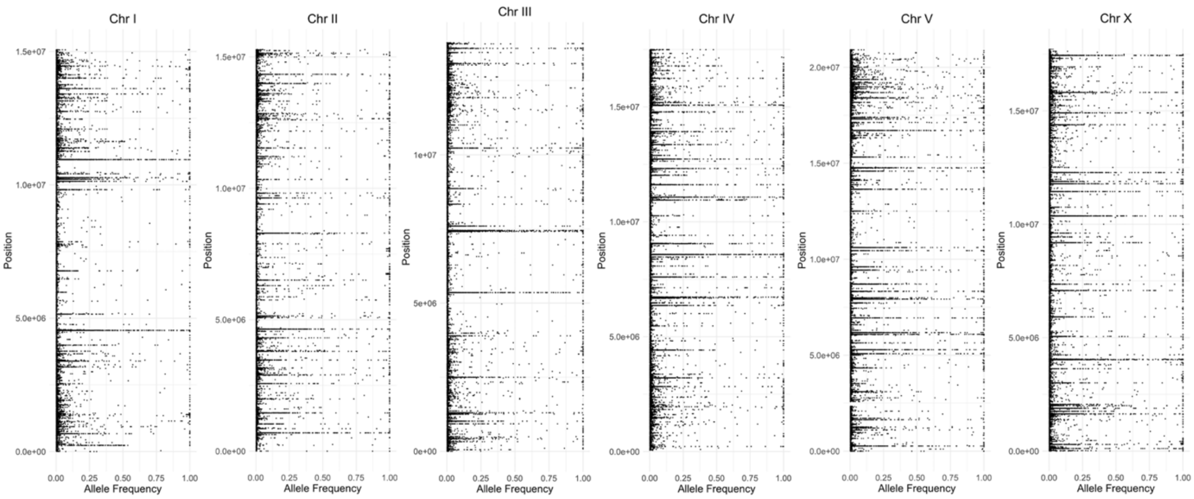
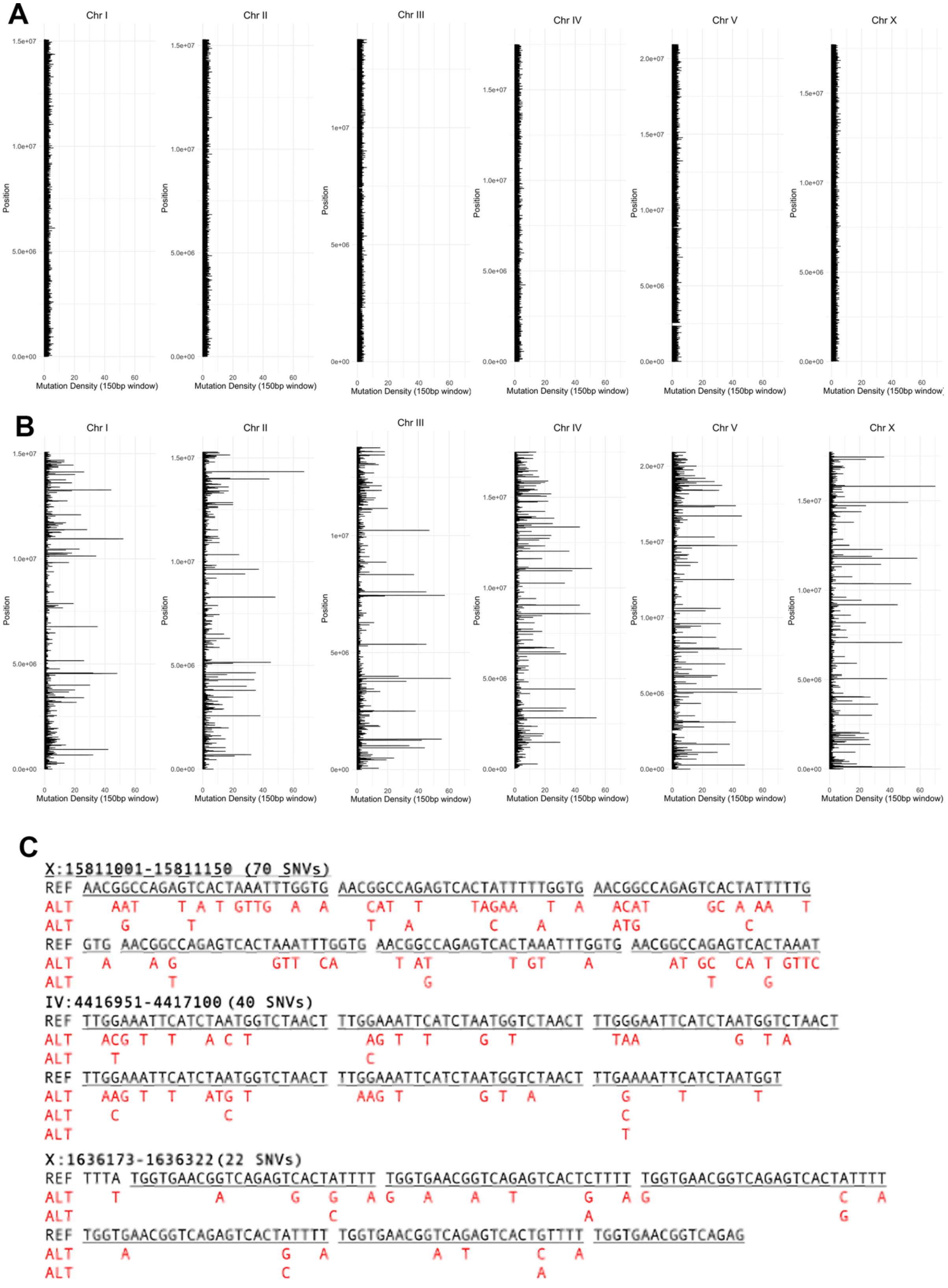


Fig 2.6. Overview of allele frequency of predicted SNVs in groups 1 to 6.
The allele frequency of all predicted SNVs was plotted alone x-axis with their chromosome location on the y-axis. Every chromosome contained ~16,000 unique SNVs.



```
X:15811001-15811150 (70 SNVs)
REF AACGGCCAGAGTCACTAAATTTGGTG AACGGCCAGAGTCACTATTTTTGGTG AACGGCCAGAGTCACTATTTTTG
ALT    AAT    T A T GTTG  A  A    CAT  T     TAGAA   T  A   ACAT      GC A AA   T
ALT     G      T                  T   A        C    A       ATG           C
REF GTG AACGGCCAGAGTCACTAAATTTGGTG AACGGCCAGAGTCACTAAATTTGGTG AACGGCCAGAGTCACTAAAT
ALT   A    A G          GTT  CA      T AT        T GT    A        AT GC  CA T GTTC
ALT          T                          G                             T     G
IV:4416951-4417100 (40 SNVs)
REF TTGGAAATTCATCTAATGGTCTAACT TTGGAAATTCATCTAATGGTCTAACT TTGGGAATTCATCTAATGGTCTAACT
ALT   ACG T  T   A C T            AG T  T     G  T          TAA          G  T A
ALT    T                          C
REF TTGGAAATTCATCTAATGGTCTAACT TTGGAAATTCATCTAATGGTCTAACT TTGAAAATTCATCTAATGGT
ALT   AAG T  T   ATG T           AAG T        G T  A         G     T       T
ALT    C           C                                         C
ALT                                                          T

X:1636173-1636322(22 SNVs)
REF TTTA TGGTGAACGGTCAGAGTCACTATTTT TGGTGAACGGTCAGAGTCACTCTTTT TGGTGAACGGTCAGAGTCACTATTTT
ALT    T          A       G   G   A G   A    A   T       G   A G                    C   A
ALT                           C                          A                          G
REF TGGTGAACGGTCAGAGTCACTATTTT TGGTGAACGGTCAGAGTCACTGTTTT TGGTGAACGGTCAGAG
ALT     A                G   A           A  T       C   A
ALT                      C                          A
```

Fig 2.7. Visualization of the density of SNVs on chromosomes.

The total SNVs were counted within a 150-bp sliding window from the beginning to the end of the chromosomes. (A) The MMP served as a non-selection control. No obvious SNVs enrichment was found on the chromosomes. (B) The AID-*elt-7* showed that 714 150-bp peaks had equal or more than 20 unique SNVs. The majority of the peaks were located in the intergenic region. (C) The repeat sequence motif was underlined in the 150-bp window. All the predicted mutations were labeled with red.

			G1-6	PCR Seq	MMP	SIFT	Allele Frequency (%)					
DNA Pos.	AA Pos.	Protein	AA Change	AA Change	AA Change		G1	G2	G3	G4	G5	G6
4258	17	degron	W/* tGg/tAg				0	0	1.31	0	0	1.48
4259	17	degron	W/* tgG/tgA				0	0	2.17	0	0	0
4322	38	degron	P/A ccG/ccA				0	3.94	0	0	0	0
4467	40	ELT-7	Q/* Caa/Taa				0	0	0	0.23	0	1.31
4471	41	ELT-7	N/S aAt/aGt			0 DEL.	1.00	0.53	0.82	0.22	1.07	1.35
4489	47	ELT-7			P/S Ccg/Tcg	0.53 TOL.	0	0	0	0	0	0
4548	67	ELT-7	Q/* Cag/Tag				0.24	0	0	0.82	2.15	1.09
4557	70	ELT-7	Q/* Caa/Taa	Q/* Caa/Taa			2.76	0	0.47	0	0.48	0
4633	95	ELT-7			R/K aGa/aAa	0.43 TOL.	0.	0.23	0.24	0	0.25	0
4692	115	ELT-7		Q/* Caa/Taa			0	0	0	0	0	0.23
4734	129	ELT-7	Q/* Cag/Tag				0	0.25	2.30	0	0	0
4765	139	ELT-7	R/H cGt/cAt			0.02 DEL.	0	0	1.75	0	0	0
4777	143	ELT-7	C/Y tGc/tAc			0 DEL.	0	0	1.00	0	2.46	0
4786	146	ELT-7	C/Y tGc/tAc			0 DEL.	0	0	0	1.70	0.52	0.26
4810	154	ELT-7	W/* tGg/tAg				0	0	0	0.54	0	1.03
4813	155	ELT-7	R/H cGt/cAt			0 DEL.	0	2.20	0	0	0	0
4828	160	ELT-7	G/E gGg/gAg	G/E gGg/gAg		0 DEL.	0	0	3.41	0	0.54	0.53
4840	164	ELT-7	C/Y tGc/tAc			0 DEL.	0	0.78	0	0.84	0	1.13
4841	164	ELT-7		C/* tgC/tgA			0	0	0	0	0	0
4843	165	ELT-7	N/S aAt/aGt			0 DEL.	1.04	0.52	1.57	1.09	0.28	0.59
4844	165	ELT-7	N/K aaT/aaG			0 DEL.	0	1.56	1.35	1.58	0.30	0.88
4845	166	ELT-7	A/T Gct/Act			0 DEL.	0	0	0	0	0	3.11
4846	166	ELT-7	A/V gCt/gTt			0 DEL.	5.85	0	0	0.23	1.51	0.28
4849	167	ELT-7	C/Y tGc/tAc	C/Y tGc/tAc		0 DEL.	0	0	0.26	2.25	0.29	0
4875	176	ELT-7		V/M Gtg/Atg			0.27	0	0	0	0	0
4881	178	ELT-7	R/C Cgc/Tgc	R/C Cgc/Tgc		0 DEL.	1.99	0	0	0	0	0
4884	179	ELT-7	P/S Ccg/Tcg			0.01 DEL.	0	2.31	0	0	0	0

4902	185	ELT-7	Q/* Cag/Tag				0	0	0	2.24	0	0
4908	187	ELT-7	P/S Cca/Tca			0.33 TOL.	0.00	2.76	2.68	0	0	0
4917	190	ELT-7		R/* Cga/Tga			0	0.28	0	0	0	0
4918	190	ELT-7	R/Q cGa/cAa			0 DEL.	0.	0.28	5.33	1.46	0	0
4976	11	emGFP	T/T acA/acT				84.69	83.85	85.60	84.84	87.11	85.66
4985	14	emGFP	V/V gtT/gtC				84.37	83.01	85.91	85.17	87.94	86.00
4991	16	emGFP	I/I atA/atT				83.91	84.04	85.83	85.00	87.55	85.37
5007	20	emGFP	G/R Gga/Aga				13.04	0	6.76	0	0	0.61
5008	20	emGFP	G/E gGa/gAa				0	3.15	0	0	0	0.63
5046	33	emGFP	G/R Gga/Aga				0	0	0	4.27	2.70	0
5047	33	emGFP	G/E gGa/gAa				0	1.69	0.53	3.00	0	6.99
5053	35	emGFP	G/E gGa/gAa				0	0	0	0	3.16	1.55
5067	40	emGFP	G/E gGa/gAa				0	0	0	1.74	0	0

Table 2.3. All predicted mutations in *elt-7* transgene compared to PCR-sanger sequencing.

The DNA position corresponded to the *elt-7* transgene construct, pCC1FOS, and the amino acid position was counted from the start codon of each protein. SIFT predicted the effect of protein function from amino acid substitutions, providing a score that is either 0.0 deleterious (Del.) or 1.0 tolerated (Tol.).

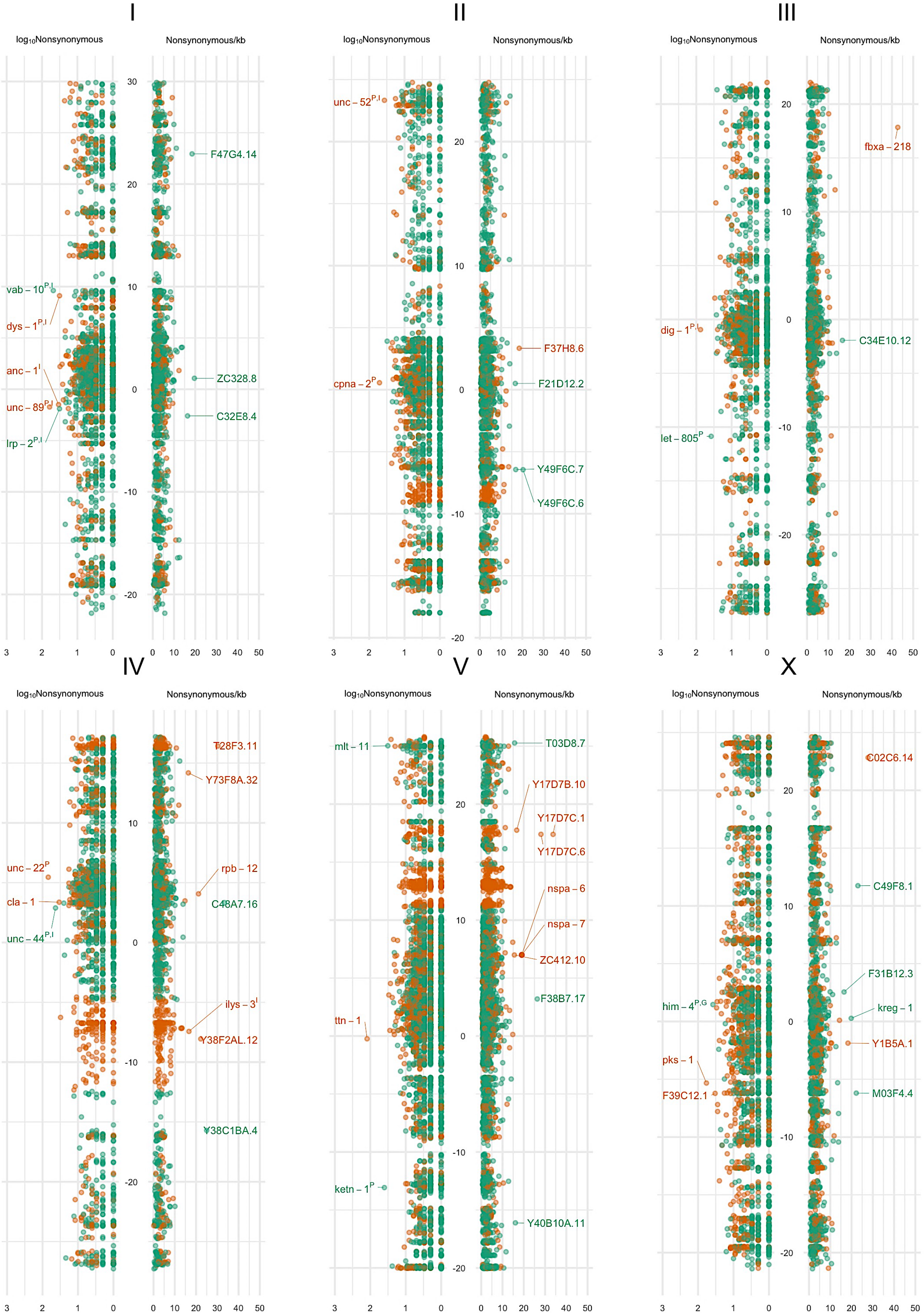

I
log$_{10}$Nonsynonymous
Nonsynonymous/kb
30
20
10
0
-10
-20
vab – 10^{P,I}
dys – 1^{P,I}
anc – 1^{I}
unc – 89^{P,I}
lrp – 2^{P,I}
F47G4.14
ZC328.8
C32E8.4
3
2
1
0
0
10
20
30
40
50
II
log$_{10}$Nonsynonymous
Nonsynonymous/kb
20
10
0
-10
-20
unc – 52^{P,I}
cpna – 2^{P}
F37H8.6
F21D12.2
Y49F6C.7
Y49F6C.6
3
2
1
0
0
10
20
30
40
50
III
log$_{10}$Nonsynonymous
Nonsynonymous/kb
20
10
0
-10
-20
dig – 1^{P,I}
let – 805^{P}
fbxa – 218
C34E10.12
3
2
1
0
0
10
20
30
40
50
IV
log$_{10}$Nonsynonymous
Nonsynonymous/kb
10
0
-10
-20
unc – 22^{P}
cla – 1
unc – 44^{P,I}
T28F3.11
Y73F8A.32
rpb – 12
C48A7.16
ilys – 3^{I}
Y38F2AL.12
Y38C1BA.4
3
2
1
0
0
10
20
30
40
50
V
log$_{10}$Nonsynonymous
Nonsynonymous/kb
20
10
0
-10
-20
mlt – 11
ttn – 1
ketn – 1^{P}
T03D8.7
Y17D7B.10
Y17D7C.1
Y17D7C.6
nspa – 6
nspa – 7
ZC412.10
F38B7.17
Y40B10A.11
3
2
1
0
0
10
20
30
40
50
X
log$_{10}$Nonsynonymous
Nonsynonymous/kb
20
10
0
-10
-20
him – 4^{P,G}
pks – 1
F39C12.1
C02C6.14
C49F8.1
F31B12.3
kreg – 1
Y1B5A.1
M03F4.4
3
2
1
0
0
10
20
30
40
50

Fig 2.8. The number of non-synonymous mutations versus mutation density in the MMP dataset.
Every protein-coding gene on the chromosomes was displayed as their total number of non-synonymous mutations on the left and the normalized mutation density per kb on the right. Genes equal to or greater than 30 non-synonymous mutations were labeled on the left, and those with mutation density equal to or greater than 15 mutant/kb were labeled on the right. Gene labeled with red color indicating at least one nonsense mutation was presented. The gene expression pattern was labeled with superscript I, indicating intestinal expression, and P, indicating pharyngeal expression.

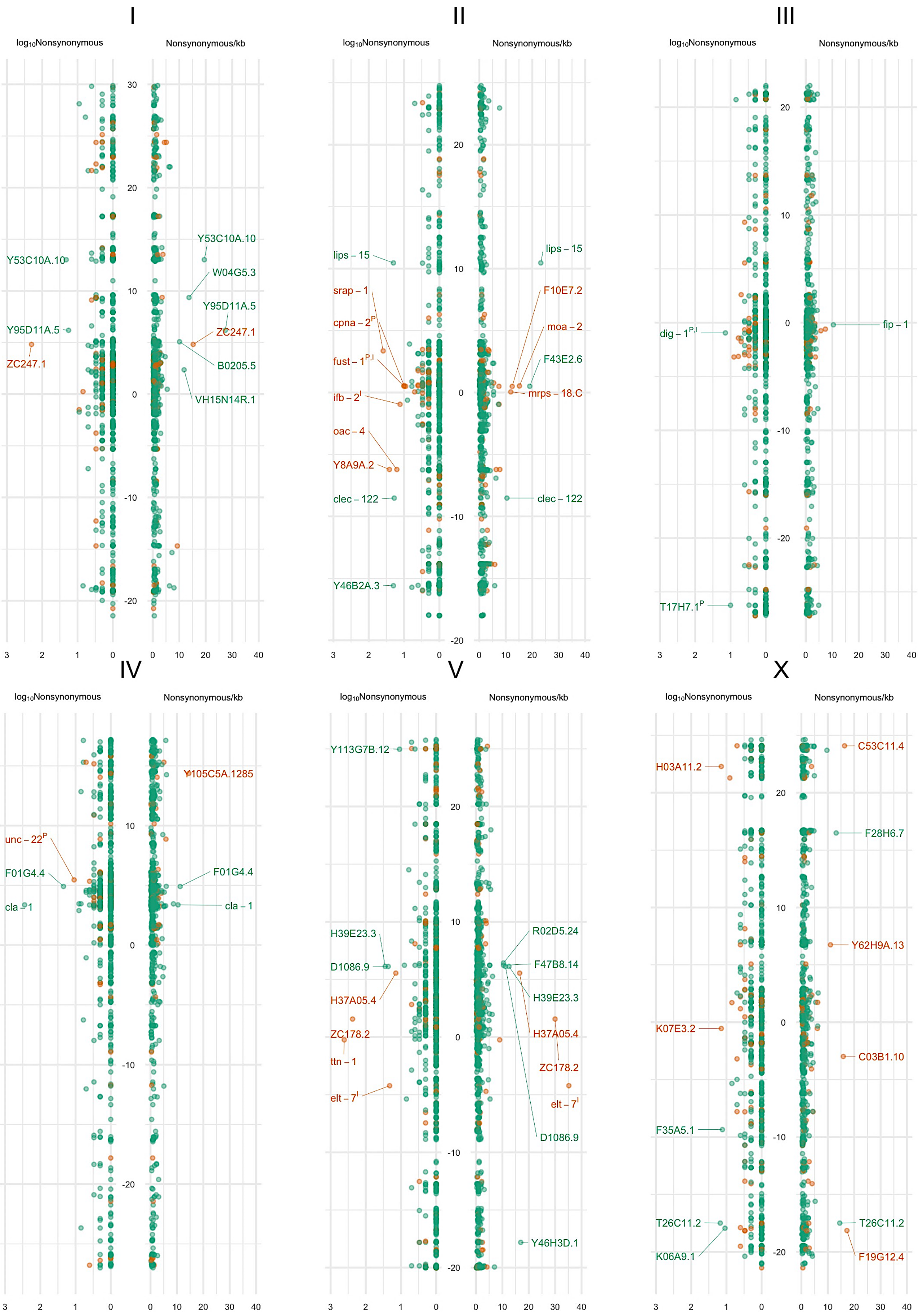

I
log10Nonsynonymous
Nonsynonymous/kb
Y53C10A.10
W04G5.3
Y95D11A.5
ZC247.1
B0205.5
VH15N14R.1
II
lips – 15
srap – 1
cpna – 2P
fust – 1P,I
ifb – 2I
oac – 4
Y8A9A.2
clec – 122
Y46B2A.3
F10E7.2
moa – 2
F43E2.6
mrps – 18.C
III
dig – 1P,I
T17H7.1P
fip – 1
IV
unc – 22P
F01G4.4
cla – 1
Y105C5A.1285
V
Y113G7B.12
H39E23.3
D1086.9
H37A05.4
ZC178.2
ttn – 1
elt – 7I
R02D5.24
F47B8.14
Y46H3D.1
X
H03A11.2
K07E3.2
F35A5.1
T26C11.2
K06A9.1
C53C11.4
F28H6.7
Y62H9A.13
C03B1.10
F19G12.4
3
2
1
0
10
20
30
40
-10
-20

Fig 2.9. The number of non-synonymous mutations versus mutation density in the AID dataset.

Every protein-coding gene on the chromosomes was displayed as their total number of non-synonymous mutations on the left and the normalized mutation density per kb on the right. Genes equal to or greater than 10 non-synonymous mutations were labeled on the left, and those with mutation density equal to or greater than 10 mutant/kb were labeled on the right. Worth to notice that on chromosome V, *ttn-1* was the gene with the highest number of non-synonymous mutations but down-ranked after the normalization, and *elt-7* was up-ranked to the highest mutation density gene after normalization. Gene labeled with red color indicating at least one nonsense mutation was presented. The gene expression pattern was labeled with superscript I, indicating intestinal expression, and P, indicating pharyngeal expression.

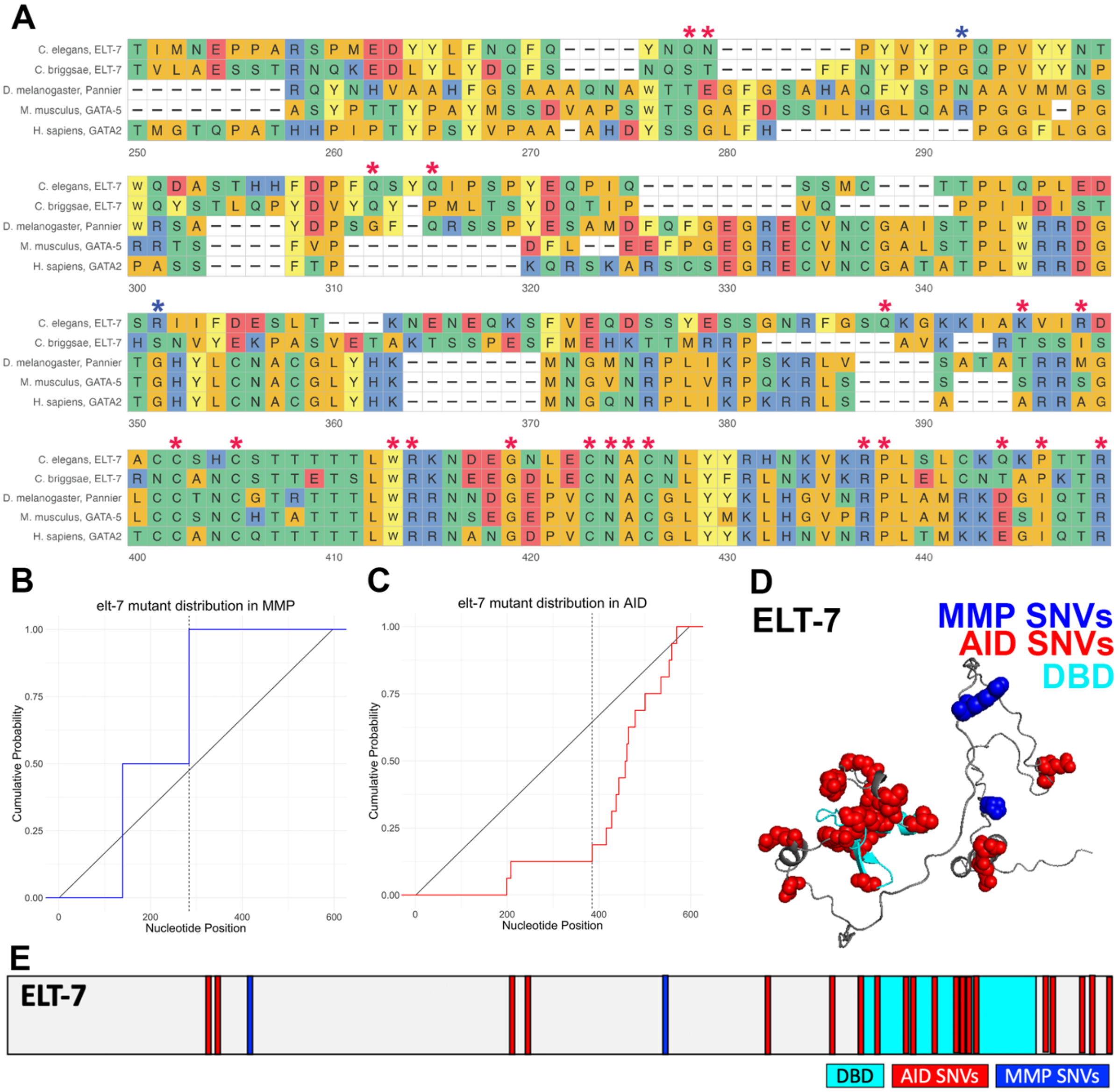


Fig 2.10. Predicted *elt-7* transgene mutations concentrated on conservated residues and in the DNA binding domain.

(A) ELT-7 sequence was aligned from *C. elegans* (ELT-7) to *H. sapiens* (GATA2). The predicted mutations in the AID-*elt-7* screening were mostly found on the conserved residues highlighted by red asterisks. Blue asterisks labeled the ELT-7 mutations in the MMP. (B, C) The KS test evaluated the distribution bias of ELT-7 mutations. The diagonal line indicated a uniform reference distribution, and the blue (MMP) and red (AID) lines correspond to an empirical distribution function of their ELT-7 mutations. The Dash line was the greatest distance for the KS statistic. (D, E) The locations of ELT-7 mutations were displayed in a 3D and 2D structure.

			MMP				**G1-6**			
	WBGene ID	Gene	#Mut	Mut Freq	DelRatio (%)	ks-pval	#Mut	Mut Freq	DelRatio (%)	ks-pval
1	WBGene00015981	elt-7	2	3.35	0.00	0.45	21	35.18	86.36	0.00
2	WBGene00008396	D1086.9	6	2.91	12.50	0.66	23	11.16	21.21	0.20
3	WBGene00002054	ifb-2	4	2.31	28.57	0.97	13	7.50	40.00	0.70
4	WBGene00016173	fust-1	5	3.71	0.00	0.23	10	7.42	22.22	0.00
5	WBGene00010711	K09B11.4	0	0.00	0.00	NA	2	4.66	50.00	0.14
6	WBGene00007135	cdk-12	4	1.81	40.00	0.98	8	3.63	66.67	0.09
7	WBGene00012156	ebp-2	2	2.22	0.00	0.73	3	3.33	100.00	0.29
8	WBGene00008870	F15H9.1	3	2.34	33.33	0.63	4	3.12	66.67	0.83
9	WBGene00005355	srh-138	1	1.02	0.00	NA	3	3.06	33.33	0.48
10	WBGene00017560	F18C5.5	1	1.49	0.00	NA	2	2.99	50.00	0.88
11	WBGene00013434	Y66D12A.8	1	1.36	28.57	NA	2	2.72	100.00	0.15
12	WBGene00010838	M03C11.1	1	0.88	0.00	NA	3	2.63	33.33	0.37
13	WBGene00018638	F49E8.7	3	1.96	20.00	0.99	4	2.61	50.00	0.74
14	WBGene00006658	twk-3	2	1.74	0.00	0.89	3	2.60	33.33	0.84
15	WBGene00020158	T02C5.1	0	0.00	0.00	NA	3	2.54	50.00	0.18
16	WBGene00019386	irld-42	2	1.64	25.00	0.65	3	2.46	50.00	0.30
17	WBGene00000711	col-138	1	1.18	50.00	NA	2	2.36	100.00	0.32
18	WBGene00010230	rpac-19	0	0.00	0.00	NA	1	2.30	100.00	NA
19	WBGene00019370	K03H6.5	1	0.77	50.00	NA	3	2.30	83.33	0.52
20	WBGene00021157	Y4C6B.3	1	0.72	0.00	NA	3	2.17	100.00	0.75
21	WBGene00007016	mdt-15	4	1.71	14.29	0.54	5	2.13	30.00	0.01
22	WBGene00022193	ppfr-4	1	1.02	33.33	NA	2	2.03	50.00	0.42
23	WBGene00011701	srab-17	1	1.01	0.00	NA	2	2.02	100.00	0.52
24	WBGene00005361	srh-145	1	1.00	0.00	NA	2	2.00	100.00	0.78
25	WBGene00005439	srh-231	1	0.98	0.00	NA	2	1.97	50.00	0.32
26	WBGene00012013	ugt-54	1	0.63	0.00	NA	3	1.89	66.67	0.76

27	WBGene00021827	dnc-6	0	0.00	0.00	NA	1	1.84	100.00	NA
28	WBGene00005065	sra-39	1	0.89	50.00	NA	2	1.77	100.00	0.46
29	WBGene00021894	Y54G2A.32	0	0.00	0.00	NA	1	1.77	100.00	NA
30	WBGene00015628	best-6	0	0.00	0.00	NA	2	1.74	50.00	0.65
31	WBGene00009116	snx-27	1	0.58	0.00	NA	3	1.73	33.33	0.65
32	WBGene00000931	dao-5	4	1.37	25.00	0.15	5	1.71	30.77	0.39
33	WBGene00010197	F57C2.5	0	0.00	0.00	NA	2	1.70	33.33	0.55
34	WBGene00009223	mdt-28	0	0.00	0.00	NA	1	1.64	100.00	NA
35	WBGene00044069	hat-1	0	0.00	0.00	NA	2	1.62	50.00	0.23
36	WBGene00020383	tin-44	1	0.78	0.00	NA	2	1.56	100.00	0.08
37	WBGene00003968	peb-1	1	0.75	0.00	NA	2	1.50	100.00	0.52
38	WBGene00002127	inx-5	1	0.74	0.00	NA	2	1.49	33.33	0.63
39	WBGene00000288	cal-4	0	0.00	0.00	NA	1	1.41	100.00	NA
40	WBGene00007746	C26D10.6	0	0.00	0.00	NA	2	1.30	100.00	0.69
41	WBGene00006997	zyg-12	1	0.43	0.00	NA	3	1.29	4.17	0.58
42	WBGene00015369	ugt-51	1	0.63	0.00	NA	2	1.27	100.00	0.20
43	WBGene00013900	ugt-18	1	0.62	0.00	NA	2	1.25	50.00	0.86
44	WBGene00011076	scav-5	1	0.62	50.00	NA	2	1.24	100.00	0.43
45	WBGene00013558	Y75B8A.25	2	0.79	0.00	0.68	3	1.19	50.00	0.08
46	WBGene00021311	thoc-5	1	0.56	0.00	NA	2	1.11	50.00	0.45
47	WBGene00008676	oac-15	1	0.51	33.33	NA	2	1.01	50.00	0.92
48	WBGene00011503	gcc-2	1	0.50	0.00	NA	2	1.01	50.00	0.02
49	WBGene00022807	srab-25	0	0.00	0.00	NA	1	0.99	50.00	NA
50	WBGene00002202	kin-19	0	0.00	0.00	NA	1	0.97	100.00	NA
51	WBGene00020188	T03F1.6	0	0.00	0.00	NA	1	0.96	100.00	NA
52	WBGene00000687	col-113	0	0.00	0.00	NA	1	0.96	100.00	NA
53	WBGene00008546	gfat-1	1	0.46	0.00	NA	2	0.92	50.00	0.53

54	WBGene00017923	F29B9.8	0	0.00	0.00	NA	1	0.87	100.00	NA
55	WBGene00011292	allo-1	0	0.00	0.00	NA	1	0.83	50.00	NA
56	WBGene00002990	lin-1	0	0.00	0.00	NA	1	0.75	66.67	NA
57	WBGene00004231	ptr-17	1	0.37	23.08	NA	2	0.74	50.00	NA
58	WBGene00003130	map-2	0	0.00	0.00	NA	1	0.71	100.00	NA
59	WBGene00017157	tyra-2	0	0.00	0.00	NA	1	0.71	33.33	NA
59	WBGene00007918	sphk-1	0	0.00	0.00	NA	1	0.70	16.67	NA
59	WBGene00013531	Y73F8A.26	0	0.00	0.00	NA	1	0.70	50.00	NA
59	WBGene00015477	attf-4	0	0.00	0.00	NA	1	0.63	50.00	NA
59	WBGene00006896	ver-3	1	0.27	20.00	NA	2	0.54	50.00	0.38

Table 2.4. Candidate genes with higher deleterious ratio and lower KS test p-value in the AID compared to the MMP.

Genes with a higher deleterious ratio and lower KS test p-value in the AID compared to the MMP were selected and ranked based on the mutation frequency. KS test failed to perform when only one mutation existed indicating as “NA”.

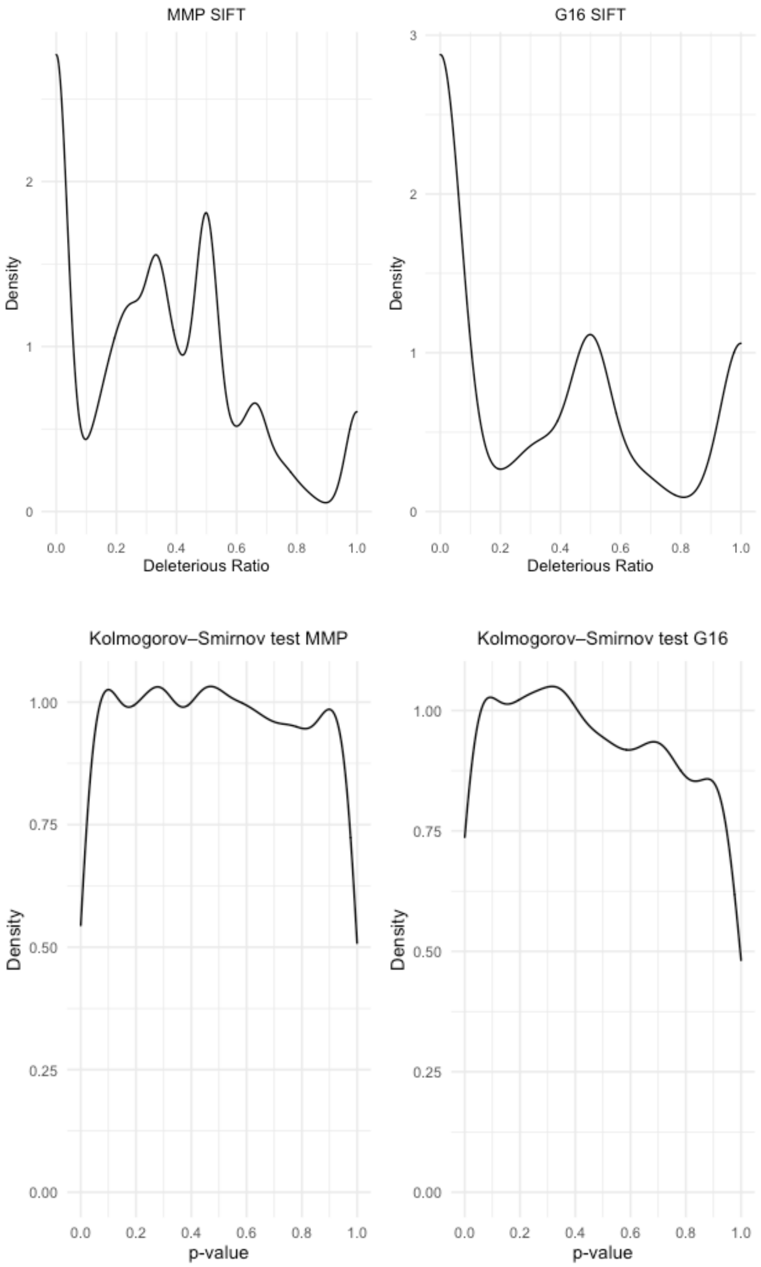


Fig S2.1. The MMP dataset showed a wild distribution of the deleterious ratio (SIFT) and KS p-value compared to the AID dataset.

(A) The MMP dataset showed a more average distribution of deleterious ratio compared to the AID-*elt-7* dataset. Of a total of 11039 genes in the MMP, 2845 (25.72%) of them had less than 20% deleterious, 7429 (67.30%) were between 20% to 80%, and 763 (6.92%) were greater than 80%. In contrast, more genes in the AID-*elt7* were highly deleterious. Of a total of 1135 genes, 187 (16.55%) were greater than 80% deleterious ratio, 590 (52.02%) were between 20% to 80%, and 356 (31.43%) were less than 20%. (B) Mutations in the AID-*elt-7* dataset showed a more non-random distribution, with 94 (8.20%) genes having a p-value lower than 0.05, whereas MMP had 583 genes (5.27%).

Chapter Three

CDK-12 is required for ELT-7-mediated Td and development arrest

Summary

CDK-12, one of the top candidate genes from the AID-*elt-7* pooled mutant sequencing data, was identified and confirmed its requirement in ELT-7-mediated Td and development arrest by complementation test and rescue experiment. In addition, the CDK-12 mutant allele, *w58*, may reduce its kinase activity, leading to decreased transcription of the *elt-7* transgene to allow animals to escape ELT-7-mediated Td.

Introduction

An overview of Cyclin-dependent kinase (CDK) 12 function.

The most prominent role of CDK12 is in the transcription elongation through regulating the phosphorylation of RNA polymerase II C-terminal domain (CTD) heptad repeat ($Y_1S_2P_3T_4S_5P_6S_7$) with its partner, cyclin K (CCNK). CDK12/CCNK has demonstrated the capability of phosphorylating Ser 2 and Ser 5, and lesser extent, Ser 7 of RNA polymerase II CTD. Unlike CDK9, a transcription elongation factor functions similarly to CDK12, but functions at the 3' end of the transcription. CDK12 was found most active at the 5' end for transcription. Indeed, loss of CDK9 results in a global transcription defect. In contrast, the loss of CDK12 leads to premature termination of long transcripts with intronic polyadenylation, resulting in mRNA isoforms with different coding sequences or truncated transcripts that may disrupt its function (Choi et al., 2020; Emadi et al., 2020). For example, CDK-12 is required for *C. elegans* larval development for L1 to L2 stage progression. Loss of CDK-12 decreases SL2 trans-splicing gene expression and ultimately leads to an L1 arrest phenotype (Cassart et al., 2020). Loss of CDK12 also leads to delayed G1/S progression, increased replication stress, and DNA repair defects due to decreased DNA repair gene expression and has been considered a hallmark of carcinogenesis in human cells (Manavalan et al., 2019). CDK12 also directly regulates RNA splicing by interacting with splicing factors or indirectly by the phosphorylation of RNA polymerase CTD. Protein interaction analysis, such as mass spectrometry, found that CDK12 interacted with several RNA-binding proteins in the exon junction complex (EJC), SR splicing factors (SRSFs), and the CDC5L/PPPF19 complex, suggesting its role in RNA splicing regulation (Choi *et al.*, 2020). Loss of function of CDK12 in cancer cell lines showed alternative splicing patterns in long transcripts. Interestingly, only ~3.5% of the long transcripts were regulated by CDK12, suggesting an additional gene-specific factor may be involved (Tien et al., 2017). Finally, CDK12 showed its role in chromatin regulation in *Drosophila*. Deletion of cdk12 causes euchromatin to turn into heterochromatin on the X chromosome, resulting in reduced neuronal gene expression in the brain and a defective courtship learning process (Pan et al., 2015). These studies showed that CDK-12 has various roles in regulating cellular function, gene expression, and epigenetic regulation.

Results

***cdk-12*, a potential causal gene, was identified and confirmed the requirements in ELT-7 mediated development arrest and Td in the AID and HS systems.**

Although *cdk-12* was in sixth place on the candidate gene list, we decided to purchase *cdk-12* based on its relatively high deleterious ratio and subjective interest (Fig 3.1). Variants calling on the pooled mutant sequencing data predicted 5 missense mutations, 2 nonsense mutations, and 1 splicing variant mutation in the *cdk-12* (Table 3.1). The allele frequency of the 7 mutations in the coding sequence was examined for the followed-up *cdk-12* cloning in each sequencing group. D403V (*w58*) was identified in group 1, and P619L (*w59*) and Q622stop (*w60*) were identified in group 3 by PCR sequencing. W484stop showed high allele frequency in groups 1 and 2; however, this allele failed to identify in PCR sequencing, possibly due to a false-positive variants prediction or the nature of this nonsense mutation unfavored animal growth. In conclusion, 3 of the 7 predicted CDS mutations were identified, suggesting a successful SNVs prediction.

Interestingly, unlike knock-out or knock-down *cdk-12* disrupted *C. elegans* development due to the lack of SL2 trans-splicing genes expression. Animals showed no sign of development arrest or delayed in *w58,* in which the kinase domain is mutated, suggesting that *w58* may be a weak hypermorph or a neomorph mutation (Fig 3.2A). The complementation test and rescue experiment were performed to confirm the requirement of CDK-12 in ELT-7-mediated Td. A null allele of *cdk-12* (*tm3846*) was used for the complementation test and showed failed complement with *cdk-12 (w58)*, suggesting the suppressor of development arrest is CDK-12 (Fig 3.2B). The rescue experiment also confirmed it as *cdk-12(w58)* animals resistance development arrest, but this phenotype was reversed when introducing a wild-type *cdk-12* transgene (*ckSi6*) (Fig 3.2C). These findings indicated that CDK-12 was the causal gene required for the ELT-7-mediated Td.

w58 not only suppresses ELT-7-mediated Td and development arrest in the AID-*elt-7* system but also partially suppresses it in the HS *elt-7* system. For example, almost 100% of animals were development arrest with a minimum of 3 mins HS *elt-7* overexpression; however, with the introduction of *w58*, 20% of animals remained developed and propagating (Fig 3.3). *w58* showed a relatively subtle effect in preventing the development arrest of the HS-elt-7 system; however, it may be due to the different copy numbers of the *elt-7* transgene in the two systems. In conclusion, CDK-12 is required for the ELT-7-mediated Td in both the AID and HS systems.

Decreased expression of *elt-7* transgene in the *w58* strain.

One of the major functions of *cdk-12* is in regulating transcription elongation. Mutation or deletion of *cdk-12* often led to lower expression of the long transcript genes. Although the *elt-7* transgene was a short transcript (~1455 bp) with no intron in the construct, we still decided to examine *elt-7* transgene expression in *w58*. JR3904 and *w58* were synchronized at

L1 and then placed on the NGM plate to allow ectopic ELT-7 accumulation. Interestingly, *elt-7* transgene was expressed 50% lower in the *w58* stain compared with JR3904, suggesting the possibility that the resistance of development arrest and Td phenotype may be due to the low expression of *elt-7* transgene (Fig 3.4). Indeed, the *elt-7* dosage-dependent development arrest and Td have been shown in the AID-*elt-7* system. The transgene dosage-dependent effect on the development arrest was further examined based on the copy number of *elt-7* transgene and the *TIR1* transgene since TIR1 has been reported to degrade degron-tagged protein without the presence of IAA (Hills-Muckey et al., 2022; Martinez et al., 2020). The AID-*elt-7* strain, JR3904, males were crossed with sperm-depleted N2 and CA1202 hermaphrodites and allowed the F1 progeny to develop on the NGM plates. The percentage of F1 animals that grow into adults was counted and compared with *w58* animals. Not surprisingly, heterozygous *elt-7* transgene animals showed partial resistance to development arrest, reconfirming the *elt-7* dosage-dependent effect of Td. Furthermore, homozygous TIR1 animals have a greater probability of developing into adults compared to those heterozygous animals, indicating a TIR1 dosage-depend baseline protein degradation on degron-tagged ELT-7. The development of heterozygous *elt-7* transgene with homozygous TIR1 animals was similar to the *w58* animals, suggesting that the main function of CDK-12 in intestinal Td is through regulating the *elt-7* transgene expression (Fig 3.5). Altogether, these results suggested that CDK-12 may not be directly involved in the ELT-7-mediated Td, instead regulating the general gene transcription, such as the *elt-7* transgene.

Explore the potential of inducing intestinal Td outside of the pharynx and gonad.

CDK-12 showed its requirement for ELT-7-mediated Td; thus, we aimed to test its sufficiency in the process, specifically whether CDK-12 was sufficient to trigger intestinal Td outside the pharynx and uterus in post-mitotic cells. Overexpression of *cdk-12* and *elt-7* was done under the HS promoter with three times HS treatments with the interval of 3 hours at L1 or early L4. However, no noticeable difference was found between overexpress *elt-7* alone and *elt-7/cdk-12* double (data not shown). The preliminary result suggests that *cdk-12* is insufficient to cause ectopic intestinal Td outside of the pharynx and uterus. However, it is possible that CDK-12 functions as a facilitator of intestinal Td; therefore, it is worth investigating the Td efficiency under weak ELT-7 expression conditions such as overexpressing CDK-12 with a short HS period or single-copy of *elt-7* transgene.

Discussion

Potential role of CDK-12 in ELT-7-mediated Td

elt-7 transgene expression showed a significantly decreased in the *cdk-12 (w58)* strain, suggesting that the primary role of CDK-12 was to facilitate gene expression, specifically the *elt-7* transgene in this case. However, CDK-12 is also involved in gene expression regulation through RNA splicing; therefore, it is possible that some of the genes required for intestinal Td, especially long transcripts, rely on CDK-12 activity. Indeed, ~3000 transcripts were upregulated and showed alternative splicing patterns in the HS *elt-7* system, suggesting that CDK-12 may be required for expressing Td genes. A mutant CDK-12, such as *w58*, may reduce those genes' expression and lead to unsuccessful or delayed Td. Interestingly, two out of seven (29%) predicted *cdk-12* mutations were outside the conserved kinase domain, implying that there might be protein interaction between CDK-12 and other factors that dictated the process or through a different mechanism, such as RNA splicing regulation. In conclusion, CDK-12 was required for *elt-7* transgene expression consist its prominent role in transcription elongation. CDK-12 may reduce Td gene expression through defective transcription elongation caused by decreased kinase domain activity or incorrect RNA splicing caused by reduced protein interaction with splicing factors on the C terminal domain.

Material and Methods

C. elegans incubation and genetics

The strains used in this study were: JR3642 *(wIs125[hsp-16-2::ELT-7 hsp-16–41::ELT-7]; rrIs01[elt-2p::lacZ::GFP + unc-119(+); kcIs6 [ifb-2p::IFB-2::CFP] (IV]),* JR4535 (*wIs125[hsp-16-2::ELT-7 hsp-16–41::ELT-7]; wIs169[hsp-16-2::CDK-12c hsp-16–41::CDK-12c]; rrIs01[elt-2p::lacZ::GFP + unc-119(+); kcIs6 [ifb-2p::IFB-2::CFP] (IV]*), JR4458 (*ieSi57[eft-3p::TIR1::mRuby::unc-54 ''UTR + Cbr-unc-119(+)] II; eft-3p::degron::ELT-7::EMGFP::unc-54 3’UTR kcIs6 [ifb-2p::IFB-2::CFP] (IV); him-5 (e1490) (V)*), JR4511 (*ieSi57[eft-3p::TIR1::mRuby::unc-54 ’'UTR + Cbr-unc-119(+)] II; cdk-12 (w58) (III); wIS167[eft-3p::degron::ELT-7::EMGFP::unc-54 3’UTR] kcIs6 [ifb-2p::IFB-2::CFP] (IV); him-5 (e1490) (V)* 5x backcross with JR4458), JR4519 (*cdk-12 (w58) (III); wIs125[hsp-16-2::ELT-7 hsp-16–41::ELT-7]; rrIs01[elt-2p::lacZ::GFP + unc-119(+]); him-5 (e1490) (V)]),* JR4514 *(ieSi57[eft-3p::TIR1::mRuby::unc-54 ’'UTR + Cbr-unc-119(+)] II; cdk-12 (tm3846)/qC1[dpy-19(e1259) glp-1(q339)] nIs189 (III); wIS167[eft-3p::degron::ELT-7::EMGFP::unc-54 3’UTR] kcIs6 [ifb-2p::IFB-2::CFP] (IV); him-5 (e1490) (V))*

Identify *cdk-12* mutant alleles

PCR was performed with a similar protocol mentioned in chapter II. In general, DNA samples from *AID-elt-7* screening lines, typically group 1 to group 3, were prepared either by single worm PCR protocol or directly extracted from the frozen glycerol stocks. For extracting lysate from the stock, a small amount of frozen glycerol stock was transferred into a PCR tube, centrifuged, and aspirated ~10 µl supernatant to a new PCR tube. Each sample was then added with 10 µl of lysis buffer supplemented with proteinase K (10mM Tris pH 8.0 (Fisher), 50mM KCl (Fisher), 2.5mM $MgCl_2$ (Fisher), 0.45% Igepal (Sigma), 0.45% Tween 20 (Sigma), 0.01% gelatin (Sigma), and 2mg/mL: proteinase K (GoldBio)) and mixed and frizzed at -80 °C for 15 mins, followed by the worm lysis PCR program of 65 °C for 1 hour then 95 °C for 20 mins. Lysis samples were diluted in a 1:10 ratio for PCR reaction.

Complementation test

3 days old sperm-depleted hermaphrodites JR4514 were crossed with R4511 male on the IAA plate overnight, the next day (Day 0), hermaphrodites were singled to the NGM plate to lay eggs overnight and then removed the next day (Day 1). The development of F1 progenies was imaged and measured on Day 3.

qRT-PCR

The single worm qRT-PCR protocol from the Snell Lab was followed (Ly et al., 2015). In short, single worms were picked and washed in dNase/rNase-free water and then lysed in 1 µL worm lysis buffer (5mM Tris pH 8.0 (Fisher), 0.5% Triton X-100 (Sigma), 0.5% Tween 20 (Sigma), 0.25mM EDTA (Fisher), and 1mg/mL: proteinase K (GoldBio)) in a 0.2 mL PCR tube with 65 °C incubation for 10 mins followed by 85 °C incubation for 1 min. Lysates were then immediately used for cDNA synthesis with the Maxima H Minus cDNA synthesis kit (Fisher) following the instruction manual. 1 μL of cDNA synthesis mixture was added to the lysate and incubate at 25 °C for 10 mins, followed by 55 °C for 30 mins, and finally at 85 °C for 5 mins. cDNA samples were diluted to a final volume of 20 μL. Quantitative real-time PCR was performed using CFX Opus 96 Real-Time PCR System (BioRad). Each 10 μL reaction contains 1 μL of the cDNA sample, primers, and SsoAdvanced Universal SYBR Green Supermix (BioRad). The thermal cycling cycle was followed based on BioRad manufactory instructions with 98 °C for 30 secs, followed by 39 cycles of 98 °C for 5 sec and 60°C for 30 sec, and finally, a melt curve analysis of 65 °C with 0.5 °C increments at 2-5 sec/step. The data were analyzed using the standard $2\text{-}^{\Delta\Delta CT}$ method and normalized by *act-1*.

Transgenic *cdk-12* overexpress line

The *cdk-12* isoform c CDS was amplified from L1 N2 cDNA samples and cloned into vectors pPD49.78 and pPD49.83 with restriction enzyme digestions and then injected into JR3642 with co-injection marker pCFJ90 (*myo-2p::mcherry*). The low transmission line was selected and exposed to Stratagene UV crosslinker (Stratalinker) under a power setting of 300. Lines that showed a high transmission rate were selected and maintained as the transgenic stock.

Overexpressing *cdk-12* and *elt-7*

Synchronized L1s were fed on OP50 NGM plates for 3 hours and harvested in M9 buffer. About ~200 animals were then transferred to 0.2 mL PCR tubes and HS with a thermocycler setting of 1 min at 25 °C, followed by various times at 33 °C, and finally 1 min at 25 °C. HS animals were transferred to NGM plates and the number of animals was counted on Day 0 and Day 3.

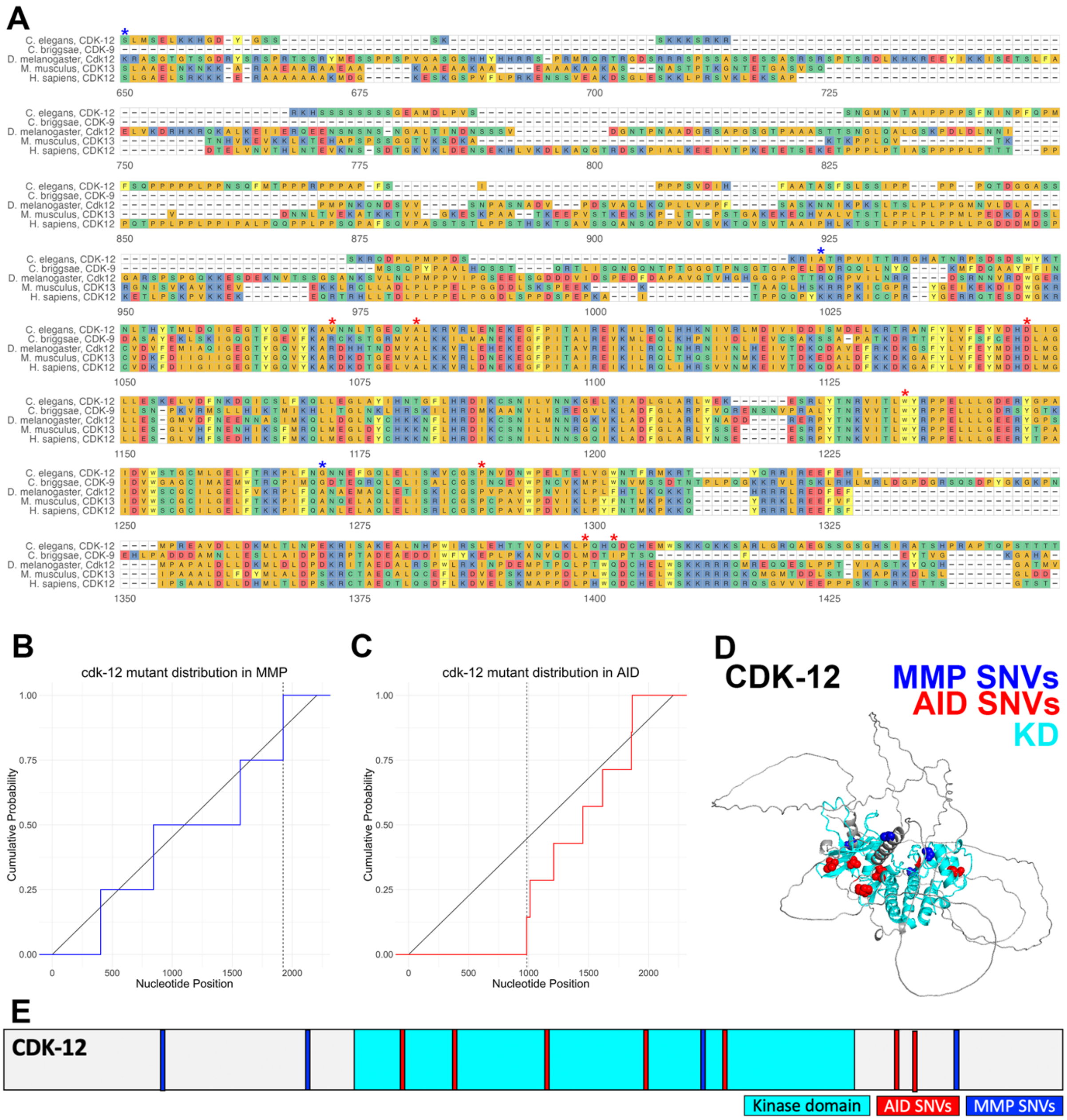
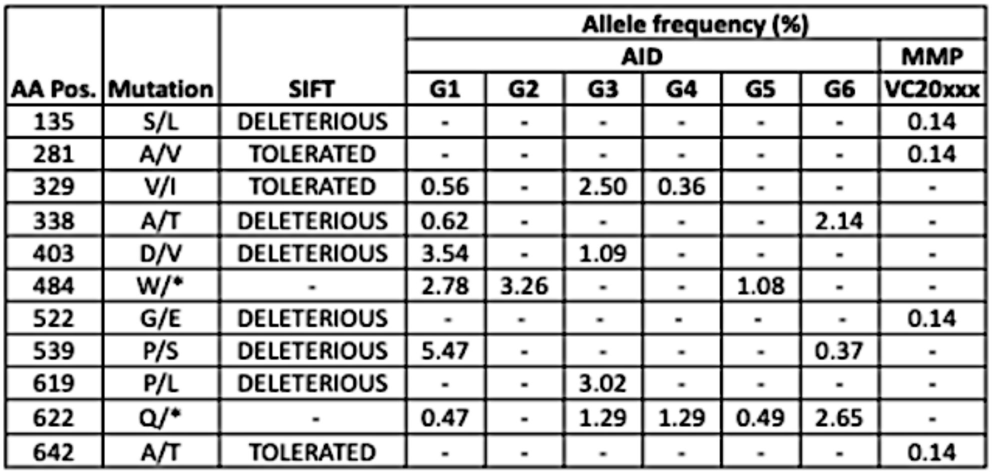


Fig 3.1. Predicted *cdk-12* mutations enriched in the kinase domain.

(A) CDK-12 sequence was aligned from *C. elegans* (CDK-12) to *H. sapiens* (CDK12). The red asterisks highlighted the predicted CDK-12 mutations in the AID-*elt-7* screening. Blue asterisks labeled the CDK-12 mutations in the MMP. (B, C) The KS test evaluated the distribution bias of CDK-12 mutations. The diagonal line indicated a uniform reference distribution, and the blue (MMP) and red (AID) lines correspond to an empirical distribution function of CDK-12 mutations. The Dash line was the greatest distance for the KS statistic. (D, E) The locations of CDK-12 mutations were displayed in a 3D and 2D structure.



			Allele frequency (%)						
			AID						MMP
AA Pos.	Mutation	SIFT	G1	G2	G3	G4	G5	G6	VC20xxx
135	S/L	DELETERIOUS	-	-	-	-	-	-	0.14
281	A/V	TOLERATED	-	-	-	-	-	-	0.14
329	V/I	TOLERATED	0.56	-	2.50	0.36	-	-	-
338	A/T	DELETERIOUS	0.62	-	-	-	-	2.14	-
403	D/V	DELETERIOUS	3.54	-	1.09	-	-	-	-
484	W/*	-	2.78	3.26	-	-	1.08	-	-
522	G/E	DELETERIOUS	-	-	-	-	-	-	0.14
539	P/S	DELETERIOUS	5.47	-	-	-	-	0.37	-
619	P/L	DELETERIOUS	-	-	3.02	-	-	-	-
622	Q/*	-	0.47	-	1.29	1.29	0.49	2.65	-
642	A/T	TOLERATED	-	-	-	-	-	-	0.14

Table 3.1. Predicted *cdk-12* mutations were mostly located in conserved residues causing the deleterious outcome of the protein function in the AID-*elt-7*.

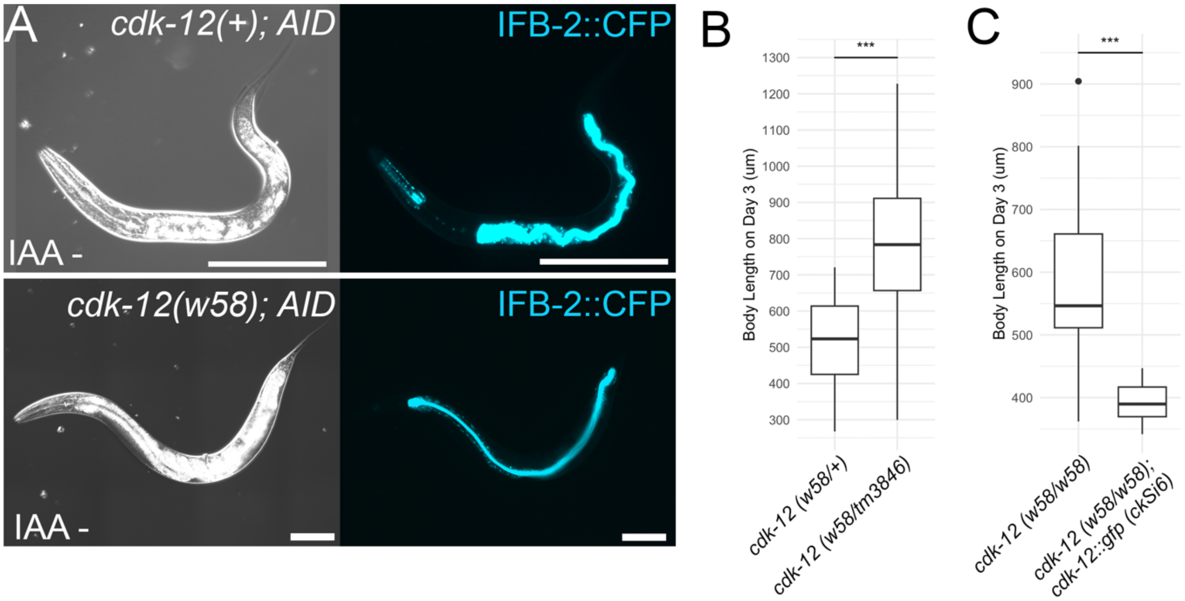


Fig 3.2. CDK-12 is required for development arrest triggered by ELT-7 overexpression. (A) The AID-*elt-7* system was used for ELT-7 overexpression in the whole animal when IAA was absent in the environment. Pharynx-to-intestine Td was observed in the *cdk-12* wild-type animals and arrest at L1/2 stages. However, Td was absent in *cdk-12 (w58)* animals and grew into adults. Scale bar = 100 µm. (B) The requirement of CDK-12 for development arrest phenotype caused by ELT-7 overexpression was examined by complementation test on *cdk-12 (w58)* and a *cdk-12* deletion allele (*tm3846*). Normal development was observed in heterozygous *w58/tm3846,* indicating a failed complementation. Each group contained 94 to 100 animals. (C) The development arrest-resistant phenotype of the *cdk-12 (w58)* animals was reversed by introducing a wild-type *cdk-12* transgene. Each group contained 14 to 20 animals. Significant differences (p<0.001, Student's t-test) between the two groups were indicated by a (***).

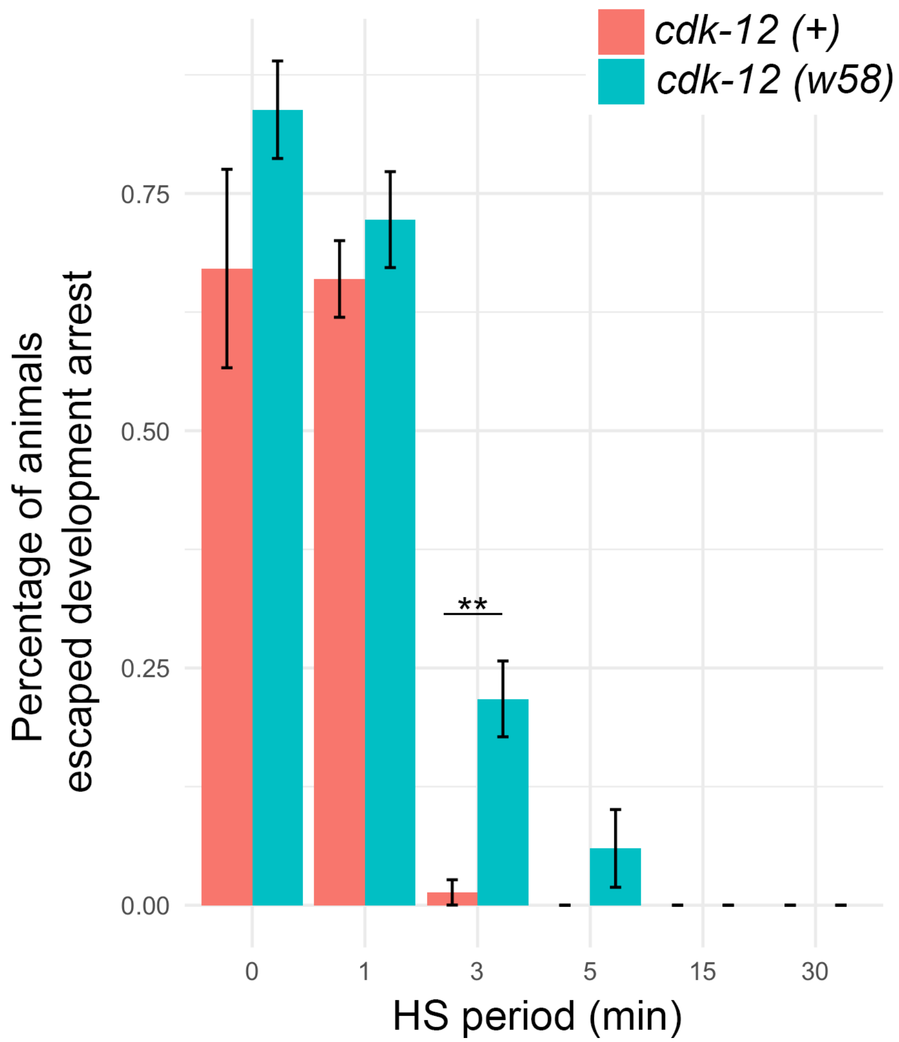


Fig 3.3. *cdk-12(w58)* partially suppresses development arrest caused by ELT-7 overexpression in the HS system.

Synchronized L1 animals were fed on OP50, followed by HS in the thermocycler at various times. The number of heat-shocked animals that develop into adults was counted on Day 4 after HS and presented as a percentage. Significant differences ($p<0.01$, Student's t-test) between the two groups were indicated by a (**) at the 3 mins HS. Each group contains 64-118 animals.

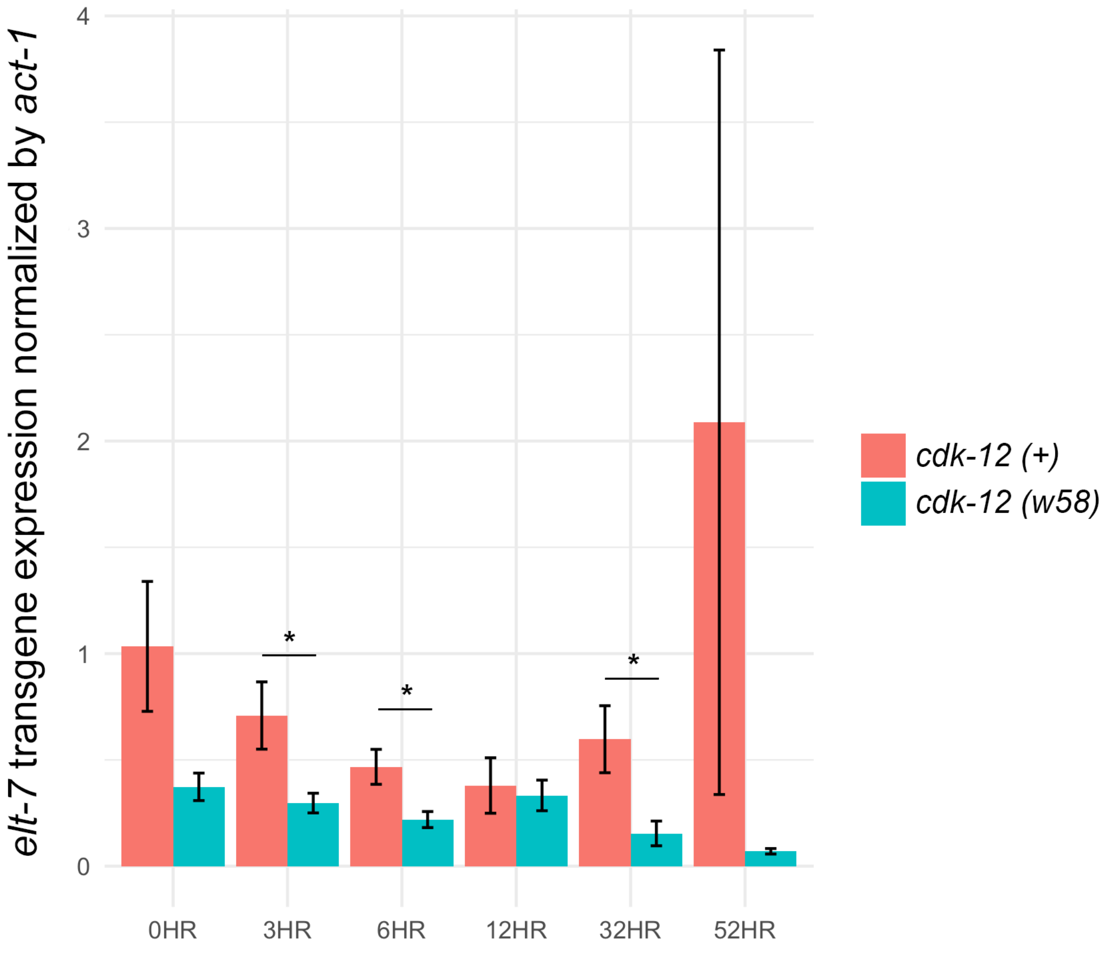
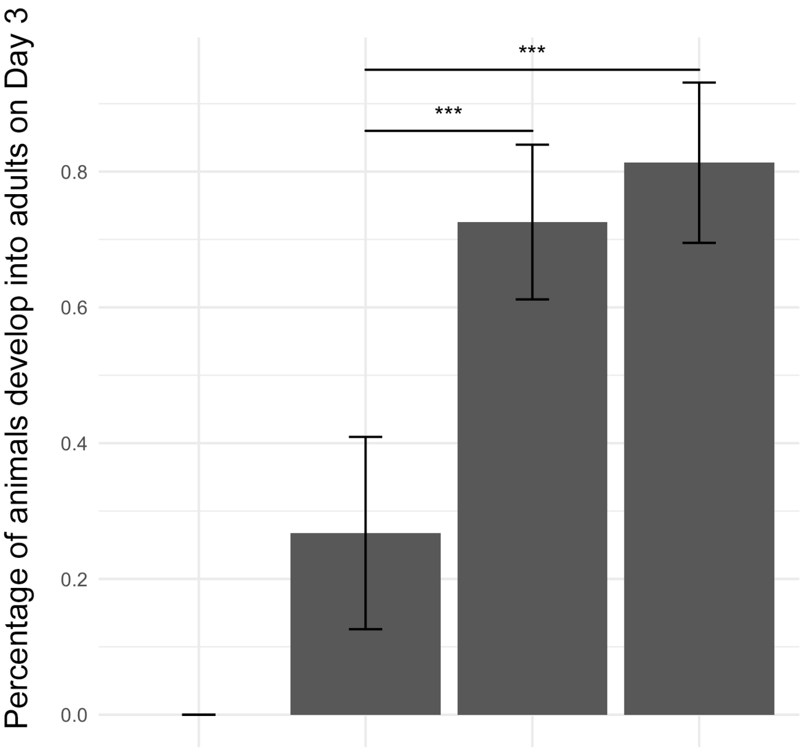


Fig 3.4. Reduced expression of the *elt-7* transgene in the *cdk-12(w58)* animals. Synchronized L1 animals were fed on regular NGM plates before harvesting at various times. Total mRNA was isolated from animals and cDNA was synthesized for qRT-PCR analysis. The expression of elt-7 transgene was decreased in *cdk-12(w58)* animals in most of the groups, and the significant differences (p<0.05, Student's t-test) between the two groups were indicated by a (*). Each group contains 3 animals, and the expression of *elt-7* transgene was normalized by *act-1*.



TIR1	+/+	+/-	+/+	+/+
degron::elt-7	+/+	+/-	+/-	+/+
cdk-12	+/+	+/+	+/+	*w58/w58*

Fig 3.5. Reduced expression of the *elt-7* transgene allows normal development similar to *cdk-12(w58)* animals.

3 Days old sperm-depleted N2 and CA1202 hermaphrodites were mated with JR3904 males and the percentage of F1 progeny that developed into adults was counted. A single copy of elt-7 transgene animals develops similarly to the w58 animals, suggesting an ELT-7 dosage-dependent Td effect. The significant differences (p<0.001, Student's t-test) between the two groups were indicated by a (***). Each group contains 11-54 animals with five experimental replicates.

Chapter Four

Future Directions & Conclusion Remarks

Future directions

Examining *elt-7* transgene expression in *cdk-12 (w59*) and *cdk-12 (w60*).

Two *cdk-12* mutations, a missense mutation *w59* and a nonsense mutation *w60*, have been isolated in group 3 through PCR sequencing. Interestingly, both alleles were located on the C-terminal end of the protein outside of the predicted kinase domain. CDK-12 is known for regulating transcription elongation through phosphorylating RNA polymerase CTD Ser 2, Ser 5, and Ser 7. CDK-12 is also known for regulating RNA splicing and chromatin modification directly or indirectly. It would be interesting to dissect the nature of *w59* and *w60* affecting CDK-12 function. Assuming the kinase activity remains intact in *w59* and *w60*, the plausible explanation will be that *w59* and *w60* suppressed intestinal Td by reducing correct RNA splicing in genes required for Td. Several experiments are required to answer such questions. 1) The causality of *w59* and *w60* must be confirmed by complementation test with *cdk-12* null mutation *tm3846* and *w58* and by rescue experiment with wild-type *ckd-12* transgene (*ckSi6*). 2) The *elt-7* transgene expression has to be confirmed by qRT-PCR. The result will help us speculate how *w59* and *w60* affect CDK-12 function and decide the follow-up experiment. 3) The RNA splicing pattern can be examined by qRT-PCR with genes having a specific isoform that up-regulated in the ELT-7-mediated Td process under wild-type CDK-12, *w59*, and *w60* conditions. Transcriptomic analysis can be applied for the same purpose and may reveal gene networks required for terminal intestinal Td.

Examining *elt-7* transgene expression in HS-*elt-7* system with *cdk-12* mutation.

The *elt-7* transgene in *w58* has shown decreased expression in the AID system. Therefore, confirming the *elt-7* transgene expression in the HS system is also necessary. Indeed, *w58* only partially suppressed the development arrest phenotype in the HS system with a short HS period (3-5 mins) and failed with a more extended period (15-30 mins). This finding cohorts with the idea that the ELT-7-mediated Td is an ELT-7 dosage-dependent event suggesting that there are too many ELT-7 being produced with a long period HS.

The decreased elt-7 transgene expression also allows a more sensitive suppressor of intestinal Td selection can be conducted in the HS system. Therefore, it may be possible to incorporate w58 in the HS system for screening genes that play a subtle role in Td with forward genetic method or RNAi knock-down.

Reduced *elt-7* transgene expression is sufficient for escaping Td and development arrest

Intestinal Td has been shown to be sensitive to the ectopic ELT-7 level in the AID system. Single-copy of *elt-7* transgene animals tend to develop into later stages compared to two copies of the transgene. Indeed, ~26% of single-copy of *elt-7* and *TIR1* transgene animals could develop into adults with a minimal brood size (~3-5 embryos) inside the animals. In addition, single-copy of *elt-7* transgene with two copies of *TIR1* transgene animals showed relatively normal development, with a few animals having a sick phenotype. This finding suggested that the main function of CDK-12 in ELT-7-mediated Td is through regulating the *elt-7* transgene expression.

Several questions remain unclear about the nature of *w58*: 1) The development of *w58* animals should be affected when the transcription decreased globally; however, no significant development difference has been noticed in the *w58* animals. 2) The *elt-7* transgene does not impose a CDK-12 target transcribe since it only has ~1455 bp with no intron. It would be interesting to investigate the global transcription changes in *w58* animals and examine genes in the proximity of the *elt-7* transgene to see if there is a regional effect.

The potential role of ***fust-1*** transcription regulator

fust-1 is the top fourth candidate gene and has impressive performance in mutation density and the KS test. Although the deleterious ratio is relatively low in *fust-1* AID-*elt-7* data, none of the 5 mutations of *fust-1* in the MMP data were deleterious, suggesting a selection of the suppressor of intestinal Td was imposed on *fust-1*. FUST-1 is predicted to function as a transcription elongation regulator in *C. elegans*, and the human homolog, RNA-binding protein FUS (*FUS*) functions in various cellular processes such as transcription regulation, RNA splicing, RNA transport, and DNA repair. These functions are very similar to the identified intestinal Td suppressor CDK-12; therefore, it is necessary to confirm the *elt-7* transgene expression in *fust-1* mutant animals first. It would be interesting to study whether CDK-12 and FUST-1 function on the same pathway.

The potential role of ***D1086.9***

D1086.9 is the top second candidate gene with a good performance on all three analyses. *D1086.9* is a nematode-specific gene only present in *Caenorhabditis* and *Pristionchus* with a CDS length of 2061 nucleotides and a protein length of 686 amino acids. However, no known function or motif has been predicted on D1086.9. Gene enrichment analysis showed that D1086.9 is mainly expressed in cephalic sheath cells and dopaminergic neurons. Mutation of D1086.9 in these cells, specifically in the pharynx, may prevent cells from undergoing intestinal Td, or partly retaining their function of stimulating pharyngeal pumping. It would be interesting to closely examine Td at the cellular level to identify D1086.9-affecting cells after confirming the causality of D1086.9 in ELT-7-mediated Td. In addition, since D1086.9 is expressed in numerous cells, it would be interesting to monitor the pharyngeal pumping rate in those mutants.

Conclusion Remarks

Forward genetics has been a reliable method of studying gene function. The unbiased mutagenesis approach allowed researchers to identify novel genes and pathways that generally would not be considered of. However, identifying the causal gene responsible for the phenotype of interest has been challenging and time-consuming. In this study, we aimed to develop a novel pipeline of identifying candidate genes through whole genomic sequencing on pooled mutant samples isolated from a large-scale forward genetic screening. Almost a hundred unique SNVs were identified from the pooled 660 mutant lines. The sequencing information was evaluated by three bioinformatic analyses: the mutation density, the deleterious ratio, and the mutation distribution. A positive experimental control, *elt-7* transgene, was utilized to evaluate the accuracy of these analyses and has proven to be true. Using our analysis pipeline, a causal gene, *cdk-12*, has been identified and showed its requirement for the ELT-7-mediated Td. In addition, at least two other candidate genes also showed their potential to regulate such Td process and the investigation is ongoing.

The pooled sequencing approach we developed has several advantages that traditional methods cannot provide. (1) The prediction of candidate genes was through computational analysis; therefore, genetic manipulation such as backcrossing or outcrossing is not required. The analysis pipeline can also be used for selecting non-viable phenotypes or for animals unable to perform genetic crossing. (2) The analysis pipeline greatly tolerates potential system failure for a false-positive result. Indeed, the mutation in the *elt-7* transgene would typically consider an experimental design flaw since it disrupts the ELT-7 overexpression system. In addition, roughly one-third of the mutant lines carry such mutations, which can be problematic in traditional mapping methods. However, our analysis pipeline identified those artificial results, including *elt-7* transgene mutations that would disrupt the system and the *ifb-2* mutations caused by the multiple copies of *ifb-2* transgene. (3) It is a fast and cost-efficient method. Since genetic manipulation is not required for our analysis pipeline, the screening process can be done relatively quickly with our developed analyses. The expanse of sequencing pooled mutant lines was also much lower than sequencing one animal at a time, with a 10-fold lower cost. (4) The strategy of the pooled analysis pipeline utilized the fundamental understanding of how genetic selection functions and the nature of mutations affecting protein function. This method can be universally applied to every organism, but ideally, to animals with large brood sizes that can support the initial screening.

The pooled mutant sequencing method was applied to identify the suppressor of development arrest caused by ELT-7-mediated Td. CDK-12, a protein function in various cellular processes, was identified and confirmed as the requirement for intestinal Td. The preliminary results suggested that CDK-12 may suppress the intestinal Td by reducing *elt-7* transgene expression. However, the decreased expression of *elt-7* transgene alone was not sufficient to completely block Td in heterozygous, one-copy of *elt-7* transgene animals. This result suggests CDK-12 may suppress intestinal Td through other mechanisms, such as RNA splicing or chromatin modification. Interestingly, another candidate gene, *fust-1*, also functions similarly to CDK-12 in regulating transcription elongation and RNA splicing. It would be valuable to investigate the detailed mechanism of these genes in ELT-7-mediated Td and identify their potential overlapping role.

References

Al Abbar, A., Ngai, S.C., Nograles, N., Alhaji, S.Y., and Abdullah, S. (2020). Induced Pluripotent Stem Cells: Reprogramming Platforms and Applications in Cell Replacement Therapy. Biores Open Access *9*, 121-136. 10.1089/biores.2019.0046.

Bar-Ziv, R., Frakes, A.E., Higuchi-Sanabria, R., Bolas, T., Frankino, P.A., Gildea, H.K., Metcalf, M.G., and Dillin, A. (2020). Measurements of Physiological Stress Responses in C. Elegans. J Vis Exp. 10.3791/61001.

Cassart, C., Yague-Sanz, C., Bauer, F., Ponsard, P., Stubbe, F.X., Migeot, V., Wery, M., Morillon, A., Palladino, F., Robert, V., and Hermand, D. (2020). RNA polymerase II CTD S2P is dispensable for embryogenesis but mediates exit from developmental diapause in C. elegans. Sci Adv *6*. ARTN eabc1450

10.1126/sciadv.abc1450.

Choi, S.H., Kim, S., and Jones, K.A. (2020). Gene expression regulation by CDK12: a versatile kinase in cancer with functions beyond CTD phosphorylation. Exp Mol Med *52*, 762-771. 10.1038/s12276-020-0442-9.

Domingo, E., Garcia-Crespo, C., Lobo-Vega, R., and Perales, C. (2021). Mutation Rates, Mutation Frequencies, and Proofreading-Repair Activities in RNA Virus Genetics. Viruses *13*. 10.3390/v13091882.

Egusa, H., Sonoyama, W., Nishimura, M., Atsuta, I., and Akiyama, K. (2012). Stem cells in dentistry--part I: stem cell sources. J Prosthodont Res *56*, 151-165. 10.1016/j.jpor.2012.06.001.

Emadi, F., Teo, T., Rahaman, M.H., and Wang, S. (2020). CDK12: a potential therapeutic target in cancer. Drug Discov Today *25*, 2257-2267. 10.1016/j.drudis.2020.09.035.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154-156. 10.1038/292154a0.

Ewe, C.K., Sommermann, E.M., Kenchel, J., Flowers, S.E., Maduro, M.F., Joshi, P.M., and Rothman, J.H. (2022). Feedforward regulatory logic controls the specification-to-differentiation transition and terminal cell fate during Caenorhabditis elegans endoderm development. Development *149*. 10.1242/dev.200337.

Fay, D.S. (2013). Classical genetic methods. WormBook, 1-58. 10.1895/wormbook.1.165.1.

Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell *128*, 635-638. 10.1016/j.cell.2007.02.006.

Graf, T. (2011). Historical origins of transdifferentiation and reprogramming. Cell Stem Cell *9*, 504-516. 10.1016/j.stem.2011.11.012.

Hills-Muckey, K., Martinez, M.A.Q., Stec, N., Hebbar, S., Saldanha, J., Medwig-Kinney, T.N., Moore, F.E.Q., Ivanova, M., Morao, A., Ward, J.D., et al. (2022). An engineered, orthogonal auxin analog/AtTIR1(F79G) pairing improves both specificity and efficacy of the auxin degradation system in Caenorhabditis elegans. Genetics *220*. 10.1093/genetics/iyab174.

Hirai, H., Karian, P., and Kikyo, N. (2011). Regulation of embryonic stem cell self-renewal and pluripotency by leukaemia inhibitory factor. Biochem J *438*, 11-23. 10.1042/BJ20102152.

Hodgkin, J., and Barnes, T.M. (1991). More is not better: brood size and population growth in a self-fertilizing nematode. Proc Biol Sci *246*, 19-24. 10.1098/rspb.1991.0119.

Jarriault, S., Schwab, Y., and Greenwald, I. (2008). A Caenorhabditis elegans model for epithelial-neuronal transdifferentiation. Proc Natl Acad Sci U S A *105*, 3790-3795. 10.1073/pnas.0712159105.

Jiang, Y.F., Jiang, Y., Wang, S., Zhang, Q., and Ding, X.D. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. Bmc Bioinformatics *20*. ARTN 556 10.1186/s12859-019-3164-z.

Jopling, C., Boue, S., and Belmonte, J.C.I. (2011). Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. Nat Rev Mol Cell Bio *12*, 79-89. 10.1038/nrm3043.

Keller, G. (2005). Embryonic stem cell differentiation: emergence of a new era in biology and medicine. Genes Dev *19*, 1129-1155. 10.1101/gad.1303605.

Koboldt, D.C. (2020). Best practices for variant calling in clinical sequencing. Genome Med *12*, 91. 10.1186/s13073-020-00791-w.

Kumar, S., and Subramanian, S. (2002). Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A *99*, 803-808. 10.1073/pnas.022629899.

Li, Q.F., Wang, N., Sui, C., Mao, H.D., Zhang, L., and Chen, J.H. (2022). PacBio single molecule real-time sequencing of a full-length transcriptome of the greenfin horse-faced filefish Thamnaconus modestus. Front Mar Sci *9*. ARTN 1028231 10.3389/fmars.2022.1028231.

Ly, K., Reid, S.J., and Snell, R.G. (2015). Rapid RNA analysis of individual Caenorhabditis elegans. Methodsx *2*, 59-63. 10.1016/j.mex.2015.02.002.

Lynch, M. (2010). Evolution of the mutation rate. Trends Genet *26*, 345-352. 10.1016/j.tig.2010.05.003.

Manavalan, A.P.C., Pilarova, K., Kluge, M., Bartholomeeusen, K., Rajecky, M., Oppelt, J., Khirsariya, P., Paruch, K., Krejci, L., Friedel, C.C., and Blazek, D. (2019). CDK12 controls

G1/S progression by regulating RNAPII processivity at core DNA replication genes. Embo Rep *20*. ARTN e4759210.15252/embr.201847592.

Mark, M., Rijli, F.M., and Chambon, P. (1997). Homeobox genes in embryogenesis and pathogenesis. Pediatr Res *42*, 421-429. 10.1203/00006450-199710000-00001.

Martinato, F., Cesaroni, M., Amati, B., and Guccione, E. (2008). Analysis of Myc-induced histone modifications on target chromatin. PLoS One *3*, e3650. 10.1371/journal.pone.0003650.

Martinez, M.A.Q., Kinney, B.A., Medwig-Kinney, T.N., Ashley, G., Ragle, J.M., Johnson, L., Aguilera, J., Hammell, C.M., Ward, J.D., and Matus, D.Q. (2020). Rapid Degradation of Caenorhabditis elegans Proteins at Single-Cell Resolution with a Synthetic Auxin. G3-Genes Genom Genet *10*, 267-280. 10.1534/g3.119.400781.

McGhee, J.D. (March 27, 2007). The C. elegans intestine (The C. elegans Research Community).

McMillan, D.R., Xiao, X., Shao, L., Graves, K., and Benjamin, I.J. (1998). Targeted disruption of heat shock transcription factor 1 abolishes thermotolerance and protection against heat-inducible apoptosis. J Biol Chem *273*, 7523-7528. 10.1074/jbc.273.13.7523.

Meng, S., Chanda, P., Thandavarayan, R.A., and Cooke, J.P. (2017). Transflammation: Innate immune signaling in nuclear reprogramming. Adv Drug Deliv Rev *120*, 133-141. 10.1016/j.addr.2017.09.010.

Miller, J.L., and Grant, P.A. (2013). The role of DNA methylation and histone modifications in transcriptional regulation in humans. Subcell Biochem *61*, 289-317. 10.1007/978-94-007-4525-4_13.

Narayan, S., Bryant, G., Shah, S., Berrozpe, G., and Ptashne, M. (2017). OCT4 and SOX2 Work as Transcriptional Activators in Reprogramming Human Fibroblasts. Cell Rep *20*, 1585-1596. 10.1016/j.celrep.2017.07.071.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814. 10.1093/nar/gkg509.

Pan, L., Xie, W., Li, K.L., Yang, Z., Xu, J., Zhang, W., Liu, L.P., Ren, X., He, Z., Wu, J., et al. (2015). Heterochromatin remodeling by CDK12 contributes to learning in Drosophila. Proc Natl Acad Sci U S A *112*, 13988-13993. 10.1073/pnas.1502943112.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. Nat Rev Cancer *11*, 761-774. 10.1038/nrc3106.

Riddle, M.R., Spickard, E.A., Jevince, A., Nguyen, K.C., Hall, D.H., Joshi, P.M., and Rothman, J.H. (2016). Transorganogenesis and transdifferentiation in C. elegans are dependent on differentiated cell identity. Dev Biol *420*, 136-147. 10.1016/j.ydbio.2016.09.020.

Romito, A., and Cobellis, G. (2016). Pluripotent Stem Cells: Current Understanding and Future Directions. Stem Cells Int *2016*, 9451492. 10.1155/2016/9451492.

Spickard, E.A., Joshi, P.M., and Rothman, J.H. (2018). The multipotency-to-commitment transition in Caenorhabditis elegans-implications for reprogramming from cells to organs. FEBS Lett *592*, 838-851. 10.1002/1873-3468.12977.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676. 10.1016/j.cell.2006.07.024.

Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., et al. (2013). The million mutation project: a new approach to genetics in Caenorhabditis elegans. Genome Res *23*, 1749-1762. 10.1101/gr.157651.113.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145-1147. 10.1126/science.282.5391.1145.

Thowfeequ, S., Myatt, E.J., and Tosh, D. (2007). Transdifferentiation in developmental biology, disease, and in therapy. Dev Dyn *236*, 3208-3217. 10.1002/dvdy.21336.

Tien, J.F., Mazloomian, A., Cheng, S.G., Hughes, C.S., Chow, C.C.T., Canapi, L.T., Oloumi, A., Trigo-Gonzalez, G., Bashashati, A., Xu, J., et al. (2017). CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. Nucleic Acids Res *45*, 6698-6716. 10.1093/nar/gkx187.

Torres Cleuren, Y.N., Ewe, C.K., Chipman, K.C., Mears, E.R., Wood, C.G., Al-Alami, C.E.A., Alcorn, M.R., Turner, T.L., Joshi, P.M., Snell, R.G., and Rothman, J.H. (2019). Extensive intraspecies cryptic variation in an ancient embryonic gene regulatory network. Elife *8*. 10.7554/eLife.48220.

Tsonis, P.A., and Del Rio-Tsonis, K. (2004). Lens and retina regeneration: transdifferentiation, stem cells and clinical applications. Exp Eye Res *78*, 161-172. 10.1016/j.exer.2003.10.022.

Wei, Z., Yang, Y., Zhang, P., Andrianakos, R., Hasegawa, K., Lyu, J., Chen, X., Bai, G., Liu, C., Pera, M., and Lu, W. (2009). Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. Stem Cells *27*, 2969-2978. 10.1002/stem.231.

Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods *16*, 1297-+. 10.1038/s41592-019-0617-2.

Zhang, L., Ward, J.D., Cheng, Z., and Dernburg, A.F. (2015). The auxin-inducible degradation (AID) system enables versatile conditional protein depletion in C. elegans. Development *142*, 4374-4384. 10.1242/dev.129635.