UNIVERSITY OF CALIFORNIA, SAN DIEGO

Dissecting the Worm Intestinal Genetic Network with Quantitative Measurement and Modeling

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in

Bioinformatics and Systems Biology

by

Chia-Yi Wu

Committee in charge:
Professor Scott A. Rifkin, Chair
Professor Jeff Hasty, Co-Chair
Professor Serge Belongie
Professor Andrew Chisholm
Professor Arshad Desai
Professor Roy Wollman

2016
The Dissertation of Chia-Yi Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

________________________________________

________________________________________

________________________________________

Co-Chair

Chair

University of California, San Diego

2016
TABLE OF CONTENTS

SIGNATURE PAGE ................................................................. iii
TABLE OF CONTENTS .............................................................. iv
LIST OF FIGURES ................................................................. vii
ACKNOWLEDGEMENTS ......................................................... ix
VITA .................................................................................... xii
ABSTRACT OF THE DISSERTATION .......................................... xiii

Chapter I. Introduction .......................................................... 1
  1.1 Understanding the role of gene expression noise in developmental process and phenotypic variation .................................................................................. 1
  1.2 The *C. elegans* intestinal specification network is a perfect system to study gene expression variation and phenotypic variation ............................................. 3
  1.3 Conservation of these intestinal specifying genes among nematodes ................... 5

Chapter II. Aro: a machine learning approach to identifying single molecules and estimating classification error in fluorescence microscopy images ........................................ 7
  2.1 Introduction ............................................................................ 7
  2.2 Implementation ....................................................................... 8
  2.3 Results and Discussions .......................................................... 13
  2.4 Conclusions ........................................................................... 15
  2.5 Acknowledgements .................................................................... 16

Chapter III. Gene Expression Variability and Robustness in the *C.
**Chapter IV.** A Computational Model of *C. elegans* Intestinal Specification Network

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>30</td>
</tr>
<tr>
<td>4.2 Methods</td>
<td>31</td>
</tr>
<tr>
<td>4.3 Results and Discussions</td>
<td>36</td>
</tr>
<tr>
<td>4.4 Conclusions</td>
<td>41</td>
</tr>
<tr>
<td>4.5 Acknowledgements</td>
<td>44</td>
</tr>
</tbody>
</table>

**Chapter V.** Evolution of the Intestinal Specification Network: *C. remanei, C. briggsae and C. elegans*

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>45</td>
</tr>
<tr>
<td>5.2 Methods</td>
<td>46</td>
</tr>
<tr>
<td>5.3 Results</td>
<td>47</td>
</tr>
</tbody>
</table>
5.4 Discussions and Conclusions........................................................................................................52

Chapter VI. Figures and Tables......................................................................................................55
Chapter VII. References................................................................................................................78
LIST OF FIGURES

Figure 1 The survival rates of the strains examined in this study------------------------55

Figure 2 GUIs in Aro -------------------------------------------------------------56

Figure 3 Flowchart of the analysis pipeline of Aro ----------------------------------57

Figure 4 Calibration curves based on bagged probabilities --------------------------58

Figure 5 The algorithm performs well even as data quality degrades-------------------59

Figure 6 Comparison of spot identification and classification methods---------------60

Figure 7 Expression of elt-2, elt-7, end-1 and end-3 at embryo level------------------61

Figure 8 The end-1, elt-7, elt-2 expressions at cell level-------------------------------62

Figure 9 Comparison of elt-7 expression levels and variability ------------------------63

Figure 10 Comparison of elt-2 and elt-7 variability-----------------------------------64

Figure 11 Comparison of elt-2 expression in end-3/- and end-1/- ---------------------65

Figure 12 Comparison of elt-2 expression in end-3/- and end-1/- with elt-7 RNAi -----66

Figure 13 Loss of elt-7 increases elt-2 gene expression variability. -------------------67

Figure 14 Delay of elt-2 expression is observed in some end-1/- experiments. --------68

Figure 15 Saturation of gene regulation can be observed for end-1 and elt-2 -----------69

Figure 16 Comparison between simulated results and experimental results------------70

Figure 17 The timing of switching point and the ELT-2 level in each strains-----------71

Figure 18 Predicting the survival rates of end-3/-, and end-1/-;end-3/- (MED-) -------72

Figure 19 Maximum merged image of Cbr-end-3 and CBG11404 labeled embryo ---------73

Figure 20 The total number of end-3.1 and end-3.2 expression in C. briggsae----------74
Figure 21 Gene expressions in *C. briggsae, C. remanei* and *C. elegans*------------------------75

Figure 22 Comparison of *end-1* in *C. elegans end-3-/-* and *C. remanei*.-------------------76

Figure 23 *elt-7* expression in *C. remanei* is significantly lower than *Cel-elt-7*.----------77
ACKNOWLEDGEMENTS

I am very fortunate to have a very supporting family and have encountered many friends and mentors who have guided and supported thought out my academic life.

First, I would like to thank my Ph.D adviser, Scott Rifkin. We both came to UCSD in the same year. I am really grateful that he took me as his first graduate student. I still remembered those days we spent doing experiments together and am truly grateful to have become the first member in Rifkin lab. Scott had been very generous in giving me opportunities to go outside of the Rifkin lab to explore my academic interests and to interact and collaborate with different people. I have had chances to broaden my academic relationships with people in this field by going to many conferences. This has truly helped me in not being restricted by what I learn or people I know in UCSD. Aside from the generosity of opportunities for me, Scott has been a great mentor in both life and academic experiences. He has particularly strengthen my training in evolutionary biology and guided me through various hypotheses and models of evolutionary biology.

Second, I would like to thank my mentors in the field of quantitative biology, Jeff Hasty, Alex Hoffman, Lev Tsimring and Roy Wollman for giving me very useful advice on mathematical modeling, imagine processing and microscopy. They have all helped me in my training in quantitative biology, encouraged me to think outside the box of biology and inspired me to think of biological reactions more with numbers and equations. Quantitative biology has always been the reason I came to the states to study
and it has inspired me so much in thinking biological experiments in an approach that is very different from my biological background and I am really fortunate and truly grateful to have been guided by these masters in this field here in UCSD.

Aside from my mentors in quantitative biology, I would also like to thank my advisers in the field of biology, Morris Maduro, Arshad Desai and Andrew Chisholm. I would especially thank Dr. Maduro for his passion in our collaboration and generosity in sharing their deletion strains with us. Without this collaboration, I would not have been able to complete my study on this network with the level of details. I'm really grateful that Dr. Desai could step in as one of my committee members in the last minute. Dr. Desai and Dr. Chisholm both have given me very valuable suggestions in biological questions, which I'm really grateful for.

At the end of my acknowledgments, I would like to thank my family, especially my parents, and two of my best friends, Minxi Rao and Sidney Kuo. I really appreciate that my parents have always supported me to study abroad and be so far away from home to pursue something I am truly interested in. I can never be fortunate enough to have Minxi Rao and Sidney Kuo as my best friends and lab mates. They have supported me through my worst days and shared happiness with me through my best days along my 6 years here.

There are still so many more people I would like to thank, all my past and preset lab mates, my previous adviser and many more. I can never become who I am or achieve what I have achieved here without the help, advice and love from all these
people.

Chapter 2, in full, is a reprint of the materials as it appears in Aro: a machine learning approach to identifying single molecules and estimating classification error in fluorescence microscopy images in BMC Bioinformatics 2015 by Allison Chia-Yi Wu and Scott A. Rifkin. Scott A. Rifkin is the corresponding author. The dissertation author was the primary investigator and author of this paper.

Chapter 3 and 4 together are being prepared for submission for publication of the material and may appear as Allison Chia-Yi Wu, Morris Maduro, Scott A. Rifkin. “The response of elt-2 in C. elegans intestinal specification network reveals the mechanism for buffering the gene expression noise.” The dissertation author was the primary investigator and author of this paper. Morris Maduro has provided the deletion strains needed for part of the experimental design. Scott Rifkin was the corresponding author.
VITA

2006 – 2009 Undergraduate Researcher, National Taiwan University

2009 Bachelor of Science, National Taiwan University

2009 – 2015 Graduate Researcher, University of California, San Diego

2011 FQRNT Merit Short-term Research Fellow

2016 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS


How a biological system copes with its intrinsic noise and extrinsic noise has become a major question in the field of quantitative biology. In particular, developmental and evolutionary consequences of gene expression noise has merged as an active area of study in the last few years. Many genetic network motifs and molecular mechanisms have been found to tune or control gene expression variation in
different biological systems. Some of the pioneer studies in quantitative biology have discovered regulatory network motifs critical to developmental processes. How gene expression variation leads to phenotypic variation and how phenotypic variation can lead to improved species fitness. To date, few studies have focused on how gene expression variation leads to phenotypic variation in developmental processes. In our study, a quantitative method, single-molecule FISH, is coupled with mathematical modeling to examine how gene expression variation can be propagated through the *C. elegans* intestinal development network and how gene expression variation during developmental processes translate into phenotypic variation.

This quantitative approach has been applied to other *Caenorhabditis* species. Differences in expression profiles of orthologous genes in the intestinal development network were observed. Taken in the context of a partially conserved network topology, expression differences were added to the model and yielded insights into mechanisms behind developmental decision making in formation of intestines in *C. briggsae* and *C. remanei*. 
Chapter I. Introduction

1.1 Understanding the role of gene expression noise in developmental process and phenotypic variation

The functional roles and developmental evolutionary consequences of gene expression noise have just begun to be appreciated in the last few years. Gene expression is stochastic in both prokaryotes (Elowitz et al., 2002) and eukaryotes (Raser and O’Shea, 2004a) and can be inherently variable. How a biological system copes with its intrinsic noise and extrinsic noise has become a major question in the field of quantitative biology. Many genetic network motifs and molecular mechanisms have been discovered to tune or control gene expression variation in different biological systems (Hao et al., 2013; Ji et al., 2013; Raser and O’Shea, 2004a; Singh et al., 2012). While some of the pioneer studies in quantitative biology have discovered the regulatory network motifs in developmental processes (Chopra and Levine, 2009; Gibb et al., 2009; Jiang et al., 2000; Shvartsman et al., 2008), some other studies are looking at how gene expression variation leads to phenotypic variation and how phenotypic variation could lead to potential better fitness of the species (Ansel et al., 2008; Burga and Lehner, 2013; Espinosa-Soto et al., 2011; Singh and Weinberger, 2009). Very few of these studies focus on how gene expression variation leads to phenotypic variation in developmental processes. In our study, we look at how gene expression variation during developmental processes can be translated into phenotypic variation in *C. elegans* intestinal specification network.

In worms, fluorescence live cell imaging has enabled direct real-time observation of gene expression, but it has traditionally been limited by the need to over-
express the fluorescent reporters in the target organism. This hinders studies on genes with low expression levels under normal conditions since the over-expression might disrupt the physiology of the organism and may not be representative of the normal function. In this study, we use a variant of fluorescent in situ hybridization (FISH) to circumvent this problem. While single-molecule FISH requires fixed specimens and thus loses the advantage of real-time observation, it allows us to quantify gene expression at the single-cell level with single-molecule resolution without the need for transgenes (Raj et al., 2008). Using this method combined with mathematical modeling, we are able to address specific questions on the role of gene expression noise and the dynamics of developmental networks in multi-cellular organisms without disrupting the system with over-expressed genes. This approach in worms has been used in the study on the vulva development in *C. elegans* (Barkoulas et al., 2013; Ji et al., 2013) and a network motif with interlinked positive and negative feedback loops in the Wnt signaling pathway has been found to control the gene expression variability of a key gene, *mab-5*, that specifies the vulva development.

In our study, instead of looking for a certain network topology that controls gene expression variability, we find alternative mechanisms that can be adopted by the organism to maintain the robustness of the gene expression with simple circuits that do not have intricate nested positive and negative feedback loops.
1.2 The C. elegans intestinal specification network is a perfect system to study gene expression variation and phenotypic variation

The C. elegans intestinal specification network is a small genetic regulatory network with major players such as skn-1, pop-1, pal-1, med-1/2, end-1, end-3, elt-7 and elt-2 (Maduro and Rothman, 2002). In our study, we focus on skn-1, med-1/2, end-1, end-3, elt-7 and elt-2 and cell fate specification after EMS stage. This network has a unique regulatory hierarchy that has at least 2 genes at each developmental stage to determine the cell fate after the specification of EMS cell. First, skn-1 is unevenly distributed to the P₁ cell and its daughter cells (Bowerman et al., 1992, 1993) and together with its activation of med-1/2, these two genes specify the P₁ → EMS cell fate. In the next stage, meso-endoderm differentiation separates the cell fates of two daughter cells of EMS cells. In this stage, med-1/2 activates end-3 and end-1. The expressions of end-3 and end-1 then secure the cell fate of the endoderm cells in E² stage (Coroian et al., 2006; Lin et al., 2009; Maduro and Rothman, 2002; Maduro et al.; Zhu et al., 1997). In the last layer of this network, elt-2 and elt-7 synergistically determine the intestinal development in the last 8-E cell stage (Fukushige et al., 1998; Sommermann et al., 2010).

The small number of genes makes it possible for us to use single-molecule FISH to study the expression profile of each of them in details. All these genes are transcription factors that turn on the gene expression of their targets. The structure of this genetic network is shown in Figure 1. This network is a good system to study the...
relationship between gene expression variation and phenotypic variation because different perturbations, such as deletions or mutations of parts of this network lead to different levels of penetrance. For example, skn-1 is a maternal gene which is expressed in the mom and the mRNA is deposited in the embryos (Bowerman et al., 1992, 1993; Lin et al., 2009). The zygotic expression of skn-1 does not turn on until very late in the embryonic stage and it is the maternally deposited mRNAs that determine the intestinal specification in early embryonic stage. Different alleles of skn-1 mutation has different levels of penetrance in their intestinal and pharyngeal development (Bowerman et al., 1993). Thanks to the small number of genes involved, we can can assay strains with single mutations, one gene per strain, and even with conditions similar to double mutation for certain cases via RNA interference. Figure 1 shows a chart of mutant strains we examined and their different levels of phenotypic variation.

The simplicity of the network structure also makes it a straightforward system for us to study how gene expression variation is propagated in the network and how it is translated to phenotypic variation. There are only two positive feedback loops, autoregulation of elt-2 and possibly of elt-7. The whole network is a feed-forward network. In addition, the availability of deletion strains and ease of RNAi allow us to isolate each gene-gene interaction. For example, the end-1-/-;elt-7(RNAi) effectively narrows the number of inputs to elt-2 to just END-3. All these features of this network make this system a great system to study how gene expression variation can be translated to phenotypic variation.
1.3 Conservation of these intestinal specifying genes among nematodes

Inter-species comparisons of genetic regulatory networks (GRNs) often lead to the identification of conserved GRN motifs and cis-regulatory sites (Miranda-Dominguez et al., 2014; Yuh et al., 2002). These previous studies mostly compare the network structure through comparative genomics and functional transgenetics (Coroian et al., 2006), which inserts genes from different species into a closely-related and well-studied species to see if the same function can be carried out. Comparative studies have shown that many of the major players in this network is conserved in other species within this genus, especially in two of the closely related species we focus on, *C. briggsae* and *C. remanei* (Husson et al., 2009; Maduro, 2009; Mitreva et al., 2005; Rudel and Kimble, 2001; Stein et al., 2003). These two species diverged from *C. elegans* about 80-110 million years ago (Cho et al., 2004). Searches through both *C. briggsae* and *C. remanei* genome result in at least one ortholog for every gene involved in this network in *C. elegans*. While *elt-2*, *elt-7*, and *end-1* all have only one ortholog in *C. briggsaei* and *C. remanei*, *end-3* and *med* genes appear to have undergone multiple duplications: there are 2 *end-3* paralogs present in *C. briggsae* (*Cbr-end-3.1* and *Cbr-end-3.2*) while the *meds* have four paralogs in *C. briggsae* and seven in *C. remanei*. Two of the *Cbr-med* orthologs and 5 out of seven *Cre-meds* can complement the loss of *Cel-med-1,2* when expressed as transgenes (Coroian et al., 2006). Both of *Cbr-end-3* paralogs appear to have conserved functions in *C. briggsae* (Maduro et al., 2005). In addition, *C. briggsae* has a remarkably similar embryonic development to *C. elegans*.
(Zhao et al., 2008).

With all these previous studies, we are confident to assume GRN structures similar to *C. elegans* exist in both of these species. In our study, we apply the same quantitative approach we used in *C. elegans* to *C. briggsae* and *C. remanei* to derive detailed expression profiles of each of these genes. By assuming similar network structures, we are able to extend our mathematical model to these species to ask two questions: First, are the roles of the major players conserved in these two species? Second, what are the alternative mechanisms that evolved to maintain the robustness to the gene expression variation?
2.1 Introduction

In the last decade, a host of new technologies for tagging and visualizing individual molecules have yielded unprecedented quantitative insight into the spatial and temporal dynamics of fundamental biological processes as varied as ligand-receptor interactions at the cell surface (Sako et al., 2000), protein localization to synaptic junctions (Dani et al., 2010), and incomplete penetrance (Raj et al., 2010).

For example, the ability to visualize mRNA transcripts at the single molecule level without transgenic methods has led single-molecule fluorescence in situ hybridization (smFISH) to be widely used in studying gene expression in various organisms (Barkoulas et al., 2013; Bumgarner et al., 2012; Ji et al., 2013; Neuert et al., 2013; Raj et al., 2008, 2010; Sako et al., 2000). Recently this technique has been pushed to image up to 32 genes simultaneously with the promise of increasing this number still more (Lubeck and Cai, 2012). These microscopy-based techniques rely primarily on fluorescent proteins or dyes that are bound to the molecule of interest and appear as a bright, roughly Gaussian spot. Background fluorescence can be considerable for some of these techniques, including smFISH (Femino et al., 1998; Raj et al., 2008), which makes distinguishing signal from noise an image processing challenge. However, a statistically principled, automated, and robust method for analyzing the images and classifying local intensity maxima as signal or noise, and estimating the accuracy and variability of these classifications has not been developed. This problem is acute since highly
sensitive microscopy methods like smFISH are ideally suited for quantitatively studying stochastic variation in gene expression and other molecular processes within a population.

We have extended a machine-learning pipeline for identifying, localizing, and counting biologically meaningful intensity maxima in 3D image stacks (Rifkin, 2011) both by improving the initial spot classification and, crucially, by providing a way to both estimate the quality of the data and generate an interval estimate for the number of molecules in it. We have tested it extensively on the challenging case of wide-field epifluorescence smFISH image stacks of nematode embryos where there can be substantial background fluorescence, and it also works on other samples like yeast and mammalian cell culture where the signal to noise ratio is more favorable. Unlike other commonly used methods (McIsaac et al., 2013; Mueller et al., 2013; Raj et al., 2010), this software does not rely on arbitrary or user-defined parameters and cutoffs, but instead recognizes and classifies individual mRNA spots by measuring several features of local intensity maxima and classifying them with a supervised random forest classifier (Breiman, 2001), it is a spot-centric approach as compared with approaches that involve thresholding an entire image (McIsaac et al., 2013; Mueller et al., 2013; Raj et al., 2010).

2.2 Implementation

Machine learning has been remarkably successful in a variety of classification and prediction tasks (Flach, 2012; Hastie et al., 2001). As with all supervised machine
learning techniques, our pipeline trains a classifier based on a curated training set and then applies this classifier to new data. Our implementation includes a GUI to create the training set and a GUI for review and revision of the final classification (Figure 2). This review GUI also allows the user to retrain the classifier incorporating any corrections. (If a dataset only consists of a few image stacks, the GUIs used for either the training or review could be used to manually curate the images without the need for machine learning). The software currently uses the random-forest implementation provided in the MATLAB Statistics Toolbox.

The first step for processing either a training image or a new image is to identify all local intensity maxima (spots) within an image stack and rank them in descending order by their background corrected intensities (Figure 3). These are then sequentially fit to 2D Gaussian surfaces until the mean squared error from the fits are persistently less than a cutoff value below which local maxima are empirically found never to be true signal spots. This cutoff is set to be extremely conservative because its function is simply to save time and memory by removing the majority of local maxima that represent noise in an image stack. The heart of the pipeline is a random forest classifier [19] – an ensemble of decision trees built from bootstrapped training sets – which has been shown to produce highly accurate classifications in a wide variety of applications (Booth et al., 2014; Caruana et al., 2008; Chen and Ishwaran, 2012; Cutler et al., 2007; Tüselmann et al., 2015; Verikas et al., 2011).

Our training GUI allows a user to view spots from a subset of image stacks in the
dataset, generate a manually curated training set by classifying them as true signal spots or noise, and build a forest of decision trees based on the features calculated from each spot (Figure 2). We find that a training set consisting of a few hundred positive and negative examples is sufficient for stable classification. For each tree, the algorithm selects a bootstrapped sample from this training data. Each split in the tree is based on a randomly chosen subset of the statistics, and the tree is grown according to pre-specified stopping criteria. The leaves can be, but are not necessarily, comprised of a single class. At the end of training, the user has a bagged ensemble of decision trees.

To classify a new local maximum, the program runs the statistics for the putative spot through each tree to a terminal leaf. The proportion of training spots in this leaf that are manually classified as good can be used to estimate the probability that the new local maximum is a true signal spot. Although such probabilities are known to be inaccurate for single decision trees (Provost and Domingos, 2002), using an ensemble of bagged trees improves the probability estimate, and so we average these proportions for a single candidate spot across all the trees in the forest to estimate a preliminary probability that it is a true spot (Biau, 2012; Kruppa et al., 2013; Malley et al., 2012; Provost and Domingos, 2002). However, these preliminary probabilities do not necessarily reflect the long-run frequency of a spot with particular features being classified as signal or noise (Bostrom, 2008; Gebel and Weihs, 2007; Niculescu-Mizil and Caruana, 2005; Zadrozny and Elkan, 2001).

In order to calibrate these preliminary probability estimates and transform them
into more accurate probabilities, we use empirical data derived from thousands of training spot examples curated by different people. We bin this data by the preliminary probability estimate and then count the number of true spots in each bin. We fit a sigmoidal function (Niculescu-Mizil and Caruana, 2005; Platt, 1999) to the plot of the proportion of true spots against the probability estimate, and we use this function to transform the preliminary probability estimate of a local maxima being a true spot (derived from averaging across the decision trees) to an empirical probability estimate based on curated data (Figure 4). This calibration curve is remarkably similar for different users and different datasets, and users can create their own calibration curves based on their own training sets. If the calibrated probability is greater than 50%, the local maximum is classified as a true spot. The user can then review and edit the classification using the review GUI and has the option to add any corrections to the training set, re-train the random forest, and re-run the classification (Figure 2). At any time the user can remake the random forest based on the augmented training set and rerun the classification.

The calibrated probability reflects uncertainty in the classification of any particular spot, and, consequently, can be used to measure the uncertainty in the count of the number of true spots in an image. A local intensity maximum with a particular preliminary probability can be thought of as a sample of size 1 from the population of all candidate spots with the same preliminary probability, of which some fraction (the calibrated probability) are true spots. We would like to estimate a confidence interval for
the count of true spots in an image. The width of the confidence interval is a measure of
the quality of an image because it will largely be driven by the fraction of spots for
which the user himself or herself would be ambivalent, based on how he or she has
classified similar spots in the training set.

To construct the interval estimate, we conduct a set of $n$ Bernoulli trials with
variable probabilities (also known as Poisson trials) (Feller, 1971) where $n$ is the number
of local maxima tested in an image. The variable probabilities ($p_k$) are based on the
calibrated probability estimates for each local maximum (indexed by $k$) (see Additional
file 1 for details). The number of good outcomes ($X_k = 1$) in this set of Poisson trials is a
simulated estimate for the number of transcripts in the image.

$$T \sim \left[ \sum_{k=1}^{n} \left( X_k \left| P\{X_k = 1\} = p_k, P\{X_k = 0\} = (1 - p_k) \right. \right) \right]$$

By rerunning this model 1000 times, we can derive a confidence interval for the
total spot number ($T$). This interval will be tight for high quality images and will widen
as image quality degrades.

Estimating the variance of random forest and other bagged predictions is still an
open problem, in part because the variance is comprised of both (a) sampling variance
from training on a limited set of data and (b) Monte Carlo effects arising from a finite
amount of bootstrapping (Duan, 2011; Sexton and Laake, 2009; Wager et al., 2013).
Because we can empirically calibrate random forest probabilities in our classification
task, we can take advantage of standard probability theory to construct an interval
estimate.
2.3 Results and Discussions

One key difficulty with evaluating image-based, molecule counting methods is that there is not an independent way to count the number of molecules in the specimen. For smFISH (as well as other techniques) it has been experimentally established (McIsaac et al., 2013; Raj et al., 2008) that, with the exception of transcriptional foci, spots in these images do represent single, fluorescently-labeled, diffraction-limited molecules. We can, however, use artificially generated data to investigate how well our method performs in the face of background noise.

In order to avoid making arbitrary assumptions about the structure of background noise, we used three 3D image stacks from actual specimens without any transcripts as the background. The background therefore consists of both autofluorescence and any diffuse fluorescence due to unbound probes that were not removed by washing. To generate signal, we sprinkled point sources of a specified magnitude throughout a blank image stack of the same size as the background stack, convolved them with a point spread function based on typical microscopy parameters, added the background and signal stacks together, and then blurred them with a Gaussian filter. The spots in these images look very much like actual data.

To test our method, we used artificial images based on one background to construct random forests and evaluated the false positive and false negative rates for the images based on the other two backgrounds. The method performed robustly (Figure 5), with the area under the ROC curve well above 90% for realistic signal intensities and
spot densities (Figure 5). The width of the confidence intervals increased at lower signal
to noise levels, but otherwise was a fairly constant fraction of the total spot count. The
software can reliably distinguish spots that touch, particularly if the local intensity
maxima are separated by at least two intervening pixels. However, when the local
mRNA density is too high, it is even impossible for humans to distinguish individual
spots. Under these circumstances an intensity and regression-based approach to
estimating transcript levels, while noisy, may be the only option (Lubeck and Cai, 2012;
Tan and van Oudenaarden, 2010).

A few unsupervised algorithms have been used to automatically count the
number of spots in smFISH (McIsaac et al., 2013; Mueller et al., 2013; Raj et al., 2010;
Thompson et al., 2002; Trcek et al., 2012; Zenklusen et al., 2008) Two of them (Mueller
et al., 2013; Raj et al., 2010) use a watershed method based on intensity to find a range
of intensities over which the number of connected components in the image is
insensitive to intensity thresholding. This number is taken as an estimate of the spot
number. However, when the expression level is high, spots are often clustered, and out-
of-focus light gives a higher local background that can vary across an image. Under
these common circumstances these methods underestimate the true signal (often because
neighboring spots are lumped together as one) while the method described here
performs consistently well with few or many spots in the image, even if they touch (see
above). FISH-Quant (Mueller et al., 2013) further analyzes the connected components,
but its performance can be very sensitive to user-defined global parameters when the
background signal is high (Figure 6).

Another approach (Thompson et al., 2002; Trcek et al., 2012; Zenklusen et al., 2008) has the primary goal of spot localization and starts by identifying individual candidate spots after intensity thresholding a 2D maximum projection, correcting for local background, and fitting them to 2D Gaussians. It then removes purported duplicates and thresholds a measure of the intensity of the entire spot to distinguish signal from noise. Because our algorithm also starts directly from the local maxima, it also works robustly for images with high or inhomogeneous backgrounds. However, it uses the 3D image, not a maximum projection, and is able to resolve clustered spots. Furthermore, while our local-maxima-centric approach uses a similar method for localization, its primary goal is robust classification without setting semi-arbitrary thresholds. The supervised learning process and the GUI allow the user to manually curate the classification of individual spots, and then feed these corrections back into the classification algorithm. This is particularly useful for low quality images, allowing the user to overrule the algorithm for spots on the boundary between signal and noise.

2.4 Conclusions

As the throughput of microscopy-based single-molecule techniques increases, robust image processing techniques will be ever more crucial. We present a machine-learning-based pipeline for identifying and classifying fluorescently labeled molecules in 3D image stacks that performs well under conditions where other algorithms fail. The software (called Aro Spot Finding Suite after *Arothron hispidus*) includes MATLAB
GUIs to generate the training set and review the classifications and a detailed manual with examples. The ability to infer biological meaning from a quantitative imaging experiment depends upon extracting reliable measurements from images. For single molecule imaging, our software uniquely provides a way to measure this reliability.

2.5 Acknowledgements

Chapter 2, in full, is a reprint of the materials as it appears in Aro: a machine learning approach to identifying single molecules and estimating classification error in fluorescence microscopy images in BMC Bioinformatics 2015 by Allison Chia-Yi Wu and Scott A. Rifkin. Scott A. Rifkin is the corresponding author. The dissertation author was the primary investigator and author of this paper.
Chapter III. Gene Expression Variability and Robustness in the *C. elegans* Intestinal Specification Network

3.1 Introduction

The *C. elegans* intestinal specification network consists of an interlinked set of feed-forward loops consisting pairs of homologous genes that have similar but not entirely redundant functions. Knocking out one of the genes in a pair usually yields a more severe phenotype than knocking out the other. For example, the *end-3* deletion allele ok1448 has 7% penetrance (Bénédetti et al., 1994) while the *end-1* deletion allele ok558 has 0% penetrance (Zhu et al., 1997). *elt-7* and *elt-2* are even more divergent. While *elt-2* deletions are lethal and lead to larvae that die of incomplete development of the gut (Fukushige et al., 1998; Sommermann et al., 2010), a *elt-7* deletion does not lead to any defects in the embryos (Sommermann et al., 2010). Furthermore, a previous study also discovered that the embryos with MED gene deletions develop incomplete intestines (Coroian et al., 2006; Maduro et al.). These studies provide qualitative evidence that suggest that the regulatory roles of these genes are organized hierarchically (Maduro and Rothman, 2002; Maduro et al., 2005, [CSL STYLE ERROR: reference with no printed form.]) but only a few quantitative studies (Nair et al., 2013; Raj et al., 2010) have been done to look at the detailed regulatory mechanisms that provide robustness to this network. Raj et al. (Raj et al., 2010) uncovered the relationship between gene expression variation and phenotypic variation in *skn-1* mutants, suggested a possible window during which *end-1* regulates *elt-2* leading to bimodal *elt-2* expression possibly boosted by the auto-positive feedback loop of *elt-2*. On the other hand, the other quantitative study suggested a that the gene expression is
coordinated with but not dependent on cell division in this network (Nair et al., 2013).

We use single-molecule fluorescence in situ hybridization (smFISH) to investigate the levels and variation of expression profiles of each gene within this network under various perturbations such as RNAi treatment and gene deletions. Aside from the single deletions we mentioned above, we also treat single deletions strains such as end-1 -/- and end-3 -/- with elt-7 RNAi which almost completely knock-down elt-7 expression, creating a condition close to double knock-outs. With end-1 -/-; elt-7 (RNAi) and end-3 -/-; elt-7 (RNAi), we are able to narrow the number of inputs to elt-2 to one and determine the strength of their regulation on elt-2 effectively. We also acquired an end-1,3(MED-) strain that has the MED-binding sites removed from the end-1 and end-3 promoters from the Maduro Lab. This strain suffers greatly from incomplete development of intestines (Maduro et al.) but also effectively leaves single activation input to end-3 from SKN-1.

Quantitative investigation of all these conditions revealed several mechanisms that C. elegans uses to buffer the noise during development and gave us insight into how the intestinal specification decision is made in C. elegans.

3.2 Methods

3.2.1 Single-Molecule Fluorescence in situ Hybridization in C. elegans and Imaging

We followed the single-molecule Fluorescence in situ Hybridization (smFISH) protocol described in Raj, et al (2010) (Raj et al., 2010). The C. elegans embryos were synchronized at the L1 stage and seeded on NGM plates with OP50. 48-52 hours later,
the young adults were collected, fixed with fixation buffer (4% formaldehyde, 1X PBS) and dehydrated in 70% ethanol. After dehydration in 70% ethanol for 6 hours, we hybridized the embryos with fluorescent DNA probes in 10% hybridization buffer (2X SSC, 1% dextran sulfate, 10% formamide) for 12-18 hours. 6 probes were designed and optimized, which include skn-1::Atto488, med-1/2::Cy5, end-3::Atto647N, end-1::TMR, elt-7::Alexa594 and elt-2::Atto647N. Taking advantage of the temporally non-overlapping expression patterns of end-3 and elt-2, we stain both end-3 and elt-2 at the same time in each experiment. After hybridization, embryos were washed in 10% wash buffer (2XSSC, 10% formamide) before imaging.

For imaging fluorescent smFISH spots, we use a Nikon TiE epifluorescence microscope with a Lumen 200 Pro as the fluorescence light source. We took around 30 slices of 0.3-4 μm z-section in 5 different fluorescence channels and 1 transluminescence channel at each position, which usually included 2-4 embryos. In each experiment, images of at least 200 embryos were taken.

3.2.2 Using Aro to analyze smFISH images

For each channel of each experiment, we chose at least 200 good spots and 200 bad spots as the initial training set for training the random forest classifier described in Chapter 2. Due to the heterogeneity in the background signal of C. elegans embryos, we find that picking a training set for each batch and each channel results in the best classification results. The classification results usually reach their best possible results in 3-4 round of training processes, even for experiments with low image quality. The
quality of image is reflected in the precision interval of total number for each embryo.

3.2.3 Deletion strains

The reference wild type strain used in this study was N2. The deletion strains we used were end-3-/-(ok1448) and end-1-/-(ok558) from CGC and end-1(ok558) end-3(ok1448) V; irSi[end-1(MED-)]II; irSi[end-3(MED-)] from Maduro lab described in (Maduro et al.). The end-1(ok558) end-3(ok1448) V; irSi[end-1(MED-)]II; irSi[end-3(MED-)] would be abbreviated as end-1,3(MED-) for simplicity, which has functional end-1 and end-3 with MED binding site removed from their promoters. All these strains were grown on E. coli OP50 and handled according to standard methods.

3.2.4 elt-7 RNAi experiment

We constructed the elt-7 RNAi plasmid by inserting a 350-nt long C. elegans elt-7 coding sequence (TGT GTA CTA CTC CAT TAC AAC CAC TTG AAG ATT CCA GAA TTA TTT TTG ATG AAA GCT TGA CGA AAA ATG AAA ATG AGC AAA AAT CCT TTG TGG AGC AAG ATA GCA GTT ACG AAT CAT CTG GAA ACC GAT TTG GTT CAC AGA AGG GAA AGA AAA TTG CAA AAG TAA TTC GTG ATG CGT GCT GCT CAC ACT GCT CAA CAA CTA CAA CCA CAC TTT GGC GTA AAA ATG ATG AAG GGA ATC TTG AGT GCA ATG CTT GCA ACT TGT ATT ACC GTC ACA ACA AAG TGA AAC GCC CGT TAT CCC TCT GCA AGC AGC CAA CTA CTC GAA AGC GGC GTC AAG CAA AAA AAG AAT A) into L4440 described in (Fire et al., 1998). HT115 bacteria was used to express the elt-7 dsRNA from this plasmid for bacterial feeding. One single colony of the bacterial feeding strain was
picked each time and grown in LB media with ampicillin before being seeded to the NGM with 50 ug/ml Carbenicillin and 1 mM IPTG (RNAi plates). Before seeding the plates with RNAi feeding strains, we grew the RNAi feeding strain to log phase (OD600=0.4) from the overnight culture in LB with ampicillin and 1mM IPTG. The bacteria was then concentrated to 10 fold of the original concentration and seeded onto 10 cm RNAi plates. The worm strains being treated were synchronized at L1 stage and grown on normal NGM plates seeded with OP50 for at least 20 hours. The L2-L3 worms were then transferred to RNAi feeding plates. This step was to ensure that RNAi does not take effect in the development of L1 larvae. The L2-L3 worms were then transferred to RNAi plates seeded with \textit{elt-7} dsRNA expressing bacteria and treated for 36-40 hours. We then collected, fixed and prepared the young adults for single-molecule FISH imaging according to method described in Section 3.2.2.

3.2.5 Nuclei counting and conversion of nuclei number to minutes after fertilization

The developmental timing of each embryo was determined by manually counting the DAPI staining of nuclei in each embryo. The accuracy was expected to be $\pm 1$ nuclei in embryos with fewer than 50 nuclei.

3.2.7 Clustering RNA spots to derive cell level expression

The cell level expression and number of E cell expressing RNA transcripts were roughly predicted by an algorithm written in MATLAB. This algorithm used k-means clustering to cluster the three dimensional locations of all mRNA spots in all the channels together. The number of clusters is optimized to the value that resulted in the
lowest total intra-cluster distance. The centroids of the clusters were observed to be close to the actual locations of nuclei. Each cluster then represents one E cell and the number of mRNA spots in each cluster represents the number of mRNA transcripts that were transcribed in this cell. An expressing E cell was defined as a cluster that has more than 10 spots of a particular gene.

3.3 Results and Discussions

3.3.1 end-1 expression reveals the regulatory relationships that maintain robustness to gene expression variation in wild type.

*end-1* expression at the embryo level plateaus from 100-150 minutes after fertilization (Figure 7), while at the cell level, a strict decrease of *end-1* is observed after entering the 4E stage (Figure 8A). This indicates that *end-1* mRNA degradation becomes dominant after entering 4E stage. In other words, any upstream regulators of *end-1*, such as END-3, SKN-1, and MED-1/2, activates *end-1* at a much lower rate after entering 4E stage. This finding corresponds to a previous observation of the duration of protein expression of these regulators: SKN-1 is last observed in 8-cell stage and is no longer detectable by immunostaining in 12-cell stage (Bowerman et al., 1993). END-3 reaches its maximal level at the end of 1E cell stage and the beginning of 2E stage (Maduro and Rothman, 2002; Maduro et al., 2005). Despite the deletion of MED binding sites on both *end-1* and *end-3* promoters or the deletion of *end-3*, the maximal level of *end-1* remains the same for *end-1-/-, end-3-/- and *end-1-/-;end-3-/-*(MED-), while the variability is increased in deletion strains (Figure 8A), suggesting that even without the inputs from MED or END-3, part of the population is still able to express
the normal level of *end-1* observed in the wild type. The fact that the maximal level in wild type does not equal to the summation of the maximal levels observed in *end-3-/-* and *end-1-/-;end-3-/-*(MED-1) indicates that *elt-2* expression is not a simple linear combination of the activating inputs from each activator. There could be a few possibilities. First, there might be interactions, such as cooperation, among the activators of *elt-2* so that the activation inputs from these activators are not independent from each other. Second, even if each activator could activate *elt-2* independently, there is a threshold for the total number of the proteins of the activators. Or it could be a combination of the two mentioned above. For the first hypothesis to hold, there

3.3.2 Decision of cell fate is made autonomously in individual cells in early embryonic stage.

The expression of *elt-2* at cell level reaches a constant level in 8-E cell stage in wild type and *end-1-/-* while many of embryos fail to express *elt-2* to the same level in *end-3-/-* and *end-1;end-3*(MED-/-) strains. As shown in the upper row of Figure 8B, we also find that the cell-to-cell variation of *elt-2* expression among E cells that are expressing *elt-2* is relatively small compared to other genes in 8-E cell stage in each strain. That is, for each cell that is expressing *elt-2* expresses *elt-2* at similar level, suggesting that whether each cell can produce the maximal level of *elt-2* is determined in individual cells autonomously. The number of E cells expressing *elt-2* (lower row of Figure 8B) further confirms this idea that intestinal specification might be determined in early embryonic stage. This observation combined with the ELT-2::GFP expression pattern in late embryonic stage (Maduro et al.) provides strong evidence that the
intestinal specification in *C. elegans* cannot be described as an “all-or-none” phenomenon at embryo level but is a decision made in individual cells in early embryonic stage.

### 3.3.3 elt-7 may have a functional role in constraining elt-2 expression variability.

Despite having similar regulators as *elt-2*, *elt-7* has a very different expression profile. *elt-7* is turned on at the same time as *end-1* in wild type animals, which indicates that END-3 plays a more important role in turning on *elt-7* than END-1. The significant delay of *elt-7* activation in *end-3-/-* at the end of 1-E stage (Figure 9) and no delay of *elt-7* activation in *end-1-/-* at the same stage (Figure 9) further shows the dominant role of END-3 in turning on *elt-7* at the right timing from the end of 1E stage. The variability of *elt-7* does not have significant changes in either *end-3-/-* or *end-1-/-* in any developmental stage, suggesting that even though END-3 and END-1 can regulate the *elt-7* expression level, they might not be able to constrain the variability of *elt-7* (Figure 9).

Compared to *elt-2* and *end-1* expression in wild type, *elt-7* expression is more variable after reaching maximal level (Figure 10). This susceptibility to noise might suggest a redundant functional role of *elt-7* but the difference between phenotypes of *elt-2(0)* and *elt-2-/-;elt-7-/-* suggest a cryptic functional role of *elt-7* (*Sommermann et al., 2010*): The *elt-2(0)* animals hatch with an almost fully developed but obstructed intestine and all of the mutant animals die of obstructed guts, whereas the animals with the *elt-2-/-;elt-7-/-* genotype died of having only dispersed gut cells. In *end-3-/-*, we
observed a significantly lower mean expression level of elt-2 with higher variability. While a sight delay in elt-2 expression in \textit{end-1/-} can be observed in beginning of the 4E stage, the mean level and coefficient of variation are not significantly changed in \textit{end-1/-} (Figure 11). We then examined the elt-2 expression in \textit{end-1/-;elt-7(RNAi)} and \textit{end-3/-;elt-7(RNAi)}. The RNAi treatment almost completely knock-down the elt-7 expression entirely (Data not shown). In \textit{end-3/-;elt-7(RNAi)}, elt-2 expresses with the lower expression level and higher variability than in \textit{end-3/-} (Figure 12) but we do not observe any difference in the survival rate compared to \textit{end-3/-}. In \textit{end-1/-;elt-7(RNAi)} worms, we observed that the mean expression level of elt-2 is significantly lower than wild type and the variability of elt-2 is significantly increased. Even though we observed that all the embryos hatch and some embryos are able to get to the maximal level of elt-2 expression eventually at the later stage, according to JH Rothman, none of these worms survive after L1 stage. This might suggest that only having elt-2 reaching the maximal level at the end of 8E stage might not be enough to determine the fate of intestinal development.

One other interesting finding from comparing these combinations of deletion and RNAi conditions is that elt-7 seems to be able to constrain the elt-2 variability. The loss of elt-7 can significantly increases the coefficient of variation in the gene expression of elt-2 in wild type, \textit{end-3/-} and \textit{end-1/-} in 8E cell stage (Figure 13). This provides yet another evidence that elt-7 is not just a regulator with redundant functions but plays an important role in constraining elt-2 especially under perturbations such as \textit{end-3/-} or
end-1-/-.

3.3.4 Low expression of elt-7 leads to delayed expression of elt-2.

In Section 3.3.3, we mentioned that it is observed that elt-2 has a lower mean expression level in end-1-/-;elt-7(RNAi) but part of the population was still able to reach the maximal level at the end of 8E stage. It is a very surprising finding considering in end-1-/-;elt-7(RNAi) situation, the only activator of elt-2 left in this system is end-3. end-3 has a very short and transient expression that ends before 2E stage. Even though we do not have quantitative experimental data to show when END-3 disappears from the system, it is still surprising that END-3 alone is enough to activate elt-2 to its maximal level.

This delayed response of elt-2 is also observed in end-1-/- experiments that have noisy end-3 expression (Figure 14B). end-1-/- has been reported to have no significant phenotype in intestinal development. Therefore, the expression of the genes in this network was expected to be similar to wild type. While this expectation is observed in end-3, elt-7 and elt-2 expression in some experiments (Figure 14A), we notice a delay of elt-2 expression in part of the population in some of the experiments (Figure 14B). This delayed expression of elt-2 is only observed in experiments that have variable expression of end-3. These experiments with variable end-3 were done with end-1-/- worms that have lived at room temperature for at least 10 generations after being thawed. Whether the variable end-3 is a result of accumulated end-1-/- mutation effect over generations still needs to be further confirmed but this observation provides another
layer of information of how \textit{elt}-7 and \textit{end}-3 regulate \textit{elt}-2 and how \textit{elt}-2 is transcribed.

Due to the fixation step of single-molecule FISH, we are unable to trace back the \textit{end}-3 expression in an embryo where \textit{elt}-2 expression is observed but we can correlate the \textit{elt}-7 expression with \textit{elt}-2 in each embryo. Surprisingly, the embryos with delayed \textit{elt}-2 expression also has lower expression of \textit{elt}-7, which together with the data from \textit{end-1-/-;elt-7(RNAi)} leads to an interesting possibility that even though \textit{end}-3 alone might be able to drive \textit{elt}-2 expression to its maximal level, it might not be sufficient to activate \textit{elt}-2 at the same rate and \textit{elt}-7 expression is necessary for \textit{elt}-2 to be expressed at the normal rate. Whether this change of \textit{elt}-2 activation rate leads to defect phenotypes is the next question we would like to ask in next the chapter using mathematical modeling method.

3.3.5 Saturation of expression is observed in \textit{end-1} and \textit{elt}-2 in wild type.

Comparison of the expression of \textit{end}-1 and \textit{elt}-2 among \textit{end-3-/-}, \textit{end-/-} and \textit{end-1-/-;end-3-/-(MED-)} with wild type often shows an interesting pattern recurring in this network: while the maximal levels in wild type can mostly still be reached in part of the population in \textit{end-3-/-} or \textit{end-1-/-}, the gene expression variation is increased in these knock-outs (Figure 15).

These expression patterns suggest that the expression of these genes often reaches maximal level in the majority of the population of wild type animals but perturbations of the network, such as loss of one of the activators, might lead to expression below the maximal level. This might be caused by several possibilities. One
possibility is that the number of activating proteins might often outnumber the number of binding sites available on the promoters of these genes so the expression of these genes in wild type is in a saturated state. By being in a saturated state, the gene can have more activating proteins in the system than it actually needs to reach maximal level and still maintain a consistent maximal level. On the other hand, in this kind of system, removing one or a few of its upstream regulators from the system might not lead to the unsaturated state, thus providing robustness to loss of regulators. Nevertheless, when this kind of system falls into an unsaturated state, the target gene being regulated becomes directly correlated with the number of its regulators. If the number of regulators happens to be variable, its target gene expression would become variable too. This susceptibility to gene expression variation allows the variation to be propagated or amplified to the downstream gene.

In our system, the expression of end-1 and elt-2 seem to fit the pattern of this kind of system. A few observation that leads to this hypothesis is: First, the maximal level of these three genes in end-3-/-, end-1/- and wild type stays the same, which means losing one of the regulators does not necessarily lead to an unsaturated state of expression in the each embryo and having all of the regulators saturates the expression. Second, the randomness seems to not only be propagated but also be amplified through the hierarchy of the network, which leads to more penetrant phenotypes in end-3/- and end-1;end-3(MED-) than end-1/- or elt-7/-.

From these observations, we propose a hypothesis that each of these genes has a threshold for the total number of their
activators and this threshold is often securely reached in wild type animals.

3.4 Conclusions

In this chapter, we examined the gene expression in many different conditions (Figure 1) with a highly quantitative method, single-molecule FISH, which allows us to examine these expressions at a resolution that no one has done before. We directly observed the saturation of expression of end-1, elt-7 and elt-2. This observation leads to a hypothesis that can be further tested with mathematical modeling in Chapter 4. With elt-7 RNAi, we discovered possibilities for the functional roles of elt-7: First, instead of being a redundant regulator of elt-2, ELT-7 does contribute to constraining elt-2 gene expression variation. Second, even though END-3 alone is enough to activate elt-2 to its maximal level, without ELT-7, it cannot activate elt-2 at the same rate and whether this leads to the phenotype of dead L1 larva in end-1-/-;elt-7-/- would be the next question that we should ask.

3.5 Acknowledgements

Chapter 3 and 4 together are being prepared for submission for publication of the material and may appear as Allison Chia-Yi Wu, Morris Maduro, Scott A. Rifkin. “The response of elt-2 in C. elegans intestinal specification network reveals the mechanism for buffering the gene expression noise.” The dissertation author was the primary investigator and author of this paper. Morris Maduro has provided the deletion strains needed for part of the experimental design. Scott Rifkin was the corresponding author.
Chapter IV. A Computational Model of *C. elegans* Intestinal Specification Network

4.1 Introduction

The major disadvantage of single-molecule FISH is that it requires fixing the samples. This step makes live-imaging impossible using this method, which also contributes to the difficulty of interpretation of a time-course data. In our worm system, we use nuclei number as a way to determine the developmental stage of the embryo and reconstruct the data through this pseudo-time line. However, a pseudo time line can never equate with a real time line. There could be many different ways of connecting the dots on the pseudo time line. To better interpret the data we derived from this method, we develop a computational model that allows us to test possible hypotheses behind this data with pseudo time lines and to help answering the question of how gene expression variability is translated into phenotypic heterogeneity.

Previous studies in the field of theoretical and synthetic biology have probed into the potential of regulatory networks in regulating the gene expression variability theoretically and analytically. Negative feedback loops can suppress the gene expression variability (Austin et al., 2006; Becskei and Serrano, 2000) while positive feedback loops can lead to multimodal or “switch-like” behaviors (Acar et al., 2005; Ferrell, 2008; Singh and Weinberger, 2009; Tsai et al., 2008) in synthetic endogenous systems. In contrast to the simplicity of the synthetic circuits, the endogenous circuits are usually embedded in densely interlocked networks with multiple positive feedback loops and negative loops. The effect of such interlocked feedback loops on gene expression variability has been previously described in *C. elegans* Wnt signaling.
In the *skn-1* network, there are only two positive auto-feedback loops for *elt-2* and *elt-7* but the expression of genes in this network shows robustness in wild type strains and only shows variability in mutant strains (Raj et al., 2010). By combining our single-molecule FISH data with modeling, we identified 2 special features of this network that maintain the robustness of gene expression with only two auto-positive feed back loops. First, the limited number of binding sites on the promoters of these genes can be easily saturated by the number of GATA factors present in the cells in the wild type. Therefore, the gene expression can be robust to gene expression variation inherent in the mRNA transcription. Second, the transcription rate of the major output of this network, *elt-2*, is comparatively slow to other genes and a slow transcription rate allows the gene expression to be robust to the variation of the abundance of its regulators. With the help of this model, we are also able to discover the mechanism that underlies the correlation of gene expression variation and phenotypic variation.

### 4.2 Methods

#### 4.2.1 A mathematical model of genetic network with thresholds for multiple activators

This computational model is a deterministic model using hill functions to describe the chemical reactions in this system. This model consists of 12 components, simulating both mRNA and protein dynamics of all the 6 genes, *skn-1*, *med-1/2*, *end-3*, *end-1*, *elt-7* and *elt-2*, in the network. We assume the transcription rate and translation
rate are mostly the same for all the genes except for a few exceptions, such as *elt-2* and *skn-1*. We have experimental data for both *skn-1* mRNA by single-molecule FISH and SKN-1 protein by immunostaining. All the proteins produced in this system are GATA factors and some of them share similar protein structures. There is no evidence for us to suspect that these GATA factors might have different protein synthesis or degradation rate. Therefore, we simplified the model by assuming the translation, transcription rates, mRNA degradation rates and protein degradation rates are the same.

We have discussed the possibility that the expression of *elt-2*, *end-1* and *elt-7* can reach a saturated state and what the possible hypotheses are behind this phenomenon in chapter 3.3.4. We have discussed that the more likely hypothesis behind the current experimental data is that for *end-1* and *elt-2* the inputs from their different GATA activators add to each other linearly with a threshold on the total number of proteins. These thresholds are the terms 'T' in the Equation (4) and Equation (6) for the transcription of *end-1* and *elt-2*. These thresholds are determined by comparing the summation of total protein numbers of the activators in the simulation and we start with a threshold that 80%-90% of simulations for wild type can pass at maximal level and less than 10% of the simulations for mutants can pass at maximal level. For example, for *end-1*, in the wild type, about 90% of the total number of END-3 and MED-1/2 pass the threshold of 100 but only less than 10% in *end-3/-, end-1/-* and *end-1/-;end-3/- (MED-)* can pass this threshold at maximal level.
\begin{align*}
(1) \frac{dx_1}{dt} &= -kd_1 x_1 \\
(2) \frac{dx_2}{dt} &= ks_2 \frac{Ka_{7,7} x_7}{(1 + Ka_{7,7} x_7)} - kd_2 x_2 \\
(3) \frac{dx_3}{dt} &= ks_3 \frac{Ka_{7,7} x_7}{(1 + Ka_{7,7} x_7)} + \frac{Kd_8 x_8}{(1 + Ka_{8,3} x_8)} - kd_3 x_3 \\
(4) \frac{dx_4}{dt} &= ks_4 \frac{Ka_{7,4} x_7}{(1 + Ka_{7,4} x_7)} + \frac{Kd_{9,4} min((x_8 + x_9), T_1)}{(1 + Ka_{9,4} min((x_8 + x_9), T_1))} - kd_4 x_4 \\
(5) \frac{dx_5}{dt} &= ks_5 \frac{Ka_{9,5} x_9}{(1 + Ka_{9,5} x_9)} + \frac{Ka_{10,5} x_{10}}{(1 + Ka_{10,5} x_{10})} + \frac{Ka_{11,5} x_{11}}{(1 + Ka_{11,5} x_{11})} - kd_5 x_5 \\
(6) \frac{dx_6}{dt} &= ks_6 \frac{Ka_{12,6} x_6}{(1 + Ka_{12,6} x_{12})} + \frac{Kd_{11,4} min((x_8 + x_9 + x_{11}), T_2)}{(1 + Ka_{11,4} min((x_8 + x_9 + x_{11}), T_2))} - kd_6 x_6 \\
(7) \frac{dx_7}{dt} &= ks_7 x_1 (t - d > 0) - kd_7 x_7 \\
(8) \frac{dx_i}{dt} &= ks_i x_{i-6} - kd_i x_i (i \geq 8)
\end{align*}

\begin{itemize}
  \item $x_1$: skn-1, $x_2$: med-1/2, $x_3$: end-3, $x_4$: end-1, $x_5$: elt-7, $x_6$: elt-2
  \item $x_7$: SKN-1 protein, $x_8$: MED-1/2, $x_9$: END-3, $x_{10}$: END-3, $x_{11}$: ELT-7, $x_{12}$: ELT-2
\end{itemize}

T1: activation threshold for end-1

T2: activation threshold for elt-2

d: the start time of 1E stage. (skn-1 expression is repressed in 0E stage)

4.2.2 Adding randomness to the model and reconstructing the gene expression at embryo level

To construct a model to study gene expression variation, one would assume that a model with stochastic differential equations should be used to capture the stochastic nature of gene expression (Paulsson, 2004). However, our focus of our model was not
on the stochastic nature of the mRNA transcription process but on how the gene expression variation can be propagated through this system with thresholds for activator proteins and how this gene expression variation leads to phenotypic variation. Therefore, we did not resort to a real stochastic model but instead, we artificially added randomness to a deterministic ODE model keep this model simple. Our assumption is that as long as we can recapitulate the mean level and variability of gene expression in experimental data from all different mutants and wild type, this model would be sufficient for our purpose to study how the gene expression variation affects phenotypic variation.

In this system, we can observe a basic level of variation of gene expression for all the genes in wild type even without any perturbations (Figure 7), which means that this system is inherently noisy. Where does this variation come from and how should we introduce it into the model? There could be several sources of this variation, such as environmental changes, intrinsic noise in the transcription processes or stochasticity during cell divisions. All our data in this study was done in a controlled environment so we could assume that the noise introduced by environmental changes should be minimized. This leaves us with intrinsic noise in the transcription processes or stochasticity during cell divisions.

mRNA transcription has been described to be stochastic in previous studies (Elowitz et al., 2002; Kærn et al., 2005; Raser and O’Shea, 2004b; Swain et al., 2002) and there is no reason for us to think that our genes in this network can be exempt from
this phenomenon. Therefore, we choose to artificially add the randomness into the transcription rate by adding a random multiplier from a uniform distribution between 0 to 1 on mRNA transcription rate of each gene.

As for stochasticity during cell divisions, we take the timing of each E cell division into account. At each E cell division, all the proteins and mRNA molecules in the system go through a bernoulli process that each molecule has half of the probability of going into one of the daughter cells. By following the cell division processes, we are also able to track the expression in mother and daughter cells and eventually able to reconstruct the whole embryo expression level.

4.2.3 Fitting parameters to gene expression variability and gene expression level at cell level

There are 29 parameters that need to be fitted to experimental data in our model. One might think there are too many parameters to fit at once. This is where we take advantage of the simplicity of this network: no feedback loops until the last two genes. This means that changes in the downstream genes does not affect its upstream genes and thus we are able to fit the parameters for upstream gene first. For example, we focus on fitting skn-1 to experimental data first without worrying that the changes in elt-2 would affect the expression of skn-1.

Another advantage we had when fitting our parameter was the availability of experimental data in different mutants. In mutants like end-1/-;end-3/- (MED-), we can narrow the inputs to end-3 to only skn-1 and this allows us to fit the parameters between end-3 and skn-1 independently from any parameters needed between med-1/2 and end-3.
To make sure our model can recapitulate both the gene expression level and variability, we define the objective function as the proportion of the overlapping region of the 95% inter-percentile range between 2.5% to 97.5% of the trajectories of the gene expressions of 200 simulated embryos with the same 95% inter-percentile range of the experimental data. This objective function describes not only how much the simulated results recapitulate the gene expression level but also the variability of gene expression. The model parameters are derived through maximizing the objective function through constrained nonlinear optimization implemented in MATLAB. The parameters were fitted to gene expression at cell level first and then we reconstruct the embryo level through combining the cell level data by following the cell division processes.

4.3 Results and Discussions

4.3.1 The model is able to recapitulate the gene expression variability.

Our objective function calculates how much the gene expression variability is captured in the model in a fast but rough way to speed up the fitting process. After getting the optimized parameters, we further tested whether the simulated results really recapitulate the mean level and variability by bootstrapping the 200 simulated results and the experimental data.

In Figure 16, we find that the model recapitulates the wild type expression level and variability very well. The mean level is not significantly different from the experimental data and the coefficient of variation is also not significantly different from experimental data for 2E and 4E stage but not for 8E stage. In *end-1*, the model is able
to capture the variability but would not be able to capture the level from mid-4E to mid-8E stage. This inconsistency might be because the Savitzky-Golay filter we used to smooth the mean level gives more lighter weight to data points that are more distant from the majority of the data. In *end-3/-*, even though the statistics shows that the model is significantly different from experimental data throughout the time *elt-2* is expressed, the model is actually just slightly shifted earlier than the experimental data. The variability of *elt-2* in *end-3/-* is also recapitulated in the model.

These differences between the model and experimental data might be hard to be captured by our objective function which only calculates the overlapping area between model and experimental data. In the future, we might want to use objective functions that do calculate the coefficient of variation to better describe the features of simulated data. However, if we do similar hypothesis testing by bootstrapping this way for each parameter set, it might significantly slow down the speed of parameter fitting. The objective function we used seemed to be able to capture most of the characteristics we want in a fast way.

**4.3.2 Slow *elt-2* transcription rate is necessary for the model to recapitulate the correlation between *elt-7* and *elt-2* expression.**

The synthesis and degradation rates of mRNA were originally assumed to be the same for all the genes to simply the model as mentioned in Chapter 4.3.1. However, this assumption is immediately challenged with the observation that *elt-2* only starts being expressed at mid-2E stage (100 minutes). With the same transcription rate, *elt-2* should be expressed at around the same timing as *elt-7* since they have the same two major
inputs, END-1 and END-3. While \textit{elt-7} is expressed around the same time as \textit{end-1}, \textit{elt-2} is not expressed until later. The two most intuitive speculation is: 1) \textit{elt-2} is repressed until the end of 2E cell stage. 2) \textit{elt-2} is not repressed but has a particularly slow transcription rate compared to the rest of the other genes.

We tested both of the hypotheses and both could achieve similar objective scores (0.741 for hypothesis 1 and 0.752 for hypothesis 2). That is, both could recapitulate the expression level and variability of \textit{elt-2}. However, if we consider the correlation of \textit{elt-2} an \textit{elt-7} in \textit{end-1/-} (Figure 14C), we immediately find that only one of the hypotheses fits the experimental data (Figure 16).

A low transcription rate may allow the gene to integrate noisy inputs from its upstream regulators. This feature would be advantageous for a gene that is important for decision making. We suspect that this might be the case for \textit{elt-2}. This low transcription rate might allow \textit{elt-2} to integrate inputs from multiple regulators, making \textit{elt-2} more robust to transcriptional noise propagated from upstream.

\textbf{4.3.2 The timing of ELT-2 becoming its own major regulator may determine the cell fate.}

In the wild type animals, \textit{elt-2} has abundant activating proteins in the system, such as \textit{end-3}, \textit{end-1} and \textit{elt-7}, throughout the E cell lineage. Both END-1 and ELT-7 last to 8E stage. END-1 degrades after entering 8E stage while ELT-7 lasts longer because of its auto-positive feedback loop. \textit{elt-2} has its own auto-positive feedback loop but when does this feedback loop become the major input activating \textit{elt-2}? Does the
timing of this 'switching point' matter? Does the ELT-2 level at the 'switching point' make a difference? Theoretically, one would imagine that \textit{elt-2} will need to have ELT-2 to be at certain level at this switching point to be able to keep activating \textit{elt-2} throughout the intestinal development since after 8-E stage, most of the regulators of \textit{elt-2}, END-1 and END-3, will no longer exist and then \textit{elt-2} will have to rely on its own positive feedback loop and ELT-7.

If we define the influence factor of \( n \rightarrow k \) (\( n \) as the activator and \( k \) as the target gene), as the part of hill function that describes the activation from \( n \rightarrow k \) (Equation 9), we can determine how much influence ELT-2 has on its own gene compared to all its other regulators at each time point. The 'switching point' by definition is the time that ELT-2 becomes its own major regulator and therefore would be the time point when the summation of influence factor from END-1, END-3 and ELT-7 starts to become smaller than the influence factor from ELT-2 itself. While it is not possible to examine this idea with experimental method, we examined the simulation results to find when or whether this switching point occurs during the embryonic stage we study and if it occurs, what the level of ELT-2 is.

\[
(9) \text{IF}_{n\rightarrow k} = \frac{(Ka_{n,k}x_n)^{h}}{(1+(Ka_{n,k}x_n)^{h})}
\]

In Figure 17, we plotted the ELT-2 levels at the switching point in the first E cell in each embryo and the color represents the ELT-2 level at the end of the 8E stage. We can find that in most of the embryos in wild type, the switching point does not occur before entering 8E stage and the chances of switching point to occur before entering 8E
stage increase in the mutants that are experimentally verified to have higher penetrance. We further examined all the eight E cells of each embryo and look for cells that have switching point occurring after entering the 8-E cell stage. If we define these cells as cells that survived and developed into gut cells and count the number of surviving cells at the end of 8E stage (Figure 18), we can find that the proportion of embryos that have 8 surviving E cells for each strain. 96% of the wild type simulated embryos have 8 expressing E cells, 89% for end-1-/-, 91.5% for end-3-/- and 79% for end-1-/-;end-3-/- (MED-). While the end-3-/- and end-1-/-;end-3-/- (MED-) have survival rates that are roughly correlated to experimental data (92% for end-3-/- and 75% for end-1-/-;end-3-/- (MED-)). We do notice there is an inconsistency with wild type and end-1-/-, the model is fitted to both data with delayed elt-2 expression and the data with wild-type-like expression. We never examined the phenotype of these embryos that lived in 25C for at least 10 generations. Whether this higher penetrance in end-1-/- reflects a possible higher penetrance would be an interesting next question to ask in the future.

This discovery from simulation provides an hypothesis of the timing of cell fate decision. In the early embryonic stage we examined, all the elt-2 upstream regulators secure elt-2 expression to the end of the 8E stage. These multiple regulators with overlapping transcription trajectories protect elt-2 from transcriptional noise and prevents ELT-2 from becoming its own major regulator before it reaches the expression level that can consistently turn on elt-2.
4.4 Conclusions

Many previous studies have discussed different network motifs that allow the biological system to be robust to intrinsic or extrinsic noise (Brandman and Meyer, 2008; Ji et al., 2013; Justman et al., 2009; Tsai et al., 2008). The intestinal specification network does not have complicated interlinked feedback loops as discussed in Wnt pathway in vulva development by Ji et al. (Ji et al., 2013) or in cell cycle oscillators (Tsai et al., 2008). This network does not have many motifs that could potentially make the network robust to gene expression noise, such as interlinked positive and negative feedback loops. One might be able to argue that the two positive feedback loops on elt-7 and elt-2 might lead them into a bistable state and become robust to expression noise. However, we did not observe bistable states in either elt-7 or elt-2 expression (Figure 7). How can such a network still be robust to intrinsic noise?

With such a network structure, two parallel and interlinked feed-forward loops, we argue that maybe the robustness to transcriptional noise in this network is not maintained through certain network motif but through mechanisms we proposed in Chapter 3 and Chapter 4.

4.4.1 Robustness to transcriptional noise can be maintained with simple genetic feed-forward loops with activation thresholds.

In Chapter 3, we discussed that there might be a threshold for the total number of proteins of the regulators for end-1 and elt-2 and these thresholds are saturated by the abundance of activating proteins in wild type. We further implemented it in our model to show that this hypotheses could indeed possibly lead to the expression profiles we
This threshold allows *elt-2* and *end-1* to maintain a consistent level even when there are more activating proteins in the system than they actually need to reach the level. There will never be 'too much' activators in the system. Therefore, multiple regulators ensure that *elt-2* would reach maximal level and even when some of the regulators are taken out of the system, part of the population might still be able to reach the saturated level because the number of proteins in the system is still above the threshold.

What could lead to this kind of system? One simple possibility is that there are only limited binding sites available for activator proteins to bind on these promoters. Therefore, after all the binding sites are filled by these proteins, no more activation input can come in and thus the gene expression reaches a saturated state. In our case of *elt-2*, all of its regulators, *end-1*, *end-3* and *elt-7* are GATA factors. These proteins are predicted to bind to the same or very similar binding sties so it would not be surprising that these proteins would end up competing with each other for the same binding sites. It would also not be surprising that in wild type the binding sites are all filled with these GATA factor proteins.

Even though this system might be robust to intrinsic noise that would commonly be encountered by the wild type, this kind of system might be vulnerable to perturbations that take out upstream genes of the network. Taking out genes like *med-1/2* or *end-3* that are more upstream makes the expression of *end-1* in an unsaturated
state. When it is in an unsaturated state, the expression reflects the gene expression variation of its upstream regulators and thus the noise is propagated to the next gene. This also provides an explanation to why the more upstream the perturbation happens, the more severe the phenotype is.

4.4.2 *elt-2* may maintain robustness to transcriptional noise through low transcription rate.

A low transcription rate allows the gene to have more time to integrate noisy inputs over a period of time while a high transcription rate could reflect the variation of the inputs more faithfully. One could imagine a high transcription rate might be beneficial when the mRNA product is needed very soon or when it is more important for the product to reflect the level of inputs more faithfully. However, in the cases for genes that have important decision making function, like *elt-2*, it might be more beneficial to make sure *elt-2* can reach a level that can secure the success of intestinal specification.

We did not measure *elt-2* mRNA transcription rate directly and there is no study that measured the mRNA transcription rate of any of the genes in this network. However, our model predicts that a low mRNA transcription rate of *elt-2* could lead to expression profiles that we observed. A slow transcription rate is required to reproduce the *elt-2-elt-7* correlation pattern we see. In the cases when *elt-7* is low, the major activator for *elt-2* is END-3. A slow transcription of *elt-2* allows *elt-2* to integrate END-3 over a longer period of time and eventually integrate enough inputs from END-3 and increase expression to eventually activate itself through its own positive feedback loop. This provides one possible explanation of how *elt-2* can be expressed much later than...
most of its regulators.

4.5 Acknowledgements

Chapter 3 and 4 together are being prepared for submission for publication of the material and may appear as Allison Chia-Yi Wu, Morris Maduro, Scott A. Rifkin. “The response of elt-2 in *C. elegans* intestinal specification network reveals the mechanism for buffering the gene expression noise.” The dissertation author was the primary investigator and author of this paper. Morris Maduro has provided the deletion strains needed for part of the experimental design. Scott Rifkin was the corresponding author.
Chapter V. Evolution of the Intestinal Specification Network: 
C. remanei, C. briggsae and C. elegans.

5.1 Introduction

Previous studies and chapters have shown that the most downstream gene in the 
*C. elegans* *skn-1* network, *elt-2*, determines the onset of gut development, while the upstream *elt-7, end-3, end-1,* and *med-1/2* constrain *elt-2* expression variability. In both 
*C. briggsae* and *C. remanei*, there has been several gene copy number changes in this 
network: there are 5 functionally similar *med* paralogs in *C. remanei*, 4 *med* paralogs in 
*C. briggsae* and two orthologs of *end-3* in *C. briggsae*. These potentially functionally 
similar homologs may add extra connections to this intestinal specification network. We 
wondered how and why these redundant genes are preserved, whether these seemingly 
redundant connections are really redundant, and whether these connections add to or 
strengthen specific dynamical network properties, such as buffering of noise or 
environmental variation or feedback strength.

In this study, we used single-molecule fluorescence *in situ* hybridization 
(smFISH) to measure the expression patterns of genes, including *end-3* orthologs, *end-1, elt-7* and *elt-2*, in the intestinal specification network in these three species, *C. remanei, C. briggsae,* and *C. elegans*. Through quantitative measurement and 
mathematical modeling in *C. elegans*, we have learned how the same network structure 
could buffer against the internal gene expression variation. We assume that similar 
regulatory interactions exist in these other species. For example, we assume that most 
of the *med* orthologs activate *end-3* and *end-1* in *C. briggsae* and *C. remanei* and both of 
the *end-3* orthologs activate *end-1* and *elt-2* in *C. briggsae*. This assumption comes
from the functional studies in which researchers found that some of these orthologs are able to rescue *C. elegans* strains with its orthologous genes deleted (Coroian et al., 2006). With this assumption in mind, we could apply our knowledge from *C. elegans* to these two other species to infer possible mechanisms behind the expression profiles. We performed the single-molecule FISH in *C. briggsae* and *C. remanei* to investigate the gene expression variability in intestinal development. Our investigations into the dynamical roles these orthologs play in intestinal development in different nematode species reveal the effects of gene duplication on dynamical network properties.

5.2 Methods

5.2.1 *Caenorhabditis* strains used and maintenance

The *C. briggsae* strain used is AF16 from CGC and there are two *C. remanei* strains used in this study: SB146 from CGC and JU1184 (a gift from Marie-Anne Felix). JU1184 is a variant of PB4641 made to be susceptible to RNAi by feeding.

5.2.2 Single Molecule Fluorescence in situ Hybridization in *C. remanei*

We developed probes for Cre-end-3, Cre-end-1, Cre-elt-7 and Cre-elt-2 using sequences from WormBase. We coupled Cy5 to Cre-end-3 probes, TMR to Cre-elt-1, A594 to Cre-elt-7 and Cy5 to Cre-elt-2. The single-molecule FISH protocol was mostly similar to the protocol for *C. elegans* described in Section 3.2.1 except for a minor changes: The *C. remanei* strains tend to lay eggs in slightly earlier stage so we collected and fixed embryos 2-4 hours earlier than we collected *C. elegans,*
5.2.3 Single Molecule Fluorescence in situ Hybridization in *C. briggsae*

*C. briggsae* has very similar life history compared to *C. elegans* so there was no modification needed for the single-molecule FISH protocol. The sequences of *Cbr-end-3.1* and *Cbr-end-3.2* are very similar. The two probe sets share 7 common probes together but they have 14 unique probe for *Cbr-end-3.1* and 9 unique probes for *Cbr-end-3.2*. We had optimized these two probe sets to be able to distinguish these two genes in the same experiment by confirming that the 7 common probes do not form any visible mRNA spots and with the unique probe sets, we are able to visualize distinct spots of *Cbr-end-3.1* and *Cbr-end-3.2* (Figure 19). Five probes are made and optimized for *C. briggsae*: *Cbr-elt-2::cy5*, *Cbr-elt-7::a594*, *Cbr-end-1::Atto488*, *Cbr-end-3.1::TMR* and *Cbr-end-3.2::Cy5*.

5.3 Results

5.3.1 The endoderm specification and intestinal specification in *C. briggsae*

The two homologs of *end-3* in *C. briggsae*, *Cbr-end-3.1* and *Cbr-end-3.2*, can be distinguished and measured in *C. briggsae* (Figure 19). While the expression of *Cbr-end-3.2* is only half of the amount of the expression of *Ce-end-3*, the expression of *Cbr-end-3.1* is comparatively low (Figure 20A). In a previous study by Lin et al. (Lin et al., 2009), they discovered that the expression of *Cbr-end-3.1/3.2* disappeared entirely in *Cbr-pop-1* RNAi condition, which effectively turns the E cells into MS cells, indicating that both POP-1 and SKN-1 are required for the determination of E cell lineage. In *C. elegans*, embryos with *Cel-pop-1* RNAi, the MS cells are turned into E cells when POP-
1 no longer exists to suppress endoderm specification genes in MS cells, showing that only SKN-1 is required for a cell to determined as a E cell. Therefore, they proposed an “AND” gate between SKN-1 and POP-1 in *C. briggsae* but an “OR” gate is proposed in *C. elegans*. With our quantitative data, we revisited the theory of “AND” gate in *C. briggsae*. The total expression of *Cbr-end-3.1* and *Cbr-end-3.2* is only half of the amount of *Cel-end-3*. If our model of activation threshold in *C. elegans* also holds true for *Cbr-end-3*, a lower expression of *end-3* might only allow it to barely pass the threshold. Therefore, *Cbr-end-3* might not be able to afford to lose inputs from either POP-1 or SKN-1, making it more vulnerable to loss or decrease of upstream inputs, for example, the case in *Cbr-pop-1*(RNAi). Lin et al. (Lin et al., 2009) proposed that the most likely explanation of this phenotype differences is the change in how the combinatorial inputs from SKN-1 and POP-1 is integrated through the *cis*-regulatory element on the *end-3* promoter. In our study, the expression of *end-3* is no longer just all-or-none but specific counts of individual mRNA transcripts, which allows us to propose an alternative hypothesis: There might not be a logic change for integrating the inputs from POP-1 and SKN-1 after all. What is changed in *C. briggsae* might just be the activation strength from POP-1 and SKN-1. The activation from both of them together is weaker in *C. briggsae* than in *C. elegans*, leading to the lower expression of *end-3*. Therefore, the “AND” gate observed might possibly be the result of this lower expression of *Cbr-end-3* and thus *Cbr-end-3* being susceptible to losing upstream inputs. One might argue that Lin et al. did not observe any expression of *Cbr-end-3.1/Cbr-end-
3.2 in both Cbr-skn-1(RNAi) and Cbr-pop-1(RNAi) with RNA FISH. Our single-molecule FISH method is more sensitive than traditional FISH so if our hypothesis holds true, it is possible that we could still detect very low level of expression of end-3 in both cases. This would be a case how this approach of single-molecule FISH could be more powerful but more experiments with Cbr-pop-1 RNAi will need to be done to further confirmed this hypothesis.

Interestingly, Cbr-end-1 is expressed at the same time as Cel-end-1 but has an approximately 1.5 fold expression level at the maximal plateau compared to Cel-end-1. However, it decreases faster than Cel-end-1 and drops down to almost 0 during 100-150-nuclei stage (Figure 20B). If our assumption of regulatory interactions holds, that means Cbr-end-1 has only Cbr-end-3.1, Cbr-end-3.2, Cbr-skn-1 and the med orgholgs as its upstream inputs. In C. elegans, SKN-1 proteins disappear before 2E stage, med-1/2 expresses mainly in the beginning of 1E stage and the expression of end-3 spans from mid-1E stage to mid-2E stage. Even though we do not have experimental data for the existence of END-3 and MED-1/2, we know that MED-1/2 expression can not last much longer than the mRNA expression of end-3. Therefore, the activation from MED-1/2 or SKN-1 only exists in 1E and maybe the beginning of 2E stage. After entering 2E stage, the main activation input for end-1 would come from END-3 and END-3 is the only regulator that sustains the expression of end-1 till 150 nuclei stage. Therefore, a higher but shorter expression of end-1 might be a result of higher activation in 1E or 2E stage but less activation in later 4E stage. Both lower expression of end-3 and faster
degradation of END-3 could lead to a shorter end-1 expression. There are two more med orthologs in the C. briggsae genome. How much they might add to activation of Cbr-end-1 is unknown but it is possible that together with SKN-1 and END-3, MED can activate higher expression of end-1 in early stage.

These pieces of evidence together show that even with a similar network structure, the C. briggsae network might exhibit a different endoderm specification mechanism as a result of differential expressions of endoderm (end-1 and end-3) and mesoderm (med) specification genes. The med might play a more dominant role in the tug-of-war in mesoderm versus endoderm specification than it does in C. elegans because of the lower expression of end-3. A decrease in upstream activation, such as POP-1, will possibly result in a decrease in expression of both med and end-3, leading to loss of expression of Cbr-end-3 and shorter or even non-existent expression of Cbr-end-1. The remaining MED in the end takes over the role of cell fate decision and turns the E cells into MS cells when there is no POP-1 to suppress MS genes that determine the MS cell lineage.

Despite the differences in expression of the endoderm specification genes, higher maximal level of expression of end-1 and lower total expression of end-3s, the expression patterns of Cbr-elt-2 and Cbr-elt-7 is very similar to Cel-elt-2 and Cel-elt-7. This further corresponds to the theory of saturation of expression for both elt-2 and elt-7.
5.3.3 Endoderm and intestinal specification in *C. remanei*

Unlike *C. elegans* and *C. briggsae*, we were unable to detect any expression of Cre-end-3 by single-molecule FISH in any of the stages we observe. This is consistent with an observation from Morris Maduro that they were not able to express Cre-end-3 in *C. elegans*. If we take the single-molecule FISH data as accurate and assume that Cre-end-3 expression is non-existent in *C. remanei*, the other remaining endoderm specifying genes are Cre-end-1 and Cre-meds. Noticeably, the Cre-end-1 expression profile exhibits a similar short expression as in *C. briggsae*, dropping sharply around 100-nuclei stage (Figure 21). With no expression of end-3, it is not surprising that Cre-end-1 has a short expression compared to Cel-end-1 and does not have higher maximal plateau expression than Cel-end-1. An interesting comparison to Cel-end-1 expression in end-3/- reveals that without the activation from end-3, Cre-end-1 has significantly higher variability in 4E stage in which Cre-end-1 starts to go down but has a similar variability in 2E stage when both Cre-end-1 and Cel-end-1 reaches the plateaus (Figure 22).

The two intestinal specification genes, Cre-elt-7 and Cre-elt-2, are also measured and compared to the other two species. While Cre-elt-7 exhibits significantly lower expression level with similar variability compared to Cel-elt-7 throughout all the developmental stages(Figure 23), Cre-elt-2 has a remarkably different expression pattern. The correlation is even more significant (R=0.73104 in *C. remanei* and R=0.24329 in *C. elegans*) than the correlation of Cel-elt-2 and Cel-elt-7 in *C. elegans* end-3/- (Error: Reference source not found). This observation is yet another evidence
of the loss of Cre-end-3 since now both Cre-elt-2 and Cre-elt-7 are under the regulation of the same gene, Cre-end-1. In a female-male species like C. remanei, they usually suffer from a lot of inbreeding depression in the lab. It is observed that even without any perturbation, lots of embryos die from defects in pharyngeal or intestinal development. The low expression of Cre-elt-2 observed in a lot of embryos might also be a result of this inbreeding depression.

CRE-ELT-2 expression does not start from 50-nuclei stage but start at the same time as Cre-elt-7 and Cre-end-1. This suggests that the particularly low transcription rate predicted by the model in C. elegans might not hold for C. remanei. Instead of having low transcription rate for elt-2, Cre-elt-2 might have the same transcription rate as all the rest of the genes. A low transcription is reported to function as a buffer for the upstream regulatory noise so a low transcription rate for an important gene for cell fate decision might be an evolutionary advantage. However, when does this low transcription rate emerge during evolution? Which one, the higher or lower transcription rate, is the more universal state within this genus? This would be an interesting topic to pursue in the future when more different species within this genus can be examined.

5.4 Discussions and Conclusions

By assuming some conservation of the network structure, we are able to make sense of differences among species using what we learned in C. elegans through quantitative method. One might argue that this assumption is far fetched. We are not trying to say that the network structure must be the same for these two species but we
can try to make sense of what lies behind these expression profiles more by this assumption. There are two major conclusions that develop from this assumption.

5.4.1 *C. briggsae* and *C. remanei* rely on different genes to constrain gene expression variation of *end-1*.

MED GATA factors and *end-3* have duplicated several times in the history of *Caenorhabditis*. MED GATA factors play an important role in the mesoendoderm cell fate decision and they not only affect endoderm specification genes but also affect the pharyngeal development. While *end-3* also had been duplicated throughout the history, surprisingly they are either not expressed or have lower total expression compared to their counterparts in *C. elegans*. On the other hand, *end-1*, as the other endoderm specifying gene, is expressed in all three species and even higher in *C. briggsae*. Compared to *end-3* expression, *end-1* expression is less transient and more robust to upstream changes. The regulatory dynamics among all these three genes determines the cell fate of the daughter cells of the EMS cell. Both *C. remanei* and *C. briggsae* seem to adopt a mechanism that relies more on early upstream regulators such as SKN-1, MED family or POP-1, instead of END-3 to constrain the expression variation of *end-1*.

5.4.2 The role of *elt-2* in *C. remanei*

While we have discussed that lower transcription rate allows the gene to have more time to integrate fluctuating inputs, what could be the advantage of having a higher transcription rate? Is there any advantage for *Cre-elt-2* to have transcription rate that might be higher than *Cel-elt-2*? By having higher transcription rate, the gene might
be able to reflect the upstream gene expression more faithfully and preserves the gene expression variability. The gene expression of \textit{Cre-elt-2} and \textit{Cre-elt-7} both reflect the gene expression variability in \textit{Cre-end-1}. Considering how \textit{Cre-end-1} expression has a shorter expression period, is it possible that by having \textit{Cre-elt-2} and \textit{Cre-elt-7} turning on early, \textit{C. remanei} might not need \textit{end-1} expression in later stage of the 4E stage and the 8E stage.

This observation leads to two questions: First, is \textit{Cre-elt-2} still the major regulator that determines intestinal development in \textit{C. remanei}? If \textit{Cre-elt-2} and \textit{Cre-elt-7} have such highly correlated expression, embryos with low \textit{Cre-elt-2} also have low \textit{Cre-elt-7}. How many of these embryos with low expression of \textit{Cre-elt-2} and \textit{Cre-elt-7} die of defects in gut development? Is it possible that there are other genes that specifies intestinal development in \textit{C. remanei}? Second, how much is this variability in \textit{Cre-elt-2} affected by the inbreeding pressure in \textit{C. remanei}? These questions might be answered with manipulation of gene expression of \textit{Cre-elt-2} and \textit{Cre-elt-7} in \textit{C. remanei} in the future.
Figure 1  The survival rates of the conditions examined in this study. From top left to right, end-1;end-3(MED-), end-3(MED-), end-1(MED-), end-1/-. From the bottom left to right, end-3/-;elt-7(RNAi), end-1/-;elt-7(RNAi), skn-1/-. 

*Intestinal differentiation* 

Survival Rate: 75% or 50%  
Survival Rate: 100%  
Survival Rate: 100%  
Survival Rate: 100%  

Survival Rate: 92%  
Hatching Rate: 100% (0% L2 larvae)  
Survival Rate: 92%  
Worms that form guts: 20%, 54%, 21%  

Chapter VI. Figures and Tables
Figure 2 GUIs in Aro. A. The training GUI. The left plot is a $16 \times 16$ square of pixels from the image on the right with local maxima (candidate spots) marked in blue. The user has the option to designate the maxima as signal, noise and add them to the training set or to skip them. B. The reviewing GUI. The left plot is a grid containing each identified local maximum ranked by its vote. Blue outlines mark signal spots; yellow boxes mark noise spots. The user has the option to correct classifications and retrain the classifier.

The image on the right shows the context for the spot currently in focus (red outline).
Figure 3 Flowchart of the analysis pipeline with details of the automated steps.
Figure 4 Calibration curves based on bagged probability estimates are robust to individual curation differences. A large corpus of manually curated training spots was binned by the bagged probability estimate derived from averaging the probability estimates for a local maximum from each tree in the ensemble. The calibration curves are constructed by fitting a sigmoidal function to the plot of bin centers (0 to 1 by steps of 0.1) versus the proportion of curated good spots in that bin. Calibration curves based on bagged probability estimates are less susceptible to curation differences between individuals than ones constructed by majority rules voting.
Figure 5 The algorithm performs well even as data quality degrades. We tested the software on artificially constructed images of varying quality with realistic signal intensities, spot densities, and background noise based on real data. Spot density is measured by average distance of a pixel to an artificial spot in an image and is shown in the inset. Signal intensity on the x-axis is the average pixel intensity at the centers of the random spots minus the mean pixel intensity of the image divided by the standard deviation of the pixel intensities in the image. A sample of spots from the background 1 images were used to train the classifier.
Figure 6 Comparison of spot identification and classification methods. A. The upper left is a maximum merge projection of an smFISH image from a C. elegans embryo for which 488 signal spots were counted by hand (using the GUIs described here). A green rectangle highlights a section of the image. The other three images show spots identified in this green rectangle by FISH-Quant (upper right), the threshold-picking method (lower right), and the method described here (Aro: lower left). The number of signal spots identified in the embryo by the various methods are noted in the lower right of each image. Circles mark the locations of identified spots and are color coded by z-slice. Arrows point to representative areas depicting the tendency of threshold method to identify a large high intensity region comprised of several spots as a single spot. B. A plot of manually counted spot number (x-axis) and estimated spot number (y-axis) by Aro, threshold-picking, and FISH-Quant across 28 C. elegans embryos. Both FISH-Quant and threshold-picking tend to underestimate the true number of spots (particularly at higher spot counts for the threshold method) while our Aro machine learning method performs well across a range of spots numbers. Spearman correlations (r) between the true and estimated spot number are listed for each method. Both Aro and threshold-picking perform significantly better than random on this dataset. Interval estimates are depicted for Aro. Neither FISH-Quant nor threshold-picking provides a way to estimate error.
Figure 7 Expression of *elt-2, elt-7, end-1* and *end-3* at embryo level in different deletion strains.
Figure 8 The end-1, elt-7, elt-2 expressions at cell level in wild type, end-1-/-, end-3-/ and end-1-/-;end-3-/-(MED-). The color represents the number of E cells of that embryo. Each dot represents the mean cell level expression of the gene and the error bar represents the standard deviation of the cell level expression among the cells that are expressing the gene. (A) end-1 and elt-7 expression at cell level in different deletion strains. end-1 expression decreases after 100 minutes and disappear around 150 minutes. elt-7 has a slower decrease rate. (B) (Upper Row) elt-2 expression at cell level in different strains. (Lower Row) The green circles represent the number of E cells that are supposed to exist at that developmental stage. The blue dots represent the total number of E cells that are expressing elt-2. The cells that are expressing elt-2 are defined as the cells that have at least 10 transcripts of elt-2.
Figure 9 Comparison of elt-7 expression levels and variability in wild type, end-3-/- and end-1-/-.

In end-3-/-, there is a slight but significant delay of activation of elt-7 compared to wild type from the end of 1E stage to mid-4E, until it reaches the same maximal level as wild type (Upper left). On the other hand, in end-1-/-, we did not observe a significant delay of elt-7 during the same time period of activation but a significant decrease of elt-7 in 8E stage is observed (Lower left). The variability of elt-7 in end-3-/- is not significantly different from wild type in any stages (Right Column). The green patch on the top of each figure marks the alpha cut-off (alpha=0.05) for a two-sided test. The faded blue or red curves represent smoothed mean curves of the statistics of bootstrapped data for each corresponding data set. Each curve is smoothed with a Savitzky-Golay filter with a span width of 100. The bright blue or red curves are the smoothed mean curves of the statistics of the experimental data. The bootstrapped samples are generated by bootstrapping data points within the same developmental stage. The bootstrap sample size is 1000 but for purpose of plotting efficiency, we only plot 500 samples.
Figure 10 Comparison of *elt-2* and *elt-7* variability. Both *elt-2* and *elt-7* reach maximal level around 60 minutes after activation and *elt-7* has significantly higher variation after reaching the maximal level. The bootstrapped data is generated the same ways as mentioned in Figure 9. *elt-2* and *elt-7* are shifted by their expression activation time, 100 minutes for *elt-2* (the end of 2E stage) and 44 minutes (start of 1-E stage) for *elt-7*. The alternative hypothesis is *elt-7* has a higher expression variability than *elt-2* so instead of a two-sided test, we use a one-sided test to calculate the p-value and the alpha=0.05, which corresponds to the green patch on the top.
Figure 11 Comparison of *elt-2* expression in *end-3/-* and *end-1/-*. 
Figure 12 Comparison of elt-2 expression in end-3/-;elt-7(RNAi) and end-1/-;elt-7(RNAi). The figure is represented the same way as it was described in Figure 9. Bootstrap sample size is 1000 and alpha=0.05 for a two-sided test. Without elt-7, elt-2 expression is significantly delayed in both end-1/-;elt-7(RNAi) and end-3/-;elt-7(RNAi) and the mean expression level is consistently lower after being activated. The variability of elt-2 is also significantly increased in end-1/- when elt-7 is absent.
Figure 13 Loss of elt-7 increases elt-2 gene expression variability. Losing elt-7 increases the gene expression variation of elt-2 in wild type (Left), end-3-/- (Middle) and end-1-/- (Right) significantly during mid-4E to 8E stage, during which elt-2 plateaus.
Figure 14 A) Comparison of the expression of end-3, elt-2 and elt-2 in experiments with end-1-/- animals that are thawed only 2-3 generations before the fixation. end-3 expression is less variable and both elt-2 and elt-7 are similar to their counterparts in the wild type. B) Comparison in experiments with end-1-/- animals that are grown at room temperature for more than 20 generations. end-3 is more variable and there is a delay of elt-2 expression. C) The elt-2 expression in end-1 animals layered with elt-7 expression as the color of the spots. The color represents the elt-7 expression.
Figure 15 Comparison among gene expression of *end-1, elt-7* or *elt-2* in wild type, *end-1-/-* and *end-3-/-* indicates that the expression of these genes are saturated in wild type.
Figure 16: Comparison between simulated results and experimental results in wild type, end-1/- and end-3/-.
Figure 17 The timing of switching point and the ELT-2 level at switching point in each cell in different strains.
Figure 18 The proportion of embryos that have 8 surviving cells corresponds to the survival rate of \textit{end-3-/-} and \textit{end-1-/-;end-3-/-(MED-)}
Figure 19 Maximum merge image of a portion of Cbr-end-3 (green) and CBG11404 (magenta) labeled 24-cell embryo. The magenta and green spots do not co-localize together, indicating that they recognize different transcripts.
Figure 20 The total number of end-3.1 and end-3.2 expressions is less than half of Cel-end-3 (Upper Row). Cbr-end-1 has a higher maximal level but shorter expression period compared to Cel-end-1.
Figure 21 Gene expressions in *C. briggsae, C. remanei, and C. elegans.*
Figure 22 Comparison of end-1 in *C. elegans* end-3-/- and *C. remanei*. The left figure shows that end-1 expression in *C. remanei* drops significantly after entering the 4E stage compared to end-1 expression in *C. elegans* end-3-/. The right figure shows that the coefficients of variation of end-1 are not significantly different when both reaches plateaus in 2E stage but the differences become significant in 4E stage.
Figure 23 *elt-7* expression in *C. remanei* is significantly lower than *Cel-elt-7* but the variability of *elt-7* are not significantly different compared to *Cel-elt-7*. 
Chapter VII. References


Tsai, T.Y.-C., Choi, Y.S., Ma, W., Pomerening, J.R., Tang, C., and Ferrell, J.E. (2008). Robust, Tunable Biological Oscillations from Interlinked Positive and Negative Feedback


