UNIVERSITY OF CALIFORNIA,
IRVINE

Comparative genomics of *Steinernema*

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biology

by

Marissa Giovanna Macchietto

Dissertation Committee:
Assistant Professor Ali Mortazavi, Chair
Assistant Professor Olivier Cinquin
Professor Brandon Gaut
Professor Anne Calof
Professor Ken Cho

2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, committee, friends, and family for all of their support and guidance through the past five years of my PhD. Chapter 2 was published with the permission of BioMed Central. The text of this chapter is a reprint of the material as it appears in *Genome Biology*.

CURRICULUM VITAE

Marissa Macchietto

Education:

2011 – 2016	University of California, Irvine Ph.D.: Developmental and Cell Biology
2007 – 2011	University of California, Irvine Bachelor of Science: Biological Sciences

Positions:

2012 – present	Graduate Student Researcher Ali Mortazavi Laboratory Department of Developmental and Cell Biology University of California, Irvine
2011 – 2012	Graduate Student Researcher Mathematical, Computational, and Systems Biology Program University of California, Irvine

Awards:

2015	1st place in Developmental and Cell Biology retreat poster competition
2012	UCI Center for Complex Biological Systems Opportunity Award
2008–2011	Dean's Honor List

Teaching Experience:

2016	Mathematics and Science (COSMOS) Module 8: Genes, Genomes, and Biocontrol
2015	Teaching Assistant for Developmental and Cell Biology lab
2015	Teaching Assistant for California State Summer School for Mathematics and Science (COSMOS) Module 8: Genes, Genomes, and Biocontrol
2014-6	Mentor to Dristi Angdembey, Negar Heidapour, and Bryan Rodriguez Undergraduate Laboratory Assistants University of California - Irvine
2014	Teaching Assistant for Scientific Writing (Bio100)
2014	Teaching Assistant for California State Summer School for Mathematics and Science (COSMOS) Module 8: Genes, Genomes, and Biocontrol
2014	Teaching Assistant for Intro to Personalized Medicine (D132)
2013	Mentored four high school student researchers from California State Summer School for Mathematics and Science (COSMOS) at UCI

2013	Teaching Assistant for California State Summer School for Mathematics and Science (COSMOS) Module 8: Genes, Genomes, and Biocontrol
2013	Teaching Assistant for Scientific Writing (Bio100)
2012	Teaching Assistant for Genetics (Bio97)
2012	Teaching Assistant for California State Summer School for Mathematics and Science (COSMOS) Module 3: Tissue and Tumor Biology and Mathematical/ Computer Modeling
2012	Mentor to Kristen Park -- high school student researcher at UCI
2012	Biology Tutor during the UCI Mathematical, Computational, and Systems Biology Summer Boot Camp

Talks:

2016	Invited Seminar Speaker for the Department of Nematology at UC Riverside "Comparative transcriptomics of *Steinernema* embryonic development" (talk)
2016	Invited Speaker for Developmental and Cell Biology Retreat – "Comparative transcriptomics of *Steinernema* embryonic development" (talk)
2016	Evolutionary Biology of *Caenorhabditis* and Other Nematodes – Cold Spring Harbor, NY "Comparative transcriptomics of *Steinernema* embryonic development" (talk)
2013	UCI Center for Complex Biology Systems Annual Retreat – "Fitness tradeoffs in experimentally evolved *E. coli*" (talk)

Poster presentations:

2015	Southern California Systems Biology Conference "Genomic analysis of *Steinernema*: Insights into insect parasitism, intragenus and intergenus evolution" (poster)
2015	International Plant and Animal Genome Conference (PAGXXIII) "Genomic analysis of *Steinernema*: Insights into insect parasitism, intragenus and intergenus evolution" (poster)
2014	UCI Center for Complex Biology Systems Annual Retreat – "Genomic analysis of Hox genes in five *Steinernema* genomes." (poster)
2013	19th International *C. elegans* Meeting at UCLA – "Genomic analysis of *Steinernema*: Insights into insect parasitism, intragenus and intergenus evolution" (poster)
2013	UCI Center for Complex Biology Systems Annual Retreat – "Genomic analysis of *Steinernema*: Insights into insect parasitism, intragenus and intergenus evolution" (poster)
2012	Annual International Conference of Intelligent Systems for Molecular Biology – HiTSeq Conference –"Time-course of Pu.1 binding in erythroid differentiation" (poster)

2012	UCI Center for Complex Biology Systems Annual Retreat- "Timecourse of Pu.1 binding in erythroid differentiation" (poster)

Graduate Coursework:

2015	Principles in Genomics
2015	Introduction to Biostatistics: An ICTS Research Methods Short Course
2014	International Course in Automated Functional Annotation and Data Mining (Jan. 2014 – CIBNOR, La Paz, Mexico)
2012	Responsible Conduct in Research
2012	Systems Developmental Biology
2012	Computational Systems Biology
2012	Mathematical and Computational Biology II (Partial Differential Equations)
2012	Systems Cell Biology
2011	Mathematical and Computational Biology I (Ordinary Differential Equations)
2011	Biophysics of Molecules and Molecular Machines
2011	Critical Thinking in Systems Biology

Skills:

- Programming in Python, UNIX, and R
- *de novo* genome assembly (PacBio and Illumina)
- *de novo* transcriptome assembly
- NextGen Sequencing Assays: ChIP-seq and RNA-seq
- ChIP-seq and RNA-seq data analysis
- qPCR, PCR
- *in situ* hybridization
- Nematode, tissue, and cell culture

Language Skills:

English (fluent)
Italian (basic)
French (basic)

Publications:

1. **Macchietto, M.**, Serra, L., Angdembey, D., Heidapour, N., Rodriguez, B., El-Ali, N., Mortazavi, A. Comparative transcriptomics of *Steinernema* embryonic development. (Manuscript in preparation)

2. Zeng, W., Ramirez, R., Jansen, C., **Macchietto, M**., Jiang, S., Conesa, A., Mortazavi, A. The landscape of global long-range interactions in mouse and human erythroid cells mediated by YY1, GATA1 and CTCF during differentiation**.** (Manuscript in preparation)

3. Dillman, A.*****, **Macchietto, M.***, Finlinson, C.F., Rogers, A., Williams, B., Antoshechkin, I., Lee, M.M., Goodwin, Z., Lu, X., Lewis, E.E., Goodrich-Blair, H., Stock, S.P., Adams,

B.J., Sternberg, P.W., Mortazavi, A. 2015. Comparative genomics of *Steinernema* reveals deeply conserved regulatory networks in nematodes. *Genom Biol*. *16*(1), 200.

4. Srinivasan, J*., Dillman, A.*, **Macchietto, M**., Heikkinen, L., Lakso, M., Fracchia, K.M., Antoshechkin, I., Mortazavi, A., Wong, G., Sternberg, P.W. 2012. The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle. *Genetics*. PMID: 23410827.

ABSTRACT OF THE DISSERTATION

Comparative genomics of *Steinernema*

By

Marissa Macchietto

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2016

Professor Ali Mortazavi, Chair

Nematodes comprise one of the most diverse bilaterian phyla, having colonized nearly every imaginable ecological niche on earth. They are major parasites of plants, animals, and humans, despite sharing a relatively conserved body plan. The *Steinernema* genus comprises over 70 characterized species that are lethal parasites of insects, which have different foraging strategies and host ranges, and are distantly related to the model organism *C. elegans*. To better understand the evolution of parasitism and development in nematodes, we sequenced and analyzed the genomes as well as transcriptomes of five key members of the *Steinernema* genus (*S. carpocapsae, S. scapterisci, S. monticolum, S. glaseri,* and *S. feltiae*). In chapter 2, using available ecological and molecular data, we explore genomic differences likely to be involved in insect parasitism, particularly in host-range and specificity of these five species. We find surprising gene family evolution of proteases, protease inhibitors, proteolytic cascade proteins, GPCRs, transposon and retroviral content, and even protein-protein interaction domains, many of which correlate excitingly with known differences in host range and specificity among these parasites. The combination of multiple closely related genomes in a non-*Caenorhabidtis* clade and accompanying deeply sequenced transcriptomes allows for powerful comparisons to other

genera such as *Caenorhabditis*. In particular, comparisons in gene expression at defined stages show surprising plasticity of timing across one-to-one orthologous genes in the five genomes when compared to *C. elegans*. Our conservation analysis shows that approximately 20 Mb are conserved across the *Steinernema* species, with 5.1 Mb of this comprising non-coding regions. Our analysis of the conserved non-coding regions combined with stage-specific gene expression data reveals that a limited number of regulatory motifs are associated with conservation of stage-specific ortholog expression in *Steinernema* and *Caenorhabditis*, which suggests that several underlying gene regulatory relationships controlling development are conserved in the two genera. In chapter 3, we investigate embryonic development in *Steinernema* by comparing the expression of orthologous genes at eleven different embryonic stages of two *Steinernema* species with two *Caenorhabditis* species. We found that zygotic transcription initiates at different developmental stages in each species, with the *Steinernema* species initiating transcription at earlier developmental stages than *Caenorhabditis*. Surprisingly, we also found that gene expression conservation during development is highest at the later embryonic stages than at the earlier ones, indicating that ortholog expression divergence across distantly related species follows a funnel-shaped model in contrast to the hourglass model of nematode development that has been previously proposed. Thus, this work provides novel insight into embryonic development across distantly related nematode species and demonstrates that the mechanisms controlling early development are more diverse than previously thought.

CHAPTER 1

Introduction

Nematodes

Nematodes are remarkable and diverse unsegmented roundworms that originated during the Precambrian or Cambrian explosion over 500 million years ago (Onstad et al., 2006; Sudhaus et al., 2008). They comprise one of 35 animal phyla called Nematoda, and they are closest evolutionarily to other molting invertebrate phyla such as nematomorphs, tardigrades, and the better-known arthropods, which include animals such as insects, crustaceans, and arachnids. Although over 25,000 nematodes species have been described to date, it is estimated that there are more than 1 million species in existence, making nematodes potentially as speciose as arthropods (Hart et al., 2008; Hugot et al., 2001; Lambshead,1993; Lambshead, 2004).

Interestingly, most people will go through their lives without ever even knowing what nematodes are because most species are microscopic, typically reaching lengths that are on the scale of 0.5-2 mm. They can be found living in soil, fresh water, and saltwater, and associated with plants, insects, livestock, and humans among other things. A nematode population study found that a cubic meter of soil contained 30 million worms from 105 different species (Yeates, 1979). Nematodes that are larger than 2mm are not frequently found in soil environments, but instead are found associated with animals as parasites. *Ascaris lumbricoides*, a human parasitic nematode that causes ascariasis and is transmitted through contaminated water, can reach up to 19 inches inside its human host, which is 240x longer than the lengths of common soil dwelling nematodes. However, other hosts have even larger nematodes infecting them. The largest known nematode is *Placentonema gigantissima,* which was discovered in a sperm whale and is recorded to be over 8 meters long (Gubanov, 1951).

Despite the differences in sizes and locations they inhabit, the general body plan of nematodes is highly conserved across species. They are essentially muscular tubes that have a

mouth, a digestive tract, an anus, and reproductive structures. They also have a simple nervous system composed of several hundred neurons organized as a nerve ring and ventral nerve cord as well as a secretory-excretory system. They are pseudocoelomates that do not have a circulatory system, and their respiration occurs by gaseous diffusion through their exoskeletons. Even though nematodes have the same general body shape, there are differences in external as well as internal morphological features between species, and nematologists distinguish nematodes based on the structure of their mouthparts and other anatomical features such as their gonads. Some nematodes have mouthparts designed for sucking up bacteria. These are the microbivore nematodes that feast exclusively on bacteria. Some have been characterized to have long piercing mouthparts (stylets), such as the plant parasite *Xiphinema Index*. This long stylet allows it to access nutrients from deep within the plant's roots. Others have a tooth or multiple teeth, such as the predatory nematode, *Pristionchus pacificus*. *P. pacificus* can switch between two developmental modes, characterized by two different types of mouth parts depending on environmental conditions (Kiontke et al., 2010; Sommer et al., 2013). It will produce one type of mouthpart to feed on bacteria when there is plenty of food around, whereas under starvation conditions it will use its tooth to kill and to feed on nematodes from different species.

Some nematodes such as *Panagrellus redivivus* (Srinivasan et al., 2013) and *Caenorhabditis elegans* have been characterized as "free-living", meaning that they feed on bacteria, fungi, other nematodes, protozoa, or other small organisms. However, a large number of the described species that have been characterized are parasites of a wide range of organisms. The potential for negative impact on human health and agriculture makes parasitic nematode research a high priority, as they have the potential to affect our population as a whole by influencing the resources we depend on. It is estimated that 1.7 billion individuals are infected

with parasitic nematodes worldwide leading to several well-known, but neglected tropical diseases such as lymphatic filariasis, onchocerciasis, schistosomiasis, soil transmitted helminthiases (STH) and trachoma (neglected tropical disease statistics for 2016 – WHO: http://www.who.int/gho/en/). Some of these nematodes are transmitted fecal-orally to uninfected individuals from infected individuals or other infected vertebrates, such as livestock (pigs) or pet dogs and cats. Others have more complex life cycles where they must spend part of their lives in the soil before becoming infective to humans or they are transmitted by a vector organism to humans, such is the case for the nematode *Onchocerca volvulus,* which is transmitted through the bite of the black fly and causes "river blindness" (Hall et al., 1999). Nematodes, such as the plant parasitic nematodes, can also impact human resources. Plant parasites can devastate the growth of crops by boring into the roots and creating large nodules in the root system or by leeching nutrients from the roots. A survey of crop loss in 35 US states reported that nematodes were responsible for up to a 25% loss in crops annually (Koenning et al., 1999) and global cost estimates of 80 billion dollars annually (Handoo, 1998). However, not all parasitic nematodes are "bad" for humans. Some are beneficial, such as the entomopathogenic ("insect pathogenic") nematodes (EPNs). EPNs are nematodes that are highly pathogenic to many types of insects and typically kill within a couple days after infection. Because of their potent insect-killing capabilities, they are employed as a biological control agent for crop-eating insects, and present a safer alternative to chemical insecticides. In this chapter, I will briefly introduce many features of nematodes, from their diversity to their genomes to their development, and then transition to covering the model system of this dissertation, the nematodes from the insect-pathogenic genus *Steinernema*. *Steinernema* nematodes are widely used in agricultural applications to control insect pests, but prior to this thesis, very little was known about their genes, genomes, and

development. Several questions were guiding our study, such as: Which genes contribute to *Steinernema* parasitism? How similar is the timing of development between *Steinernema* species and *C. elegans*? How similar is development between *Steinernema* species and *C. elegans* at the molecular level? Is the nematode body plan conserved at the molecular level across nematode species?

Nematode diversity

The large number of species in Nematoda shows a surprising amount of phylogenetic diversity given how extremely conserved their body plans are. Nematoda is divided into twelve monophyletic clades based on the sequence of the small rRNA subunit (Holterman et al., 2006). Clades 3-12 are closely related to each other and fall under the class Chromadoria, while clades 1 and 2 belong to the more basal classes Enoplia and Dorylaimia respectively (Figure 1). Members of all of these clades have diversified to inhabit every ecological niche imaginable and have interesting adaptions both in terms of environment and behavior. Species such as *Cryonema crassum* (clade 5) live at the freezing point inside of holes in the arctic ice (Tchesunov and Riemann, 1995), while others such as Stilbonematinae (clade 5) live at the redox-boundary of sulfur-rich marine sediment and are coated in ectosymbiotic sulfur-oxidizing bacteria to help them survive (Nussbaumer et al., 2004). *Steinernema tami* (clade 10) has dimorphic sperm, which is not a characteristic of other *Steinernema* species. While males of other *Steinernema* species such as *S. feltiae* produce chains of motile monomorphic spermatozoa (5um/spermatozoa), *S. tami* produces motile megaspermatozoons (30-35um) that are coated with immobile microspermatozoa (3um) that are attached through gap junctions (Yushin et al., 2007). There is also quite a lot of diversity in reproductive strategies used in each clade of the phylum.

Nematodes have been found to reproduce using hermaphroditic (self-fertilization – sexual reproduction), gonochroistic (male-female sexes – sexual reproduction), and parthenogenic (mostly clonal – asexual reproduction) mechanisms, which suggest that different species have specific reproductive strategies adapted for their particular niches. Because there is so much diversity across nematodes, this raises the question the extent of gene expression conservation during their development, and which genes are contributing to their adaptions.

The nematode *C. elegans* as a model organism

Sydney Brenner selected the free-living nematode *C. elegans* in the 1970s as a model organism for the study of metazoan development with a particular interest in neurodevelopment. *C. elegans* was chosen for this purpose because of its transparency, ease of culture, fast generation time, small nervous system, amenability to genetic manipulation and convenience of genetic analysis. Another significant asset is its deterministic mode of development. Every cell in the *C. elegans* adult has been traced from the single-cell embryonic stage to the final tissue that it is a part of. Additionally, each *C. elegans* adult hermaphrodite produces exactly 959 somatic cells, and approximately one-third of them (302) develop into neurons that form ~7,500 synaptic connections, which have also all been mapped (White et al., 1986). *C. elegans* has two orders of magnitude fewer neurons than other model organisms, such as the fruit fly *Drosophila melanogaster*, which has 250,000 neurons, or a larval zebrafish, which has 100,000 neurons (Hinsch et al., 2007). Its deterministic development coupled with its low overall cell number makes *C. elegans* ideal for determining the functions of each of its cells and for determining the contributions of each cell to the development of the organism. Studies in *C. elegans* have helped us to understand much about development in nematodes, and as a result, *C. elegans* has been

treated as a representative for the nematode phylum. However, other studies (Schierenberg, 2006; Voronov et al., 1998), as well as the work presented in chapters 2 and 3, reveal that *C. elegans* may not be representative of all nematodes, and they highlight the importance of studying other nematodes.

Nematode development

Nematode development consists of an embryonic stage, followed by four larval stages (L1-L4), and an adult stage. During larval development, the nematode must shed its exoskeleton, also known as the cuticle, so that it may grow to the next larval stage. If growing conditions are not favorable due to a lack of food or overcrowding, L1-stage nematodes can enter an alternative developmental program to turn into long-lived dauers or infective juveniles (IJ) rather than transitioning from the L2 stage to the L3 stage. Dauer/IJ worms are in a state of lower metabolic activity and can live for months without food and under harsh environmental conditions. When conditions return back to propitious levels, the dauers/IJs can emerge from their "hibernation" and resume growth to reach adulthood. Once adults are sexually mature, they will begin to produce oocytes (hermaphrodites and females) and sperm (males). Hermaphrodites such as *C. elegans* also produce sperm in the L4 stage prior to making oocytes.

Embryonic Development

Scientists have studied nematode embryonic development in the larger human parasite *Ascaris lumbercoides* since the 1800s. However, most of what is known today about nematode embryonic development at the molecular and genetic level has been found in *C. elegans* since the 1970s, and so herein a discussion of the features of embryonic nematode development in *C. elegans* will be followed by a comparison to development in other nematodes.

Embryonic development in nematodes is comprised of two main phases: 1) proliferation and 2) organogenesis/morphogenesis. In the first phase, rapid cell proliferation occurs to establish the cell numbers in the worm, and in the second phase, the cells are differentiated into the final tissues of the first larval stage. Development commences with the fertilization of an oocyte by a sperm to produce a zygote. This results in the restoration of the diploid genome number and a polarization of cytoplasmic determinants ("P granules") to the posterior end of the zygote. After the first cleavage division, two cells are produced with the posterior cell containing more P granules than the anterior cell. The two cells can be distinguished by their size and position in the embryo. The anterior cell is larger and is referred to as the "AB" cell, while the posterior cell is smaller and referred to as the "P" cell. P granules are ribonucleoprotein complexes that likely function to direct the P-cell to become the germ line progenitor (Strome and Wood, 1983). With each successive division of the P cell, one of the daughters will maintain the program to produce the germ-line, while the other will be programmed by the cellular environment around it to form other cell types and tissues that are needed. Maternally deposited proteins and transcripts that are asymmetrically distributed in the zygote pattern the zygote to promote immediate fate specification of the cells. The maternal transcript and protein products also guide the embryo through the first several stages of embryogenesis. Maternal products also repress zygotic transcription. It is not until the 4-cell stage in *C. elegans* that zygotic transcription becomes activated (Edgar et al., 1994, Baugh et al., 2003). Once transcription begins, some of the first zygotic products to be produced are miRNAs that target the maternal mRNAs for degradation and proteins to target maternal proteins for degradation (Tadros et al., 2009). This promotes the transition from maternal control to zygotic control. However, not all maternal products are destroyed immediately, and some will remain in the embryo through

gastrulation (26-350-cell) (Baugh et al., 2003). Studies have found that embryonic development can proceed up to the 100-cell stage before aborting when zygotic transcription is inhibited (Edgar et al., 1994). By the 8-cell stage, one cell is already destined to give rise to the entire endoderm or gut of the worm (E-cell), while its sister cell is specified to produce the mesoderm (MS-cell) or muscle of the worm. Gene regulatory networks governing the early specification of E-cell lineage have been mapped out for *C. elegans* (Maduro et al., 2002).

In *C. elegans*, gastrulation begins at the 26-cell stage and ends at approximately the 350-cell stage. During this period, cells migrate to their new locations in the embryo to form the three distinct germ layers: the endoderm (gut), mesoderm (muscle, pharynx), and ectoderm (epithelium, neurons). This involves the movement of a few cells of the E-lineage and M-lineage from the ventral periphery to the cavity inside the embryo called the blastocoel (Sulston et al., 1983). Additional cell divisions occur with subsequent cell movement to form the early digestive tract. After gastrulation completes (~350 cells), cell proliferation continues until 558 undifferentiated cells are formed (von Ehrenstein and Schierenberg, 1980; Wood 1988), after which morphogenesis begins ("lima bean" stage). The morphogenesis stages (lima bean, comma, 1.5-fold, 2-fold, moving) are easier to distinguish because the embryo has more distinct features. The first muscle contractions are detected between the 1.5-fold to 2-fold stages, and pharyngeal pumping occurs in the moving stage.

Comparative embryonic development in nematodes

Studies of the early cell lineages and blastomere arrangements in a small handful of other nematodes found major similarities to *C. elegans,* which incorrectly led to the assumption that embryogenesis is highly conserved across all nematodes (Boveri, 1899; Müller, 1903).

Additional analyses of nematode development in many branches of the phylum uncovered significant variations at some of the most crucial steps of embryonic development and in how cell lineages are specified, such as the number of cells that migrate during the process of gastrulation and in the timing of gastrulation (Schierenberg, 2006). In 26-cell *C. elegans* embryos, two gut precursor cells and several mesodermal progenitors migrate from the embryo ventral periphery into the center of the embryo to give rise to the gut and muscle tissues respectively in later embryonic stages (Sulston, 1983). Nematodes from clades 2-12 have also been characterized to undergo gastrulation in this way (Schierenberg, 2006) (See phylogeny in Figure 1). However, gastrulation in other nematodes can happen as early as the 8-cell stage, as is the case for *Plectida* (clade 5), or as late as the 64-cell stage, as it does for *Triplonchida* (clade 1) (Lahl et al., 2003; Schierenberg, 2005). Interestingly, gastrulation in *Triplonchida* is more reminiscent of gastrulation in other animals outside of nematodes, suggesting that it is a more ancient group (Malakhov, 1994; Aleshin et al., 1998). Clade 1 nematodes appear to undergo development very differently than nematodes from the other clades. All nematode clades (1, 3-12) except clade 2 undergo asynchronous cleavages divisions and produce cells of different sizes during early embryogenesis (Voronov et al., 1998). In contrast, clade 1 nematode early cell divisions are synchronous and produce blastomeres that are indistinguishable from each other by size, position, and appearance (Voronov et al., 1998). Enoplian blastomeres do not exhibit determinate cell lineage patterns as in other nematode clades (Voronov et al., 1998). For example, the endoderm precursor cell, which is established at the 8-cell stage in *C. elegans* forms at a particular position in between the MS-cell and the P-cell. However, in clade 1 nematodes, the endoderm precursor can be derived from one of multiple blastomeres, and this indeterminate mode of development is reminiscent of other bilaterians. Clade 2 early development also varies

slightly from the clades 3-12. Even though the first division produces two cells of different size, these cells are not homologous to the AB and P cell in the nematodes of clades 3-12 because their fate is not yet set. At the 4-cell stage, the daughters of these two different blastomeres mix and match to form the AB progeny and the P progeny (Voronov et al., 1998). In addition, the endoderm cell forms from the AB lineage in these nematodes, while it forms via the P lineage in clade 3-12 nematodes (Drozdovskii, 1975, Voronov et al., 1998).

Another major variation in embryonic development was found in *Ascaris*, a clade 8 human parasitic nematode that originally seemed to develop identically to *C. elegans*. During *Ascaris* embryonic development, all blastomeres, except those that form the germ line, lose a fraction of their non-coding chromatin (~56%) through a process called chromosome diminution (Boveri, 1887; Davis et al., 1979), which may be a way of maintaining genetic balance in somatic cells and giving a selective advantage to the germ cells (Davis et al., 1979). Chromosome diminution is also used by the clade 10 parasitic nematode *Strongyloides papillosus* for the purpose of sex determination (XX, XO) (Nemetschke et al., 2010), where male offspring are produced through selective elimination of one of the X chromosomes through a mechanism that is yet to be elucidated.

There are also widespread differences in the time length of nematode embryonic development. *C. elegans* (clade 9) early cleavage divisions take approximately 0.5 hours to complete at 20°C, while *Enoplus brevis* (clade 1), a saltwater nematode, takes 4-5 hours to complete each of these early divisions at the same temperature (Voronov et al., 1998). Whether this has adaptive value or is the result of developmental drift, we expect that this variation in the cell division dynamics of development will be reflected as changes in embryonic gene expression to some extent. The differences during development between distant nematode

species raise the question about how conserved gene expression is during nematode development and whether genes that are shared across all species are expressed at the same time points during development.

Comparative genomics in nematodes

In the past two decades, comparative genomics has flourished thanks to the sequencing of many draft and complete genomes belonging to all types of organisms, and it has rapidly changed our understanding of how the genomes of these organisms have evolved. Nematodes have proven to be a fruitful model system for studying genome evolution because of the vast diversity of species in the phylum, and because they have relatively small genome sizes (50-200 Mb) making them cost-effective for sequencing. Following the publication of the *C. elegans* genome in 1998 (*C. elegans* Sequencing Consortium, 1998), efforts were organized to sequence additional genomes to leverage more information from the *C. elegans* genome. The hermaphrodite *C. briggsae* was sequenced as a companion genome for *C. elegans* in 2003 (Stein et al., 2003). Using the *C. briggsae* genome, *C. elegans* gene models were able to be validated and corrected, and conserved regions were determined across their DNA sequences. The study also found that there was a remarkable rate of intrachromosomal rearrangement between them. Comparisons found more than 4,000 chromosomal rearrangement events, many of which were local inversions and transpositions, but there were also some cases of between-chromosome translocations (Parkinson et al., 2004). This is more than 4x the number of rearrangements between human and mouse (Zhao et al., 2004). *C. elegans* and *C. briggsae* were the only two hermaphroditic *Caenorhabditis* species at the time and had long been assumed to be sister species, but comparative genomics of both species and other *Caenorhabditis* genomes revealed

them to be distantly related and to have evolved hermaphroditism independently (Kiontke et al., 2004). In addition, an analysis in *Caenorhabditis* resolved the relationships of the species to each other, showing that *C. briggsae*, *C. remanei*, and *C. brenneri* are more closely related to each other than to *C. elegans* within the "elegans supergoup" and that *C. castelli* (*C. sp.* 12), *C. angaria* (PS1010) and *C. drosophilae* are more closely related to each other than to *C. elegans* and are found in another more distant group called the "drosophila supergroup". Sequence analysis across *Caenorhabditis* species also revealed massive intron loss within the genus and very little intron gain (Kiontke et al., 2004). Of the ~12,000 orthologous gene pairs between *C. elegans* and *C. briggsae*, there were twice as many orthologs genes with *C. elegans*-specific introns than *C. briggsae*-specific introns (Kiontke et al., 2004).

Comparative analyses extending to other nematode genomes led to several unexpected findings. Parkinson et al. set out to sequence and assemble expressed sequence tags (ESTs) from a variety of nematode species (~30) in order to characterize the diversity of genes in nematodes, or "nematode genespace" (Parkinson et al., 2004; Mitreva et al., 2005). When scientists had sampled gene diversity in bacteria, they encountered a case of diminishing returns, where sequencing an increasing number of species resulted in decreasing amount of new gene information. However, what this study found is that nematode genespace increased linearly with the sequencing of new species. Each nematode contributed 30-70% new genes, indicating that the genetic diversity and adaptive potential is seemingly limitless (Parkinson et al., 2004).

The analyses of multiple nematode ESTs also found orthologs that are present in other animals, but that are missing from the complete *C. elegans* genome. A good example of this is the Hox genes, which are conserved transcription factor genes important in specifying body segment identity during development (Bateson, 1984). *C. elegans* was found to have

significantly fewer Hox genes than arthropods and vertebrates, and scientists believed that this was due to nematodes being more simple or 'primitive' organisms. However, hox analyses performed on other nematodes revealed that several of the hox family genes such as *hox-3* (Hox3) and *ant-1* (Antp) did exist in these other nematodes, indicating that they were lost in the *Caenorhabditis* lineage over the course of its evolution (Parkinson et al., 2004; Aboobaker et al., 2003). In addition, evidence of horizontal gene transfer was found in a few of the plant parasitic (root-knot) nematodes of the *Meloidogyne* genus (Mitreva et al., 2005). This was a surprising finding since the occurrence of horizontal gene transfer in eukaryotes was not well documented. These comparative genomic and comparative embryonic developmental findings are interesting because they contradict the established idea that all nematodes should have similar if not identical molecular and genetics programs governing their formation due to their identical body plans. Our findings on *Steinernema* development in chapter 3 further support this idea.

Entomopathogenic nematodes from the genus *Steinernema*

Although they make up a very small portion of the overall nematode phylogenetic tree, entomopathogenic nematodes (EPNs), which are lethal parasites of insects, have some of the most interesting behaviors and interspecies relationships. EPNs exist in soil environments as infective juveniles (IJs), which are starvation-resistant and adapted to life outside of their hosts. Upon finding and entering a host, EPNs release lethal symbiotic bacteria residing in their gut and together kill their host, which allows both to feed upon the dying insect (Kaya et al., 1993). The nematodes go through several rounds of their life cycle in the host until all food is exhausted and emerge from the carcass as IJs.

The two main genera of EPN nematodes in the phylum Nematoda that are parasites of arthropods and have a symbiotic relationship with pathogenic Gram-negative enteric bacteria are *Steinernema* (clade 10) and *Heterorhabditis* (clade 9). Evidence suggests that the ancestors of these clades independently associated with these organisms approximately 375 million years ago and that they have been evolving independently of other nematode lineages since (Poinar et al., 1993). Over 80 *Steinernema* species and 30 *Heterorhabditis* species have been discovered and characterized to date. It is likely that interspecies interactions have tightly shaped and constrained the evolutionary trajectory of entomopathogenic nematode species, preventing their diversification.

The pathogenic symbiotic bacteria associated with *Steinernema* species belong to the genus *Xenorhabdus*. In general, each steinernematid is associated with its own species of *Xenorhabdus*, and many *Steinernema* species are not able to either grow on or carry the bacteria of another species. Although the nematode and bacteria cooperate to take down their insect host, the *Xenorhabdus* species and the *Steinernema* species are highly pathogenic to insects on their own. The lethal dose to kill 50% of the hosts (LD50) for *Xenorhabdus nematophila* (symbiotic bacteria of *Steinernema carpocapsae*) is < 20 cells when injected into the common laboratory waxworm host *Galleria mellonella* (Akhurst and Dunphy, 1993), while infection by approximately 10 aposymbiotic (carrying no symbiotic bacteria) *Steinernema carpocapsae* IJs results in 100% death rate in *G. mellonella* (Han et al., 2000).

In *S. carpocapsae*, 40 to greater than 100 *X. nematophila* cells are contained in a special compartment (receptacle) comprised of two specialized nematode guts cells in the anterior portion of the IJ's digestive tract (Martens et al., 2003). When an IJ enters the insect and is triggered to develop by the haemocoel, the bacteria are released from the nematode through

defecation. The bacteria then begin proliferating and secreting insecticidal toxins, proteases and lipases to kill the host and digest its tissue. Over the course of infection, they also secrete antimicrobial and antifungal metabolites to inhibit the proliferation of competing microbes and fungi (Li et al., 1995; McIrnerney et al., 1991). The production of these compounds is beneficial to the steinernematids, and when the bacteria reach higher densities within the insect, the steinernematids will begin to feed on them to drive their development. After the nematodes go through several generations and the nutrients are depleted, IJs will emerge from the insect to seek out a new host. Prior to leaving the host, pre-IJ stage nematodes will take in 1-3 bacteria to colonize the receptacle to use for future infections (Martens et al., 2003). IJs become completely sealed in a secondary cuticle, which offers them extra protection, but also prevents them from feeding. IJs are starvation resistant and can live and maintain their infectivity in moist soil for several months. How the nematode supports and maintains the bacteria during this period is unknown.

Steinernema host-finding strategies and unique jumping behavior

Steinernema species have different insect host preferences with some having broader host ranges than others. The host preferences of each species tie closely to the type of host-finding strategy that they use. The two types of host-finding strategies are cruising and ambushing, and some species use one or a combination of both. Cruisers are more active than ambushers. They cruise the soil to search for insects that reside underground (< 20 cm from the topsoil) like beetle larvae, and they can search for insects over large areas. When they detect vibrations or the presence of volatile compounds released by an insect such as CO_2 and insect pheromones (Campbell and Kaya, 2000), they move in the direction of the stimulus and crawl onto their

insect host. In contrast, ambushers like *Steinernema carpocapsae*, are stationary nematodes. They stand on their tails and lie in wait for an insect to pass by. This behavior is called nictation, and is also used by dauers and IJs of other species, such as *C. elegans*. When the ambushers sense a stimulus, they bend over and then catapult themselves in the direction of the stimulus. Thus, ambushers can only really infect insects that are above ground (top 5 cm of soil) since they cannot nictate from within the soil. Interestingly, this jumping behavior is not seen in nematodes outside of the genus *Steinernema*. This brings up questions about what the genetics behind this behavior are.

Another interesting behavior found in *Steinernema longicaudum*, is a male on male fighting to the death. Male adults developing in the same insect host engage in fights by constricting each other resulting in injuries, paralysis, and death (Zenner et al., 2014). Approximately 25% of the nematodes males were found dead when between 2 and 50 males were in the insect (Zenner et al., 2014). When more than 200 males were in the insect, this percentage dropped to 4% (Zenner et al., 2014). In addition, they found that a male's tendency to engage in fighting was higher if it passed through the IJ stage than if it developed normally through the L3 stage (Zenner et al., 2014).

Steinernema nematodes have interesting behaviors and inter-species relationships making them great models for studying parasitism, symbiosis, and development. In addition, studies of *Steinernema* could produce additional insights into development of closely related nematode parasites that impact humans such as *Ascaris lumbricoides* that are too difficult to study in a laboratory.

Theme of thesis

This thesis has two focuses. The first is to compare the genomes and transcriptomes of *Steinernema* species to non-parasitic nematodes to uncover genomic features that make these nematodes pathogenic to insects. The second focus is to compare stage-specific transcriptomes of *Steinernema* species to the more distant *Caenorhabditis* species to determine how conserved gene expression is and to explore any major differences in gene expression during the development of these species. In chapter 2, I assemble the genome sequences for five *Steinernema* species and show that they can be used to find expanded gene families that are potentially involved in their insect parasitic life styles, with several of them having conserved stage-specific gene expression. I show that the anterior Hox genes in the cluster have expanded in *Steinernema* and have identified the presence of multiple inserted conserved orthologs in the region that are not found anywhere near the Hox cluster on the *C. elegans* chromosome. I recover putative conserved regulatory motifs through mining the conserved regulatory regions of *S. carpocapsae* and *C. elegans* orthologs that were also conserved in their developmental gene expression. My co-author Adler Dillman conducted the protein domain comparative analyses and along with Byron Adams performed the phylogenomic analysis. Chapter 2 was published in Genome Biology in 2015 (Dillman and Macchietto et al., 2015). Chapter 3 follows up on cross-species gene expression conservation analyses we did in chapter 2 by focusing on embryonic gene expression. In chapter 3, I show with a high-resolution embryonic time course between two *Steinernema* species and two *Caenorhabditis* species that gene expression converges over the course of embryonic development across these distantly related species. I also show that although embryonic development takes longer in *Steinernema*, zygotic gene expression commences at an earlier developmental stage in *Steinernema* than in *Caenorhabditis*. Lastly, I

show that the neddylation pathway is likely to play a bigger role during *Steinernema* development than in *Caenorhabditis*. In chapter 4, I conclude with a brief discussion about what the next steps for *Steinernema* research are based on my findings.

Figure legends

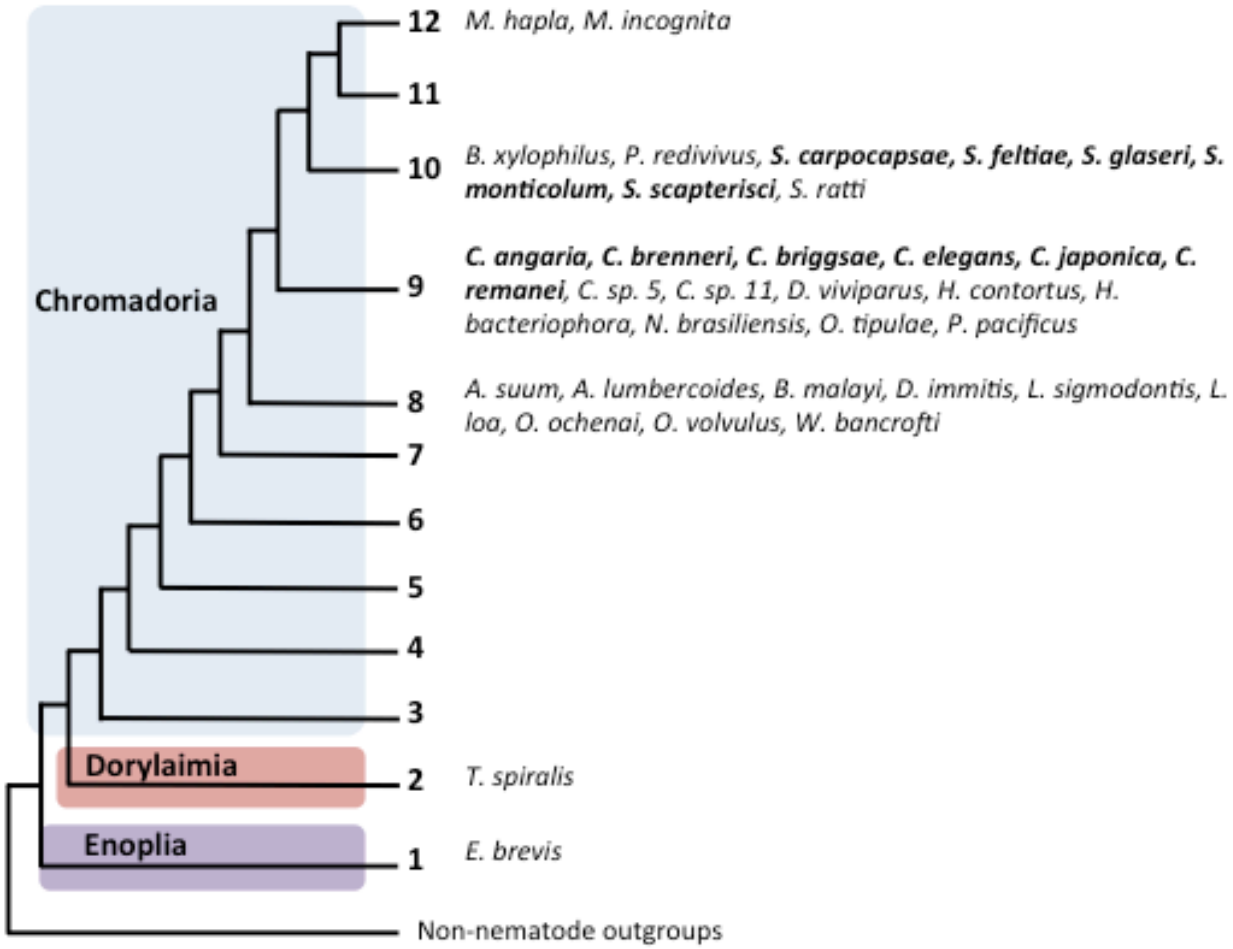


Figure 1. Nematode phylogenetic tree based on the 18S rRNA. Nematodes are assigned to one of twelve monophyletic clades (1-12) based on the sequence of their 18S rRNA (Holterman et al., 2006). Figure adapted from Dillman, 2012.

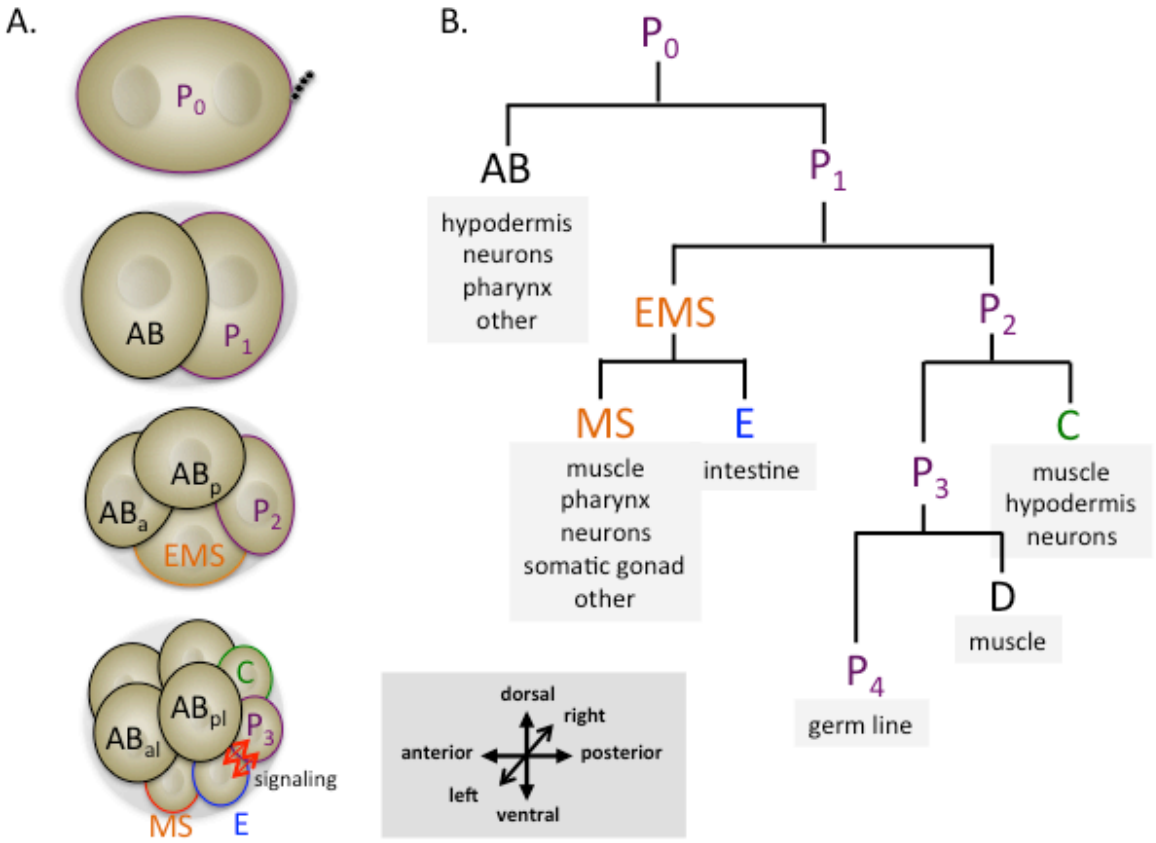


Figure 2. Early founder cells in the *C. elegans* embryo. A) Cell identities for the first three cleavages. AB_{al} and AB_{pl} refer to the left anterior and left posterior AB cells, while AB_{ar} and AB_{pr} refer to the right anterior and right posterior AB cell respectively. The compass to the right indicates the embryo axes. The AB cells are located at the anterior end, while the P cell is located at the posterior end. B) A tree showing each founder cell, their relationship to each other, and a list of tissues they give rise to. The cell names in the tree correspond to the cell names in A. Figure adapted from Rose et al., 2014.

References

1. Onstad, D.W., Fuxa, J.R., Humber, R.A., Oestergaard, J., Shapiro-Ilan, D.I., Gouli, V.V., Anderson, R.S., Andreadis, T.G., and Lacey, L.A. (2006). An abridged glossary of terms used in invertebrate pathology, 3rd Edition (Society for Invertebrate Pathology).
2. Sudhaus, W. (2008). Evolution of insect parasitism in rhabditid and diplogastrid nematodes. In Advances in arachnology and developmental biology, Volume 12, S.E. Makarov and R.N. Dimitrijevic, eds. (Vienna-Belgrade-Sofia: SASA), pp. 143–161
3. Hart, M.W. (2008). Speciose versus species-rich. Trends in ecology & evolution *23*, 660–661.
4. Hugot, J.P., Baujard, P., and Morand, S. (2001). Biodiversity in helminths and nematodes as a field of study: an overview. Nematology *3*, 199–208.
5. Lambshead (1993). Recent developments in marine benthic biodiversity research. Oceanis *19*, 5–24.
6. Lambshead, P.J. (2004). Marine nematode biodiversity. In Nematode morphology, physiology, and ecology, Z.X. Chen, Y. Chen, S.Y. Chen and D.W. Dickson, eds. (CABI), pp. 4554–4558.
7. Yeates, G. W. (1979). Soil nematodes in terrestrial ecosystems. *J Nematol*, *11*(3), 213-229.
8. Hotez, P. J., Brindley, P. J., Bethony, J. M., King, C. H., Pearce, E. J., & Jacobson, J. (2008). Helminth infections: the great neglected tropical diseases. *J Clin Invest*, *118*(4), 1311-1321.
9. GUBANOV, N. M. (1951). [Giant nematoda from the placenta of Cetacea; Placentonema gigantissima nov. gen., nov. sp]. *Dokl Akad Nauk SSSR*, *77*(6), 1123-1125.
10. Kiontke, K., & Fitch, D. H. (2010). Phenotypic plasticity: different teeth for different feasts. *Curr Biol*, *20*(17), R710-R712.

11. Sommer, R. J., & McGaughran, A. (2013). The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Mol Ecol*, *22*(9), 2380-2393.
12. Srinivasan, J., Dillman, A. R., Macchietto, M. G., Heikkinen, L., Lakso, M., Fracchia, K. M. et al. (2013). The draft genome and transcriptome of Panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle. *Genetics*, *193*(4), 1279-1295.
13. Hall, L. R., & Pearlman, E. (1999). Pathogenesis of onchocercal keratitis (River blindness). *Clin Microbiol Rev*, *12*(3), 445-453.
14. Koenning, S. R., Overstreet, C., Noling, J. W., Donald, P. A., Becker, J. O., & Fortnum, B. A. (1999). Survey of crop losses in response to phytoparasitic nematodes in the United States for 1994. *J Nematol*, *31*(4S), 587-618.
15. Handoo, Z. A. (1998). Plant-parasitic nematodes. http://www.ars.usda.gov/Services/ docs.htm
16. Holterman, M., van der Wurff, A., van den Elsen, S., van Megen, H., Bongers, T., Holovachov, O. et al. (2006). Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown Clades. *Mol Biol Evol*, *23*(9), 1792-1800.
17. Tchesunov, A.V., and Riemann, F. (1995). Arctic sea ice nematodes (Monhysteroidea), with descriptions of *Cryonema crassum* gen. n., sp. n. and *C. tenue* sp. n. Nematologica *41*, 35–50.
18. Nussbaumer, A.D., Bright, M., Baranyi, C., Beisser, C.J., and Ott, J.A. (2004). Attachment mechanism in a highly specific association between ectosymbiotic bacteria and marine nematodes. Aqua. Microb. Ecol. *34*, 239–246.
19. Yushin, V. V., Mutsuhiro, Y., Spiridonov, S. (2007). Riders on the sperm: Sperm

dimorphism and spermatozeugmata in nematodes from the genus Steinernema (Rhabditida: Steinernematidae). *Nematology*, 9(1), 61-75.

20. White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci*, *314*(1165), 1-340.
21. Hinsch, K., & Zupanc, G. K. (2007). Generation and long-term persistence of new neurons in the adult zebrafish brain: a quantitative analysis. *Neuroscience*, *146*(2), 679-696.
22. Voronov, D. A., Panchin, Y. V., & Spiridonov, S. E. (1998). Nematode phylogeny and embryology. *Nature*, *395*(6697), 28.
23. Schierenberg, E. (2006). Embryological variation during nematode development. *WormBook*, 1-13.
24. Strome, S., & Wood, W. B. (1983). Generation of asymmetry and segregation of germ-line granules in early C. elegans embryos. *Cell*, *35*(1), 15-25.
25. Edgar, L. G., Wolf, N., & Wood, W. B. (1994). Early transcription in Caenorhabditis elegans embryos. *Development*, *120*(2), 443-451.
26. Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, *130*(5), 889-900.
27. Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development*, *136*(18), 3033-3042.
28. Maduro, M. F., & Rothman, J. H. (2002). Making worm guts: the gene regulatory network of the Caenorhabditis elegans endoderm. *Dev Biol*, *246*(1), 68-85.
29. Sulston J. (1988). In "The nematode *C. elegans*" (W. B. Wood ed.) pp123-155. Cold Spring

Harbor Laboratory Press, New York.

30. von Ehrenstein G., and Schierenberg E. (1980) in "Nematodes as Biological Models" Volume 1, Zuckerman B. M. (ed.), pp 2-68. Academic Press, New York
31. Wood W. B. (1988b). In "The nematode *C. elegans*" (W. B. Wood ed.) pp215-241. Cold Spring Harbor Laboratory Press, New York.
32. Boveri, T. (1899). Die Entwickelung von *Ascaris megalocephala* mit besonderer Rücksicht auf die Kernverhältnisse. In Festschrift für C. v. Kupffer. (Jena: Gustav Fischer Verlag), pp. 383–430.
33. Müller, H. (1903). Beitrag zur Embryonalentwickelung der *Ascaris megalocephala*. Zoologica *41*, 1–30.
34. Lahl, V., Halama, C., Schierenberg, E. (2003). Comparative and experimental embryogenesis of Plectidae (Nematoda). Dev. Genes Evol. *213*, 18–27.
35. Schierenberg, E. (2005). Unusual cleavage and gastrulation in a freshwater nematode: developmental and phylogenetic implications. Dev. Genes Evol. *215*, 103–108.
36. Malakhov, V.V. (1994). Nematodes, Structure, Development, Classification and Phylogeny. (Washington, DC: Smithsonian Institution Press).
37. Aleshin, V.V., Kedrova, O.S., Milyutina, I.A., Vladychenskaya, N.S., Petrov, N.B. (1998). Relationships among nematodes based on the analysis of 18S rRNA gene sequences: molecular evidence for monophyly of Chromadorian and Secernentean nematodes. Russ. J. Nematol. *6*, 175–184.
38. Voronov, D. A., & Panchin, Y. V. (1998). Cell lineage in marine nematode Enoplus brevis. *Development*, *125*(1), 143-150.
39. Drozdovskii, E. M. (1975). [Ovum cleavage in Eudorylaimus and Mesodorylaimus species

(Nematoda; Dorylaimida) and role of cleavage in determining the composition of nematode subclasses]. *Dokl Akad Nauk SSSR*, *222*(4), 1105-1108.

40. Goldstein, B. (2001). On the evolution of early development in the Nematoda. *Philos Trans R Soc Lond B Biol Sci*, *356*(1414), 1521-1531.
41. Boveri, T. (1887). Über die Differenzierung der Zellkerne während der Furchung des Eies von *Ascaris megalocephala*. Anat. Anzeiger *2*, 668–693.
42. Davis, A. H., Kidd, G. H., & Carter, C. E. (1979). Chromosome diminution in Ascaris suum. Two-fold increase of nucleosomal histone to DNA ratios during development. *Biochim Biophys Acta*, *565*(2), 315-325.
43. Nemetschke, L., Eberhardt, A. G., Hertzberg, H., & Streit, A. (2010). Genetics, chromatin diminution, and sex chromosome evolution in the parasitic nematode genus Strongyloides. *Curr Biol*, *20*(19), 1687-1696.
44. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, *282*(5396), 2012-2018.
45. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N. et al. (2003). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol*, *1*(2), E45.
46. Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J. et al. (2004). A transcriptomic analysis of the phylum Nematoda. *Nat Genet*, *36*(12), 1259-1267.
47. Zhao, S., Shetty, J., Hou, L., Delcher, A., Zhu, B., Osoegawa, K. et al. (2004). Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res*, *14*(10A), 1851-1860.
48. Kiontke, K., Gavin, N. P., Raynes, Y., Roehrig, C., Piano, F., & Fitch, D. H. (2004).

Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc Natl Acad Sci U S A*, *101*(24), 9003-9008.

49. Mitreva, M., Blaxter, M. L., Bird, D. M., & McCarter, J. P. (2005). Comparative genomics of nematodes. *Trends Genet*, *21*(10), 573-581.
50. Bateson W. (1894). *Materials for the Study of Variation*. New York, NY: Macmillan.
51. Aboobaker, A. A., & Blaxter, M. L. (2003). Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. *Curr Biol*, *13*(1), 37-40.
52. Kaya, H. K., Gaugler, R. (1993). Entomopathogenic nematodes. *Annu. Rev. Entomol.* 38:181-206.
53. Poinar, G. O. (1993). "Origins and phylogenetic relationships of the entomophilic rhabditis, Heterorhabditis and Steinernema". *Fundamental and Applied Nematology* 16(4): 333-338.
54. Adams, B. J.; Peat, S. M.; Dillman, A. R. Phylogeny and Evolution. In *Entomopathogenic Nematodes: Systematics, Phylogeny, and Bacterial Symbionts*. Nyugen, K. B., Hunt, D. J. Brill Leiden: Boston, 2007; Vol. 5, pp 693-713.
55. Akhurst, R.J. and G. B. Dunphy (1993). Tripartite interaction between symbiotically associated entomopathogenic bacteria, nematodes and their insect hosts. *Parasites and Pathogen of Insects*, 2, 1-23
56. Han, R., & Ehlers, R. U. (2000). Pathogenicity, development, and reproduction of Heterorhabditis bacteriophora and Steinernema carpocapsae under axenic in vivo conditions. *J Invertebr Pathol*, *75*(1), 55-58.
57. Martens, E. C., Heungens, K., & Goodrich-Blair, H. (2003). Early colonization events in the mutualistic association between Steinernema carpocapsae nematodes and Xenorhabdus nematophila bacteria. *J Bacteriol*, *185*(10), 3147-3154.

58. Li, J., Chen, G., Webster, J. M., & Czyzewska, E. (1995). Antimicrobial metabolites from a bacterial symbiont. *J Nat Prod*, *58*(7), 1081-1086.

59. McInerney, B. V., Taylor, W. C., Lacey, M. J., Akhurst, R. J., & Gregson, R. P. (1991). Biologically active metabolites from Xenorhabdus spp., Part 2. Benzopyran-1-one derivatives with gastroprotective activity. *J Nat Prod*, *54*(3), 785-795.

60. Campbell, J. F., Kaya, H. K. (2000). Influence of insect associated cues on the jumping behavior of entomopathogenic nematodes (Steinernema spp.). *Behaviour*, 137(5), 591-609.

61. Zenner, A. N., O'Callaghan, K. M., & Griffin, C. T. (2014). Lethal fighting in nematodes is dependent on developmental pathway: male-male fighting in the entomopathogenic nematode Steinernema longicaudum. *PLoS One*, *9*(2), e89385.

62. Dillman, A. R., Macchietto, M., Porter, C. F., Rogers, A., Williams, B., Antoshechkin, I. et al. (2015). Comparative genomics of Steinernema reveals deeply conserved gene regulatory networks. *Genome Biol*, *16*, 200.

63. Rose, L., & Gonczy, P. (2014). Polarity establishment, asymmetric division and segregation of fate determinants in early C. elegans embryos. *WormBook*, 1-43.

CHAPTER 2

Comparative genomics of *Steinernema* reveals deeply conserved gene regulatory networks

Abstract

Background: Parasitism is a major ecological niche for a variety of nematodes. Multiple nematode lineages have specialized as pathogens, including deadly parasites of insects that are used in biological control. We have sequenced and analyzed the draft genomes and transcriptomes of the entomopathogenic nematode *Steinernema carpocapsae* and four congeners (*S. scapterisci*, *S. monticolum*, *S. feltiae*, and *S. glaseri*).

Results: We used these genomes to establish phylogenetic relationships, explore gene conservation across species, and identify genes uniquely expanded in insect parasites. Protein domain analysis in *Steinernema* revealed a striking expansion of numerous putative parasitism genes, including certain protease and protease inhibitor families, as well as fatty acid- and retinol-binding proteins. Stage-specific gene expression of some of these expanded families further supports the notion that they are involved in insect parasitism by *Steinernema*. We show that sets of novel conserved non-coding regulatory motifs are associated with orthologous genes in *Steinernema* and *Caenorhabditis*.

Conclusions: We have identified a set of expanded gene families that are likely to be involved in parasitism. We have also identified a set of non-coding motifs associated with groups of orthologous genes in *Steinernema* and *Caenorhabditis* involved in neurogenesis and embryonic development that are likely part of conserved protein–DNA relationships shared between these two genera.

Introduction

Nematodes are remarkably adept at evolving parasitic lineages with animal-parasitic and plant-parasitic lineages arising many times independently throughout the phylum (Blaxter et al., 1998; van Megen et al., 2009). To increase our understanding of the evolution of parasitism, we sequenced five species within the insect-parasitic *Steinernema* (Nematoda: Steinernematidae), an intensely studied genus used for decades in biological control against agricultural insect pests and also as a model for animal parasites (Fig. 1a,b, Table 1) (Castelletto et al., 2014; Dillman et al., 2012; Kaya et al., 1993). Unlike most other sequenced nematodes, which are either harmful or seemingly innocuous to humans, steinernematids are beneficial to humans. *Steinernema* are considered insect pathogenic or entomopathogenic nematodes because of their ability to rapidly (24–48 h) kill an insect host (Kaya et al., 1993; Dillman et al., 2012; Gaugler et al., 1990). Entomopathogenic lineages have arisen independently at least three times among nematodes (Dillman et al., 2012). Their ability to kill insects is due in part to their mutualistic association with enterobacteria of the genus *Xenorhabdus*, which are vectored by the only free-living stage in the nematodes' life cycle, known as the infective juvenile (IJ) (Additional file 1: Fig. S1) (Kaya et al., 1993; Dillman et al. 2012). Once a suitable host is found, the IJs release the bacteria inside the host, where it grows and helps kill the host by septicemia. The bacteria and host tissues provide a food source for the nematodes to mature and reproduce inside the host cadaver. Once resources are depleted, the bacteria and a new generation of nematodes (IJs) re-establish their association and emerge from the host remains to search for a new host to infect (Kaya et al., 1993; Stock et al., 2008). The bivalent symbiosis (i.e., mutualism and parasitism/pathogenesis) in this tripartite system have made steinernematids (and their bacterial symbionts) an appealing model for understanding mutualism, parasitism, host-seeking, insect immune suppression, and

subterfuge (Castelletto et al., 2014; Stock et al., 2008; Castillo et al., 2011; Hallem et al., 2011). In addition to studying parasitism, sequencing five species within a genus (congeners) allowed us to leverage comparative analyses not only within *Steinernema* but among more distantly related taxa such as *Caenorhabditis elegans*. Comparative genomics is a powerful way to understand the complexity of the developmental programs contained within a genome (i.e., promoters, enhancers, transcription-factor binding sites, and the intricate gene regulatory networks that connect transcription factors to each other and their targets (Davidson et al., 2010)). Sequencing closely related organisms for comparative analyses can facilitate the identification of genus-specific gene family expansions and functional non-coding regions of genomes. For example, decoding the developmental programs embedded within the *C. elegans* genome has been challenging, but has benefited from the sequencing of additional congener genomes. The sequencing of the *C. briggsae* genome identified over 1,200 new *C. elegans* genes and helped correct the predicted exon structure for thousands of already annotated genes, but revealed relatively little about conserved non-coding elements (Stein et al., 2003).

Results and discussion

We sequenced and assembled the genomes of five *Steinernema* species (*S. carpocapsae*, *S. feltiae*, *S. glaseri*, *S. monticolum*, *S. scapterisci*) for comparative analysis (Fig. 1, Table 1, Additional file 1: Fig. S1). We focused on sequencing *S. carpocapsae* in greater depth than the others to use it as a representative for comparative analyses with other nematode genera. The other species were chosen based on their commercial availability, their evolutionary relationships, and their varied host specificities and foraging strategies. In addition, we sequenced and assembled de novo the mRNA of the IJ-stage of each species to aid in genome annotation. Additional RNA was collected for *S. carpocapsae, S. feltiae*, and *C. elegans* at the embryonic, first larval (L1), and young adult stages for a comparative analysis of gene expression, which is discussed below. The final genome assembly sizes ranged from 80 to 90 Mb and 28,000 to 36,000 genes (Fig. 1, Table 1, Additional file 1: Fig. S2, Additional file 2: Table S1) and *S. carpocapsae* was the best-assembled genome (scaffold N50 = 299 kb) with an estimated genome completeness of 98 % (Fig. 1, Additional file 2: Table S2). Detailed methods for assembly and annotation can be found in the "Methods" section.

Phylogenetic analysis

Although taxon selection clearly influences phylogenomic accuracy (Havird et al., 2010), sequencing the genomes of multiple species in the same genus should increase confidence in our ability to recover their evolutionary history (Kubatko et al., 2007; Nadler et al., 2006; Rokas et al., 2003; Zhao et al., 2013). Current best estimates place *Steinernema* in Holterman clade 10 and thus closely related to the sequenced nematode *Bursaphelenchus xylophilus*, a plant-parasite, and the free-living *Panagrellus redivivus* (Fig. 1a) (Blaxter et al., 1998; van Megen et al., 2006

;Holterman et al., 2006). Previous attempts to recover relationships among different species of *Steinernema* resulted in several poorly resolved/supported nodes, likely due to the limited number of molecular markers used and their homoplastic and/or plesiomorphic nature (Adams et al., 2007). We evaluated the relationships among the five *Steinernema* using a supermatrix of 3,885 strictly orthologous genes (1:1:1:1:1:1), with *P. redivivus* as our out-group taxon (Fig. 1b). The relationships we recovered are strongly supported but differ from previous hypotheses in that *S. monticolum*, which was chosen for sequencing based on its hypothesized close relationship to *S. carpocapsae* and *S. scapterisci* (Nadler et al., 2006), was more closely related to *S. feltiae* than any of the other nematodes in our analysis.

Gene orthology

The predicted proteome of an organism can highlight the conserved proteins shared with other species in its phylum and genus as well as the specializations that allow each species to adapt to its environment. The predicted proteome of 28,313 *S. carpocapsae* proteins was compared to the predicted proteomes of eight other nematode species and an insect out-group: *P. redivivus*, *C. elegans*, *Pristionchus pacificus*, *Meloidogyne hapla*, *Bursaphelenchus xylophilus*, *Brugia malayi*, *Ascaris suum*, *Trichinella spiralis*, and the parasitoid wasp *Nasonia vitripennis* (Fig. 2a) (C. elegans Sequencing Consortium, 1998; Dieterich et al., 2008; Ghedin et al., 2007; Jex et al., 2011; Kikuchi et al., 2011; Mitreva et al., 2011; Opperman et al., 2008; Srinivasan et al., 2013; Werren et al., 2010). The other nematodes used in this comparison included free-living (*C. elegans* and *P. redivivus)*, necromenic (*P. pacificus*), plant-parasitic (*M. hapla* and *B. xylophilus*), and vertebrate-parasitic species (*B. malayi*, *A. suum*, and *T. spiralis*) (Fig. 1a). Most of the predicted *S. carpocapsae* genes had homologs (BLASTp e-value cut-off: 10^{-5}) in one or

more species included in this analysis; 10,350 orthology clusters included 17,653 (62.3 %) *S. carpocapsae* proteins. A total of 266 of these clusters were found exclusively in nematodes. We found that 1,633 orthology clusters included at least one protein from each of the ten taxa analyzed, 486 of which were strictly conserved at a 1:1 ratio across all taxa (Additional file 3). While most molecular phylogenetic studies of nematodes rely on one or a few genes, this set of 486 highly conserved genes is a source of characters that could increase phylogenetic resolution in future analyses (van Megen et al., 2009; Holterman et al., 2006). In this analysis, there were 10,660 orphan *S. carpocapsae* proteins (37.7 % of the proteome) that did not cluster with any other proteins in the dataset, suggesting either that they are uniquely derived within *S. carpocapsae,* or that they have evolved such disparate primary sequences that they cannot be linked to their orthologs by sequence similarity alone. Protein orthology was also evaluated using the predicted protein sets for the five steinernematids sequenced and included either *C. elegans* or *P. redivivus* as out-group taxa. In these analyses the number of *S. carpocapsae* orphan proteins changed little, from 37.7 % in the phylum-wide analysis to 32.3 % or 32.4 % respectively (Fig. 2, Additional file 2: Table S1). Of the predicted *S. carpocapsae* genes, 80.5 % had at least partial RNA-seq transcript support (Additional file 2: Table S3). It is remarkable that these putative orphan proteins consistently included more than 30 % of the predicted proteome even when examining species within *Steinernema,* whereas a detailed genomic analysis of 12 species of *Drosophila* revealed the percentage of orphan proteins ranges from 14 % to 27 % in that genus (Clark et al., 2007).

Protein family domain abundance

We then analyzed the predicted protein domains to understand which gene families have undergone expansions in *Steinernema* that may have contributed to adaptation to a parasitic lifestyle. The *S. carpocapsae* proteome was predicted to have a total of 17,518 Pfam domains from 3,256 different Pfam accession categories. The relative Pfam domain abundances of the *S. carpocapsae* genome were compared to those in the parasitic *Bursaphelenchus xylophilus*, *Brugia malayi*, and *Ascaris suum*, as well as the free-living *C. elegans* (Fig. 2b, Additional file 1: Figs S3–S6). Overall, most Pfam domains were detected at similar levels in both genomes, with some notable exceptions. For example, while most transcription factor domains showed similar prevalence in both genomes, we found the expected enrichment of C4 zinc fingers in *C. elegans* that are associated with nuclear hormone receptors, as well as a novel three-fold enrichment of the alcohol dehydrogenase transcription factor Myb/SANT-like domain in *S. carpocapsae* (Fig. 2b). The *S. carpocapsae* genome appears to be enriched in proteases, protease inhibitors, certain families of G protein-coupled receptors (GPCRs) , and fatty acid- and retinol-binding (FAR) proteins, among others (Fig. 2b, Additional file 1: Figs S3–S6). The abundance of predicted Pfam domains from *S. carpocapsae* was compared with the other four *Steinernema* species we sequenced (Fig. 2, Additional file 1: Figs S7–S10). The domain richness of certain types of proteases, protease inhibitors, and certain families of GPCRs varied widely between the different species of *Steinernema*, though some enrichments were shared, such as the greater abundance of certain protease and protease inhibitor families, and FAR proteins, which appeared in all *Steinernema* species and are discussed separately below (Fig. 2c).

Putative parasitism genes: proteases and protease inhibitors

Proteases (peptidases) are involved in a wide variety of biological functions including development, digestion, signal transduction, and immune responses (Kanost et al., 2005). Of particular relevance in these genomic analyses is the role proteases play in parasitism, such as tissue penetration and immune suppression or evasion (Abuhatab et al., 1995; Balasubramanian et al., 2009; McKerrow et al., 1990; Toubarro et al., 2009). A total of 654 peptidases were identified in the *S. carpocapsae* genome (Fig. 2, Additional file 2: Table S4). These can be broken down into five key classes based on the chemical groups that function in catalysis: aspartic (6.3 %), cysteine (19 %), metallo- (32.7 %), serine (37.6 %), and threonine (4.1 %). Because steinernematids can be lethal even without their pathogenic symbionts (Burman et al., 1982; Dunphy et al., 1985; Han et al., 2000) and proteolytic activity is higher in the excreted-secreted products in more virulent strains (Han et al., 2000; Simóes et al., 2000), proteases were among the first products examined in relation to the lethality of *Steinernema* nematodes and have been suggested to be actively pumped into host tissues by parasitic nematodes (Balasubramanian et al., 2009; Toubarro et al., 2009; James et al., 2004; Trap et al., 2000; Zang et al., 2001). Steinernematids have more predicted proteases (Fig. 2, Additional file 2: Table S4) than any other nematode sequenced to date. This correlates with the remarkably broad host ranges of many *Steinernema* species, which can infect multiple species across many insect orders in some cases, whereas other parasitic nematodes have more restricted or specialized host ranges. Breaking the proteases into subclasses highlights species-specific expansions of serine and metalloproteases among *Steinernema* species. However, the abundance of aspartic, cysteine, and threonine proteases is relatively similar across nematodes (Fig. 2c, Additional file 2: Tables S4–S7). The serine and metalloproteases are the most highly represented families in nematode

excreted-secreted products, suggesting that they play a role in parasitism (Trap et al., 2000). We found *Steinernema*-specific expansions of chymotrypsin-like (S01A), Lon-A-like (S16), and signal peptidase I-like (S26A) serine proteases and expansions of the astacin (M12A), carboxypeptidase A1-like (M14A), and the pitrilysin (M16A) metalloproteases. Whereas chymotrypsin-like and carboxypeptidase A1-like proteases were expanded in all five of the *Steinernema* spp. when compared to other nematodes, other proteases such as the Lon-A-like, signal peptidase I-like, astacin, and pitrilysin proteases were only expanded in certain species (Fig. 2c, Additional file 2: Tables S8–S11). These expansions represent putative parasitism genes and may affect the host-range and specificity of these species, influencing their ability to infect and avoid the immune response of certain potential host species. Some proteases in these expanded families have characterized roles in parasitism in *Steinernema*. For example, an S01A chymotrypsin-like protease from *S. carpocapsae* has increased expression in IJs exposed to waxworm hemolymph and suppresses waxworm prophenoloxidase activity and immune encapsulation *in vitro* (Balasubramanian et al., 2009). Additional biochemical and molecular studies are needed to understand immune suppression and evasion by steinernematids and the role proteases play in these processes.

Previous work has shown a functional role for several proteases in the parasitism of insects by entomopathogenic nematodes. For example, Toubarro et al. identified an *S. carpocapsae* serine protease that hydrolyzes extracellular matrix proteins and induces apoptosis of insect cells (Toubarro et al., 2009). Two other *S. carpocapsae* serine proteases are involved in immune subversion by inhibiting insect prophenoloxidase activity and disrupting cellular encapsulation by the insect immune response (Balasubramanian et al., 2009; Balasubramanian et al., 2010).. Also, an *S. carpocapsae* astacin is upregulated in IJs upon infection of an insect host,

suggesting a role in the infection process (Jing et al., 2010). Our findings further support the notion that certain families of proteases play a role in parasitism and may have shaped niche partitioning among the many species of insect parasites.

The virulence of parasitic nematodes is heavily influenced by not only proteases but also protease inhibitors (Zang et al., 2001; Milstone et al., 2000). In addition to the expansion of proteases, the *Steinernema* genomes show large expansions of several specific protease inhibitor families, such as the I4 serine protease inhibitor (serpin) family, the I8 chymotrypsin/elastase inhibitor family, and the I63 pappalysin-1 inhibitor family (Fig. 2c, Additional file 2: Table S12) (Rawlings et al., 2012). This genus-specific expansion in *Steinernema* species and the known role of many protease inhibitors in parasitism, particularly serpins (reviewed by Molehin *et al.* 2012), suggests that these protease inhibitors are putative parasitism genes likely used by steinernematids to successfully infect hosts. We examine stage-specific gene expression of some of these putative parasitism genes below (S26 proteases and I63 protease inhibitors). Future investigations of the expression context and biochemical function of the expanded proteases and protease inhibitors identified here in these parasitic nematodes might reveal that they facilitate the parasitism of insects and that the various expansions and retractions of these families among steinernematids influence host range and specificity.

Putative parasitism genes: fatty acid- and retinol-binding proteins

The fatty acid- and retinoid-binding protein (FAR) gene family represents another dramatic case of genus-wide expansion in *Steinernema* (Fig. 2d, Additional file 1: Fig. S11, Additional file 2: Table S13). *Steinernema* species have between 38 and 54 FAR proteins compared to 19 in *P. pacificus* and fewer in the other nematodes we examined (Additional file 1:

Fig. S11, Additional file 2: Table S13). FAR proteins are a family of lipid-binding proteins that have high binding affinities for fatty acids, retinol, and retinoic acid and are unique to nematodes (Kennedy et al., 2013). They are important in the growth, development, and reproduction of *C. elegans*, which, like most if not all nematodes, is auxotrophic for sterols. However, FAR proteins were originally discovered in vertebrate-parasitic nematodes, where, in addition to their role in growth and development, they are thought to play a key role in parasitism by functioning in the sequestration of host retinoids as well as by contributing to immune evasion or suppression, though their exact role is not understood (Kennedy et al., 2013; Garofalo et al., 2002). Although parasitism arose independently multiple times among nematodes, FAR proteins have been implicated in the parasitism of plants, invertebrates, and vertebrates across all of the parasitic lineages, suggesting that this protein family is particularly important to parasitism (Fig. 1b) (Kennedy et al., 2013; Hao et al., 2010; Iberkleid et al., 2013). We examine the stage-specific expression of FARs and explore their genome architecture below. While the function of these proteins in parasitism remains to be shown, one possibility is that they interact with eicosanoids—fatty acids involved in immunological signaling in plants, mammals, and insects (Campos et al., 2014; Lawrence et al., 2002; Stanley et al., 2006). Inhibiting eicosanoid biosynthesis has been shown to reduce the melanotic encapsulation response of insects, which is thought to be insects' primary defense against nematode parasites (Castillo et al., 2011; Carton et al., 2002). For example, *Xenorhabdus nematophila*, the insect-pathogenic symbiont of *S. carpocapsae*, has been shown to dampen the host insect immune response by inhibiting eicosanoid synthesis in infected insects, increasing the likelihood of a successful infection by *S. carpocapsae* (Park and Kim, 2003; Park et al., 2003). Thus inhibiting eicosanoid biosynthesis in hosts is one way that parasitic nematodes may suppress host immunity.

Differential gene expression analysis

We collected mRNA from the early embryonic, L1, IJ, and young adult stages of *S. carpocapsae* in biological replicates for differential expression analysis. A total of 4,557 genes were differentially expressed (DE) in *S. carpocapsae* across the time course [false discovery rate (FDR) < 1×10^{-5} and fold changes > 4×] (Fig. 3a, Additional file 2: Table S14, Additional file 4). Gene Ontology (GO) analysis of the DE stage-specific gene sets revealed enrichment for mitosis-related GO terms (1,618 genes) in the early embryonic stage. This agrees with what has been observed in *C. elegans*, for which the majority of cell divisions occur during the first half of embryogenesis (Sulston et al., 1983). DE L1 genes (954 genes) were enriched for GO terms involved in feeding and sensation, neuronal cell fate, and muscle contraction. While muscle contraction should be important for all post-embryonic stages, these particular functions might be overrepresented in the L1 stage because the cells that carry out these functions make up a greater proportion of the body mass of the organism at this stage relative to other stages. DE genes in all the post-embryonic developmental stages were associated with ribosomal constituents, translation, and growth (201 genes), reflecting the dependence of early embryos on maternal ribosomes and other translation machinery. Moreover, while cellular division occurs primarily during embryonic development and during portions of larval stages (Sulston and Horvitz, 1977), cellular growth of particular cell types occurs primarily over the duration of each developmental stage. These results show that our stage-specific gene sets capture the biologically meaningful processes occurring during these developmental stages and likely reflect processes essential for *S. carpocapsae* development. We also investigated the similarity of transcript isoform expression during development in *S. carpocapsae* and *S. feltiae* and found that a large fraction of isoform pairs, 1,377 out of 3,202 (43 %) in *S. carpocapsae* (Additional file 5), and

1,189 out of 2,333 pairs (51 %) in *S. feltiae* have diverged in their expression during development (Additional file 1: Fig. S12). We further used our data to examine the stage-specific expression of the FAR proteins, the S26 proteases, and the I63 protease inhibitor family (Fig. 4, Additional file 1: Fig. S13). In each of these protein families, a single ortholog was very highly expressed and dominated the other orthologs, and many of the orthologs had distinct stage-specific expression in the two species. We identified DE genes in each of these categories in different *S. carpocapsae* developmental stages (Fig. 4a,b), suggesting further specializations in parasitism.

Gene expression conservation across species

The expression of orthologous genes during development is known to diverge (Sinha et al., 2012; Wittkopp et al., 2012). In order to identify genes with conserved patterns of stage-specific expression across closely and more distantly related species, mRNA from the corresponding embryonic, L1, IJ, and adult stages of *S. feltiae* and *C. elegans* was collected and sequenced for comparison to *S. carpocapsae* (Fig. 3, Additional file 2: Table S14). We limited our analysis to 5,569 1:1:1 orthologs present in all three species to avoid the complications of divergent expression due to gene duplications. We used two methods for determining conserved stage-specific ortholog expression. The first method binarized stage gene expression values using a flexible threshold to sort genes into stage-specific sets. We used this method to quantify the number of orthologs that are “on” and “off” in the same developmental stages between species. We found that 79.3 % (4,416/5,569) of these orthologs had conserved expression between *S. carpocapsae* and *S. feltiae*, whereas pairwise comparisons of the expression of each *Steinernema* species to *C. elegans* showed lower overall conservation of stage-specific expression of 61–63 %

(3,504/5,569 and 3,432/5,569) (Additional file 1: Fig. S14). Nevertheless, given that the steinernematids are phylogenetically distant from *C. elegans* yet share expression of more than two-thirds of their 1:1:1 orthologous genes, these results suggest that gene expression of this core set of unduplicated genes is highly conserved among nematodes. In a separate analysis, we treated the expression of each ortholog during development in each species as a vector and calculated their cosine similarities to address whether the ortholog expression profiles parallel each other during development. We found that 1,441 out of 5,569 orthologs (25.8 %) had a conserved pattern of stage-specific expression (ortholog expression similarity > 0.95) between *S. carpocapsae* and *S. feltiae* (Fig. 3b, Additional file 6), whereas there was more divergence with *C. elegans*. Only 541 (9.7 %) orthologs were conserved in stage-specific expression between *C. elegans* and *S. carpocapsae* and 490 (8.7 %) between *C. elegans* and *S. feltiae* when all developmental stages were considered.

Using the stage-specific gene expression data, we determined the gene expression levels of 41 FARs, as well as the expression levels of a family of serine proteases and protease inhibitors (S26 and I63) that may play a role in the parasitic lifestyle of *S. carpocapsae*. We found that sets of the I63 protease inhibitors were expressed at particular post-embryonic developmental stages, with the highest expression levels occurring in the IJ (839.8 FPKM) and adult stage (239.4 FPKM) (Fig. 4a). These are the stages most important in the successful infection of an insect host and these expression data support the notion that I63 protease inhibitors are important for *S. carpocapsae* parasitism. However, most of the S26 proteases (14/17 proteases) were expressed primarily in the embryonic stage, suggesting that they are involved in development rather than the parasitism of insects by *Steinernema* (Additional file 1: Fig. S13). Additionally, we found that 39 of 41 FAR genes were primarily expressed during the

post-embryonic stages (Fig. 4b), and that about half of these genes appeared in clusters in the genome sequence. Of the eight *C. elegans* FAR genes, only *far*-1 was conserved in the steinernematids. This gene is reported as having highest expression in L3 *C. elegans* worms (Garofalo et al., 2003). We confirmed this, seeing high expression in the dauer and L1 stages (Additional file 1: Fig. S15). Among the stages we tested, *Steinernema far-1* orthologs had highest expression in L1 (Additional file 1: Fig. S16), suggesting that they function in development and not parasitism, but this remains to be tested.

Genome conservation and synteny analysis

The evolution and conservation of non-coding regions and their relationship to gene expression remains an open problem, with the central premise of comparative genomics being that conservation is one signature of potential function and functional linkage of elements with genes. We therefore aligned the sequences of the five *Steinernema* genomes globally to find such linkages and to reveal patterns of evolution in syntenic gene clusters. A genome-wide analysis of syntenic 1:1 orthologs of *S. carpocapsae* with each of the four congeners we sequenced revealed that the most closely related species pair, *S. carpocapsae* and *S. scapterisci,* had the most syntenic 1:1 orthologs, with 11,272 of 12,395 (90.9 %) 1:1 orthologs in synteny in scaffolds containing at least two syntenic orthologs, and 6,576 of 12,395 (53.0 %) 1:1 orthologs in synteny in scaffolds with ten or more syntenic 1:1 orthologs (Additional file 2: Table S15). However, the greatest stretch of syntenic 1:1 orthologs was between *S. carpocapsae* and *S. feltiae*, with 191 orthologous genes spanning a distance of 878 kb in *S. carpocapsae* and 794 kb in *S. feltiae*, which is a rather unexpected finding given the better assembly of *S. scapterisci* (scaffold N50 = 90,783 bp) compared to *S. feltiae* (scaffold N50 = 47,472 bp) (Table 1, Additional file 2: Tables

S15 and S16). A local analysis of synteny was done to investigate two noteworthy sets of genes. The first set of genes was the nematode Hox cluster that is quickly evolving in all nematodes (Aboobaker and Blaxter, 2003; Aboobaker and Blaxter, 2010). All of the core nematode Hox genes (*ceh-13*, *lin-39*, *mab-5*, *egl-5)* were found in most of the *Steinernema* assemblies (Fig. 5, Additional file 1: Fig. S17A), and an expansion was identified in the anterior portion (*ceh-13*, *lin-39*) of the Hox cluster, where the gap between *ceh-13* and *lin-39* is 19 kb in *C. elegans* and has expanded to 35–43 kb in several of the *Steinernema* genomes considered in this study (Fig. 5, Additional file 1: Fig. S17A,B). Also, approximately 15 expressed genes have become embedded between *ceh-13* and *lin-39* in *Steinernema* genomes*,* the 1:1 orthologs of which are not present anywhere near the Hox genes or each other in *C. elegans*, suggesting that the distance between Hox genes in the cluster in *Steinernema* is in the process of expanding (Fig. 5, Additional file 1: Fig. S17C–E).

The second set of genes we investigated was the family of FAR genes in *Steinernema*. We found a total of 22 out of 41 FAR genes in synteny across three distinct syntenic clusters in *S. carpocapsae*. By examining the location of the 1:1 orthologs of these genes in the other *Steinernema* species, we found that the majority of these orthologs are also syntenic in *S. scapterisci* (13/17 1:1 orthologs) and *S. feltiae* (10/12 1:1 orthologs) and that their expression during development is also conserved across the *Steinernema* species (Additional file 1: Fig. S18). Interestingly, we also saw that the most highly expressed FAR in *S. carpocapsae* has five paralogs in *S. feltiae*, with one dramatically changing its expression pattern from adult to embryonic stage, which suggests that this family is undergoing further rapid functional evolution within *Steinernema* (Fig. 4c).

Conserved non-coding networks

Non-coding cis-regulatory elements bound by transcription factors control the expression of their associated genes; two of the major goals of comparative genomics are the discovery of these elements and of the gene regulatory networks encoded by these tshared elements. We expect that genes with conserved gene expression profiles would share conserved cis-regulatory elements. Several studies of the evolution of gene expression have shown that cis-regulatory changes represent a major component (reviewed in Necsulea et al., 2012). In addition, rapid "re-wiring" of gene regulatory networks due to site turnover even between relatively closely related species in mammals and flies makes it difficult to find these cis-regulatory elements using global alignments alone (reviewed in Villar et al. 2014). Our previous experience with the small amount of non-coding sequence alignment between two distantly related species within the same genus, *C. elegans* and *C. angaria*, suggested that we would find very little directly alignable non-coding sequence between two distant genera (Mortazavi et al., 2010). We therefore postulated that, while the sets of orthologs conserved in stage-specific gene expression during *Steinernema* and *Caenorhabditis* development (Fig. 3b) are likely to be regulated by shared sets of non-coding, cis-regulatory elements, we would need to use a strategy that leverages non-coding alignability within a genus but does not require it for comparison with orthologs in a more distant genus such as *Caenorhabditis*. We filtered any conserved sequences that overlapped either gene models or transcripts assembled from our RNA-seq data in *S. carpocapsae* (Additional file 7). We found that 14.8 Mb (17.2 %) of the *S. carpocapsae* genome comprises conserved coding sequence while a further 4.5 Mb (5.2 %) comprises conserved non-coding sequence (Additional file 1: Fig. S19A). We then searched for novel regulatory motifs around nine sets of *Steinernema* orthologs with conserved expression patterns between *S. carpocapsae* and *S. feltiae* (Fig. 3c, Additional

file 1: Fig. S19B), and found 30 non-redundant motifs (Additional file 2: Table S17, Additional file 8. 24 of which matched the sequences of one of more motifs from the WormBase database (p-value < $1e^{-4}$ and e-value < 0.5) (Additional file 2: Table S18). All 30 of these motifs were mapped to the conserved non-coding regions in *S. carpocapsae* and *C. elegans* (from multiple sequence alignment of seven *Caenorhabditis* species, UCSC), revealing that they are enriched in the neighborhood of genes involved in the same biological processes (GO terms) (Fig. 6a, Additional files 9, 10). We found that the shared enriched GO terms that also involved a high percentage of 1:1 orthologs between the two species were related to processes such as neurogenesis, axonogenesis, embryogenesis, and muscle development. We further restricted ourselves to orthologous genes in *S. carpocapsae* and *C. elegans* that shared the same motifs and built three representative subnetworks of motifs-to-genes based on these GO enrichments (Additional file 2: Table S19). These networks revealed conserved associations between regulatory motifs and their target genes between the two species for genes in the core of neurogenesis/axonogenesis, embryogenesis, and muscle development (Fig. 6a–c, Additional file 1: Fig. S20). In particular, 25 regulatory motifs (degree ≥ 5) potentially regulate 92 neurogenesis genes whereas 16 overlapping regulatory motifs regulate 25 muscle development-related target genes in both *C. elegans* and *S. carpocapsae* (Fig. 6b,c, Additional files 11, 12, and 13). In order to verify that the motif-associated GO term enrichments we obtained were not due to chance, we created 100 randomized GO term sets by shuffling all of the annotated *S. carpocapsae* gene GO terms. We reassigned the GO term sets to new genes, and ran all 30 motif-associated gene sets through a Fisher's exact test using these randomized GO sets (30 motifs × 100 randomizations = 3,000 Fisher's exact tests in total). We were unable to recover GO term enrichments for any of the GO terms that comprised the neuronal, embryo, or muscle networks for any of our motifs

using randomized shuffling (0/3,000, FDR < 0.05), suggesting that the GO enrichments we identified are meaningful.

Multiple motifs from the same networks clustered together in or near some of the orthologous target genes of *S. carpocapsae* and *C. elegans*. Some of these motif clusters showed conserved order and position, whereas others showed variation in order only, position only, or variation in both between species (Fig. 6d, Additional file 1: Fig. S21). Comparative analysis of the *Steinernema* congeners led to the identification of these conserved motifs. We found them conserved in *C. elegans* and enriched near genes influencing similar biological processes in a distantly related genus. This finding suggests that they are under evolutionary selection, although their functionality remains to be tested.

Conclusion

The sequencing of multiple species of *Steinernema* enabled us to identify gene family expansions that are consistent with and likely important to the particular biology of these species as parasites; to generate new hypotheses about genes likely to be important in parasitism; to explore their genealogy more deeply than ever before and refine our understanding of their relationships to each other as well as define other phylogenetic markers that could be used in subsequent analyses; to identify stage-specific enrichment of functional gene classes; to demonstrate that the differential expression of stage-specific genes is influenced by phylogeny; to explore the evolution of the developmental control genes in the Hox gene cluster and diagnose expansion and rapid evolution of this cluster; and to identify previously unknown conserved non-coding regulatory motifs that regulate similar biological processes in distantly related organisms (Dillman et al., 2012). Our results point to a core set of conserved motifs, functioning in both *C.*

elegans and *S. carpocapsae,* that regulate similar biological processes key to proper nematode development across vast phylogenetic distance. These motifs are not detectable from direct sequence alignment between *Caenorhabditis* and *Steinernema* but can be found when analyzing genus-level conservation and using conserved gene expression and gene-motif association between orthologs. Further analysis will be required to assess whether these motifs form a phylum-wide core kernel of regulatory relationships or are restricted to the last common nematode ancestor of these two genera.

Figure legends

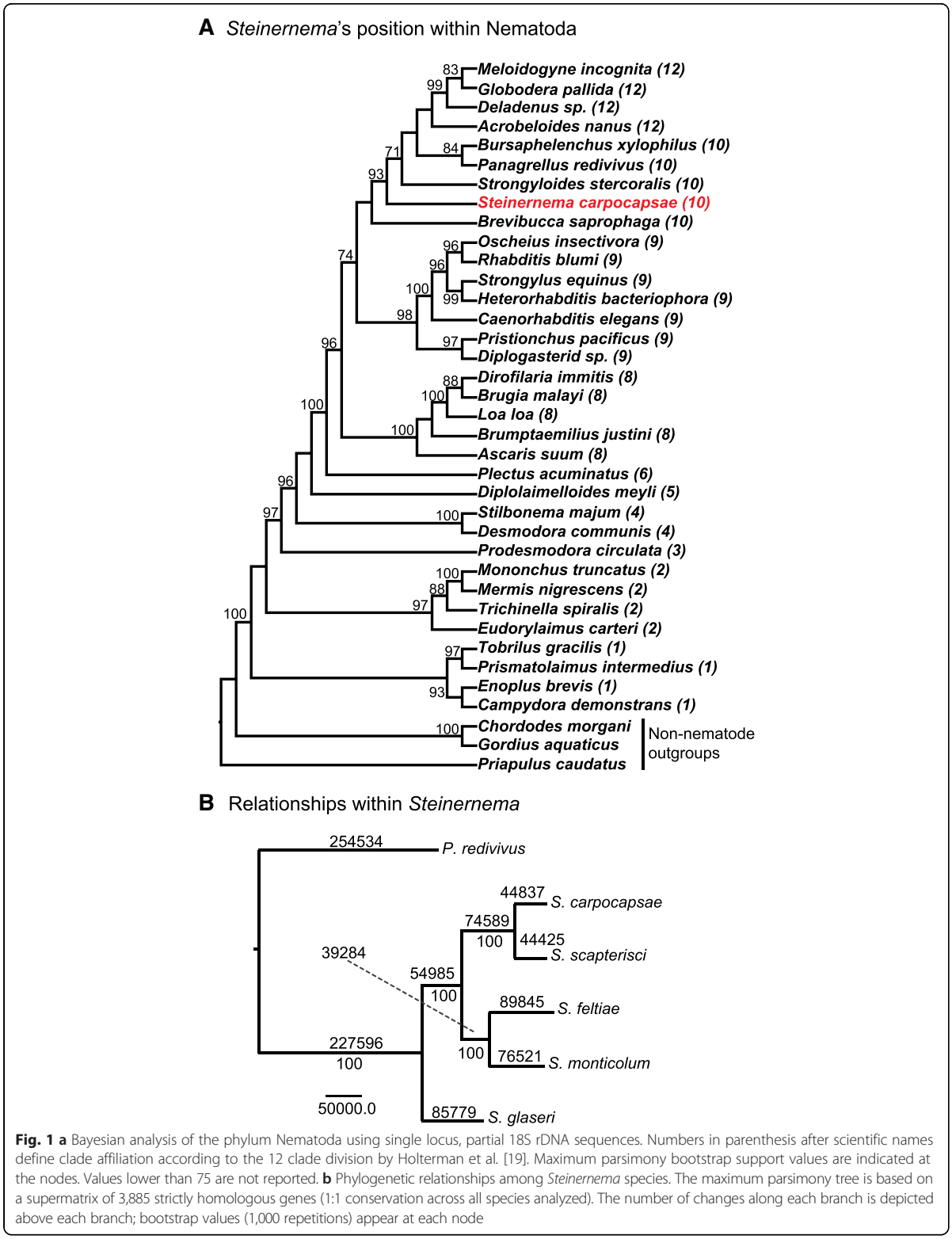
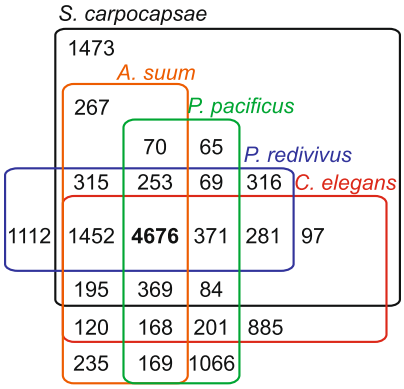


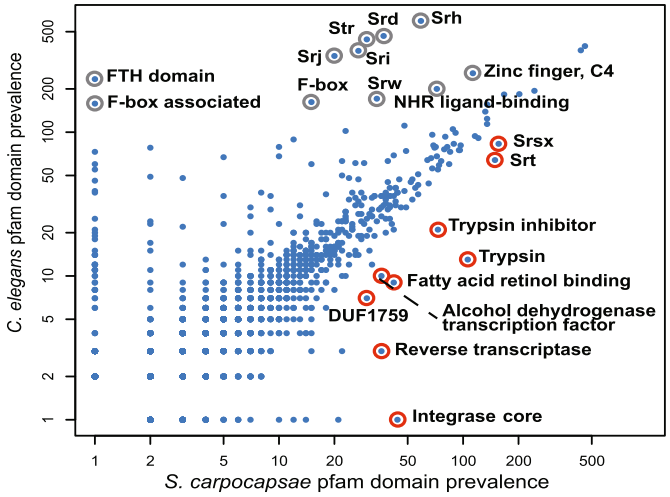
Fig. 1 a Bayesian analysis of the phylum Nematoda using single locus, partial 18S rDNA sequences. Numbers in parenthesis after scientific names define clade affiliation according to the 12 clade division by Holterman et al. [19]. Maximum parsimony bootstrap support values are indicated at the nodes. Values lower than 75 are not reported. **b** Phylogenetic relationships among *Steinernema* species. The maximum parsimony tree is based on a supermatrix of 3,885 strictly homologous genes (1:1 conservation across all species analyzed). The number of changes along each branch is depicted above each branch; bootstrap values (1,000 repetitions) appear at each node

(See figure on previous page.)
Fig. 2 a Gene orthology clusters among five sequenced species of nematodes, with 4,676 orthologous clusters being shared among all five species and 1,473 clusters being unique to *S. carpocapsae*. **b** The abundance of Pfam protein family domains in the *C. elegans* and *S. carpocapsae* genomes. The nine most enriched Pfam domains (biggest absolute difference) in *S. carpocapsae* relative to *C. elegans* are highlighted in *red* while the eleven most enriched Pfam domains in *C. elegans* relative to *S. carpocapsae* are highlighted in *gray*. **c** Select Pfam domains that are enriched in the sequenced steinernematids compared to other nematode species. **d** Protein neighbor-joining tree of the fatty acid- and retinol-binding proteins in nematodes. Monophyletic protein clades with at least one protein from each of the five *Steinernema* spp. are highlighted in *blue*. This figure illustrates both the abundance and diversity of FAR proteins among steinernematids. *EPN* entomopathogenic nematodes, *FAR* fatty acid- and retinol-binding proteins

A Gene orthology among nematodes



B Protein family domain prevalence in *C. elegans* compared to *S. carpocapsae*



C Select protein domain prevalence

Protein domains	Scar	Ssca	Sfel	Sgla	Smon	Bxyl	Pred	Cele	Ppac
Cullins	19	14	46	28	21	9	16	7	7
FAR Proteins	41	42	43	54	38	9	5	8	19
M12A Proteases	37	54	103	106	78	25	37	39	53
M14A Proteases	26	23	19	16	24	10	12	9	9
S01A Proteases	82	127	98	94	119	4	57	5	26
S26A Proteases	18	43	12	4	9	2	2	2	1
I63 Protease inhibitors	36	53	39	25	36	0	6	8	5

D Fatty acid- and retinoid-binding protein gene tree

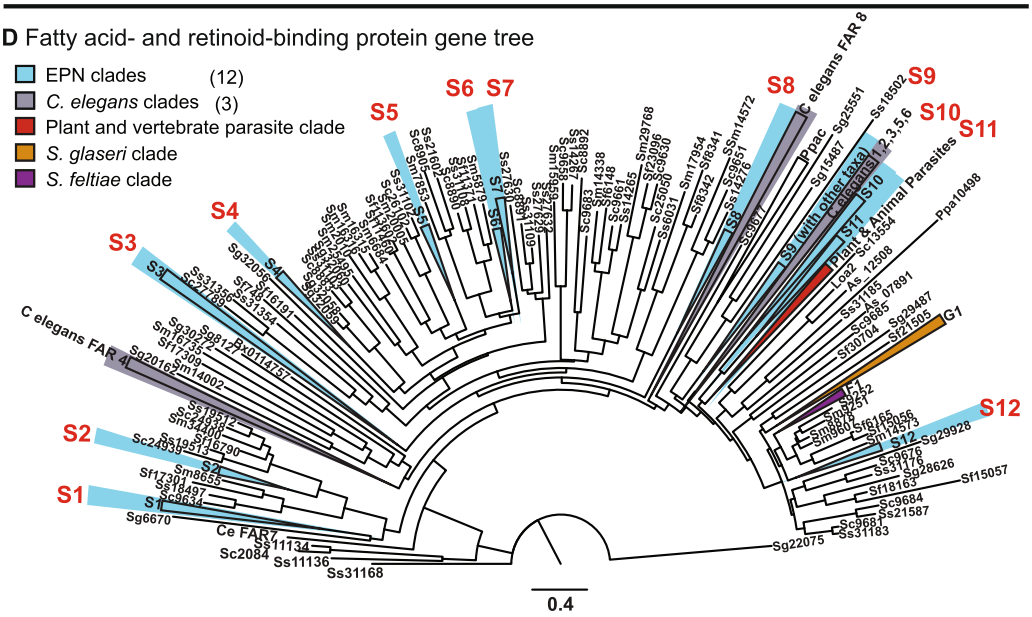


Fig. 2 (See legend on next page.)

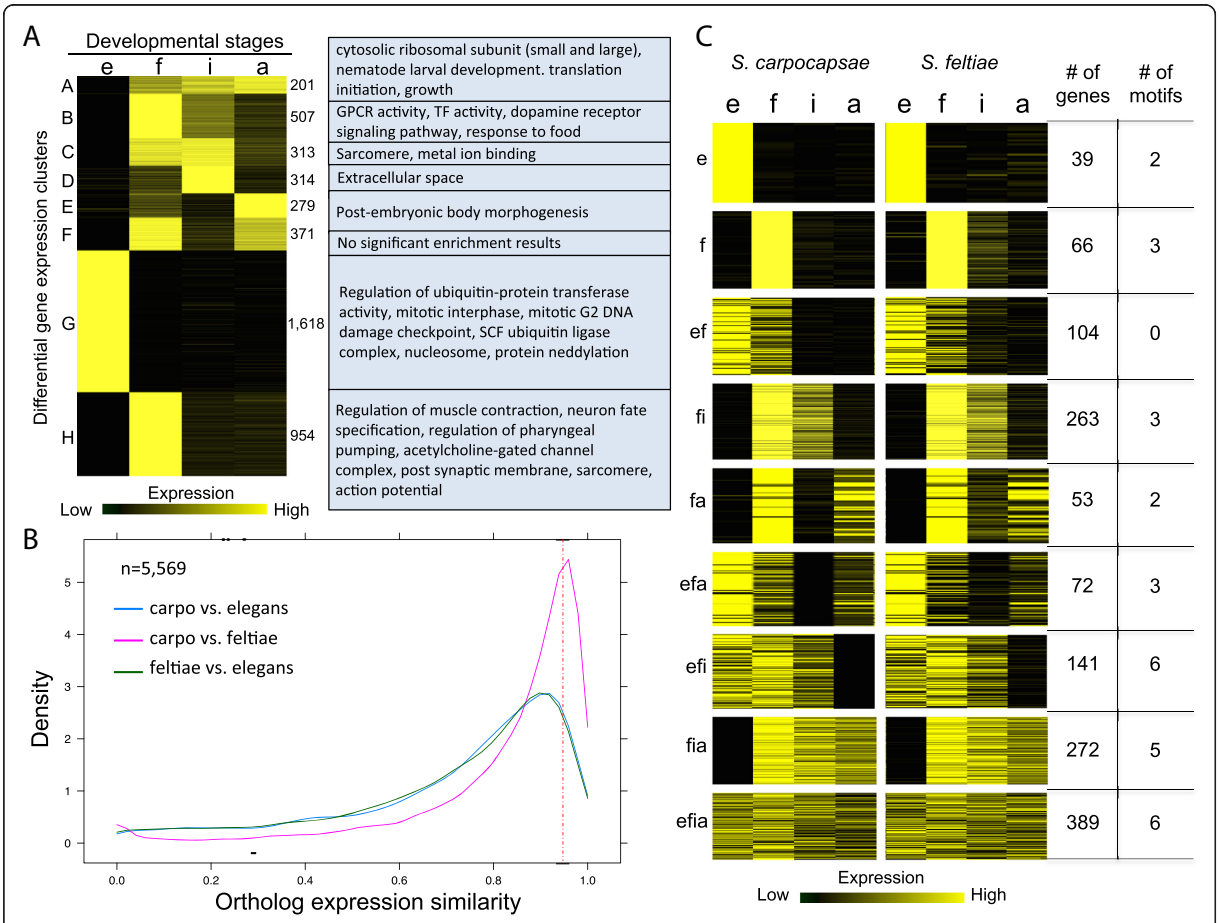


Fig. 3 a Heat map of 4,557 differentially expressed (DE) genes (FDR < 1×10^{-5}, fold change > 4×) during *S. carpocapsae* development. Gene Ontology term enrichment analysis was performed on the DE gene sets with Blast2GO (Fisher's exact test, FDR < 0.01). Gene expression for each stage for each gene was scaled so that the total expression of the row sums to 1. **b** Plot showing the distribution of gene expression profile similarities for 5,569 1:1:1 orthologs between species pairs during development. Ortholog expression (TPM) during development for each species was treated as a vector, and ortholog expression similarity was determined by calculating the cosine similarity of the vectors, where 1 corresponds to identical expression profiles, and 0 corresponds to divergent expression profiles. Orthologs with conserved stage-specific expression profiles have similarity measures > 0.95. **c** Heat map showing the ortholog expression profiles of the conserved stage-specific orthologs (cosine similarity > 0.95) in (b) in *S. carpocapsae* and *S. feltiae*. Gene expression is scaled so that the total expression across a row sums to 1 as in (a). The number of genes in each gene set and the number of significant non-redundant motifs that were derived from each gene set are shown to the *right*. *e* embryonic, *f* first larval, *i* infective juvenile, *a* adult developmental stages

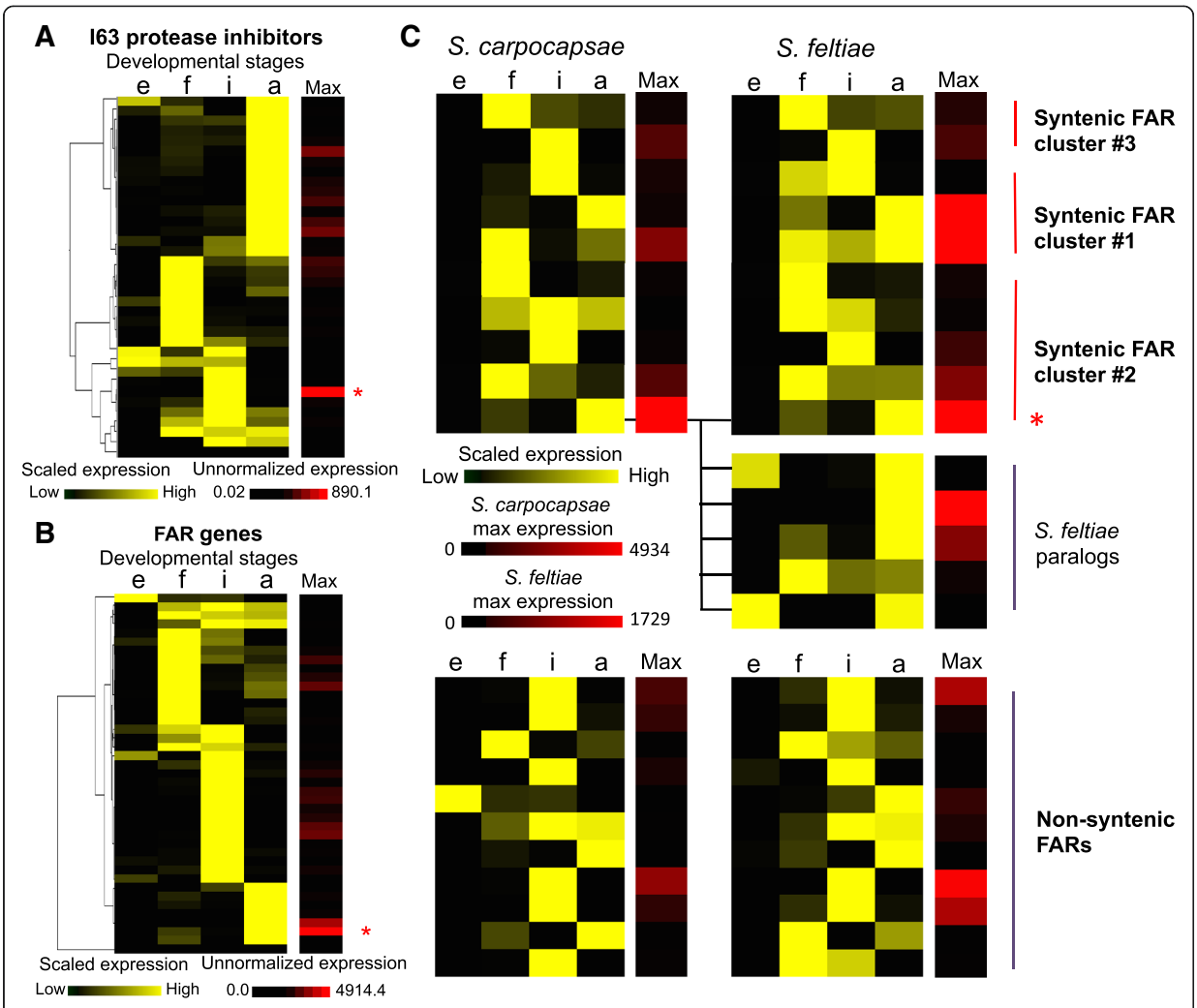


Fig. 4 Stage-specific gene expression of FAR proteins and protease inhibitors in *S. carpocapsae*. Heat map of the scaled gene expression of **a** 37 I63 protease inhibitors, and **b** 41 FAR proteins in *S. carpocapsae*. Expression of each gene was scaled between 0 (minimum expression, *black*) and 1 (maximum expression, *yellow*) across the developmental stages, and hierarchically clustered based on the expression profiles. The *Max* column shows each gene's maximum non-normalized expression value (FPKM) that was achieved during the time course. Genes that have high expression are in *red*, while genes that have low expression are in *black*. The *asterisk* indicates the gene that has the highest expression. **c** Gene expression of 22 1:1 orthologous FAR genes across three syntenic FAR clusters and the non-syntenic FARs were compared between *S. carpocapsae* and *S. feltiae*. For each FAR ortholog, gene expression was scaled between 0 (minimum expression) and 1 (maximum expression) across the developmental stages. The *Max* column shows the maximum gene expression levels of the FARs in *S. carpocapsae* and *S. feltiae*. Five *S. feltiae* paralogs of the bottom-most ortholog in FAR syntenic cluster #2 are shown separately beneath the heat map. *e* embryonic, *f* first larval, *i* infective juvenile, *a* adult developmental stages

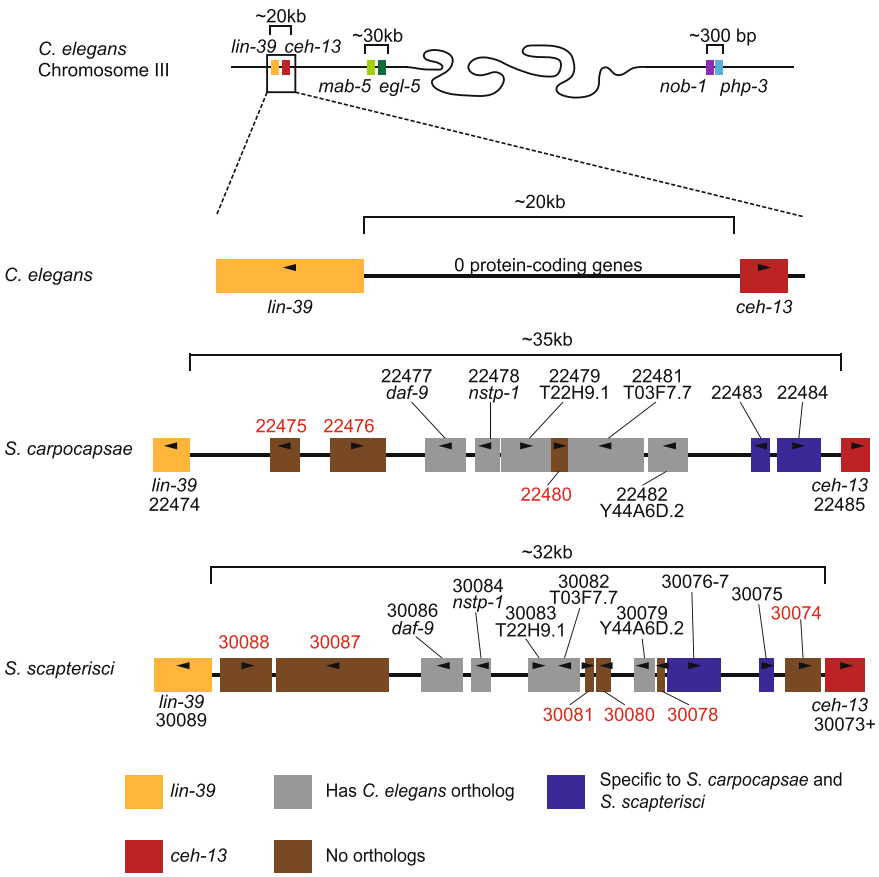


Fig. 5 Hox cluster architecture in *Steinernema*. Comparisons of the Hox clusters of *C. elegans, S. carpocapsae*, and *S. scapterisci*. Each cluster is mapped at the same scale, with the *colored boxes* representing different putative genes between the *lin-39* and *ceh-13* orthologs. Genes marked in *blue* are specific to *Steinernema*, not having orthologs in *C. elegans*. *Gray* genes have a *C. elegans* ortholog, though they are not syntenic in the nematodes compared. Genes marked in *brown* are unique, not having obvious orthologs in the other nematodes in this comparison

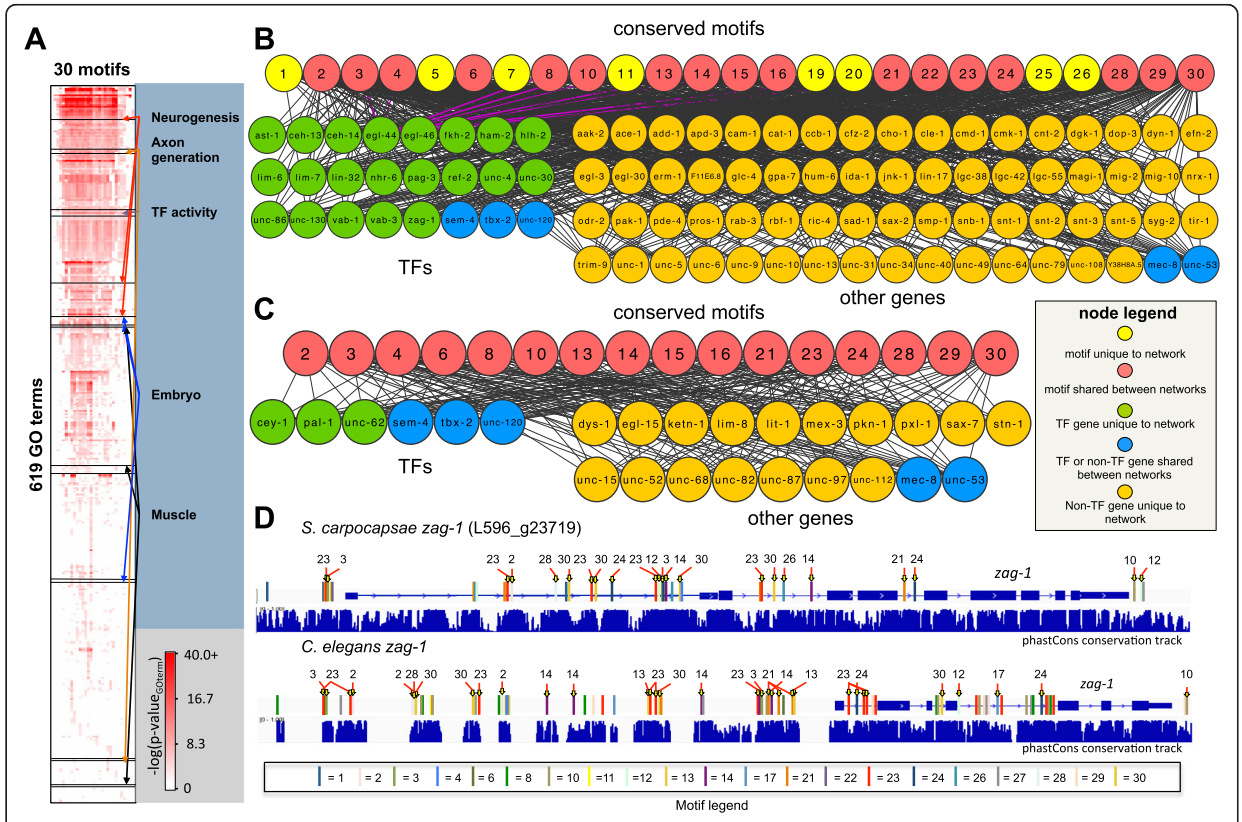


Fig. 6 Conserved non-coding networks in *Steinernema* and *Caenorhabditis*. **a** A hierarchically clustered heat map of 30 derived regulatory motifs and the GO terms that the target genes of these motifs are enriched in. Only motif-GO term associations that are shared between *S. carpocapsae* and *C. elegans* are shown. *p*-values depicted are from *C. elegans* associations. *Colored arrows* point to single GO term or groups of GO terms that belong to the four developmental categories shown. **b** A network of conserved *S. carpocapsae* and *C. elegans* motif-target gene associations related to neurogenesis GO terms. Only nodes for motifs and downstream genes with degrees ≥5 are shown in the network. **c** A network of conserved *S. carpocapsae* and *C. elegans* motif-target gene associations related to muscle GO terms. Only nodes for motifs and downstream genes with degrees ≥5 are shown in the network. **d** *zag-1* gene model in *S. carpocapsae* and *C. elegans* showing conserved motifs, and well as conserved regulatory modules (clusters of conserved motifs). Sequence conservation tracks are displayed below each gene model. Associations between *zag-1* and motifs are highlighted in *red* in the neurogenesis network. *GO* Gene Ontology

Table 1 Features of the *Steinernema* draft genomes

	S. carpocapsae	*S. scapterisci*	*S. feltiae*	*S. glaseri*	*S. monticolum*
Estimated genome size (Mb)	85.6	79.4	82.4	92.9	89.3
N50 (bp)	299,566	90,783	47,472	37,444	11,556
N90 (bp)	54,505	15,213	7,098	7,610	2,984
N10 (bp)	979,322	496,671	303,346	112,910	31,326
Number of scaffolds	1,578	2,877	5,839	7,515	14,331
GC content (Mb)	45.53	47.98	46.99	47.63	42.01
N content (Mb)	2.39	0.76	2.76	3.37	4.34
N content (%)	2.80	0.96	3.36	3.64	4.87
Maximum scaffold size (bp)	1,722,607	1,149,164	1,470,990	339,094	110,081
Number of Augustus-predicted genes	28,313	31,378	33,459	34,143	36,007
Number of Augustus-predicted transcripts	31,944	33,149	36,434	37,120	38,381
Average gene length (bp)	2,030	1,842	1,730	1,855	1,604
Average intron length (bp)	194	153	154	218	161
Average exon length (bp)	212	224	220	216	217
Average intergenic distance (bp)	1,105	723	746	930	761
Average number of exons per gene	5	4	4	4	4
Average number of introns per gene	4	3	3	3	3
GC content in coding regions (%)	51.86	51.92	51.08	53.68	46.92
Number of genes with no introns	4676	5611	6230	7521	6171
Repeat content (%)	7.46	2.75	6.70	5.34	10.49

Additional file 1: Supplementary figures and legends. (PDF 3646 kb)

Additional file 2: Supplementary tables and legends. (PDF 1760 kb)

Additional file 3: 1:1 *C. elegans* gene identifiers for the 1:1 orthologs conserved across the phylum Nematoda. These represent putative phylogenetic markers though their informative value remains to be tested. (TXT 7 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM3_ESM.txt

Additional file 4: Genes differentially expressed during *S. carpocapsae* development. Lists 4,557 DE genes (differential gene expression cutoff: FDR < 10−5 and fold change > 4×) from the eight DE gene expression clusters, and their gene expression levels (FPKM) during *S. carpocapsae* development (Fig. 3a). The far right column indicates which cluster each gene belongs to. Cluster A genes – genes with high expression in the L1, IJ, and adult stage. Cluster B genes - genes with high expression in the L1 and medium/low expression in the IJ stage. Cluster C genes – genes with high expression in the L1 and IJ stage. Cluster D genes – genes with high expression in the IJ stage. Cluster E genes - genes with high expression in the adult stage and medium/low expression in the L1 stage. Cluster F genes - genes with high expression in the L1 and adult stage. Cluster G genes - genes with high expression in the embryo stage. Cluster H genes – genes with high expression in the L1 stage. (TXT 166 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM4_ESM.txt

Additional file 5: *S. carpocapsae* isoform similarity. *S. carpocapsae* transcript isoform pairs (where both isoforms have a summed expression > 1 TPM during developmental time course), their isoform similarities (cosine similarities), and the transcript isoform expression (TPM) values of each of the isoforms during development in replicates. Isoforms that had summed expression < 1 TPM were removed from the analysis. (TXT 423 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM5_ESM.txt

Additional file 6: *S. carpocapsae* and *S. feltiae* orthologs with conserved expression (cosine similarity > 0.95). Lists 1,438 *S. carpocapsae* orthologs that have conserved expression with *S. feltiae* during nematode development. The file includes the gene ID of *S. carpocapsae* (column 1), the ortholog expression similarity (cosine similarity, column 2), the expression values (in TPM) of the *S. carpocapsae* and *S. feltiae* orthologs during the time course in replicates (columns 3–18), and the stage(s) that the ortholog expression is conserved in (column 19). (TXT 171 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM6_ESM.txt

Additional file 7: Additional Cufflinks gene and isoform annotations for *S. carpocapsae* . The file was generated by combining Cufflink's transcript annotations for four developmental stages (embryo, L1, IJ, and adult) with the Augustus-predicted gene annotations (.gtf format). Gene and isoform IDs beginning with "CUFF" were predicted by Cufflinks, whereas ones beginning with "L596_" were predicted by Augustus. The Augustus annotations here match the WormBase gene annotations for *S. carpocapsae*. (GTF 55690 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM7_ESM.gtf

Additional file 8: MEME motif position weight matrices derived from the final 30 conserved stage-specific gene sets. These motifs were derived from conserved non-coding regions ±3 kb of *S. carpocapsae* orthologs that have conserved stage-specific expression profiles during development between *S. carpocapsae* and *S. feltiae.* The motif IDs in the file are numbered from 1 to 30. In parentheses next to each motif number is the developmental stage-specific gene set the motif was derived from (e, f, ef, fi, fa, efi, efa, fia, efia) and its old MEME ID. (TXT 17 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM8_ESM.txt

Additional file 9: *S. carpocapsae* and *C. elegans* predicted regulatory motifs GO term enrichments. Thirty significant, non-redundant motifs and the 619 GO terms they are enriched in for both *S. carpocapsae* and *C. elegans*. The heat map shows the –log10(*p*-value) for each motif-associated GO term. (PNG 1053 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM9_ESM.png

Additional file 10: *S. carpocapsae* and *C. elegans* shared GO term table. Contains all the motif-associated GO (MAG) terms (FDR < 0.05) that are shared between *S. carpocapsae* and *C. elegans*. (TXT 188 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM10_ESM.txt

Additional file 11: Embryonic development network file. Contains motifs that have a conserved association with embryonic development-related genes (See "Methods" regarding motif conservation) in both *S. carpocapsae* and *C. elegans*. The first column contains the motif ID, the second column contains the edge weight, and the third and fourth columns contain *C. elegans* gene IDs in two different formats. (TXT 6 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM11_ESM.txt

Additional file 12: Neurogenesis/axonogenesis network file. Contains motifs that have a conserved association with neurogenesis-related genes (See "Methods" regarding motif conservation) in both *S. carpocapsae* and *C. elegans*. The first column contains the motif ID, the

second column contains the edge weight, and the third and fourth columns contain *C. elegans* gene IDs in two different formats. (TXT 19 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM12_ESM.txt

Additional file 13: Muscle development network file. Contains motifs that have a conserved association with muscle development-related genes (See "Methods" regarding motif conservation) in both *S. carpocapsae* and *C. elegans*. The first column contains the motif ID, the second column contains the edge weight, and the third and fourth columns contain *C. elegans* gene IDs in two different formats. (TXT 5 kb)
Link: https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0746-6/MediaObjects/13059_2015_746_MOESM13_ESM.txt

Materials and methods

Strain culturing and maintenance. *S. carpocapsae* (strain All), *S. scapterisci* (strain FL), *S. feltiae* (strain SN), *S. glaseri* (strain NC), and *S. monticolum* (Mount Jiri strain) were reared and maintained using standard methods (Kaya and Stock, 1997) (Fig. 1, Additional file 1: Fig. S1). Briefly, five last-instar *Galleria mellonella* waxmoth larvae or a single adult cricket for *S. scapterisci* (American Cricket Ranch, Lakeside, CA, USA) were placed in a 5 cm Petri dish with a 55 mm Whatman 1 filter paper acting as a pseudo-soil substrate in the bottom of the dish. Up to 250 µl containing 500–1,000 IJs suspended in water was evenly distributed on the filter paper. After 7–10 days the insect cadavers were placed on White traps (White, 1927). Waxmoth cadavers infected with *S. glaseri* were placed in a Petri dish partially filled with plaster of Paris and harvested from this, because *S. glaseri* emerge as pre-IJs that will not properly develop if they emerge directly into water (Kaya and Stock, 1997). Emerging IJs from all species were harvested, washed for 10 minutes in 0.4 % Hyamine 1622 solution (Fluka), and rinsed three times with water.

Isolation of DNA and RNA. To harvest bulk genomic DNA and IJ RNA, IJs from each species were washed in 0.4 % Hyamine, rinsed three times, and acclimated in Ringer's solution for 15–30 minutes prior to nucleic acid collection. For DNA extraction, a Wizard Genomic DNA Purification Kit (Promega) was used following the manufacturer's protocol. The genomic DNA was then treated with RNase A to remove any RNAs present in the sample. For RNA extractions, the nematodes were snap-frozen in liquid nitrogen in ~100 µL aliquots and stored at −80 °C. Worms were then freeze-thawed three or four times to break the cuticle before extracting bulk RNA. Bulk RNA was then extracted using a phenol-chloroform extraction using Trizol

(Invitrogen). This sample was treated with DNase to remove lingering DNA and then poly-A selected to isolate eukaryotic messenger RNA, reducing if not removing bacterial contamination. To isolate embryonic, L1, and adult stage-specific RNA from *S. carpocapsae*, 1,000–2,000 IJs were placed onto 10 cm lipid agar plates seeded with overnight cultures of *Xenorhabdus nematophila* (strain ATCC 19061). These cultures were allowed to grow for ~42 hours to collect young adults, or ~68 hours to collect embryos from mature gravid females. Gravid females were collected from the plates by adding enough distilled water to the cover the surface of the plates, swirling the plates by hand to lift the nematodes into suspension, and pouring them into conical tubes. These were then pelleted by gentle centrifugation and rinsed several times with distilled water until the supernatant was clear. The nematodes were placed in separate conical tubes in 5 mL aliquots, and topped off to 50 mL with bleach solution (16.6 mL of 12 % bleach, 5 mL of 1 M KOH, and 80 mL of distilled water). Eggs were harvested by bleaching the nematodes until all nematode tissue was dissolved, leaving only the eggs. These embryos were then either harvested for total RNA as described above, or they were allowed to hatch to L1s in Ringer's solution over a period of ~30 hours before harvesting the total RNA. To isolate embryonic, L1, and adult stage-specific RNA from *S. feltiae*, 1,000–2,000 IJs were placed onto 10 cm lipid agar plates seeded with overnight cultures of *Xenorhabdus bovienii* (Akhurst and Boemare ATCC 35271). These cultures were allowed to grow for ~36 hours to collect adults or ~55 hours to collect embryos from gravid females. To collect L1s, we waited until all embryos hatched, which was ~24 hours. The same bleaching procedure was followed to harvest embryos and L1s as for *S. carpocapsae*. To isolate embryonic, L1, and adult stage-specific RNA from *C. elegans* (N2 strain) worms were placed onto 10 cm nematode growth media (NGM) plates seeded with overnight cultures of *Escherichia coli* (OP50 strain). To these, 200 uL aliquots of OP50 were

added every day. Plates with lots of gravid adults were bleached according to the guide for maintenance of *C. elegans* in Wormbook (Stiernagle, 2006). The embryos were either collected to harvest embryos, placed in Ringer's solution for ~20 hours to harvest L1s, or plated on fresh NGM plates seeded with *E. coli* OP50 and collected ~47 hours later to harvest young adults.

Genomic and RNA-seq library construction. The genomic library was constructed using an Illumina Paired-End DNA Sample Preparation Kit according to the manufacturer's instructions. Briefly, 3 μg of genomic DNA were fragmented using nebulization. The fragments were end-repaired, 3′-adenylated, and ligated to Illumina's paired-end adaptors. The ligation products were size-selected on an agarose gel to yield fragments of approximate length of 350 bp. These fragments were then PCR-amplified to produce the finished library. The mate-pair or "jumping" library was prepared using an Illumina Mate Pair Library Preparation Kit v2. Briefly, 7.5 μg of genomic DNA was fragmented using a HydroShear device (Genomic Instrumentation Services, Inc.) to generate fragments of ~2.2 kb. Following end repair and biotinylation, the 2.2 kb fragment was gel-purified and circularized. Circular DNA was fragmented using a Bioruptor NGS (Diagenode, Inc.) and biotin-containing fragments were isolated using Dynabeads (Invitrogen, Inc.). The fragments were end-repaired, 3′-adenylated, and ligated to NEBNext Multiplex Adaptors (NEB, Inc.). The ligation products were PCR-amplified and size-selected using AMPure XP beads (Beckman Coulter, Inc.) to generate the finished library of approximately 450 bp in length. Genomic libraries were sequenced on an Illumina Genome Analyzer IIx sequencer in paired-end mode with the read length of 76 bp. The jumping library was sequenced on an Illumina HiSeq2000 in paired-end mode with the read length of 100 bp (Additional file 2: Table S20).

The first set of RNA samples, which was used for genome annotation, was prepared from 10 μg of total RNA, poly(A)-selected, and libraries constructed using a standard unstranded protocol (Mortazavi et al., 2010; Mortazavi et al., 2008). Libraries were quantified using a Qubit fluorometer (Invitrogen) and size distributions were verified using an Agilent Bioanalyzer and the High Sensitivity DNA Kit. These RNA-seq libraries were sequenced on the Illumina Genome Analyzer IIx sequencer in paired-end mode to a read length of 76 bp (Additional file 2: Table S21). The second set of RNA-seq samples, which was used for the gene expression analyses, was prepared from 5–30 μg of total RNA, poly(A)-selected using a Dynabeads mRNA DIRECT Kit (Invitrogen), and fragmented with a hydrolysis buffer containing magnesium ions (Mortazavi et al., 2008). Double-stranded cDNA was prepared from the mRNA fragments using Invitrogen's SuperScript Double-Stranded cDNA Synthesis Kit. During the second strand of reverse transcription, dUNTP (Applied Biosystems) was added to label the second strand (stranded protocol), and the libraries were prepared following the Myer's Lab ChIP-seq protocol version 2011 with Illumina sequencing adapters. The libraries were sequenced on either the Illumina HiSeq 2000 or the NextSeq 500 sequencer in single-end mode to a read length of 50 bp or 75 bp, respectively (Additional file 2: Table S14). Reads for RNA-seq samples used for the gene expression analysis and gene expression tables were submitted to Gene Expression Omnibus (GEO) under the accession number [GSE68588].

Genome assembly. The genomic libraries were built, sequenced, assembled, filtered, and repeat-masked as previously described (Mortazavi et al., 2010) using Velvet 1.2.07 and RepeatModeler 1.0.5, RepeatMasker 3.0.3, recon 1.70, and RepeatScout 1.0.5 (Table 1). The genomes and gene annotations are available at (WormBase Parasite).

The Whole Genome Shotgun project for *S. carpocapsae* has been deposited at DDBJ/EMBL/GenBank under the accession [AZBU00000000]. The version described in this paper is version AZBU01000000.

The Whole Genome Shotgun project for *S. feltiae* has been deposited at DDBJ/EMBL/GenBank under the accession [AZBV00000000]. The version described in this paper is version AZBV01000000.

The Whole Genome Shotgun project for *S. glaseri* has been deposited at DDBJ/EMBL/GenBank under the accession [AZBX00000000]. The version described in this paper is version AZBX01000000.

The Whole Genome Shotgun project for *S. monticolum* has been deposited at DDBJ/EMBL/GenBank under the accession [AZHV00000000]. The version described in this paper is version AZHV01000000.

The Whole Genome Shotgun project for *S. scapterisci* has been deposited at DDBJ/EMBL/GenBank under the accession [AZBW00000000]. The version described in this paper is version AZBW01000000.

Transcriptome assembly and genome annotation. IJ-stage, paired-end 75 bp, unstranded RNA-seq data sequenced to an average depth of 76 million reads for *S. feltiae*, *S. glaseri*, *S. monticolum*, and *S. scapterisci*, and embryo, L1, IJ, and adult stage data for *S. carpocapsae* were de novo assembled into expressed sequence tags (ESTs) with Oases 0.2.6 as previously described (Schulz et al., 2012), with the following options: -m 23, -M 59, -s 4, -ins_length. To annotate each genome, ESTs were mapped onto the genome with BLAT 3.4 and these used as hints for gene finding using Augustus 2.6 with *C. elegans* settings (options: --species = caenorhabditis, --gff3 = on, --alternatives-from-evidence = true, --uniqueGeneId = false, --

protein = on, --codingseq = on, --noInFrameStop = true, --UTR = on, --hintsfile) (Stanke et al, 2008). Separately, RNA-seq reads were mapped onto the genome using TopHat 1.4 (Trapnell et al., 2009) to find novel transcripts using Cufflinks 2.0.2 (Trapnell et al., 2010) (Table 1, Additional file 2: Table S22, Additional file 7), which is described in more detail in a later section below.

Filtering bacterial symbiont DNA and other bacterial DNA contaminants from genomes. Protein sequences coded by intronless Augustus-predicted genes (putative bacterial contamination) were compared to a database using blastp in Blast2GO (Conesa et al., 2005) to determine the identities of the bacterial contaminants present in the respective nematode genomes (Additional file 1: Fig. S2). Assembled bacterial genomes matching the species blast results were obtained from GenBank, and their sequences were compared to the respective nematode genomes with BLAT 3.4, and removed from the nematode assemblies when the sequence match was >94 % identical (Kent, 2002). After filtering out bacterial DNA contamination, the genome annotations were repeated for each assembly using Augustus.

Orthology analyses. To study the evolution of gene families across nematodes, we used the available predicted protein datasets from WormBase release WS225 — *Brugia malayi*, *Caenorhabditis elegans*, *Meloidogyne hapla*, *Pristionchus pacificus*, and *Trichinella spiralis* (C. elegans Sequencing Consortium; Dieterich et al., 2008; Ghedin et al., 2007; Mitreva et al., 2011; Opperman et al., 2008). We also included the *Ascaris suum* and *Bursaphelenchus xylophilus* predicted proteome datasets from WormBase release WS229 (Jex et al., 2011; Kikuchi et al., 2011). We also used the *Panagrellus redivivus* genome assembly prior to its WormBase release (Srinivasan et al., 2013). For out-group and comparative analysis we used the predicted protein

dataset of the *Nasonia vitripennis* (v1.2) genome project, obtained from the NCBI/NIH repository (Werren et al., 2010) (Fig. 2a–c, Additional file 1: Figs S3–S10). Version 1.4 of the OrthoMCL pipeline was used to cluster proteins into families of orthologous genes, with default settings and the BLAST parameters recommended in the OrthoMCL documentation (Li et al., 2003) (Fig. 2, Additional file 2: Table S1).

Protein domain analyses. To evaluate the prevalence of protein domains in the proteome of *Steinernema carpocapsae* and other species, we used the hmmscan program from the latest version of HMMER (3.0) software package, which implements probabilistic profile hidden Markov models (Finn et al., 2011). We set our threshold *E*-value criterion at 10^{-6}, so that no known false-positive matches would be detected in assigning Pfam domain identities. We ran this analysis on the proteomes mentioned above and filtered out splice isoforms from the *C. elegans* proteome.

Gene tree analyses. Some protein families were further explored by evaluating gene trees either with whole protein sequences or by protein domain sequences. To do these analyses we aligned protein sequences using MUSCLE (Edgar, 2004). Aligned protein sequences were then evaluated by distance analysis using the JTT matrix and a subsequent Neighbor-joining tree was created using the PHYLIP software package version 3.68, using the protdist and neighbor programs, and seqboot where bootstrap values were reported (Felsenstein et al., 2005) (Fig. 2d, Additional file 1: Fig. S11).

Supermatrix construction and whole genome phylogenetic analysis. The orthology analysis above resulted in 3,885 strictly orthologous genes (1:1 conservation across all steinernematid species and the out-group, *P. redivivus*). These strict orthologs were then compiled and used for

the supermatrix construction and subsequent phylogenetic analysis. Because alignment accuracy greatly influences phylogenetic analyses and an earlier study on *Steinernema* phylogeny shows that there can be greater topological variation due to different alignment construction parameters than owing to the methods used to generate the phylogenies (Nguyen et al., 2005; Simmons et al, 2011), we took a very conservative approach to generating the amino acid sequence alignments. Accordingly, each gene was first aligned separately in MAFFT v6.821b (Katoh et al., 2002). The L-INS-i algorithm was chosen because it is the most accurate setting in MAFFT for datasets containing fewer than 200 species (Katoh et al., 2002). Because this analysis incorporated more genes (3,885 per species) than can reasonably be checked by eye, we used GBlocks v0.91 (Castresana, 2000) to objectively eliminate highly divergent and ambiguously aligned regions of the transformation series (Talavera et al., 2007; Swofford, 2002; Felsenstein, 1985). Using the batch feature of GBlocks we applied strict settings: four of the six species' amino acids were required to make a conserved position for a column, five of the six species' amino acids were required to create a flank position, ten conserved amino acids were required to make a block, eight consecutive non-conserved amino acids was the maximum allowed, and all gaps were removed.

GBlocks identified sequence divergence and alignment ambiguity problems that led us to remove 14 genes from the analysis. Prior to the GBlocks analysis a supermatrix of all of the genes contained 2,924,577 amino acids; the optimized alignment was reduced to 1,320,306 amino acids, a 45 % reduction. GBlocks output was used to concatenate the individual gene files into a supermatrix.

We constructed phylogenetic trees in PAUP* v4.0b10 (Swofford, 2002) under the parsimony optimality criterion. The tree search parameters for the supermatrix were an

exhaustive parsimony search enforcing a monophyletic root. The result was a separate tree file for each gene and another for the supermatrix. We inferred nodal support by bootstrap analysis (Felsenstein, 1985) of the supermatrix in PAUP* with 500 repetitions using a heuristic search with randomized additions. The parsimony analysis of the supermatrix resulted in only one best tree (Fig. 1b). The bootstrap values were all 100 on each node, suggesting that the data provide strong support for the solution. The tree that was supported by the largest number of genes was the same tree that was the most parsimonious solution for the supermatrix (data not shown).

Analysis of genome completeness. Genome completeness was determined by clustering *S. carpocapsae*, *S. feltiae*, *S. glaseri*, *S. monticolum*, and *S. scapterisci* protein sets with a core set of highly conserved eukaryotic proteins (Core Eukaryotic Gene Mapping Approach, CEGMA) using OrthoMCL 1.4 as previously described (Srinivasan et al., 2013; Mortazavi et al., 2008; Li et al., 2003; Parra et al., 2007). The percentages of genome completeness for each species was found by dividing the number of proteins that were orthologous to CEGMA proteins by the total number of CEGMA proteins (Additional file 2: Table S2).

Gene expression analyses. Stranded, single-ended 50 bp RNA-seq reads from the embryonic, L1, IJ, and adult stages of *S. feltiae*, *S. carpocapsae*, and *C. elegans* sequenced to an average depth of 22, 30, and 33 million reads respectively were trimmed to 35 bp to remove low quality bp (Additional file 2: Table S14). Prior to read mapping, transcriptome indexes were prepared for *S. carpocapsae*, *S. feltiae*, and *C. elegans* (WS220) using the RSEM command (version 1.2.12) rsem-prepare-reference (Li et al., 2011). Reads were mapped to each respective species' annotations using bowtie 0.12.8 with the following options: -S, --offrate 1, -v 1, -k 10, --best, --strata, -m 10 (Langmead et al., 2009). Gene expression was quantified using the RSEM

command, rsem-calculate-expression, with the following options: --bam, --fragment-length-mean (Li et al., 2011). We used EdgeR to analyze genes that were DE during the developmental time course of each species, and we considered a gene to be DE if it had an FDR < 1×10^{-5} and a fold change > 4× (Robinson et al., 2010). DE genes were K-means clustered into eight clusters (Fig. 3a, Additional file 4) using Cluster 3.0 (de Hoon et al., 2004), and visualized with JavaTree View (Saldanha, 2004). The optimal K for clustering was found using the Akaike information criterion. DE gene clusters were functionally annotated using Blast2GO's Fisher's exact test (Kent, 2002).

Finding novel genes and isoforms using Cufflinks and Cuffmerge. Unstranded paired-end RNA-seq data collected from four *S. carpocapsae* developmental stages (embryo, L1, IJ, adult) were aligned to the *S. carpocapsae* genome using TopHat 1.4.0 and Bowtie 0.12.8 with the following options: -r 50, –G < annotations > (Conesa et al., 2005). Gene expression for the aligned reads was quantified with Cufflinks 2.0.2 using the following options: -u, -g < annotations>. Transcript annotations from each developmental stage were merged together with Cuffmerge (options: -g < annotations>, -s < genome>) (Additional file 7). The Cuffmerge output showed genes and transcripts that were discovered by Cufflinks but missed by Augustus. The Cufflinks annotations were used in combination with the Augustus annotations to delineate coding versus non-coding sequences in downstream analyses.

Unstranded, paired-end RNA-seq data for the IJ stage in the other species were aligned to their respective genomes using TopHat 1.4.0 and Bowtie 0.12.8 with the following options: -r 50, –G < annotations>. Cuffmerge was not used. Gene expression was quantified with Cufflinks 2.0.2 using the following options: -u, -g < annotations >.

Multiple genome alignment. Five whole repeat-masked *Steinernema* genomes were aligned using MULTIZ/TBA (multiz-tba.012109). Contigs from the best-assembled genome, *S. carpocapsae*, were concatenated together with 100 bp "N" spacers and used as a reference sequence for the alignment process. The aligned sequences were analyzed with Phast 1.2.1 (phyloFit options: --tree, phastCons options: --target-coverage 0.4, --expected-length 10, --estimate-trees, --nopostprob) to determine regions of sequence conservation across the genomes using setting for *C. elegans* as previously described (Mortazavi et al., 2010; Felsenstein and Churchill, 1996; Margulies et al., 2003; Siepel et al., 2005). PhastCons parameters were also varied around those used for *C. elegans* (Mortazavi et al., 2010), but the *C. elegans* parameters provided a good balance between small and large blocks of conservation. Conserved sequences that matched Augustus and Cufflinks coding sequences or 5′ or 3′ untranslated regions were considered conserved coding sequences, whereas sequences that mapped anywhere else were considered conserved non-coding sequences. DNA from the anterior portion of the Hox cluster (*ceh-13* and *lin-39*) in *S. carpocapsae*, *S. scapterisci*, and *S. feltiae* were also aligned using MUSSA (Kuntz et al., 2008) to find conserved regions of their DNA. MUSSA was run with a conservation window size of 30 nucleotides and a nucleotide conservation threshold of 23 nucleotides.

Gene expression conservation. To determine the degree of gene expression conservation during development between nematode species, we compared gene expression data for four developmental stages in *S. carpocapsae*, *S. feltiae*, and *C. elegans*. Two methods were used for determining conserved gene expression. The first method binarized the expression data using a flexible threshold to sort the genes into stage-specific sets (Additional file 1: Fig. S14). We examined the gene expression levels of the 1:1:1 orthologs at four developmental stages and

asked if an ortholog that was expressed above an averagely expressed gene (10 FPKM) in a particular set of developmental stages in a nematode species was expressed at least above 5 FPKM in the other nematode species in the same set of developmental stages. If the ortholog was expressed in the same set of developmental stages, it was considered conserved in stage-specific gene expression. If not, stage-specific gene expression was considered to have changed. We used this method to determine the fraction of orthologs that are "on" and "off" in the same developmental stages between species. However, to address whether their expression profiles parallel each other during development, we treated the ortholog expression calculated in transcripts per million (TPM, which is interconvertible with FPKM) during development as vectors, and calculated the cosine similarity (Fig. 3b). The cosine similarity provides a measure of similarity between a pair of vectors: a similarity measure of 1 means that the two vectors are perfectly correlated, whereas a similarity measure of 0 means the vectors are orthogonal (i.e., uncorrelated). We calculated the cosine similarities for each ortholog used in the binary method with developmental stage replicates for each species. We found that orthologs with cosine similarities > 0.95 had extremely similar expression profiles during development, so we set this to be our conservation threshold. This gave us a total of 1,441 orthologs with conserved expression profiles between *S. carpocapsae* and *S. feltiae* (Additional file 6). We sorted these orthologs into stage-specific gene sets by requiring developmental stages to contribute to at least 10 % of the total gene expression during the time course to be considered "on." Stage-specific gene sets containing more than 30 genes were used for motif finding (e, f, ef, fi, fa, efi, efa, fia, efia).

Motif discovery. Nine sets of *Steinernema* stage-specific orthologs were chosen for motif mining (Fig. 3c). Conserved non-coding regions ±3 kb or within introns of the genes were

obtained by intersecting bed coordinates for the regions upstream of these genes with the genome-wide set of conserved non-coding regions using bedtools/2.15.0 bedintersect (Quinlan and Hall, 2010). The conserved non-coding bed regions were converted to fasta sequence using bedtools getfasta, filtered for sequences >8 bp, and run through MEME 4.8.1 (settings: -minw 6 -maxw 12 -dna -nmotifs 20-50 -mod zoops -revcomp) to find recurring regulatory motif sequences (Bailey et al., 2009). We discovered 440 motifs in total across the nine gene sets and searched for them across both the *S. carpocapsae* and *C. elegans* conserved non-coding regions using FIMO with the following settings: --thresh 0.3 --qv-thresh --max-stored-scores 20000000 –bgfile --parse-genomic-coord (Grant et al., 2011). We used the WS220 gene annotations and the corresponding conserved regions for *C. elegans* from the UCSC Genome Browser (ce10/WS220: phastConsElements7way.txt) for these analyses. The conserved non-coding regions were produced for *C. elegans* by retaining conserved regions that did not intersect annotated coding regions (bedtools bedintersect, settings = -wa).

Filtering out redundant and insignificant motifs. Motifs that could not map to any conserved non-coding regions within the q-value threshold (q-value < 0.3) in either species were removed from the analysis. The remaining motif set was compared to itself with TOMTOM to identify redundant motifs, using the following settings: -min-overlap 5 -dist pearson -thresh 0.05 (Gupta et al., 2007). The redundant motif with the highest MEME e-value of the pair of matching motifs was removed from the analysis. In the end, we were left with 30 non-redundant motifs (Additional file 2: Table S17, Additional file 8).

Motif-gene association. The final set of non-redundant motifs was associated with the nearest gene models for each species, forming motif-associated gene sets using bedtools closest with the following setting: -d (Quinlan and Hall, 2010).

Novel motif comparison to WormBase motif database. The final set of 30 motif position weight matrices was compared to 5,512 motifs from WormBase (Araya et al., 2014; Gerstein et al., 2010) with TOMTOM using the following settings: -min-overlap 5 -dist pearson -evalue -thresh 1.0. Out of 30 motifs, 24 matched WormBase motifs with a *p*-value < $1e^{-4}$ and an e-value < 0.5 (Additional file 2: Table S18).

Motif conservation. GO term enrichments were determined for each *S. carpocapsae* and *C. elegans* motif-associated gene set using the Fisher's exact test in Blast2GO (Conesa et al., 2005). Motif-associated GO terms with FDRs < 0.05 and that were shared between *S. carpocapsae* and *C. elegans* were considered for the analysis (Fig. 6b,c; Additional files 9 and 10).

Conserved GO term network generation. Enriched motif-associated GOs (MAGs) shared between *S. carpocapsae* and *C. elegans* were analyzed for the number and percentage of motif-associated 1:1 orthologs shared between them. MAGs that shared 30 % 1:1 orthologs were involved in biological processes under or related to the parent terms such as neurogenesis, embryogenesis, and muscle development. Thus, we focused on GO terms related to these particular processes and generated networks by placing shared 1:1 ortholog targets from related GO terms and the putative conserved motifs that regulate them into three networks: a neurogenesis-related network, an embryonic-related network, and a muscle-related network. The supplemental figures show all the conserved motif–gene associations regardless of motif and gene node degree, while the main figures show all nodes that had degrees greater than 5 (Fig.

6b,c, Additional file 1: Fig. S20, Additional files 11, 12, and 13). The motifs and ortholog associations within these networks are conserved between *S. carpocapsae* and *C. elegans*. Motif locations around the gene models were investigated around interesting orthologs, such as *egl-44* and *zag-1*, to see if the motif sites are conserved in their location or have changed over time (Fig. 6d, Additional file 1: Fig. S21).

Randomized GO term control network. To verify that the motif-associated GO term enrichments we obtained were not due to chance, we created 100 randomized GO term sets by shuffling all of the annotated *S. carpocapsae* gene GO terms that were derived from Blast2GO. We reassigned the GO term sets to new genes that were previously annotated. Unannotated genes were not assigned a randomized GO term set. We applied Fisher's exact test to all 30 MAG sets using these randomized GO sets (30 motifs × 100 randomizations = 3,000 Fisher's exact tests in total), and the GO enrichment results for the neuronal, embryo, and muscle GO terms were analyzed. We did not recover enrichments for any GO terms associated with these terms with FDRs < 0.05.

Acknowledgements

This research was supported by grants from the US National Institutes of Health (NIH) and the Howard Hughes Medical Institute, for which PWS is an investigator. ARD was supported by NIH training grants (5T32GM007616 and 5T32HG000044). AM and MM were supported by an NIH New Innovator Award to AM (DP2 GM111100). XL was supported by the UW-Madison Graduate School research funds.

References

1. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, et al. A molecular evolutionary framework for the phylum Nematoda. Nature. 1998;392:71–5.
2. van Megen H, van Den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, et al. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology. 2009;11:927–50.
3. Castelletto ML, Gang SS, Okubo RP, Tselikova AA, Nolan TJ, Platzer EG, et al. Diverse host-seeking behaviors of skin-penetrating nematodes. PLoS Pathog. 2014;10, e1004305.
4. Dillman AR, Guillermin ML, Lee JH, Kim B, Sternberg PW, Hallem EA. Olfaction shapes host-parasite interactions in parasitic nematodes. Proc Natl Acad Sci U S A. 2012;109:E2324–2333.
5. Kaya HK, Gaugler R. Entomopathogenic nematodes. Annu Rev Entomol. 1993;38:181–206.
6. Dillman AR, Chaston JM, Adams BJ, Ciche TA, Goodrich-Blair H, Stock SP, et al. An entomopathogenic nematode by any other name. PLoS Pathog. 2012;8, e1002527.
7. Gaugler R, Kaya HK. Entomopathogenic nematodes in biological control. Boca Raton: CRC Press; 1990.
8. Dillman AR, Sternberg PW. Entomopathogenic nematodes. Curr Biol. 2012;22:R430–431.
9. Stock SP, Goodrich-Blair HG. Entomopathogenic nematodes and their bacterial symbionts: The inside out of a mutualistic association. Symbiosis. 2008;46:65–75.
10. Castillo JC, Reynolds SE, Eleftherianos I. Insect immune response to nematode parasites. Trends Parasitol. 2011;27:537–47.
11. Hallem EA, Dillman AR, Hong AV, Zhang Y, Yano JM, DeMarco SF, et al. A sensory code for host seeking in parasitic nematodes. Curr Biol. 2011;21:377–83.

12. Davidson EH. Emerging properties of animal gene regulatory networks. Nature. 2010;468:911–20.
13. Stein LD, Bao ZR, Blasiar D, Blumenthal T, Brent MR, Chen NS, et al. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. PLoS Biol. 2003;1:166–92.
14. Havird JC, Miyamoto MM. The importance of taxon sampling in genomic studies: An example from the cyclooxygenases of teleost fishes. Mol Phylogenet Evol. 2010;56:451–5.
15. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 2007;56:17–24.
16. Nadler SA, Bolotin E, Stock SP. Phylogenetic relationships of Steinernema Travassos, (Nematoda: Cephalobina: Steinernematidae) based on nuclear, mitochondrial and morphological data. Syst Parasitol. 1927;2006:161–81.
17. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003;425:798–804.
18. Zhao L, Zhang N, Ma PF, Liu Q, Li DZ, Guo ZH. Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in Poaceae. Plos One. 2013;8, e64642.
19. Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, et al. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. Mol Biol Evol. 2006;23:1792–800.
20. Adams BJ, Peat SM, Dillman AR. Phylogeny and evolution. In: Nguyen KB, Hunt DJ, editors. Entomopathogenic nematodes: Systematics, phylogeny, and bacterial symbionts,

Volume 5. Leiden-Boston: Brill; 2007. p. 693–733. Nematology monographs and perspectives.

21. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 1998;282:2012–8.
22. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, et al. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet. 2008;40:1193–8.
23. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, et al. Draft genome of the filarial nematode parasite *Brugia malayi*. Science. 2007;317:1756–60.
24. Jex AR, Liu S, Li B, Young ND, Ross SH, Li Y, et al. *Ascaris suum* draft genome. Nature. 2011;479:529–33.
25. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog. 2011;7, e1002219.
26. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, et al. The draft genome of the parasitic nematode *Trichinella spiralis*. Nat Genet. 2011;43:228–36.
27. Opperman CH, Bird DM, Williamson VM, Rokhsar DS, Burke M, Cohn J, et al. Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. Proc Natl Acad Sci U S A. 2008;105:14802–7.
28. Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, Fracchia KM, et al. The draft genome and transcriptome of *Panagrellus redivivus* are shaped by the harsh demands of a free-living lifestyle. Genetics. 2013;193:1279–95.

29. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, et al. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science. 2010;327:343–8.

30. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007;450:203–18.

31. Kanost MR, Clarke T: Proteases. In: Gilbert LI, Iatrou K, Gill S, editors. Comprehensive Molecular Insect Science. Vol 4. Oxford: Elsevier; 2005. p. 247–266.

32. Abuhatab M, Selvan S, Gaugler R. Role of proteases in penetration of insect gut by the entomopathogenic nematode *Steinernema glaseri* (Nematoda, Steinernematidae). J Invertebr Pathol. 1995;66:125–30.

33. Balasubramanian N, Hao YJ, Toubarro D, Nascimento G, Simoes N. Purification, biochemical and molecular analysis of a chymotrypsin protease with prophenoloxidase suppression activity from the entomopathogenic nematode *Steinernema carpocapsae*. Int J Parasitol. 2009;39:975–84.

34. McKerrow JH, Brindley P, Brown M, Gam AA, Staunton C, Neva FA. *Strongyloides stercoralis*: identification of a protease that facilitates penetration of skin by the infective larvae. Exp Parasitol. 1990;70:134–43.

35. Toubarro D, Lucena-Robles M, Nascimento G, Costa G, Montiel R, Coelho AV, et al. An apoptosis-inducing serine protease secreted by the entomopathogenic nematode *Steinernema carpocapsae*. Int J Parasitol. 2009;39:1319–30.

36. Burman M. *Neoaplectana carpocapsae* - toxin production by axenic insect parasitic nematodes. Nematologica. 1982;28:62–70.

37. Dunphy GB, Rutherford TA, Webster JM. Growth and virulence of *Steinernema glaseri* influenced by different subspecies of *Xenorhabdus nematophilus*. J Nematol. 1985;17:476–82.

38. Dunphy GB, Webster JM. Influence of *Steinernema feltiae* (Filipjev) Wouts, Mracek, Gerdin and Bedding DD136 strain on the humoral and hemocytic responses of *Galleria mellonella* (L) larvae to selected bacteria. Parasitology. 1985;91:369–80.

39. Han R, Ehlers RU. Pathogenicity, development, and reproduction of *Heterorhabditis bacteriophora* and *Steinernema carpocapsae* under axenic *in vivo* conditions. J Invertebr Pathol. 2000;75:55–8.

40. Simóes N, Caldas C, Rosa JS, Bonifassi E, Laumond C. Pathogenicity caused by high virulent and low virulent strains of *Steinernema carpocapsae* to *Galleria mellonella*. J Invertebr Pathol. 2000;75:47–54.

41. James ER, Green DR. Manipulation of apoptosis in the host-parasite interaction. Trends Parasitol. 2004;20:280–7.

42. Trap C, Boireau P. Proteases in helminthic parasites. Vet Res. 2000;31:461–71.

43. Zang X, Maizels RM. Serine proteinase inhibitors from nematodes and the arms race between host and pathogen. Trends Biochem Sci. 2001;26:191–7.

44. Balasubramanian N, Toubarro D, Simoes N. Biochemical study and *in vitro* insect immune suppression by a trypsin-like secreted protease from the nematode *Steinernema carpocapsae*. Parasite Immunol. 2010;32:165–75.

45. Jing Y, Toubarro D, Hao Y, Simoes N. Cloning, characterisation and heterologous expression of an astacin metalloprotease, Sc-AST, from the entomoparasitic nematode *Steinernema carpocapsae*. Mol Biochem Parasitol. 2010;174:101–8.

46. Milstone AM, Harrison LM, Bungiro RD, Kuzmic P, Cappello M. A broad spectrum Kunitz type serine protease inhibitor secreted by the hookworm *Ancylostoma ceylanicum*. J Biol Chem. 2000;275:29391–9.

47. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2012;40:D343–350.

48. Molehin AJ, Gobert GN, McManus DP. Serine protease inhibitors of parasitic helminths. Parasitology. 2012;139:681–95.

49. Kennedy MW, Corsico B, Cooper A, Smith BO. The unusual lipid-binding proteins of nematodes: NPAs, nemFABPs and FARs. In: Kennedy MW, Harnett W, editors. Parasitic nematodes: molecular biology, biochemistry, and immunology. Wallingford: CABI; 2013. p. 397–412.

50. Garofalo A, Klager SL, Rowlinson MC, Nirmalan N, Klion A, Allen JE, et al. The FAR proteins of filarial nematodes: secretion, glycosylation and lipid binding characteristics. Mol Biochem Parasitol. 2002;122:161–70.

51. Hao YJ, Montiel R, Abubucker S, Mitreva M, Simoes N. Transcripts analysis of the entomopathogenic nematode *Steinernema carpocapsae* induced *in vitro* with insect haemolymph. Mol Biochem Parasitol. 2010;169:79–86.

52. Iberkleid I, Vieira P, Engler JD, Firester K, Spiegel Y, Horowitz SB. Fatty acid-and retinol-binding protein, Mj-FAR-1 induces tomato host susceptibility to root-knot nematodes. Plos One. 2013;8:e64586.

53. Campos ML, Kang JH, Howe GA. Jasmonate-triggered plant immunity. J Chem Ecol. 2014;40:657–75.

54. Lawrence T, Willoughby DA, Gilroy DW. Anti-inflammatory lipid mediators and insights into the resolution of inflammation. Nat Rev Immunol. 2002;2:787–95.

55. Stanley D, Miller J. Eicosanoids in invertebrate immunity: an *in vitro* approach. In Vitro Cell Dev Biology Ani. 2006;42:5a–a.

56. Carton Y, Frey F, Stanley DW, Vass E, Nappi AJ. Dexamethasone inhibition of the cellular immune response of *Drosophila melanogaster* against a parasitoid. J Parasitol. 2002;88:405–7.

57. Park Y, Kim Y. Eicosanoids rescue *Spodoptera exigua* infected with *Xenorhabdus nematophilus*, the symbiotic bacteria to the entomopathogenic nematode *Steinernema carpocapsae*. J Insect Physiol. 2000;46:1469–76.

58. Park Y, Kim Y, Putnam SM, Stanley DW. The bacterium *Xenorhabdus nematophilus* depresses nodulation reactions to infection by inhibiting eicosanoid biosynthesis in tobacco hornworms, *Manduca sexta*. Arch Insect Biochem Physiol. 2003;52:71–80.

59. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol. 1983;100:64–119.

60. Sulston J, Horvitz HR. Postembryonic cell lineages of the nematode *Caenorhabditis elegans*. Dev Biol. 1977;56:110–56.

61. Sinha A, Sommer RJ, Dieterich C. Divergent gene expression in the conserved dauer stage of the nematodes *Pristionchus pacificus* and *Caenorhabditis elegans*. BMC Genomics. 2012;13:254.

62. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2012;13:59–69.

63. Garofalo A, Rowlinson MC, Amambua NA, Hughes JM, Kelly SM, Price NC, et al. The FAR protein family of the nematode *Caenorhabditis elegans* - differential lipid binding properties, structural characteristics, and developmental regulation. J Biol Chem. 2003;278:8065–74.

64. Aboobaker A, Blaxter M. Hox gene evolution in nematodes: novelty conserved. Curr Opin Genet Dev. 2003;13:593–8.

65. Aboobaker A, Blaxter M. The nematode story: Hox gene loss and rapid evolution. Adv Exp Med Biol. 2010;689:101–10.

66. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2012;15:734–48.

67. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet. 2014;1:221–33.

68. Mortazavi A, Schwarz EM, Williams BA, Schaeffer L, Antoshechkin I, Wold B, et al. Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. Genome Res. 2010;20:1740–7.

69. Dillman AR, Mortazavi A, Sternberg PW. Incorporating genomics into the toolkit of nematology. J Nematol. 2012;44:191–205.

70. Kaya HK, Stock SP. Techniques in insect nematology. In: Lacey L, ed. Manual of techniques in insect pathology. San Diego, CA: Academic Press Limited; 1997.

71. White GF. A method for obtaining infective nematode larvae from cultures. Science. 1927;66:302–3.

72. Stiernagle T. Maintenance of C. elegans. In: The *C. elagans* Research Community, eds. WormBook. 2006. doi/10.1895/wormbook.1.7.1

73. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

74. WormBase Parasite. http://parasite.wormbase.org

75. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.

76. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.

77. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-seq. Bioinformatics. 2009;25:1105–11.

78. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–U174.

79. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.

80. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

81. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.

82. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29–37.

83. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

84. Felsenstein J. PHYLIP (Phylogeny Inference Package). 36th ed. 2005.

85. Nguyen KB, Maruniak J, Adams BJ. Diagnostic and phylogenetic utility of the rDNA internal transcribed spacer sequences of Steinernema. J Nematol. 2001;33:73–82.
86. Simmons MP, Muller KF, Webb CT. The deterministic effects of alignment bias in phylogenetic inference. Cladistics. 2011;27:402–16.
87. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
88. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
89. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56:564–77.
90. Swofford DL. Phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer Associates; 2002.
91. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125:1–15.
92. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.
93. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
94. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
95. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
96. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004;20:1453–4.

97. Saldanha AJ. Java Treeview-extensible visualization of microarray data. Bioinformatics. 2004;20:3246–8.

98. Felsenstein J, Churchill GA. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol. 1996;13:93–104.

99. Margulies EH, Blanchette M, Program NCS, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. Genome Res. 2003;13:2507–18.

100. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.

101. Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, et al. Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res. 2008;18:1955–68.

102. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

103. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–208.

104. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.

105. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8:R24.

106. Araya CL, Kawli T, Kundaje A, Jiang LX, Wu BJ, Vafeados D, et al. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. Nature. 2014;512:400–U363.

107. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative Analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science. 2010;330:1775–87

CHAPTER *3*

Comparative transcriptomics of *Steinernema* embryonic development

Abstract

Cells express distinct sets of genes in a precise spatio-temporal manner during embryonic development. There is a wealth of information about embryonic development in *C. elegans*, but much less is known about embryonic development at the molecular level in nematodes from other taxa. We are interested in insect pathogenic nematodes from the genus *Steinernema* as models of parasitism and symbiosis as well as a satellite model for evolution in comparison to *C. elegans*. We determined the timing of embryonic development in two *Steinernema* species (*S. carpocapsae* and *S. feltiae*) for which we have assembled genomes, as well as for two *Caenorhabditis* species (*C. elegans* and *C. angaria*). We found that the timing between embryonic developmental stages in *Steinernema* is longer than in *Caenorhabditis*, and that the timing is also variable between the pairs of closely related species. We sequenced the transcriptomes of single embryos of each species during embryonic development at eleven specific stages (zygote, 2-cell, 4-cell, 8-cell, 24-44-cell, 64-78-cell, comma, 1.5-fold, 2-fold, moving, and L1) for comparative analysis. Single embryo transcriptomes were highly correlated within replicates and also generally highly correlated with neighboring developmental stages. Correlations between single embryos drastically decrease between the 4-cell and 8-cell stage in both *Steinernema* species, while in both *Caenorhabditis* species, a moderate decrease in correlation occurs later between the 8-cell and 24-44-cell stage. Our analysis of known *C. elegans* maternal transcripts in the four species revealed that the expression of maternal transcripts showed a discrepancy in timing between the *Caenorhabditis* and *Steinernema* species, which is indicative of differences in the maternal to zygotic transition in the two genera. We compared the temporal expression of other orthologs in *Steinernema* and *Caenorhabditis* to determine their degree of temporal conservation during development between these two taxa.

Introduction

Embryonic development in *Caenorhabditis elegans* is deterministic and is characterized by invariant cell lineages (Sulston, 1983). Studies have been done to perturb a large gamut of regulatory factors to uncover their roles in *C. elegans* lineage specification during embryonic development, and many factors have been well characterized and documented (Gerstein et al., 2010; Araya et al., 2014). However, far fewer molecular and genetic studies have been conducted on nematodes that are distantly related to *C. elegans* and comparative developmental studies across nematodes have been based primarily on observations (Schierenberg, 2006). These studies have noted and compared features of early divisions across nematodes, such as the synchronicity of the divisions, the sizes of cells produced from the divisions, the cell-cell interactions ("T" shape embryo vs "I" shape embryo after removal of egg shell) after the divisions, and when the timing of cell fate commitment occurs in them (Voronov et al., 1998; Schierenberg, 2006). Many of these developmental features segregate based on their phylogeny. For example, clade 2 nematodes have synchronous cell divisions and produce cells of equivalents sizes that are unspecified, while clades 3-12 follow asynchronous divisions and produce cells of different sizes with determined cell fates (Voronov et al., 1998). Differences in the timing between developmental stages and the occurrence of certain developmental landmarks such as gastrulation spur questions about how similar gene expression is at equivalent stages across diverse nematode species such as whether different nematodes species express the same genes at the same stages of development, how conserved is the expression of orthologous genes during development, how much of the transcriptome changes from one stage to another in a species, and how much of gene expression similarity across species depends on absolute time versus dependent on morphological stage?

Thus far molecular studies of comparative development have focused primarily on the genus *Caenorhabditis*. A comparative study of embryonic developmental gene expression was conducted across five Caenorhabditid species in order to investigate the relationship between embryonic developmental morphology and gene expression in the genus (Levin et al., 2012) in order to determine whether there are 'phylotypic' stages during embryonic development. The phylotypic stage is a stage of development where morphological variation, and thus gene expression variation, across species is minimal. They found that the time for each species to reach the same developmental stage (morphological stage) varied and found that the degree of transcriptome divergence between any two stages is dependent on time. If the timing between stages in one species took 3 hours and the timing in another took 4 hours, then the transcriptome should in theory be more divergent in the second species because the transcriptome has had more time to change in expression from the first state. They found that this generally occurred, except when two specific developmental stages were considered. Levin et al found that at the 4th division of the AB lineage (~24-cell stage) and especially at the ventral enclosure stage (~421-560-cell stage), divergence in gene expression became independent of time suggesting that the evolutionary constraints at these stages are stronger than at other development stages. Crucial developmental regulators involved in muscle and neuron tissue differentiation, and proteins containing homeobox, immunoglobulin-like, SH3, PDZ, and PH (cell-cell signaling) domains were also enriched at the ventral enclosure stage, suggesting that this stage could be the 'phylotypic' stage (Levin et al, 2012). While this study showed that time plays an important role in gene expression during development, it did not delve into the degree of ortholog expression conservation during development across the species. In addition, it also only compared closely related species that are all from the same genus. Given that nematodes are so diverse, we were

interested in investigating how gene expression varies during development across species of different genera.

While clade 10 nematodes such as *Steinernema* are thought to develop very similarly to clade 9 worms such as *C. elegans*, we found in a previous study that mixed-stage embryonic gene expression showed little conservation between *C. elegans* and *Steinernema* (Dillman et al., 2015). We were interested in whether these expression differences reflect the variations in their modes of embryonic development. Comparisons of gene expression at different developmental stages between *Steinernema* species and *C. elegans* revealed a higher degree of ortholog expression conservation between *Steinernema* species than between either of them and *C. elegans* (Figure 1). Surprisingly, the orthologs that had diverged in expression between *Steinernema* and *C. elegans* had a striking Gene Ontology (GO) enrichment for transcription factors and genes involved in pattern and cell fate specification (Fisher's exact test, FDR < 0.05) (Figure 1), which were processes and functions that we had hypothesized would have been conserved between them because of the extreme conservation of the roundworm body plan across species. To determine whether a single developmental stage contributed more to the divergence in gene expression, we removed each stage one at a time and repeated the analysis. We found that the removal of the embryonic stage resulted in few orthologs that had diverged in expression and no significant GO term results (Fisher's exact test, FDR < 0.05). This led us to reason that either 1) we had not collected a balanced "mixed stage" embryo population for our embryonic analyses for *C. elegans* and/or *Steinernema carpocapsae* and *Steinernema feltiae*, or 2) embryonic development at the level of orthologous gene expression may proceed very differently between species of these different genera. In order to answer both of these questions, we produced a high-resolution RNA-seq time course of embryonic development in *S.*

carpocapsae, *S. feltiae*, and *C. elegans* along with a more distantly related Caenorhabditid for which a genome has already been sequenced (*Caenorhabditis angaria*) (Mortazavi et al., 2010) (See phylogeny in Figure 2). In this chapter, we investigate 1) the degree of conservation of embryonic developmental gene expression between these genera and within each genus, 2) how the timing of embryogenesis varies across them, and 3) what pathways could be significantly different between them during embryogenesis.

Results

Embryonic developmental timing varies across nematodes

We imaged the embryonic development of *S. carpocapsae*, *S. feltiae*, *C. elegans* and *C. angaria* at 24°C to determine how developmental timing varies among them, and found that the *Steinernema* species take longer to develop from the 2-cell stage to the L1 stage than the *Caenorhabditids* do (Figure 3A). This increase in developmental time corresponds mainly to delayed early cleavage divisions in *Steinernema*. Specifically, the timing between the 4-cell to 8-cell and 8-cell to 24-44-cell stage is approximately 50% longer in *S. carpocapsae* and *S. feltiae* than it is in the *Caenorhabditids*.

Stage-specific transcriptomes of individual staged embryos

We investigated transcriptome changes during embryonic development of *S. carpocapsae*, *S. feltiae*, *C. elegans*, and *C. angaria* spanning 11 developmental stages (zygote, 2-cell, 4-cell, 8-cell, 24-44-cell, 64-78-cell, comma, 1.5-fold, 2-fold, moving, and L1) using RNA from individual embryos in quadruplicates (Figure 3B-C). We first asked whether orthologous genes showed conserved expression patterns over the course of embryogenesis in order to get

insights on the level of conservation of development between *Steinernema* and *Caenorhabditis* at the level of gene expression. We found that both *Steinernema* species had higher numbers of expressed genes (defined as > 1 TPM) than the *Caenorhabditis* species and that this was also true for genes that are present in a single copy across species and share ancestry (1:1:1:1 orthologs) (Figure 4A and 4B). However, the number of expressed genes (as well as 1:1:1:1 orthologs) were more comparable between genera at the later stages of embryonic development (from the comma to L1 stage), than at the earlier stages. We considered whether the larger numbers of expressed genes in *Steinernema* could be due to larger numbers of annotated genes in the *Steinernema* genomes. Interestingly, we found that the proportions of expressed genes are comparable across species: between 35-56% of all genes in *S. carpocapsae,* 40-60% of all genes in *S. feltiae,* 35-55% of all genes in *C. elegans*, and 35-48% of all genes in *C. angaria* are expressed at any given time during embryogenesis (Figure 4A). This analysis shows that the number of expressed genes and orthologs (> 1 TPM) is highly variable across the species during early embryonic development and less variable during later stages.

Single embryo correlations

Since the time between early embryonic stages is longer in *Steinernema* species, we postulated that the gene expression between pairs of early embryonic stages is potentially more divergent (less correlated) in *Steinernema* when compared to *Caenorhabditis* (Levin et al., 2012). To verify this, we calculated the Pearson's correlation coefficient between all pairs of single embryo transcriptomes for each species (Figure 5). We confirmed that 1) replicate embryo transcriptomes were highly correlated with each other, and 2) there was no contamination from embryos of other stages due to sample swaps. We found, as expected, that embryos that are more

distant in time showed lower correlations than embryos that are closer stages in all four species. However, the degree of correlation between corresponding adjacent embryonic stages showed marked differences between the two genera with *Caenorhabditis* species showing higher correlation between adjacent early embryonic stages than *Steinernema* species. In terms of the overall structure of the correlation matrices, we found similar structures between species of each genus, in contrast to the different structures observed across genera. Interestingly, the *Steinernema* correlation matrices showed a drastic decrease in transcriptome correlation (from > 0.9 to < 0.6) between 4-cell and 8-cell embryos. This substantial change in transcriptomes could reflect the earlier onset of maternal transcript degradation in *Steinernema*. These stage-to-stage transcriptome changes were less pronounced in *Caenorhabditis* because most of the early embryonic stages (from the zygote to the 4-cell) correlated so highly with each other that the stages could not be differentiated from each other globally. Because the global gene expression of the zygote and 2-cell are representative of the maternal transcriptome, we attempted to determine when zygotic transcriptional change commence and dominate. In doing so, we were able to detect a slight drop in correlation at the 8-cell stage in *C. elegans* and at the 24-44-cell in *C. angaria*. This suggests that the transcriptional landscape of *Steinernema* is changing faster than *Caenorhabditis* in the early embryo and that the onset of maternal transcript degradation is occurring at a later stage in *Caenorhabditis angaria* compared to *C. elegans*. Thus we observe both a set of within-genus differences as well as more dramatic differences between genera at the earliest embryonic stages.

Maternal *oma-1/2* dynamics during early embryogenesis

In *C. elegans* embryos, the degradation of the maternally deposited proteins and transcripts *oma-1* and *oma-2* are crucial for the activation of zygotic gene expression (Tadros et al., 2009; Stitzel et al., 2006). When phosphorylated by MBK-2, proteins OMA-1 and OMA-2 work in concert to sequester TAF-4, an important member of the RNA pol II complex, preventing it from activating zygotic gene expression (Tadros et al., 2009; Stitzel et al., 2006). We explored whether the embryonic stages at which we detect the first upregulation of zygotic gene expression across all four species coincide with downregulation/degradation of maternal *oma-1/2* transcripts (Figure 6). We investigated the orthology and expression of the *oma-1/2* gene across the four species and found that *C. elegans* underwent a triplication of an ancestral *oma* gene to produce *oma-1*, *oma-2*, and *moe-3*. Both *oma-1* and *oma-2* transcripts are highly expressed in the *C. elegans* zygote, but we found that *oma-1* is downregulated one stage earlier than *oma-2* (8-cell versus 24-44-cell). While *C. elegans* has three *oma* genes involved in oocyte maturation, we found that the other species have only a single copy of the *oma* gene that shares homology with these *C. elegans* genes. Focusing on the dynamics of these closely related *oma* genes, we find that the *oma-1/2* transcripts in *S. carpocapsae*, *S. feltiae*, and *C. angaria* are downregulated by the 2-cell, 8-cell, and 24-44-cell stage, respectively (Figure 6). We further found that more distant paralogs of the *oma* genes in all four species (*pos-1, mex-3, mex-5, mex-6, ccch-1, ccch-2, ccch-5, Y11A8C.20, dcf-13, C35D6.4, F38C2.7, Y60A9.3, Y116A8C.19*) are also strictly maternally expressed (Figure 7A-B). We further found that there are fewer *oma* paralogs in the *Steinernema* species and *C. angaria* (8 in *S. carpocapsae,* 7 in *S. feltiae,* and 5 in *C. angaria*) than in *C. elegans* (16 in *C. elegans*), indicating that these paralogs in other species may combine the roles of more than one paralog in *C. elegans*. Although we find evidence of degradation of the *oma-1/2* transcripts earlier in *Steinernema*, we lack data on when the OMA-

1/2 proteins are degraded to establish whether *oma-1/2* transcript degradation is responsible for the earlier upregulation of genes that we observe.

Genus-specific trajectories during embryonic development

To assess how gene expression of single embryos varies across species during embryonic development, we performed principal component analysis (PCA) on all of the single embryos (175) from all four species for the set of 4,156 1:1:1:1 orthologous genes (Figure 8). We found that Principle Component 1 (PC1), which accounts for 21.9% of the variance across the single embryos, separated the embryos based on developmental time (early embryos versus intermediate embryos versus late embryos). We found that PC2 (14.6%) separated embryos by the number of orthologs expressed, PC3 (8.9%) separated embryos by genus, and PC4 separated *C. elegans* and *C. angaria* embryos, but not the Steinernematids (Figure 8A-C). We tracked the developmental trajectories of each species on a plot of PC1 versus PC2, and found a clear difference between the early embryos (from the zygote to 24-44-cell stage) of *Steinernema* and *Caenorhabditis* along PC2, but observed a convergence in later embryos from the 64-78-cell stage to the L1 stage (Figure 8A). The top and bottom 100 gene loadings contributing to differences along PC3 are orthologs that have taken on very different expression profiles during development between the two genera (Figure 8B, Figure 8D). The PCA plots clearly show divergence of ortholog expression between genera at the earliest stages of development followed by convergence in expression at later stages.

Orthologous gene and transcription factor profiles during embryogenesis

A heatmap of 1:1:1:1 orthologs confirms that a set of orthologs which are expressed primarily during later embryonic development (comma to L1) shows conserved expression over the embryonic stages across all four of the species. However, we can also see that another set of orthologs, which appear to be strictly maternal in *C. elegans* and *C. angaria*, i.e. are expressed only from the zygote stage up until the early or intermediate stages (8-cell to 24-44-cell), show downregulation at earlier stages (4-cell to 8-cell) in *Steinernema*, and interestingly, are then re-expressed in later stages of development (Figure 9A). This suggests that maternal-specific and other early embryonic orthologs in *Caenorhabditis* have new, additional roles in later embryonic development in *Steinernema*. Alternatively, these orthologs may have been expressed in these later stages in ancestral species and have been lost at these time points in *Caenorhabditis*.

Another noticeable feature of the ortholog heat maps is a lack of highly expressed orthologs at the 8-cell and the 24-44-cell stages in *S. feltiae*, and to a lesser extent the 2-cell through 8-cell stages in *S. carpocapsae* when the heat maps are clustered based on expression pattern in *C. elegans*. We hierarchically clustered the 1:1:1:1 orthologs based on expression in other species and found 305 orthologs in *S. feltiae* and 403 orthologs in *S. carpocapsae* that are expressed most highly in the 8-cell and 24-44-cell stages, showing that there are orthologs expressed at these stages in the *Steinernema* species (Figure 9B).

Since transcription factors (TFs) are responsible for regulating the expression of genes during development, we suspected that the expression of transcription factors would mirror the profiles observed for the 1:1:1:1 orthologs and would also show major differences in the early and intermediate embryonic stages between the genera. We plotted the expression of TFs that are orthologous across all 4 species, 3 out of the 4 species, and 2 out of the 4 species to assess their expression profiles (Figure 10). The 253 1:1:1:1 orthologous TFs showed identical expression

dynamics as the set of all 4,156 1:1:1:1 orthologs (Figure 9A). The subset of TFs that are expressed primarily during early embryogenesis in *Caenorhabditis* show less early embryo-specificity in *Steinernema,* with these TFs most highly expressed at the 24-44-cell and 64-78-cell stages in *Steinernema*. Focusing on TFs across all species combinations, we find that the maternal and early transcription factors are species- and genus- specific. We found many TFs that were specific to *C. elegans* (159) or *C. elegans* and *C. angaria* (99) that have diverse expression profiles during the time course. The group of 159 *C. elegans*-specific TFs includes 66 nuclear hormone receptors and the GATA TFs *end-3* and *end-1* that specify the endoderm at the 8-cell and 24-44-cell stage respectively.

Focusing on TFs that are expressed in *S. carpocapsae* and one or more species but not in *C. elegans* (189 TFs), we find 26 TFs (14%) that have early embryo-specific expression. The set of 189 TFs found in *S. carpocapsae,* but not *C. elegans,* had GO enrichments, such as positive mesodermal fate specification (FDR=1.1e-5), response to retinoic acid (FDR=1.9e-5), dorsal/ventral pattern formation (FDR=1.3e-4), positive regulation of cell differentiation (FDR=1.1e-3), neuron projection morphogenesis (FDR=3.7e-3), and BMP signaling (FDR=1.8e-2) (Table 1). These results suggest that the Steinernematid specific-TFs are likely to participate in the regulation of multiple developmental processes.

Differential gene expression analysis of adjacent stages to find major expression transitions

In order to detect specific transcriptional changes between early embryos, we performed differential gene expression (DE) analyses between pairs of adjacent early developmental stages using either all of the genes within each species or the 4,156 1:1:1:1 orthologs shared between them (Figure 11 and Figure 12). In *Caenorhabditis*, very few genes or orthologs were

differentially expressed (FDR < 0.05 and fold change > 2x) between stages before 4-cell. Once the embryos reached the 8-cell stage in *C. elegans*, 972 genes became differentially upregulated relative to the 4-cell stage, consistent with our previous correlation matrix results (Figure 11, Figure 5) and with previous published results showing that zygotic expression begins at the 4-cell stage in *C. elegans* (Edgar et al., 1994, Baugh et al., 2003). In contrast, *C. angaria* showed very little change in gene expression until the 8-cell to 24-44-cell stage transition. At that point, 1,440 genes were upregulated in the 24-44-cell stage relative to the 8-cell stage, indicating that zygotic transcriptional changes are occurring at later developmental stages in *C. angaria* than in *C. elegans*. Both *Steinernema* species showed a substantial upregulation of gene expression (4,787 genes in *S. carpocapsae* and 2,938 genes in *S. feltiae*) at the 8-cell stage similar to *C. elegans*. However, both *S. carpocapsae* and *S. feltiae* show upregulation in a subset of genes prior to the 8-cell stage, in the 2-cell (541 genes) and 4-cell stages (251 genes), respectively. A Gene Ontology (GO) analysis of these early upregulated genes at 2-cell in *S. carpocapsae* found that they are enriched for terms involved in yolk granules and ubiquitination (Fisher's exact test, FDR < 0.05). We did not find any significant GO term enrichments for the early upregulated *S. feltiae* genes. It is unclear why there is an upregulation from zygote to 2-cell and then a plateau in gene expression from 2-cell to 4-cell in *S. carpocapsae*. It may be that zygotic transcription starts for a small subset of genes at an earlier stage in *Steinernema* than in *Caenorhabditis*, although additional mechanisms would need to be ruled out experimentally (see Discussion).

Gene expression dynamics during embryogenesis in individual species

We used maSigPro (Conesa et al., 2006; Nueda et al., 2014) to find gene sets that share common temporal dynamic profiles over the embryonic time course in each of our species.

maSigPro uses a two-step regression strategy in which it first identifies differentially expressed (DE) genes, and then distinct, statistically significant expression profiles. The expression levels of each gene were tested against a null model, where gene expression does not change over the time course, to determine genes that are significantly differentially expressed during embryogenesis (FDR < 0.05 in *C. elegans*, *C. angaria*, *S. carpocapsae* and *S. feltiae*) (Figure 14). Genes with similar developmental expression trajectories were clustered together to generate nine expression clusters per species. We observed four major types of expression profiles across the four nematodes: a profile that represents the maternal transcripts (clusters 1 and 2), a profile that represents the first transcripts expressed by the zygote that do not overlap with maternally deposited ones (clusters 3 and 4), a profile that represents transcription in later stages (comma to 2-fold) when morphogenesis and organogenesis are occurring (*Caenorhabditis* = clusters 5, 6, *Steinernema* = clusters 5, 6, 7), and finally a profile that represents transcription that will characterize the L1-stage worm (*Caenorhabditis* = clusters 7, 8, 9, *Steinernema* = clusters 8, 9)

In *C. elegans*, we find differences in the rates of transcript decay within the two clusters of genes that are expressed early and represent transcriptional products that were deposited in the embryo by the mother (clusters 1 and 2). *C. elegans* cluster 1 transcripts begin degrading between the 4-cell and 8-cell stages and their levels are drastically reduced by the 24-44-cell stage, while cluster 2 transcripts show a slower rate of decay, degrading linearly over time from the 8-cell stage until the comma to 2-fold stages. We also found that each of these *C. elegans* gene clusters have slightly different functional annotation enrichments, such as proteasome complex (FDR = 1.2e-33) in cluster 1 and gastrulation with mouth forming first (FDR = 8.4e-13) in cluster 2 (Table 2).

In *S. carpocapsae*, the maternal transcripts (cluster 1) appear to drastically decrease by the 8-cell stage. *S. feltiae* maternal transcripts (cluster 1) also appear to substantially decrease by 8-cell, but then they exhibit a gradual decline and are present at low levels into the 24-44-cell and 64-78-cell stage. However, in *C. angaria,* both sets of maternal transcripts (cluster 1 and 2) decline greatly by the 24-44-cell stage, but the average expression of genes in cluster 2 remains lowly expressed (at ~25 expression units) at least until the 64-78-cell stage.

Interestingly, the expression plots show that the *C. elegans* cell cycle transcripts are almost completely degraded by the comma stage, directly correlating with previous results that cell number ceases to increase by the comma stage in *C. elegans* (Figure 15) (Karabey et al., 2003). We focused on the expression of cyclin transcripts in the four species to observe how they behave and found that across all four species, cyclin b (*cyb-1*) is either the most highly expressed or one of the most highly expressed cyclins primarily during the early stages of embryogenesis (Figure 16). The time at which *cyb-1* and other cyclins degrade, however, is variable across the species. *cyb-1* transcripts degrade by the 8-cell stage in *S. carpocapsae* and by the comma stage in *C. elegans* and *C. angaria*. Interestingly, in *S. feltiae*, the *cyb-1* transcript degrades at the 24-44-cell stage, peaks at the 64-78-cell stage, and degrades again by the 2-fold stage. This is an interesting finding because it could indicate that maternal cyclin-b is succeeded by other more lowly expressed cyclins to fulfill the requirement for cell division in *S. carpocapsae*. Alternatively it could mean that cyclin protein levels are maintained for a longer periods of time in *S. carpocapsae* than the other nematodes.

maSigPro ortholog expression clustering within each genus

To determine orthologs that show significant temporal expression dynamics between the species in *Steinernema* and *Caenorhabditis*, we used maSigPro on 9,844 1:1 orthologs shared between *S. carpocapsae* and *S. feltiae* and 6,840 1:1 orthologs that are shared between *C. elegans* and *C. angaria* (Figure 17A). We found that 4,819 (48.9%) of the *Steinernema* orthologs and 4,462 (65.2%) of the *Caenorhabditis* orthologs were dynamically expressed during embryonic development (Benjamini Hochberg FDR < 0.01). These dynamically expressed genes partitioned into 9 different clusters (Figure 17B-C) based on their expression profile. Clusters 1 and 2 show the dynamics of the early orthologous embryonic or "maternal" transcripts, clusters 3 and 4 show the dynamics of early to intermediate embryonic development (8-cell to 24-44-cell or 74-78-cell), clusters 5 and 6 show the dynamics of intermediate to late genes, and clusters 7-9 show the dynamics of very late development until hatching. The clusters also represent orthologs that are higher on average in one species than another at around the same time points during development. For example, *Steinernema* cluster 1 and 2 show genes that are "high" in both *Steinernema* species very early on during embryogenesis, but it is clear that cluster 1 genes are much higher in *S. carpocapsae* than *S. feltiae*, while cluster 2 genes are higher in *S. feltiae* than *S. carpocapsae* (Figure 17B). It is interesting that for all of these clusters, except for cluster 9 in *Steinernema*, there is a fairly large difference in the magnitude of expression between species of the same genus.

Contributions of non-orthologous genes to embryonic development in *S. carpocapsae*

Given our finding that our 1:1:1:1 orthologs show more conserved expression during later development, we asked what the contribution of the other 75% of genes are to development. We focused on *S. carpocapsae* genes that share homology with at least one other *Steinernema*

species (*S. feltiae*, *S. glaseri*, *S. monticolum*, and *S. scapterisci*) but no homology to any of the *Caenorhabditis* species (*C. elegans, C. angaria, C. briggsae, C. japonica,* and *C. remanei*), and that are both expressed at an average of 10 TPM during embryonic development and have at least one replicate with expression > 50 TPM (Figure 18). These expression thresholds were set to ensure that these genes are true expressed genes and not pseudogenes. We found 5,679 genes that fit these criteria and that 1,036 genes (18.1%) are expressed between zygote to 4-cell (clusters 1 and 2), 2,272 genes (39.8%) are expressed between 8-cell and 64-78-cell (cluster 2 and 3), and the remaining 2,389 (41.9%) genes are expressed at some point between comma to L1 (clusters 4 and 5) (Figure 18). Approximately half of these *Steinernema*-only genes (2,674, 47%) have no match to proteins from any other species. Of the 3,005 genes that do have annotations, 24 are fatty-acid and retinol binding proteins, fatty-acid amide hydrolases, fatty-acid desaturases or fatty-acid elongation protein annotations, and 53 are ubiquitin E3 ligases or ubiquitin-related proteins. Another 26 are homeobox-domaining containing proteins. This could suggest alternative gene expression cascades or programs governing *Steinernema* development. Together, *Steinernema*-conserved genes that have no *Caenorhabditis* orthologs are expressed throughout embryogenesis and are likely to affect several processes during their development.

The neddylation pathway is upregulated during *Steinernema* early embryogenesis

Neddylation is a protein modification process similar to ubiquitination that is important for many functions across eukaryotes and that plays a very important in *C. elegans* embryonic development (Bosu et al., 2010). It is unclear how many targets of neddylation there are, but a well-studied set of targets is the cullin family of proteins (Bosu et al., 2010; Enchev et al., 2015). Cullins are protein scaffolds that, when neddylated, hold ubiquitin E3 ligases close to their

protein targets for ubiquitination (Enchev et al., 2015) (Figure 19A-B). Once target substrates are ubiquitinated, they can be recognized and degraded by the proteasome complex. Thus, neddylation is essential for activating cullins to target specific proteins for degradation. A genome-wide protein domain analysis across *Steinernema* and multiple sequenced nematode species revealed an expansion of the cullin domains in *Steinernema* relative to other nematodes (Dillman et al., 2015; See Figure 2C in chapter 2) with 19 and 46 cullin domains in *S. carpocapsae* and *S. feltiae* respectively compared to 5-8 cullin domains in *C. elegans*, *Drosophila*, and human (Petroski et al., 2005). This large expansion suggests an increased role for cullins in *Steinernema*. We also investigated the prevalence of cullin domains in other clade 10 species and found that there are more cullin domains in the more closely related *Panagrellus redivivus* (16 domains) than the more distantly related *Bursaphalanchus xylophilus* (9 domains). This suggests that cullins have expanded in the superfamily *Panagrolaimoidea* to which *Steinernema* and *P. redivivus* belong. Interestingly, when we analyzed the top 100 most highly expressed genes during early embryonic development in *Steinernema* (from zygote to 24-44-cell stage), we found that multiple genes critical to the neddylation pathway (*ned-8*, *ubc-12*, *rbx-1*, *csn-5*, *ula-1*, *dcn-1*) were among the top expressed genes during early embryonic development, which was not the case in *C. elegans* or *C. angaria* (not shown). We also found that *Steinernema* species show a sharp spike in expression for all neddylation pathway members (*ned-8*, *ubc-12*, *rbx-1*, *csn-5*, *ula-1*, *dcn-1*, *cand-1*) prior to gastrulation at the 8-cell to 24-44-cell stage (Figure 19B).

Because neddylated cullins help to bring substrates and ubiquitin ligases together, we were interested to see how the expression of ubiquitin pathway members changed during development. There are two ubiquitin genes, *ubq-1* and *ubq-2*, upwards of 17 ubiquitin

conjugating enzymes (*ubc-1* to *ubc-26*), and one ubiquitin-activating enzyme (*uba-1*) in *C. elegans*. We focused our analysis on the ubqituitins, ubiquitin-activating enzyme, and the expression of two ubiquitin-conjugating enzymes that are essential to embryonic development in *C. elegans* (*ubc-2* and *ubc-14*). We found that ubiquitin (*ubq-1*) expression decreases during development from zygote to L1 in *Caenorhabditis* (Figure 20). However, in *Steinernema*, the expression of *ubq-1*, which is the main ubiquitin-producing gene, peaks at the 8-cell stage and 24-44-cell stage in *S. carpocapsae* and *S. feltiae* respectively, paralleling the expression profiles of *ned-8* and other neddylation pathway components in these species. Interestingly, the expression of *ubq-1* and *ubq-2* is anticorrelated in all four species. This is very striking for *S. feltiae*, where the peak at the 8-cell stage of *ubq-1* corresponds to a dip of *ubq-2* at the same stage. The expression of the *ubq-1* gene is similar to the expression of the neddylation pathway genes, which suggests that these pathways are coordinated. However, the expression of other members of the ubiquitin pathway such as the ubiquitin activating enzymes (*uba-1*) and the E2 ubiquitin conjugating enzymes (*ubc-2* and *ubc-14*), which cleave ubiquitin for activation and ligate ubiquitin to target substrates respectively, do not show the same expression profile as neddylation and *ubq-1*. Together these results suggest a greater role for neddylation and cullin-mediated ubiquitination in early to intermediate embryonic development in *Steinernema* that could be an interesting avenue for further study.

Using single-embryo RNA-seq to resolve the developmental stages of pooled embryo RNA-seq

In our previous work, we collected mixed-stage embryo populations and found that the gene expression differences between the embryonic data sets across species of different genera

were large. We used the single-embryo data sets to determine the stages of our mixed-stage embryo populations by plotting the Pearson's correlation coefficient for the expression between each single embryo and RNA-seq sample (Figure 21). We found that *S. carpocapsae* RNA-seq datasets correlated most highly with the 8-cell to 24-44-cell stage embryos, *S. feltiae* correlated most highly with the 4-cell, 8-cell, and 24-44-cell stage embryos, and *C. elegans* correlated highly with the all stages, but most highly with the later developmental stages (comma, 1.5-fold, and 2-fold stage). This indicates that we did in fact compare embryos from later developmental stages to earlier developmental stages across species skewing our gene expression conservation results in our previous analysis. However, we found that gene expression in our high-resolution time course is indeed different during early embryonic development regardless of the previous technical staging issues.

Discussion

We generated a high-resolution single embryo RNA-seq time course that spans 11 developmental stages in four species to determine the extent of ortholog expression conservation during embryonic development across distantly related species. We found that 1:1:1:1 orthologs expressed primarily during early stages of embryogenesis in *C. elegans* had diverged in expression in *Steinernema*, while those expressed during late embryogenesis up until hatching showed greater conservation across the species. Focusing on the early stages, we found that larger transcriptional changes were occurring at earlier stages in *Steinernema* than in *Caenorhabditis*. Specifically, we found that genes were upregulated as early as the 2-cell stage in *S. carpocapsae* and the 4-cell stage in *S. feltiae* which, interestingly, also coincided with the degradation of the *oma-1/2* transcripts in these species. This evidence supports the idea that

zygotic gene expression is happening at earlier stages in *Steinernema*. However, because we do not know when the OMA-1/2 proteins are degraded and because we see very little overall maternal transcript degradation at these stages, it is unclear if zygotic transcription is actually occurring at these times. Alternatively, this upregulation could be the result of another mechanism such as differential transcript polyadenylation, which has been characterized to occur in *Xenopus laevis* during embryogenesis as a way to express maternal transcripts at the correct times during early development without requiring zygotic genome activation (i.e. active zygotic transcription) (Radford et al., 2008; Simon et al., 1992).

Interestingly, we found that all the members of the neddylation pathway are upregulated in the Steinernematids but not in the Caenorhabditids, with *S. carpocapsae* neddylation transcripts consistently peaking in expression one developmental stage before *S. feltiae*. It is very striking that all neddylation pathway members are tightly regulated in *Steinernema* at these stages that overlap gastrulation. Their coordinated expression would suggest that they are regulated by the same set of TFs. Proper protein clearance is essential prior to gastrulation. It is possible that with the expansion of the cullins, there was also a diversification of cullins and the target substrates they pair with (i.e. each cullin could have its own set of substrates). Whatever the case may be, the expression of neddylation pathway genes strongly indicates that protein degradation prior to gastrulation in *Steinernema* species is likely to be even more critical than in *C. elegans*.

A transcriptome analysis conducted by Levin et al. during the embryonic development of five *Caenorhabditis* species found that ortholog expression was constrained at several points during the middle of embryogenesis within the genus (Levin et al., 2012). They termed these points of convergence developmental milestones, and their findings are reminiscent of the highly

debated ‘phylotypic’ stage of the hourglass model of animal development. The evolutionary biologist Duboule first proposed this model in 1994, and it predicts that embryonic divergence during development follows an hourglass-like shape, where embryos of different species are most divergent at the earliest and latest stages of development, but not the middle stages of development when the body plan is being set (Duboule, 1994). For example, hox gene expression is seen as one such source of developmental constraint. Hox genes are TFs responsible for regulating body segment identity, and their expression is tightly controlled during the middle stages of development so that the proper cells and tissues are formed at the correct times (Bateson, 1984).

Our embryonic analysis has shown that expression of orthologs at later developmental time points show greater conservation than earlier ones between these species, which is less pronounced than expected in the hourglass model. The lower degree of expression conservation between 1:1:1:1 orthologs during earlier embryonic development in contrast to the later stages of development leads us to propose the funnel model of embryonic development for nematodes who are more distantly related than the ones considered in the Levin et al. study (Figure 22). In our model, gene expression variation is highest within the earliest stages (zygote to 8-cell) and lowest within the later stages (64-78-cell to L1), suggesting that there is greater developmental constraint on the expression of these later-stage orthologs. The massive variation we see in gene expression in early development is reminiscent of what nematologists have previously seen at the macroscopic level between different species, such as differences in the timing of gastrulation, AP axis specification, and when the endoderm and mesoderm cells are specified. Thus, our findings show that embryonic development is less constrained at the early stages and becomes more constrained as development progresses to a free-living L1 stage. Studying the molecular

differences at the early stages across these nematodes would be an interesting focus of future research.

Materials and Methods

Strains. *S. carpocapsae* (strain ALL) and *S. feltiae* (strain SN) were cultured and maintained according to Dillman et al. 2015. *C. elegans* (N2) were grown on Nematode Growth Media (NGM) plates seeded with OP50. *C. angaria* (PS1010) were grown on nutrient agar + 0.1% cholesterol plates seeded with OP50.

***Caenorhabditis* nematode culture and embryo isolation.** Mixed-stage populations of *C. elegans* and *C. angaria* grown on OP50 plates were collected by adding ddH_2O to the agar plates and swirling to lift the nematodes off of the plates. The nematode suspensions were poured into 15 mL conical tubes, and repeated until plates were clean. The suspensions were spun down at 2,000 RPM for 1 min, and washed twice with ddH_2O. Nematode pellets were treated for 5 min in a 5 mL solution containing 1.25 mL fresh bleach, 2.25 mL 1 M NaOH, and 1.5 mL ddH_2O in 15 mL conicals with intermittent vortexing. After the 5-minute incubation, the conical tubes were topped off with M9 buffer, spun at 2,000 RPM for 2 min, and embryo pellets were washed three times to remove traces of bleach solution.

***Steinernema* nematode culture and embryo isolation.** Approximately 10,000 *S. carpocapsae* and *S. feltiae* IJs were seeded on lipid agar plates on top of lawns of *Xenorhabdus nematophila* and *Xenorhabdus bovienii* respectively. Nematodes were grown at room temperature until gravid adults were present (3-4 days for *S. carpocapsae* and 2-3 days for *S. feltiae*), and adults were bleached to obtain embryos using the same protocol that was used for *C. elegans* above, except that the embryos were washed and collected in Ringer's solution instead of M9 buffer.

Embryonic time course. Embryos of *S. carpocapsae*, *S. feltiae*, *C. elegans*, and *C. angaria* were imaged every 5 minutes for 24 hours at 24°C on the EVOS inverted microscope (Figure 3A). *S. carpocapsae* and *S. feltiae* embryos were imaged in Ringer's solution, while *C. elegans* and *C. angaria* were imaged in M9 buffer. Time data for each stage transition was collected for at least 3 embryos. The average number of embryos collected per stage is 10 embryos. Developmental timeline was made using the timeline library in R version 3.2.3 (Bryer, 2013).

Experimental Design. We collected and sequenced single embryos at 11 embryonic stages per species (*S. carpocapsae*, *S. feltiae*, *C. elegans*, and *C. angaria*) in quadruplicates (Figure 3B). We amplified the very low quantities of mRNA from each of these individual embryos into cDNA by following Smart-seq2 protocol with minor modifications detailed below (Picelli et al., 2014) (Figure 3C). We sequenced a total of 175 single embryos; each was sequenced an average depth of 10 million reads.

Embryo collection for Smart-seq2. Pellets of embryos were resuspended in 2 mL of Ringer's solution (made with DEPC water) + 0.01% tween 20. DEPC was used in the Ringer's solution to limit RNase contamination, and tween 20 was used to prevent embryos from sticking to any surfaces. Resuspended embryos were passed through at 40 μm mesh filter into a 60mm x 15mm petri dish to remove debris. Enough Ringer's solution + 0.01% tween 20 was added to coat the bottom of the petri dish and reduce the density of the embryos so that they could easily be collected. Embryos were visualized in the dish using an EVOS inverted microscope, and single embryos were imaged and collected in 1.5 μL using a micropipette. If more than one embryo was collected, embryos were diluted further by pipetting them into 20 μL Ringer's solution + 0.01%

tween 20 on a clean slide that was pretreated with RNase ZAP or 70% ethanol. Single embryos were collected in 1.5 µL into PCR tube strip, and 2 µL of lysis buffer (18 µL 0.3% Triton-X 100 + 2 µL RNase inhibitor SIGMA), 1 µL of oligo-dT primer, and 1 µL of dNTP mix were added to each embryo. Embryos were heated to reverse secondary structure of RNA, reverse transcribed and PCR amplified according to the Smart-seq2 protocol by Picelli (Picelli et al., 2014) (Figure 3C). All embryos, regardless of embryonic stage, were amplified for 18 cycles through PCR. PCR primers were cleaned up from the embryo samples by adding a 1:1 ratio of Ampure XP beads to sample, which were both equilibrated to room temperature, incubated for 8 min, placed on a magnet, and washed with 200 µL of 80% ethanol 3 times. Beads were dried at room temperature for approximately 5 min (until the beads cracked), after which, 17.5 µL of EB was added and incubated off the magnet for 3 min. Samples were placed back on the magnet, and 15 µL of cDNA was collected for each sample. Sample cDNA concentration was quantified using the Qubit fluorometer and bioanalyzed using the Agilent 2100 Bioanalyzer to check the cDNA quality.

Single embryo library preparation and sequencing. For library preparation, 20 ng of cDNA from each sample was prepared using the regular Nextera tagmentation protocol (Gertz et al., 2012). The protocol reagents were scaled down, so that 2 µL of transposase, 10 µL of buffer, and 8 µL of cDNA (20 ng total) were used yielding a total volume of 20 µL. Transposase was cleaned up from the tagmented DNA using the QIAGEN columns as follows. Three volumes of buffer PM was added to each sample, placed into a QIAGEN spin column and spun at 13,000 RPM for 1 min. Flow through was removed, 750 µL of buffer PE (prepared with ethanol) was added, and samples were spun at 13,000 RPM for 1 min. Flow through was removed, and

samples were spun again at the previous settings to dry the columns. Columns were placed into new clean collection tubes, and 30 µL of EB prewarmed to 55°C was added to the center of each column and incubated for 1 min before spinning down at the same settings.

In a PCR tube, 30 µL of sample, 35 µL of Phusion high fidelity master mix, 2.5 µL 25 µM Nextera adapter ID XX, and 2.5 µL 25 µM Nextera adapter Ad_noMX were combined and mixed well with a pipette. Samples were spun down quickly, and amplified for 6 cycles using the PCR program with the following settings: 72 °C for 5 min, 98 °C for 30 s, [98 °C for 10 s, 63 °C for 30 s, 72 °C for 1 min] for 6 cycles, 72 °C for 5 min, and hold at 4 °C.

PCR amplified libraries were cleaned up using a 1:1 ratio of Ampure XP beads to sample, and prepared in the same way as the bead cleanup above, except that 30 µL of EB was added to the beads to resuspend the library sample, which was then collected in 27.5 µL after 2 min.

Sample library fragments were between 200-600 bps with an average size of 360 bps after the Nextera tagmentation protocol. Samples were sequenced as paired-end 43 bp on the Illumina NextSeq 500 to an average depth of ~10 million reads.

Gene expression analyses. Unstranded, paired-end RNA-seq reads for all species were trimmed to 40 bp from their 3' ends to remove low quality nucleotide sequences. Transcriptome indexes were prepared for *S. carpocapsae* (downloaded from WormBase ParaSite), *S. feltiae* (downloaded from WormBase ParaSite), *C. elegans* (WS220), and *C. angaria* using the RSEM command (version 1.2.12) rsem-prepare-reference (Li et al., 2011). Reads were mapped to each respective species' annotations using bowtie 0.12.8 with the following options: -S, --offrate 1, -v 1, -k 10, --best, --strata, -m 10 (Langmead et al., 2009). Gene expression was quantified using the RSEM command, rsem-calculate-expression, with the following options: --bam, --fragment-

length-mean (Li et al., 2011). For all analyses, gene expression was reported in Transcripts Per Million (TPM).

Orthology relationships analysis. Orthologs and paralogs were determined across the four species by blasting their protein sequences using OrthoMCL 1.4 with the default settings (Li et al., 2003). Additionally, manual annotation of orthologs and paralogs of select genes for analyses was done using WormBase ParaSite.

Differential Expression Analyses. Differential gene expression was determined using the Bioconductor package, edgeR v.3.2.4 (Robinson, 2010). The RSEM count data was used for calculating differential expression, and genes were called as differentially expressed if they had an FDR < 0.05 and a fold change > 2x. Four replicates were used per stage for the analysis, except for the 64-78-cell stage RNA-seq data for *C. angaria*, which had 3 replicates. Early adjacent stages were pair-wise compared to detect the onset of the maternal to zygotic transcription (Figure 11 and 12). Late adjacent stages were pair-wise compared to detect large transcriptional changes to identify differences across species (Figure 13).

Correlation matrices. A pseudocount of 1 TPM was added to the gene expression of each gene for all the single embryos of each species and $\log_2$ scaled. Pearson's correlation coefficient (ρ) was determined from the data using the corr() function in R version 3.2.3 (R Development Core Team, 2008).

Heat maps. Heat maps of gene expression were mean-centered, normalized, and hierarchically clustered with Cluster 3.0 and visualized using Java Treeview (de Hoon et al., 2004, Saldanha et al., 2004).

Differential temporal dynamics during development in individual species with maSigPro. 28,313 genes in *S. carpocapsae*, 33,459 genes in *S. feltiae*, 20,389 genes in *C. elegans*, 27,970 genes in *C. angaria* were run through maSigPro as single time series using their respective time course data (Figure 14A). A pseudo count of 1 was added to each gene for each sample, and the gene counts were normalized in edgeR using calcNormFactors() and cpm(). maSigPro was run with counts = TRUE setting for count-based expression. Significance threshold (p-value) was adjusted to 0.05. Significant genes were clustered into 9 expression profiles for each species.

Differential temporal dynamics during development with maSigPro. 9,844 1:1 orthologs shared between *S. carpocapsae* and *S. feltiae* and 6,840 1:1 orthologs shared between *C. elegans* and *C. angaria* were run through maSigPro (Nueda et al., 2014) as multiple time series using *S. carpocapsae*'s and *C. elegans*' time course data respectively (Figure 17). A pseudo count of 1 was added to each gene for each sample, and the gene counts were normalized in edgeR using calcNormFactors() and cpm(). maSigPro was run with counts = TRUE setting for count-based expression. Significance threshold (p-value) was adjusted to 0.01. Significant genes were clustered into 9 expression profiles for each species.

Figures and Figure Legends

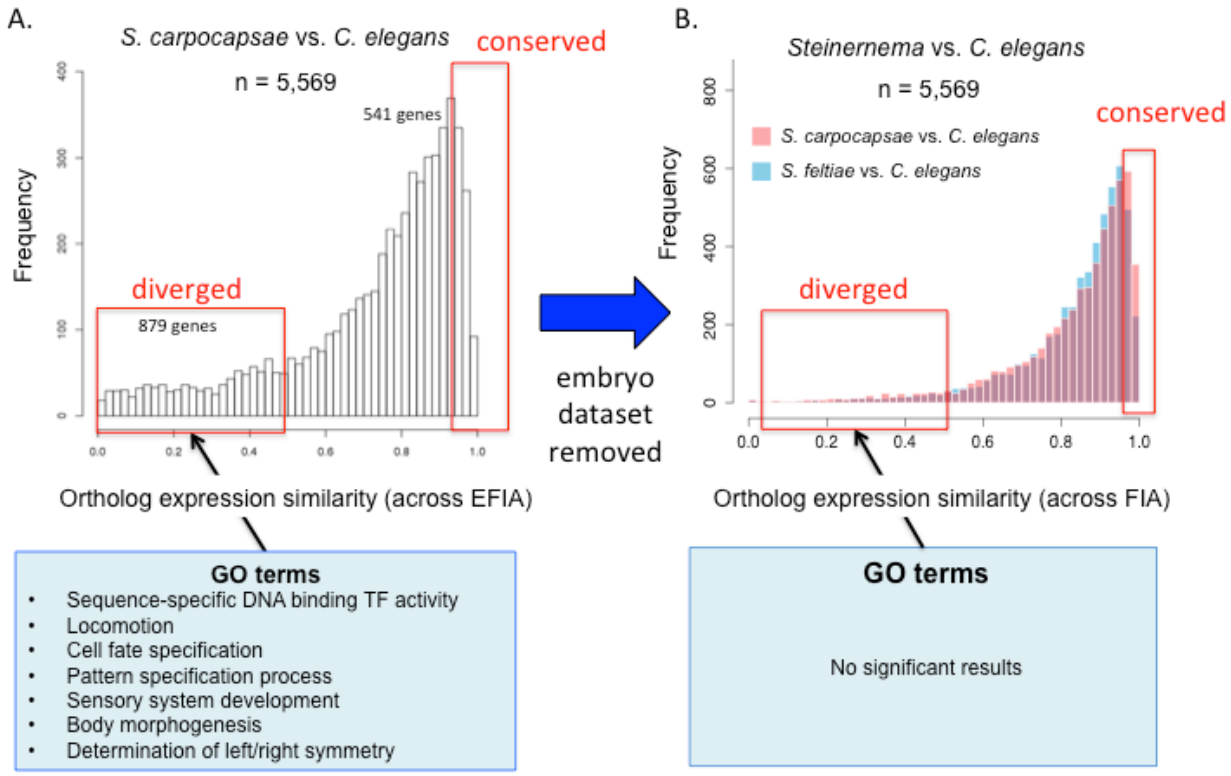


Figure 1. Potential embryonic ortholog expression divergence. A) Histogram showing the ortholog expression similarities (cosine similarity) of 5,569 1:1 orthologs across four stages of nematode development (embryo, first larva, dauer/IJ, and young adult) between *S. carpocapsae* and *C. elegans*. Orthologs with cosine similarities < 0.5 are considered to have diverged in expression, while orthologs with cosine similarities > 0.95 are conserved in expression. Gene Ontology analysis was performed on divergent orthologs. B) Histograms showing ortholog expression similarities of 5,569 1:1:1 orthologs between *S. carpocapsae* and *C. elegans* and *S. feltiae* and *C. elegans* after the embryonic datasets were removed from the analysis. Same conservation/divergence thresholds were used as above.

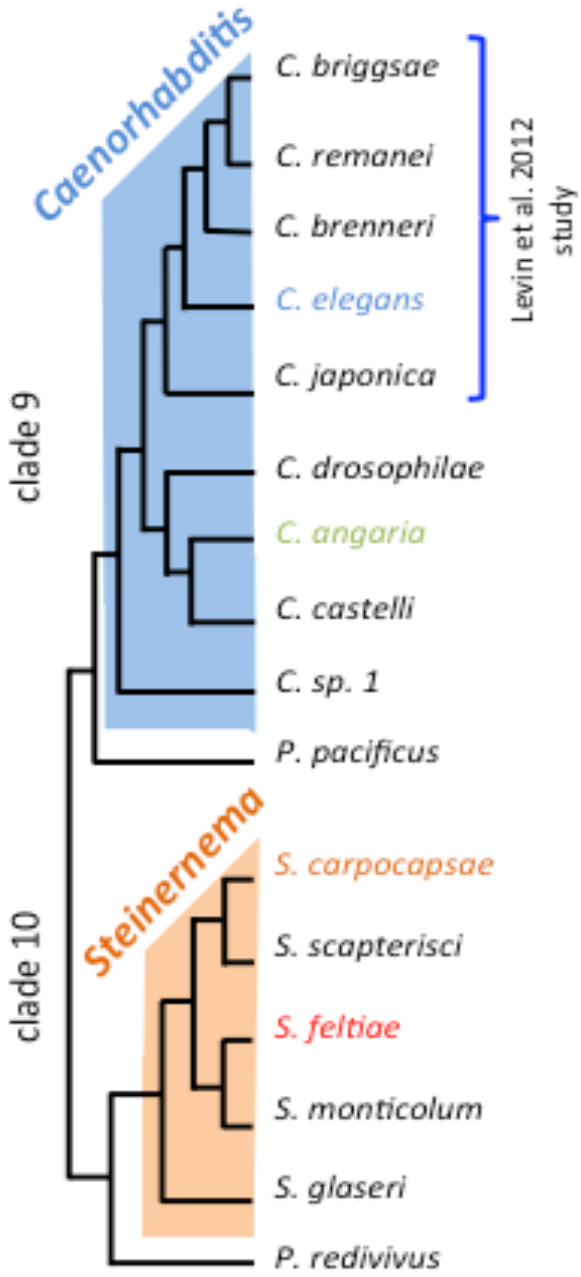


Figure 2. Phylogenetic tree showing the relationships of the nematodes in this study. Phylogenetic tree showing the relationships of the four nematodes in this study (*S. carpocapsae*, *S. feltiae*, *C. angaria*, and *C. elegans*). Several species from each genus and an outgroup species are included to highlight the evolutionary distances between the nematodes under investigation. Of note, the evolutionary distance between the Caenorhabditids in our study (*C. elegans* and *C. angaria*) is further than the distances between *C. elegans* and any of the four *Caenorhabditis* species chosen for the Levin et al. 2012 study. Branch lengths are not to scale.

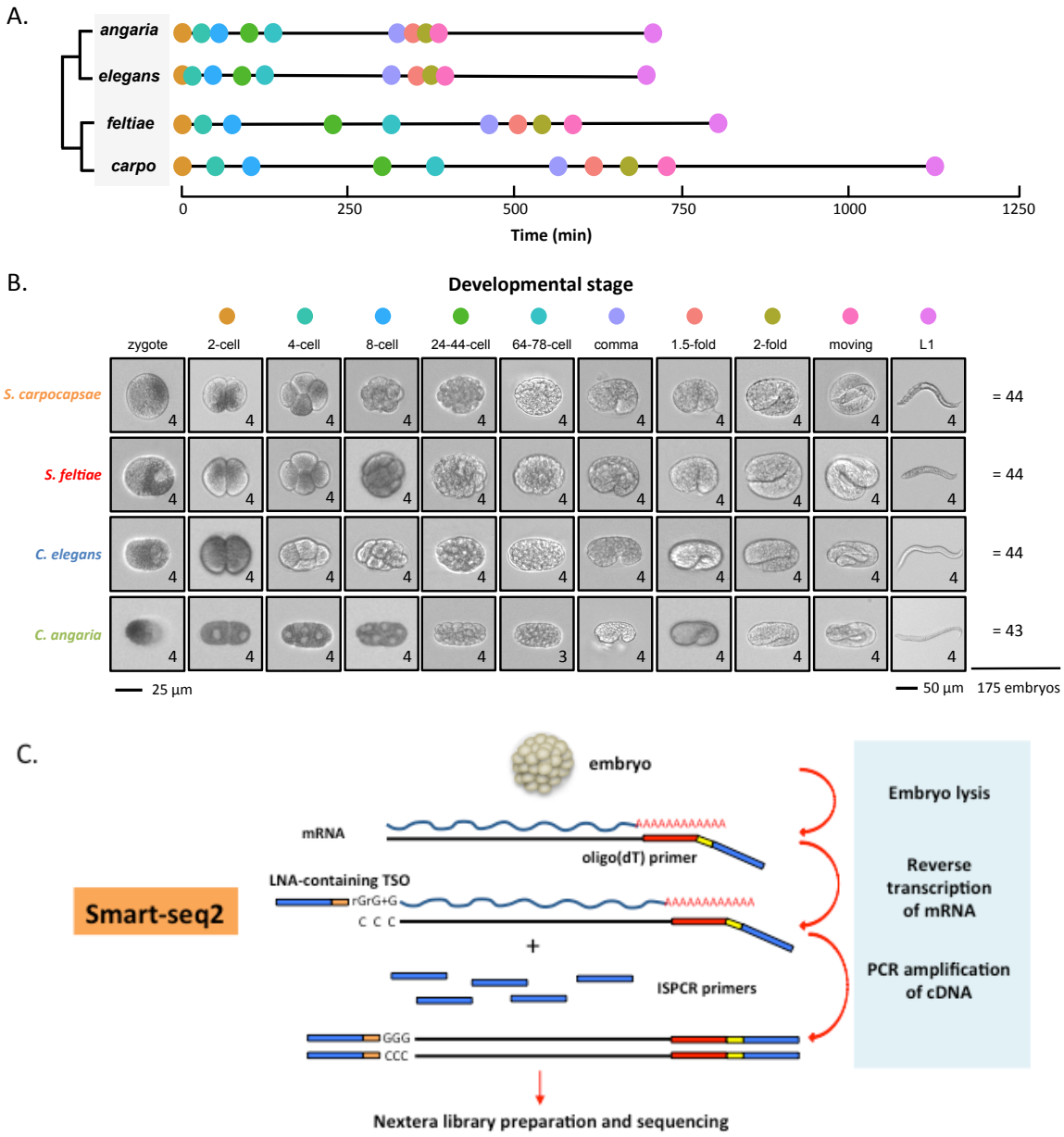


Figure 3. Timing of embryonic development at 24°C across four nematode species and the experimental design. A) Embryonic development was tracked using a time-lapse microscope for each species at 24°C with representative images of each stage shown. The timeline shows the average timing between stages based on at least 3 embryos imaged for the transition between pairs of stages. Stage key is in 3B. B) Images of the morphologies of 11 embryonic stages of two *Steinernema* and two *Caenorhabditis* species. Three to four embryos of each embryonic stage for each species were collected for single embryo RNA-sequencing with Smart-seq2. Embryos are on one scale (scale bar = 25µm) and the L1s are on another (scale bar = 50 µm). C) Smart-seq2 workflow. Single embryos were collected and lysed to extract total RNA. mRNAs were selectively reverse transcribed to produce full-length cDNA using olido(dT) primers containing PCR primers. Full-length cDNA was amplified for 18 PCR cycles and prepared into sequencing libraries.

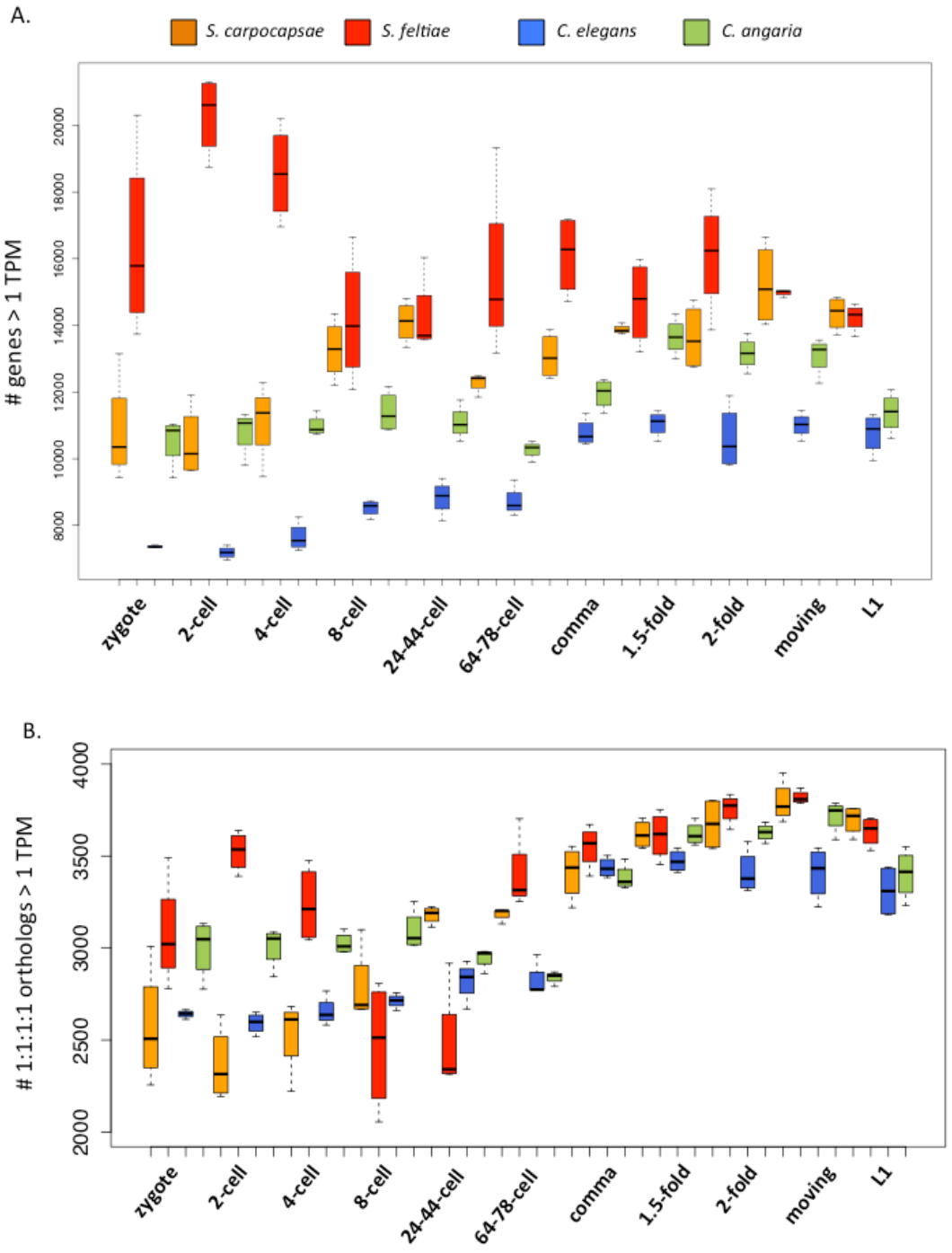


Figure 4. Numbers of expressed genes during development across four species. A) Number of genes expressed greater than 1 transcript per million (TPM) out of 28,313 genes in *S. carpocapsae*, 33,459 genes in *S. feltiae*, 20,389 genes in *C. elegans*, 27,970 genes in *C. angaria* at each embryonic stage. B) Number of 1:1:1:1 orthologous expressed greater than 1 TPM out of 4,156 orthologs shared between the four species.

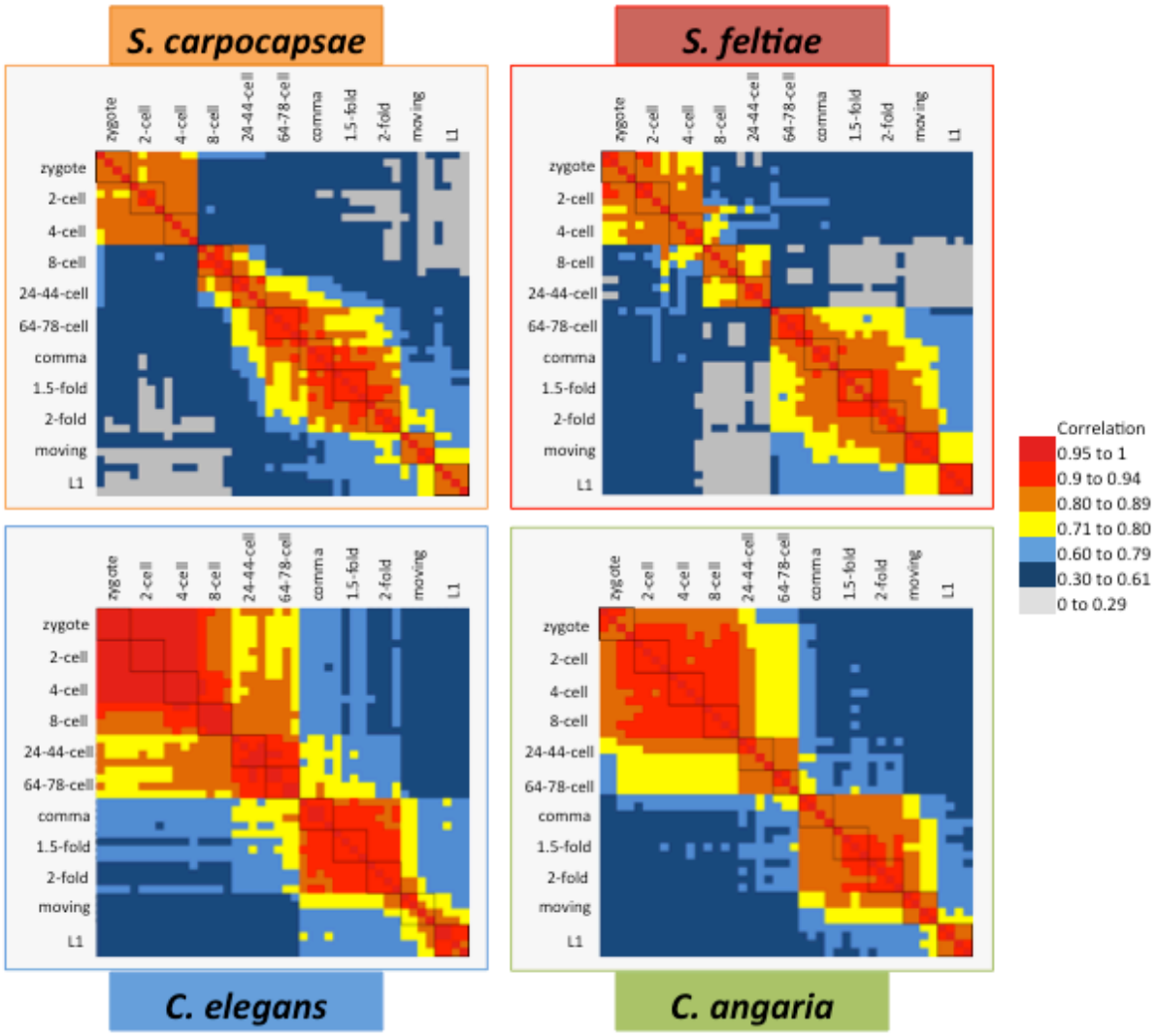


Figure 5. Transcriptome correlation across single embryos of four nematode species. All staged single embryo transcriptomes were pairwise compared to each other to determine their Pearson's correlation coefficients. Heat maps show the correlation coefficients of all the comparisons. Four replicate embryos are shown per developmental stage, except for the 64-78-cell stage in *C. angaria*, which has three replicate embryos. Red indicates almost perfect correlations (0.9 to 1), while grey indicates little to no correlation (0 to 0.3).

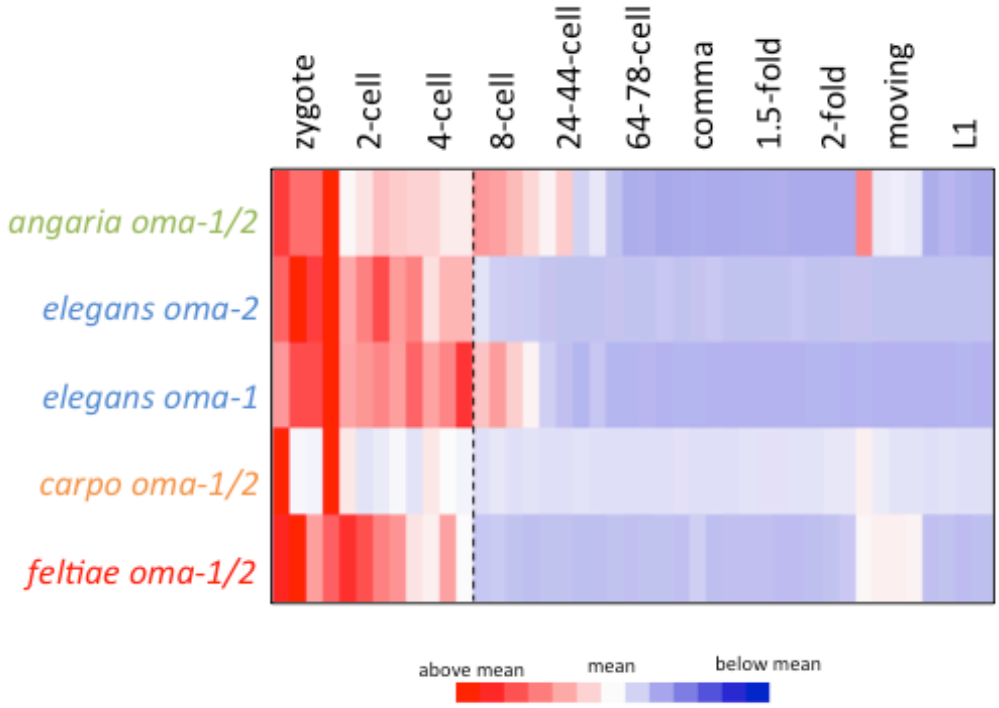


Figure 6. Degradation of maternal transcripts *oma-1/2* varies across species. Heat map of gene expression of *oma-1/2* gene. *C. elegans* has two copies of the *oma* gene, while the other species have only a single copy. Dashed line delineates the boundary between 4-cell and 8-cell replicates. All species show four embryo replicates per developmental stage, except for the 64-78-cell stage, which has three replicates.

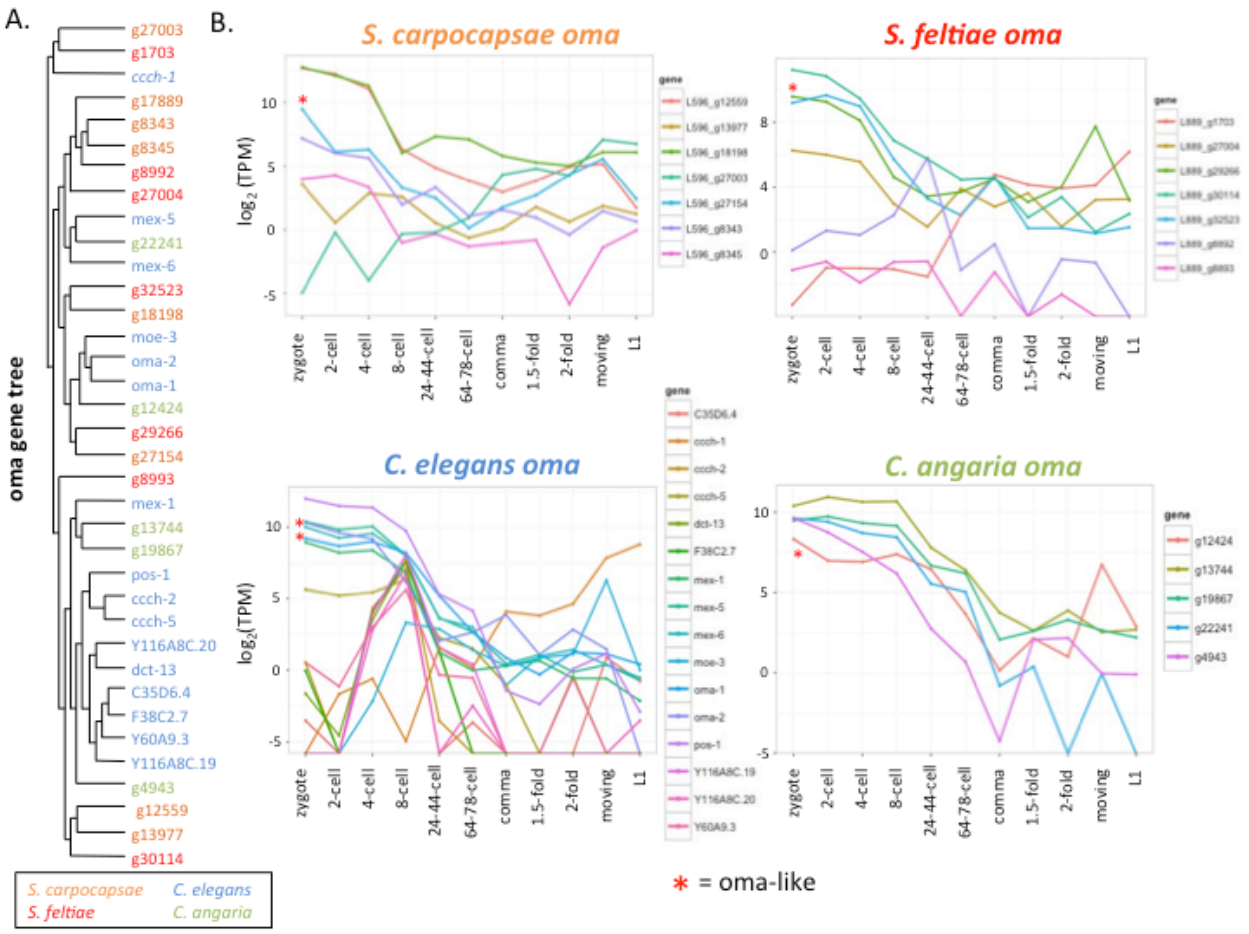
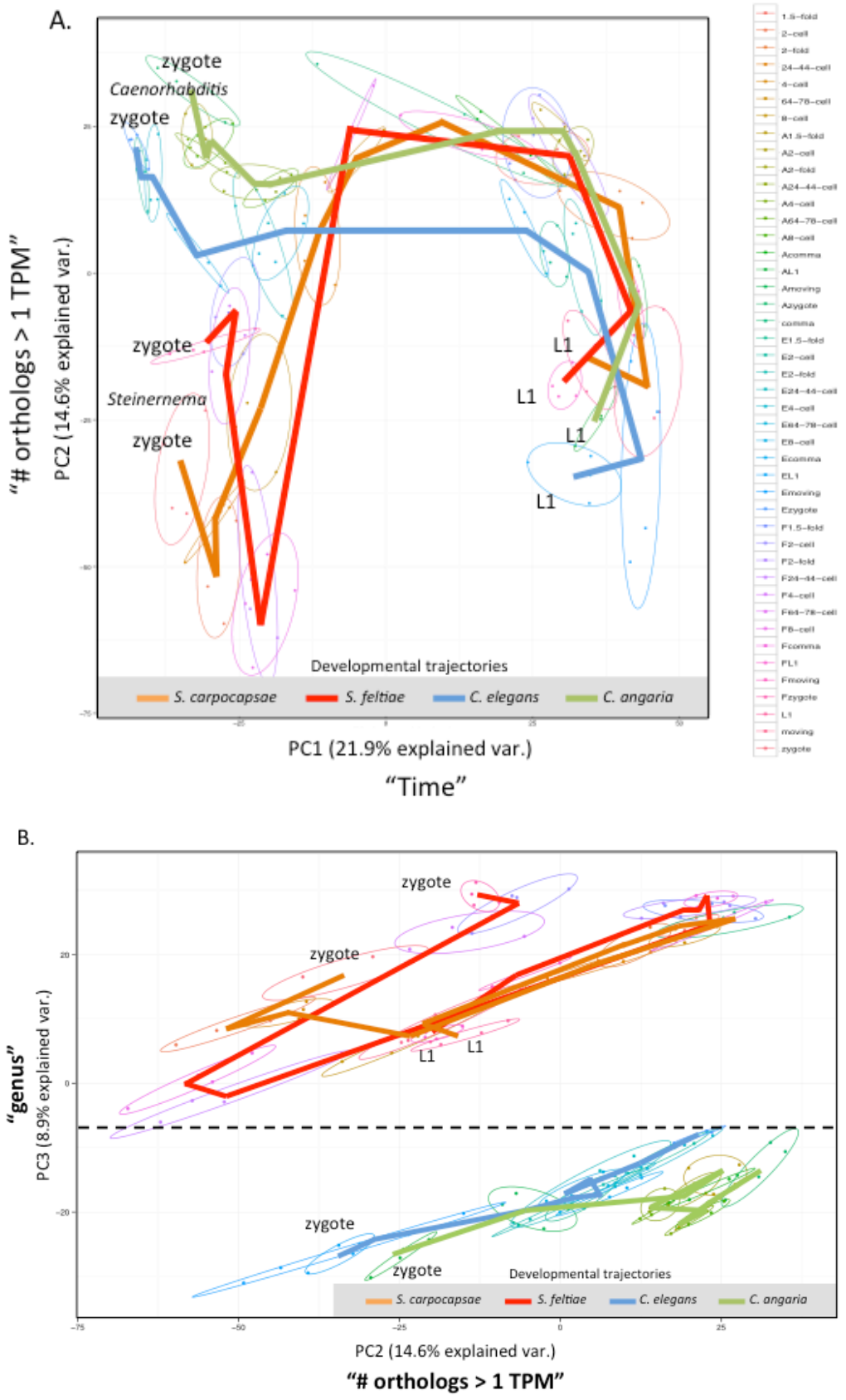
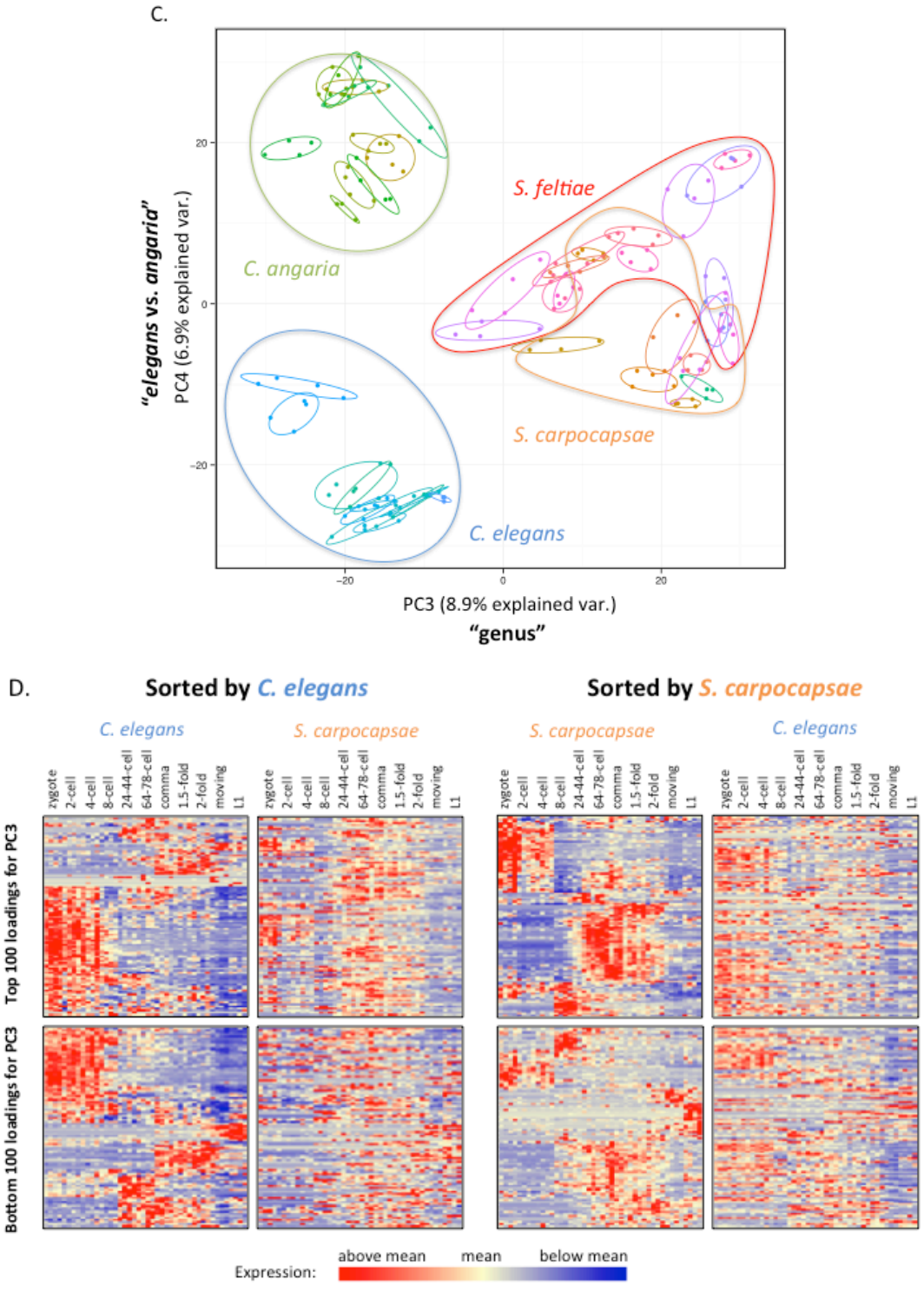


Figure 7. Oma gene relationships and expression profiles.
A) A tree of all genes in *C. elegans* (16 genes), *C. angaria* (5 genes), *S. carpocapsae* (7 genes), and *S. feltiae* (7 genes) that share sequence homology to the *C. elegans oma-1* and *oma-2* genes. Genes are color coded by species. B) Expression profiles of the *oma*-like genes in each species. Red asterisks indicate the genes that are closest in sequence identity to the *C. elegans oma-1* and *oma-2* genes. Expression (TPM) is displayed on log2 scale.

Figure 8. Principal Component Analysis of 4,156 1:1:1:1 shared orthologs between four species. A) PC1 versus PC2. B) PC2 versus PC3. C) PC3 versus PC4. The first two plots (A and B) show the developmental trajectories of each species and the clear genus-specific clustering in PC2 and PC3, but in the last plot, samples are circled by species to show the distinction between *C. elegans* and *C. angaria* from each other and the two *Steinernema* species. D) Gene expression of top and bottom 100 loadings of PC3. The expression of top and bottom 100 loadings/orthologs of PC3 for *S. carpocapsae* and *C. elegans* was sorted by *C. elegans* and *S. carpocapsae*. Gene expression is mean-centered.


A.
zygote
Caenorhabditis
zygote
zygote
Steinernema
zygote
L1
L1
L1
L1
Developmental trajectories
S. carpocapsae
S. feltiae
C. elegans
C. angaria
"# orthologs > 1 TPM"
PC2 (14.6% explained var.)
PC1 (21.9% explained var.)
"Time"
B.
zygote
zygote
L1
L1
zygote
zygote
Developmental trajectories
S. carpocapsae
S. feltiae
C. elegans
C. angaria
"genus"
PC3 (8.9% explained var.)
PC2 (14.6% explained var.)
"# orthologs > 1 TPM"


C.
S. feltiae
C. angaria
S. carpocapsae
C. elegans
"elegans vs. angaria"
PC4 (6.9% explained var.)
PC3 (8.9% explained var.)
"genus"
D.
Sorted by C. elegans
Sorted by S. carpocapsae
C. elegans
S. carpocapsae
S. carpocapsae
C. elegans
zygote
2-cell
4-cell
8-cell
24-44-cell
64-78-cell
comma
1.5-fold
2-fold
moving
L1
Top 100 loadings for PC3
Bottom 100 loadings for PC3
Expression:
above mean
mean
below mean

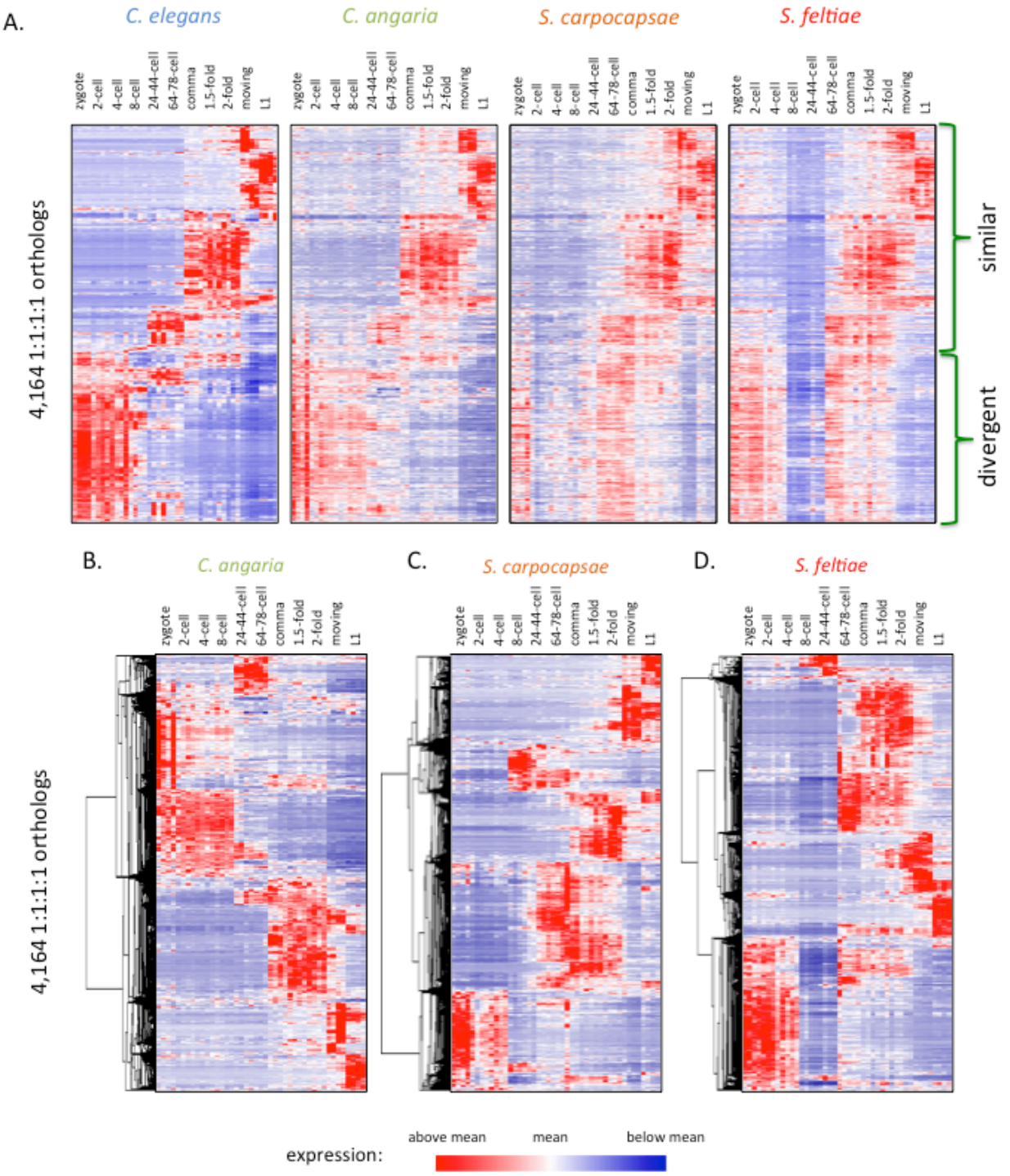


Figure 9. Heat maps of 1:1:1:1 ortholog expression during embryonic development. Gene expression (TPM – transcripts per million) during embryonic development of 4,164 1:1:1:1 orthologs was mean-centered and hierarchically clustered based on expression in A) *C. elegans,* B) *C. angaria*, C) *S. carpocapsae*, and D) *S. feltiae*.

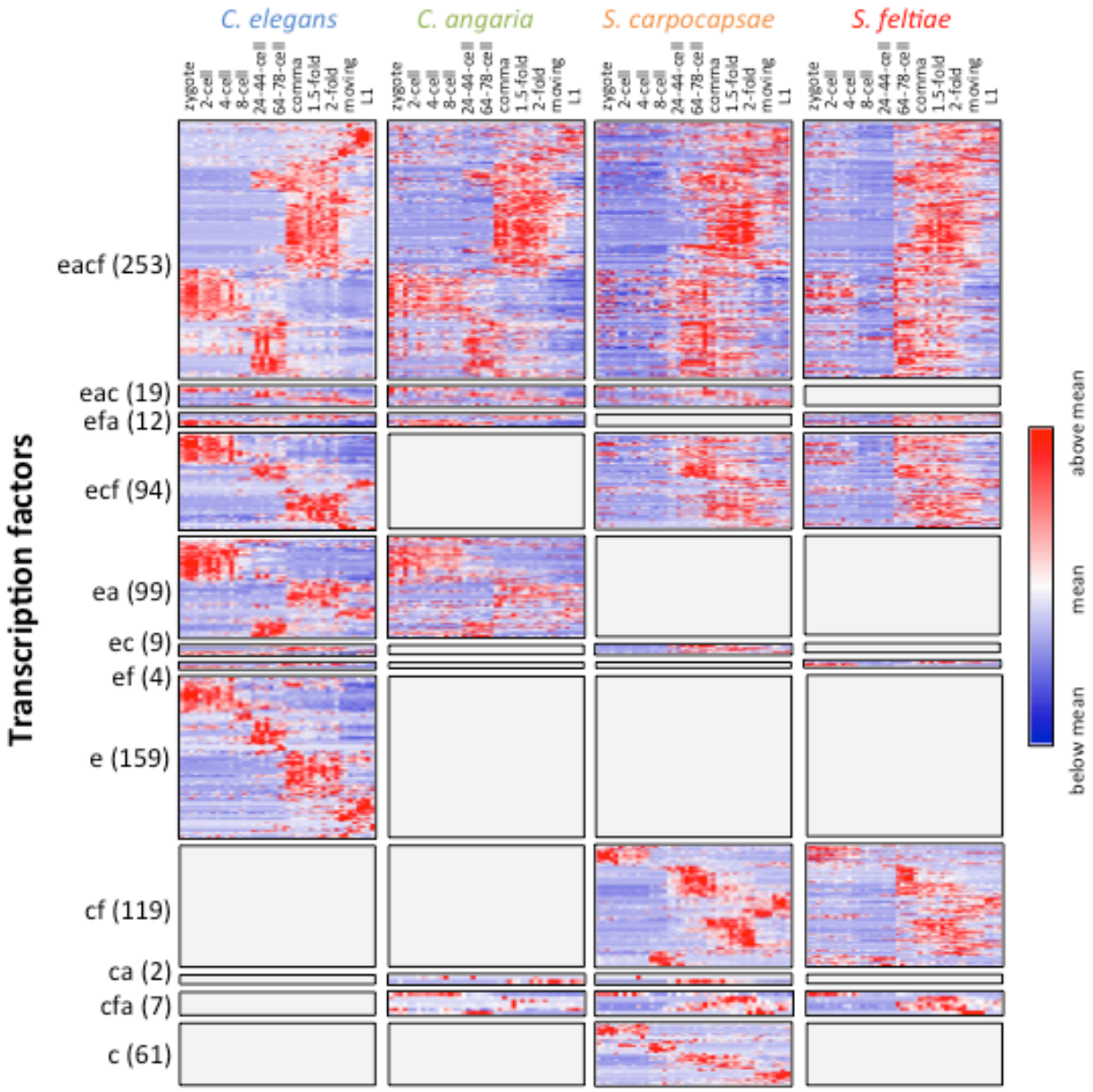


Figure 10. Transcription factor expression during nematode embryonic development. TF orthologs that are shared across all four species, 3 out of 4 species, 2 out of 4 species, *C. elegans* only and *S. carpocapsae* only were hierarchically cluster on expression in *C. elegans* and mean-centered. The letters e, c, f, and a indicate that the orthologs are present in *C. elegans*, *S. carpocapsae*, *S. feltiae*, and *C. angaria*, respectively. The numbers in parentheses indicate the number of orthologs found for the particular combinations of species.

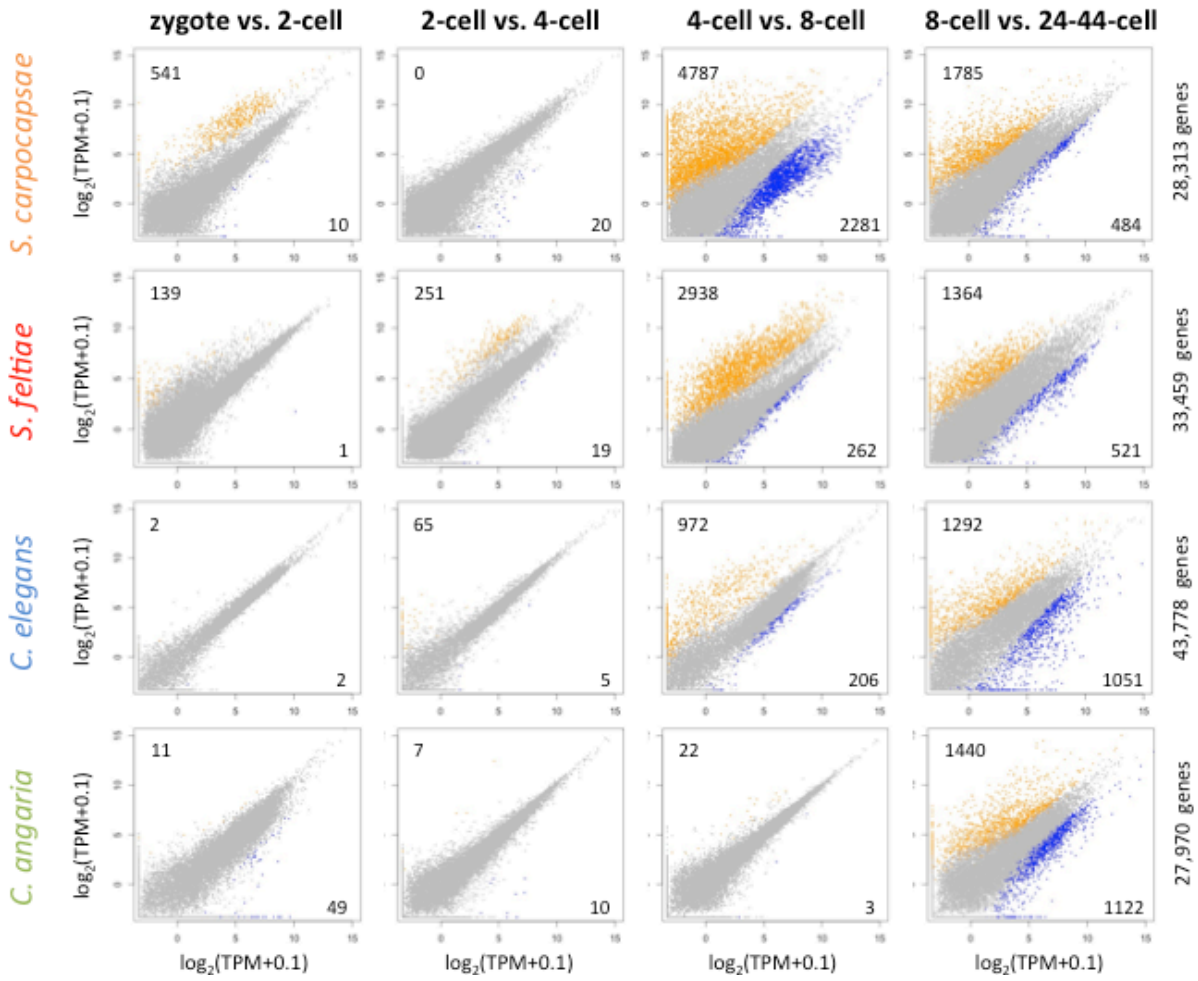


Figure 11. Differential gene expression of all genes during early embryonic development across species. Gene expression $\log_2$(TPM+0.1) of all genes was plotted for adjacent early developmental stages for all four species. The earlier stages are displayed on the x-axis and the later stages are displayed on the y-axis. Genes that are differentially expressed (FDR < 0.05 and fold change > 2X) between the stages, and are more highly expressed in the earlier stage or later stage are shown in blue and yellow respectively. Genes in gray are not differentially expressed.

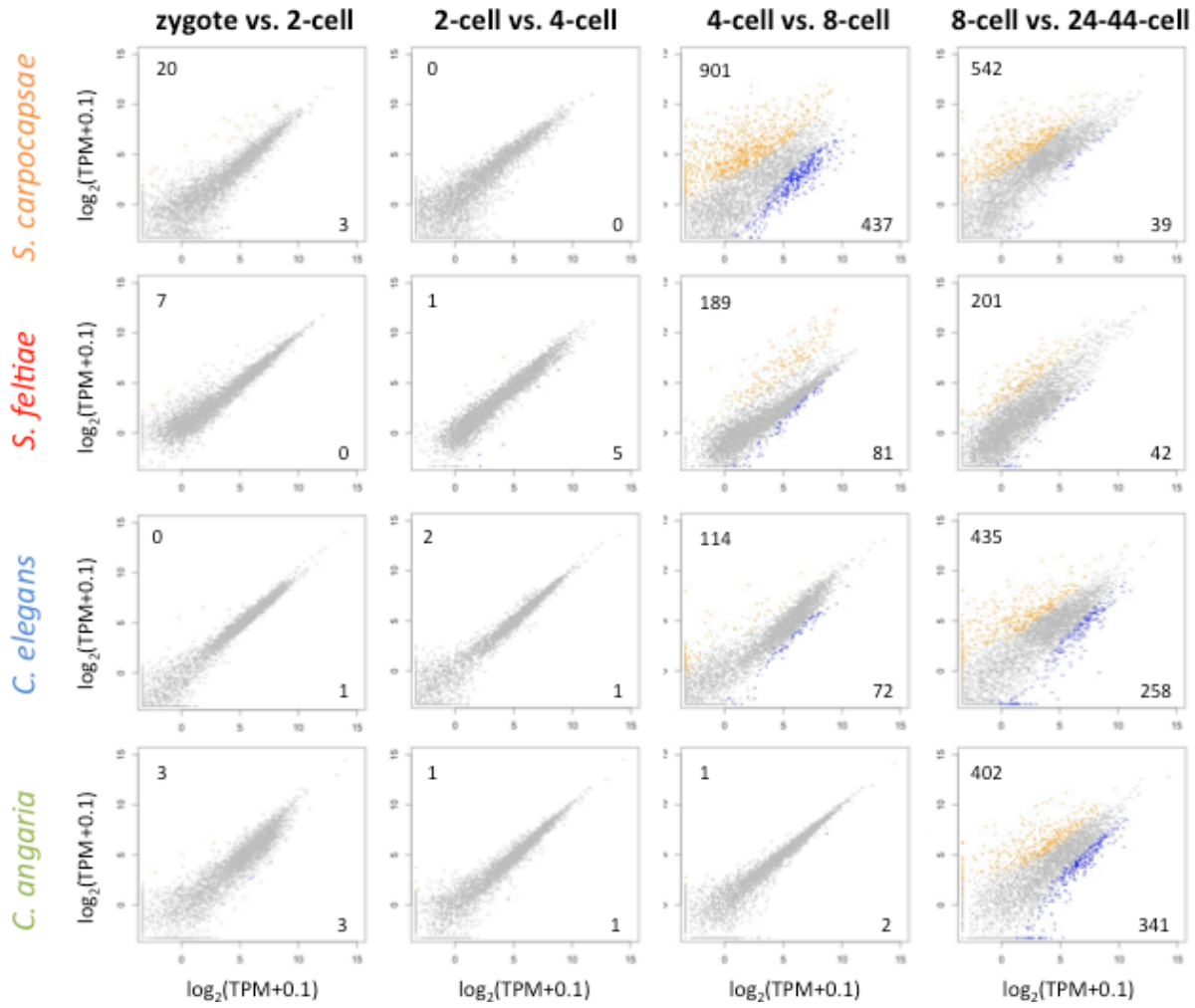


Figure 12. Differential gene expression of orthologous genes during early embryonic development across species. Gene expression $\log_2$(TPM+0.1) of 4,156 1:1:1:1 orthologs was plotted for adjacent early developmental stages for all four species. The earlier stages are displayed on the x-axis and the later stages are displayed on the y-axis. Orthologs that are differentially expressed (FDR < 0.05 and fold change > 2X) between the stages, and are more highly expressed in the earlier stage or later stage are shown in blue and yellow respectively. Orthologs in gray are not differentially expressed.

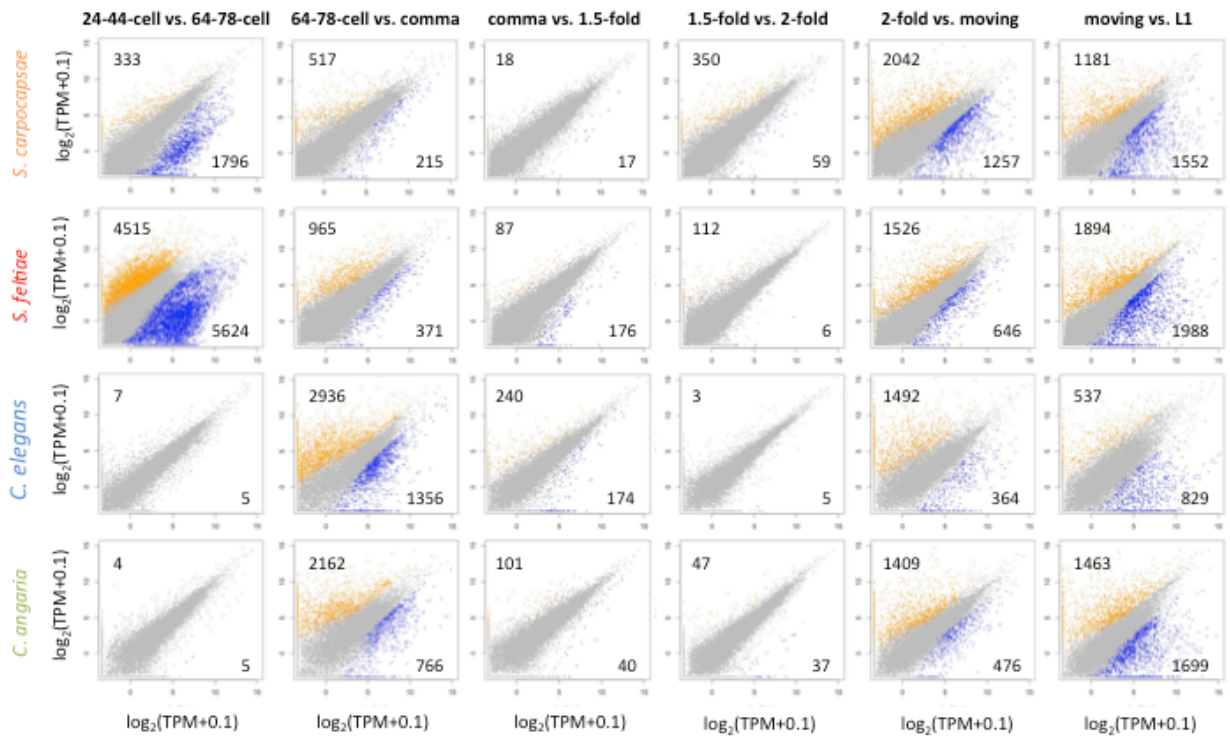


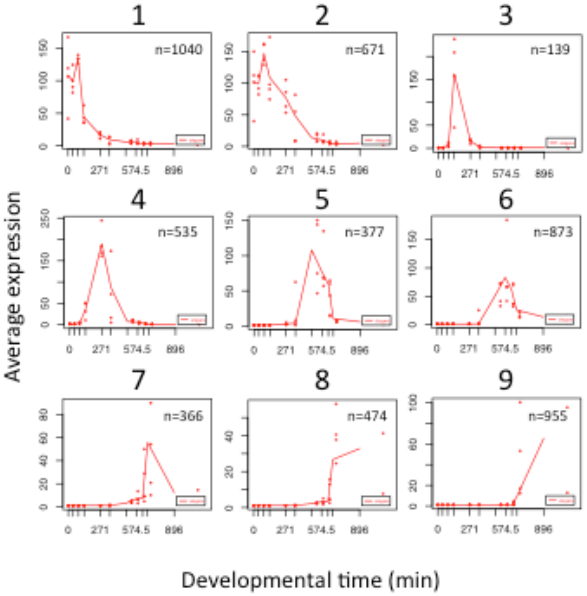
Figure 13. Differential gene expression during intermediate to late embryonic development in four species. Gene expression $\log_2$(TPM+0.1) of all genes was plotted for adjacent intermediate to late developmental stages in all four species. The earlier stages are displayed on the x-axis and the later stages are displayed on the y-axis. Genes that are differentially expressed (FDR < 0.05 and fold change > 2X) between the stages, and are more highly expressed in the earlier stage or later stage are shown in blue and yellow respectively. Genes in gray are not differentially expressed.

Figure 14. Gene expression dynamics during development using maSigPro. A) The time data for *C. elegans, C. angaria, S. carpocapsae* and *S. feltiae* used for calculating the expression profiles of genes that show statistically significant gene expression dynamics during the development in B) *C. elegans*, C) *C. angaria*, D) *S. carpocapsae*, and E) *S. feltiae*. The gene expression was plotted against the approximate time in minutes post fertilization. Genes were organized by expression into 9 clusters. The number of genes in each cluster is displayed under the time. The red lines display the average expression level of the genes in the cluster.

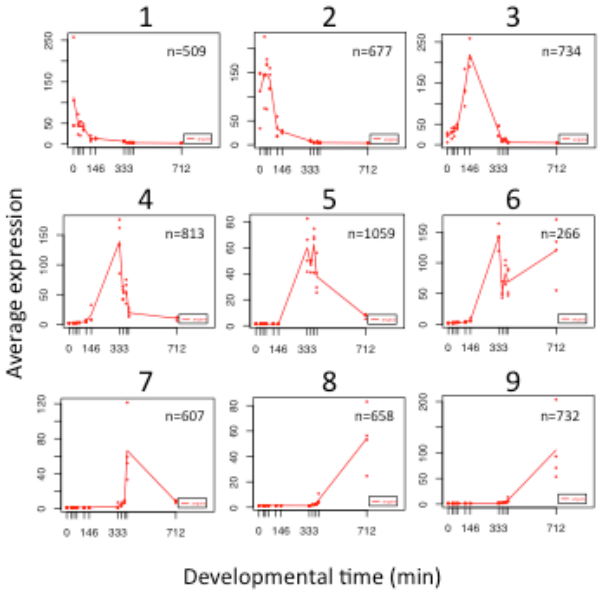
A.

stage											
min (elegans)	0	30	47	73	119	154	342	382	402	422	724
min (angaria)	0	30	60	85	129	165	352	376	396	414	732
min (carpocapsae)	0	30	82	131	329	409	593	646	700	755	1153
min (feltiae)	0	30	63	103	255	344	489	533	569	615	830
	estimated										

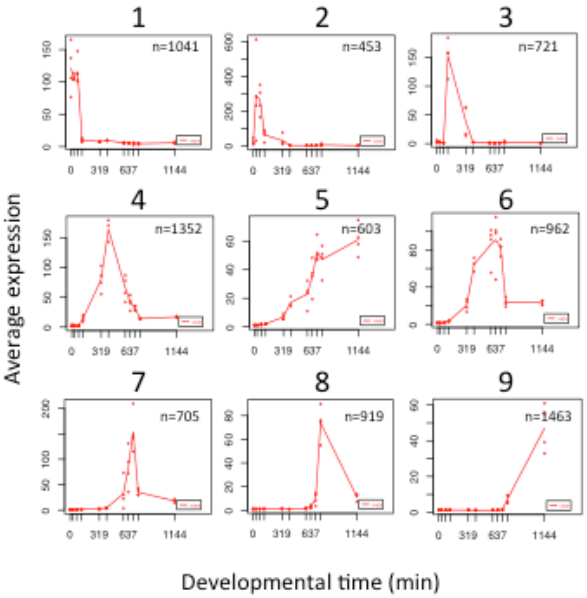
B. *C. elegans*



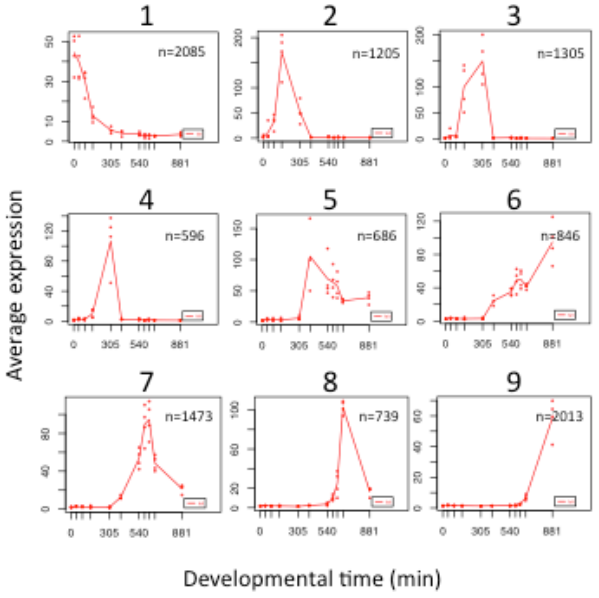
C. *C. angaria*



D. *S. carpocapsae*



E. *S. feltiae*



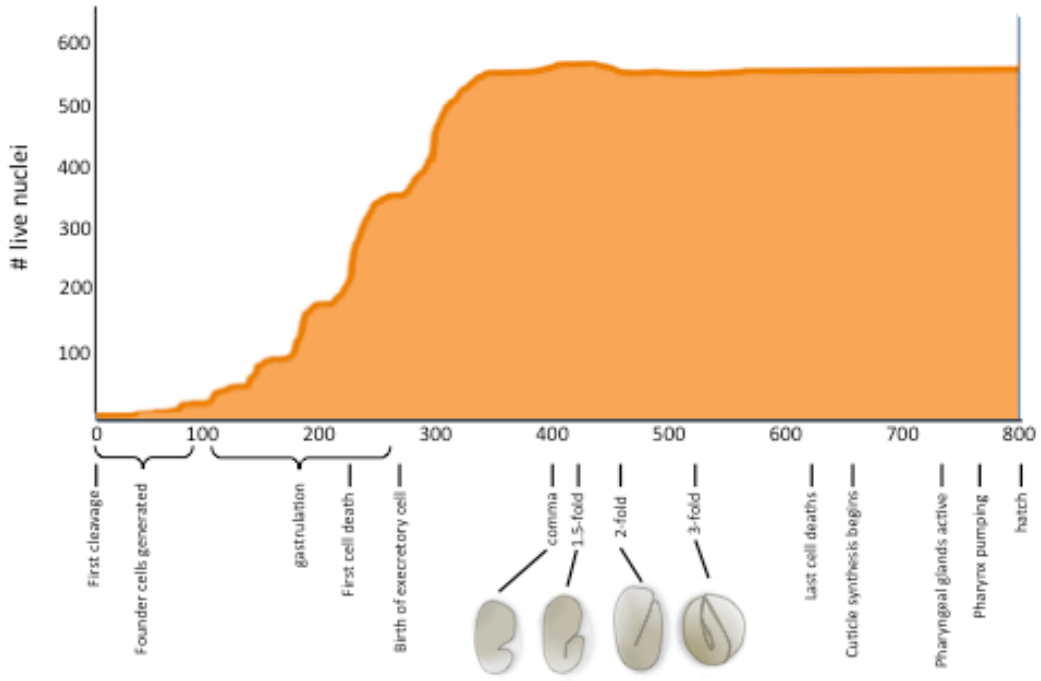


Figure 15. **Number of cell nuclei over the course of *C. elegans* embryonic development**. Figure was adapted from WormAtlas (Yusef Karabey, 2003).

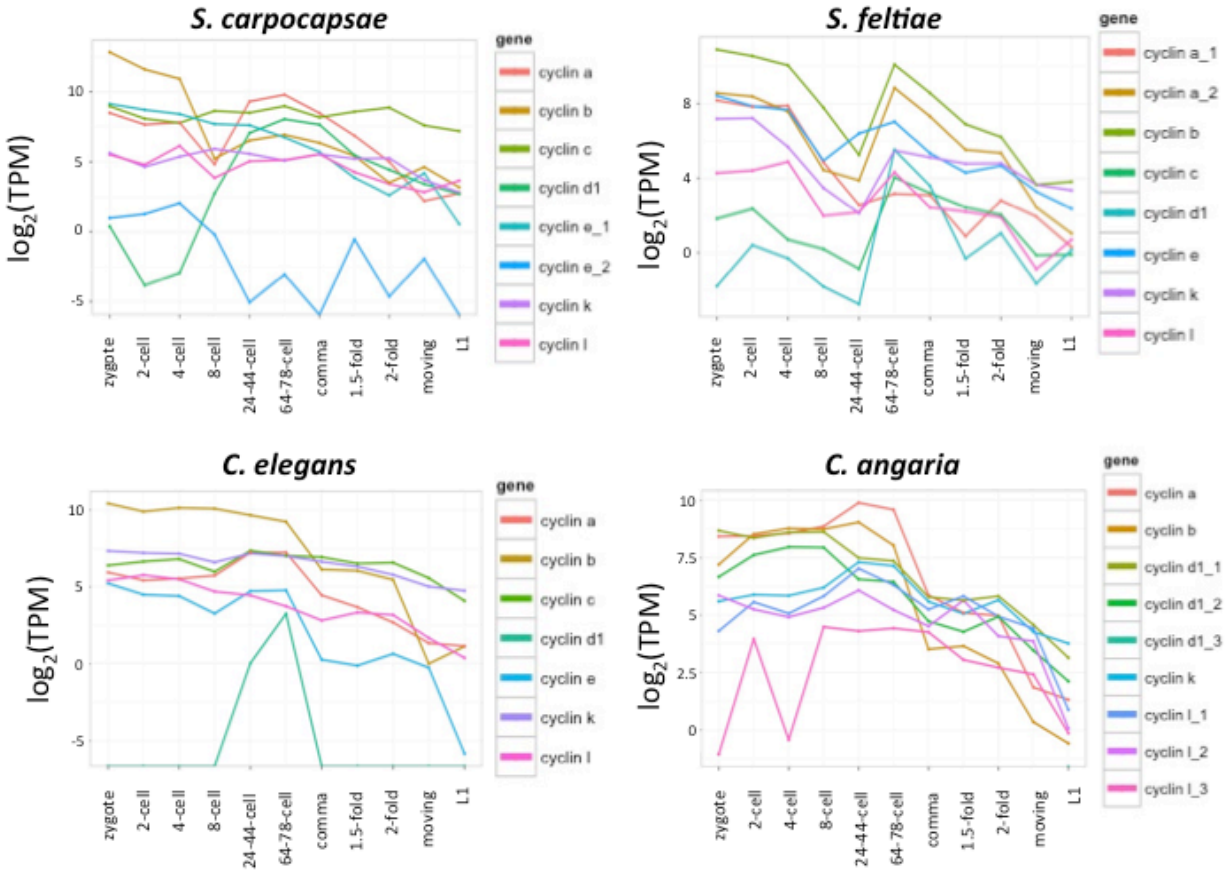


Figure 16. Cyclin family gene expression during embryonic development across species. *C. elegans* cyclin orthologs were found using WormBase ParaSite, and the expression of the genes were plotted over embryonic development for all four species. Expression (in TPM) is displayed on log2 scale.

A.

stage											
min (carpocapsae)	0	40	85	134	271	350	534	588	642	697	1095
min (elegans)	0	30	50	80	160	196	377	412	432	454	756

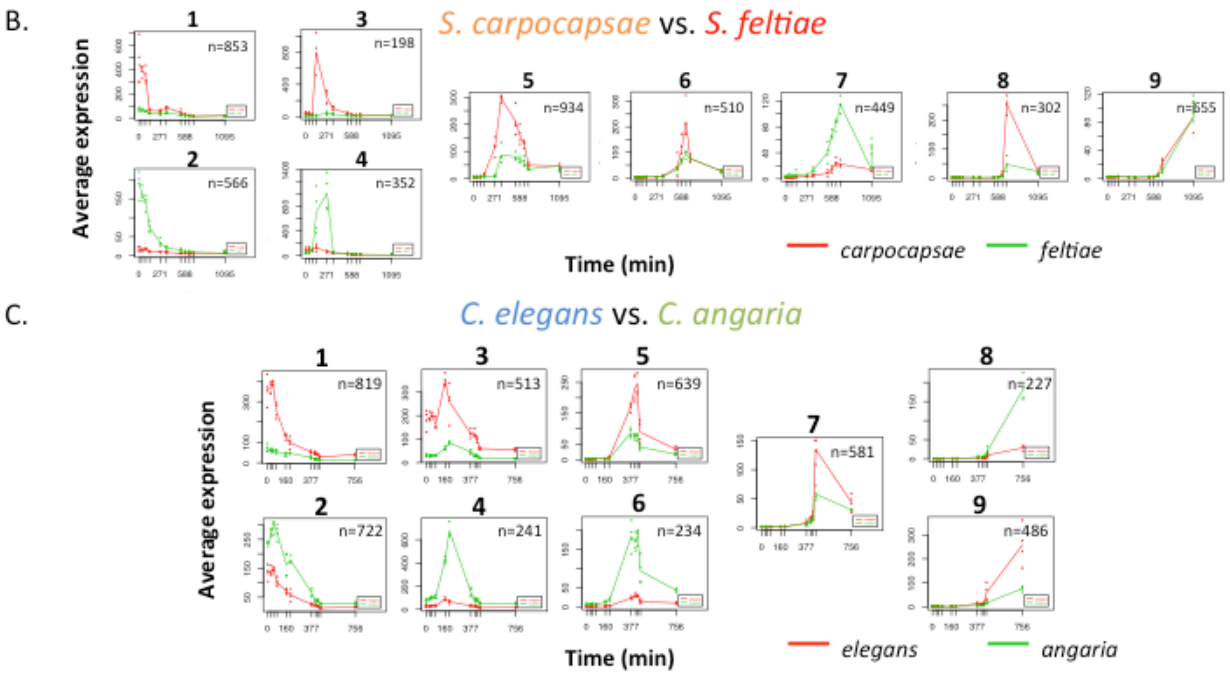


Figure 17. Gene expression dynamics in *Caenorhabditis* and *Steinernema* using maSigPro. A) The time data for *S. carpocapsae* and *C. elegans* used for calculating the expression profiles of genes that show statistically significant gene expression dynamics during the development between B) *S. carpocapsae* and *S. feltiae,* and C) *C. elegans* and *C. angaria.* The gene expression was plotted against the approximate time in minutes post fertilization. Genes were organized by expression into 9 clusters. The number of genes in each cluster is displayed within each box. The red and green lines display the average expression level of the genes in the cluster for each respective species.

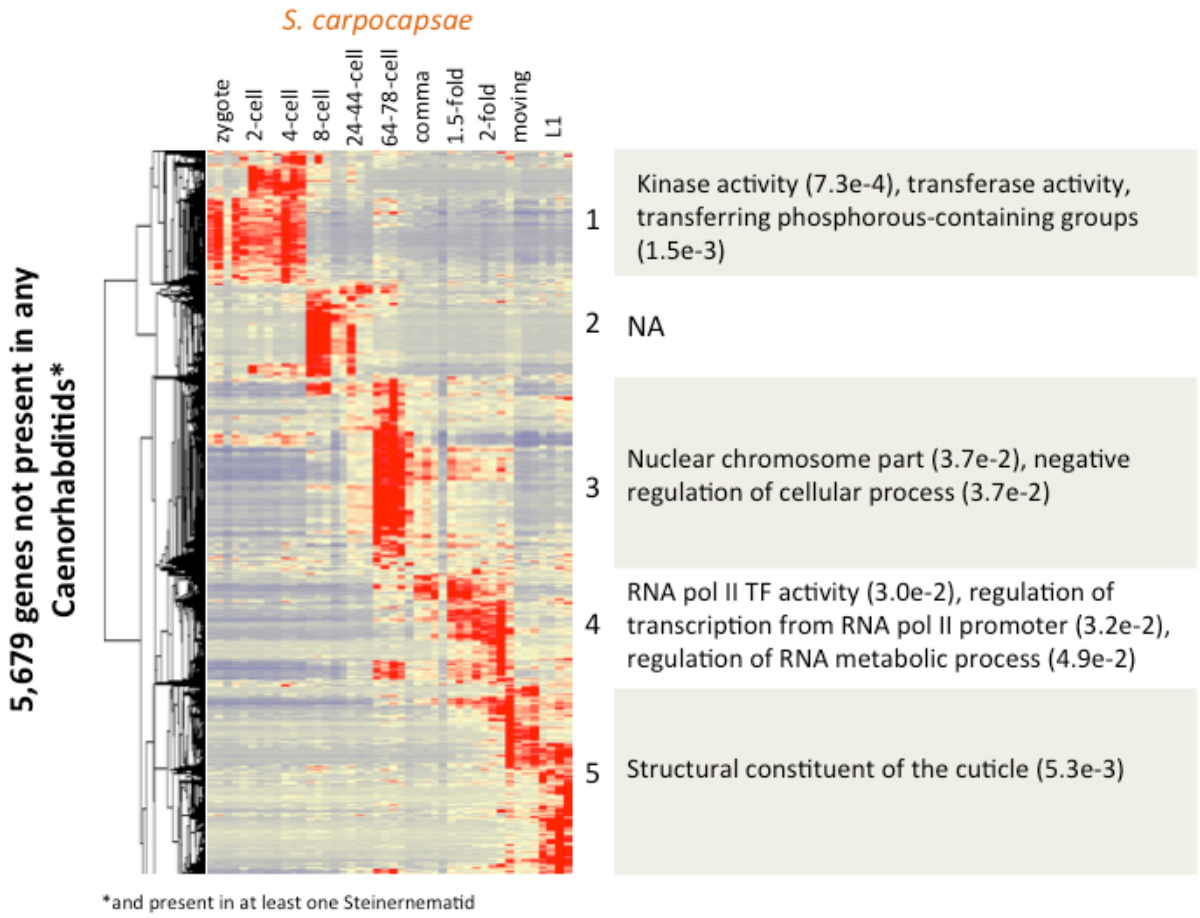
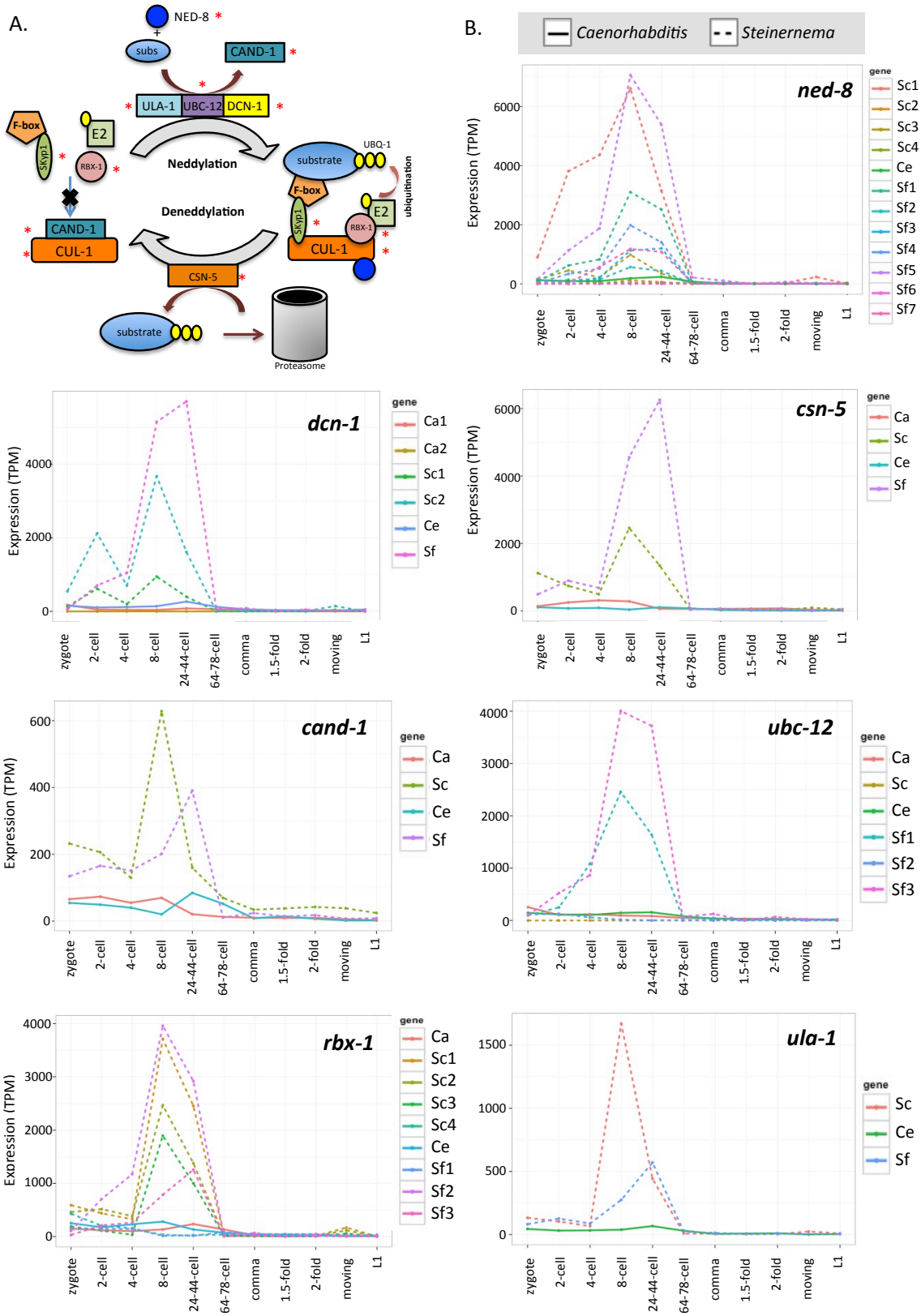


Figure 18. Expression of *Steinernema*-only genes during embryogenesis.
S. carpocapsae genes that have homologs in at least one other sequenced *Steinernema* species (*S. feltiae*, *S. scapterisci, S. monticolum*, and *S. glaseri*), but not any of six *Caenorhabditis* species (*C. elegans, C. angaria, C. remanei, C. brenneri, C. briggsae,* and *C. japonica*), and that are expressed at an average of 10 TPM and have at least one replicate expressed > 50 TPM during embryogenesis are plotted. Genes were mean-centered and hierarchically clustered. Five major clusters were discerned and Gene Ontology (GO) enrichment analyses were performed on the genes of each cluster (Fisher's exact test, FDR < 0.05). The FDRs for each GO term are reported in parentheses next to each GO term.

Figure 19. Neddylation pathway orthologs are upregulated during development in *Steinernema*. A) The neddylation and deneddylation pathways in *C. elegans*. CUL-1, a cullin protein, is in an inactive state when bound by CAND1. Neddylation of CUL-1 by a complex comprising ULA-1, UBC-12, and DCN-1 activates it so that it can bring ubiquitin ligases in proximity of their target substrates. Cullins are deactivated through deneddylation by CSN-5. Panel was adapted from Kandala, 2014. B) Ortholog expression profiles for members of the neddylation pathway. Solid profiles correspond to orthologs from *Caenorhabditis*, while dashed profiles correspond to orthologs from *Steinernema.* WormBase ParaSite paralogs are indicated with a number after the species name in the figure legends. Expression (TPM) is unscaled scale.


A.
NED-8
subs
CAND-1
ULA-1
UBC-12
DCN-1
F-box
SKyp1
E2
RBX-1
Neddylation
Deneddylation
CUL-1
CSN-5
substrate
UBQ-1
ubiquitination
Proteasome
B.
Caenorhabditis
Steinernema
ned-8
dcn-1
csn-5
cand-1
ubc-12
rbx-1
ula-1
Expression (TPM)
Exoression (TPM)
gene
Ca1
Ca2
Sc1
Sc2
Sc3
Sc4
Ca
Sc
Ce
Sf1
Sf2
Sf3
Sf4
Sf5
Sf6
Sf7
Sf
zygote
2-cell
4-cell
8-cell
24-44-cell
64-78-cell
comma
1.5-fold
2-fold
moving
L1

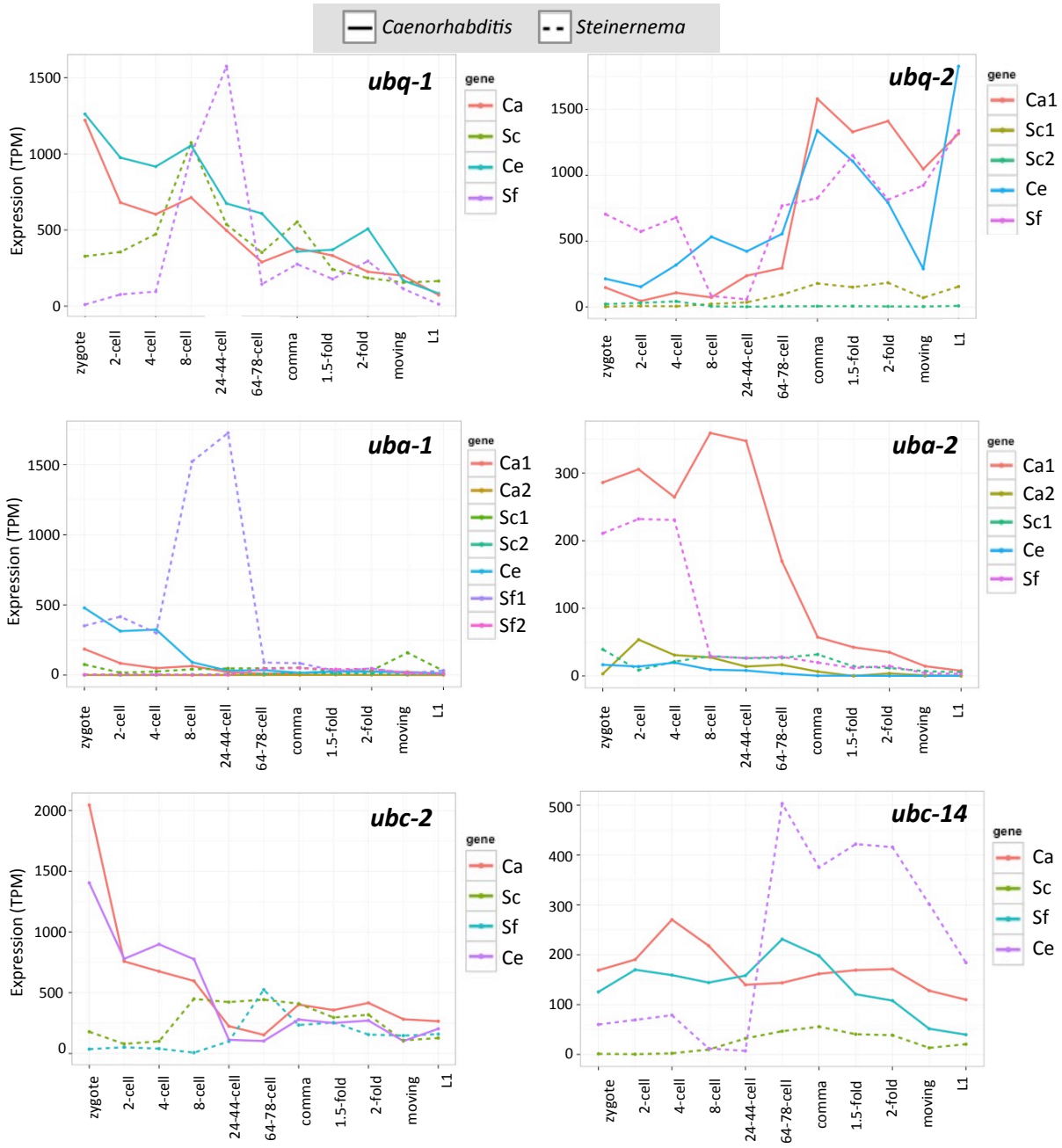


Figure 20. Expression of ubiquitin pathway members during embryogenesis.
A) Unscaled expression is plotted for two ubiquitin-coding genes, *ubq-1* and *ubq-2*, and other members of the ubiquitin pathway (*ubc-2*, *ubc-14*, *uba-1*). *ubq-1* codes for a 832aa polypeptide that gets cleaved into 11 ubiquitin proteins, while ubq-2 is a gene fusion of a ubiquitin and a 60S ribosomal subunit, and is much smaller at 128aa. *uba-1* and *uba-2* are ubiquitin-activating enzymes, and *ubc-2* and *ubc-14* are ubiquitin-conjugating enzymes and are essential for embryonic development.

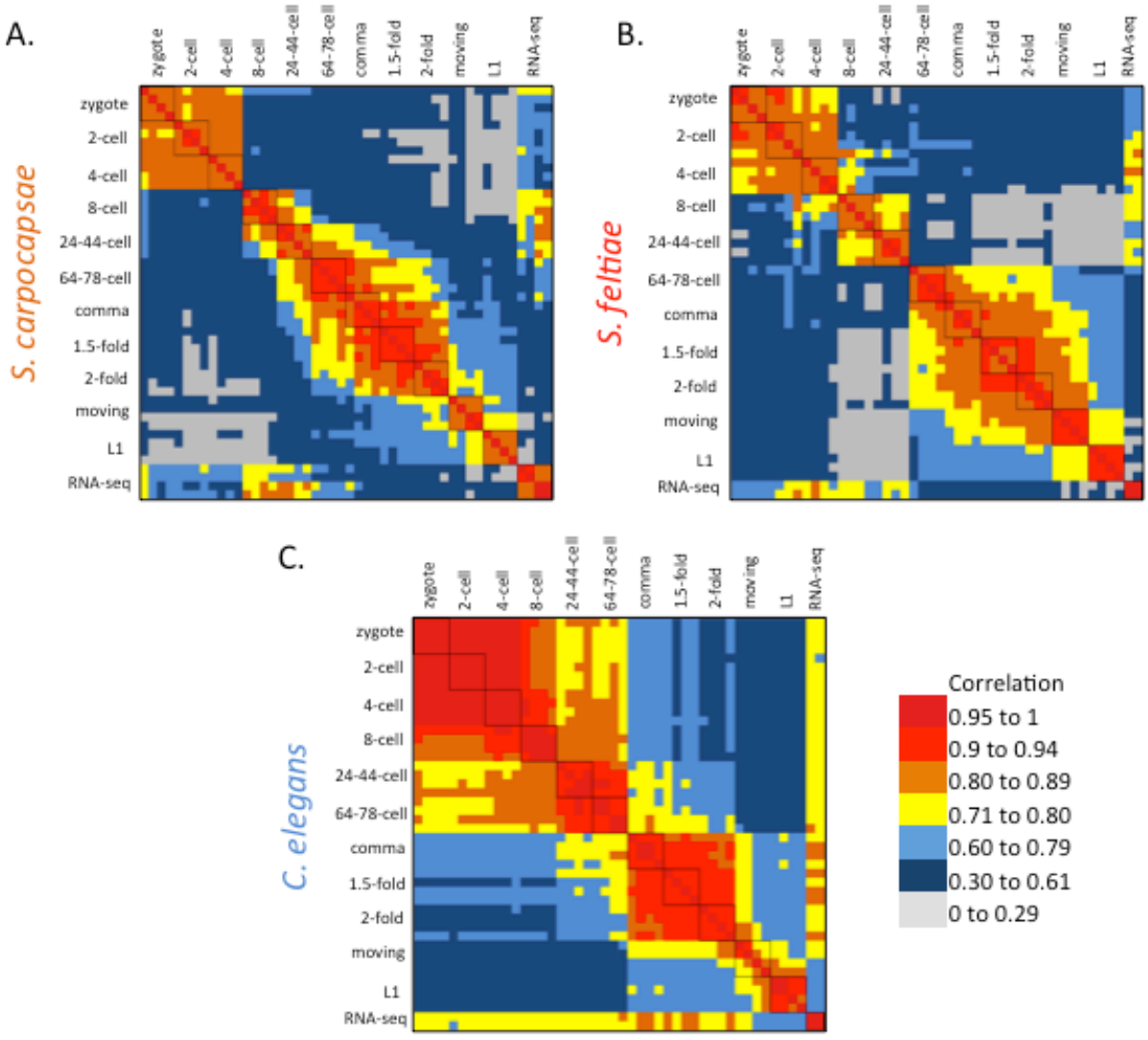


Figure 21. Correlations between published mixed-stage RNA-seq datasets and single embryo Smart-seq2 datasets. Pearson correlation matrices showing all correlations between 48 single embryos (11 developmental stages, 4 replicates per stage) and at least 2 RNA-seq mixed embryonic stage data sets for A) *S. carpocapsae*, B) *S. feltiae*, and C) *C. elegans*. The last two rows and columns of each matrix are the mixed-stage RNA-seq datasets. Black boxes denote replicates for each stage. Red and orange indicate high Pearson correlation coefficients, while grey and dark blue denote low correlation coefficients.

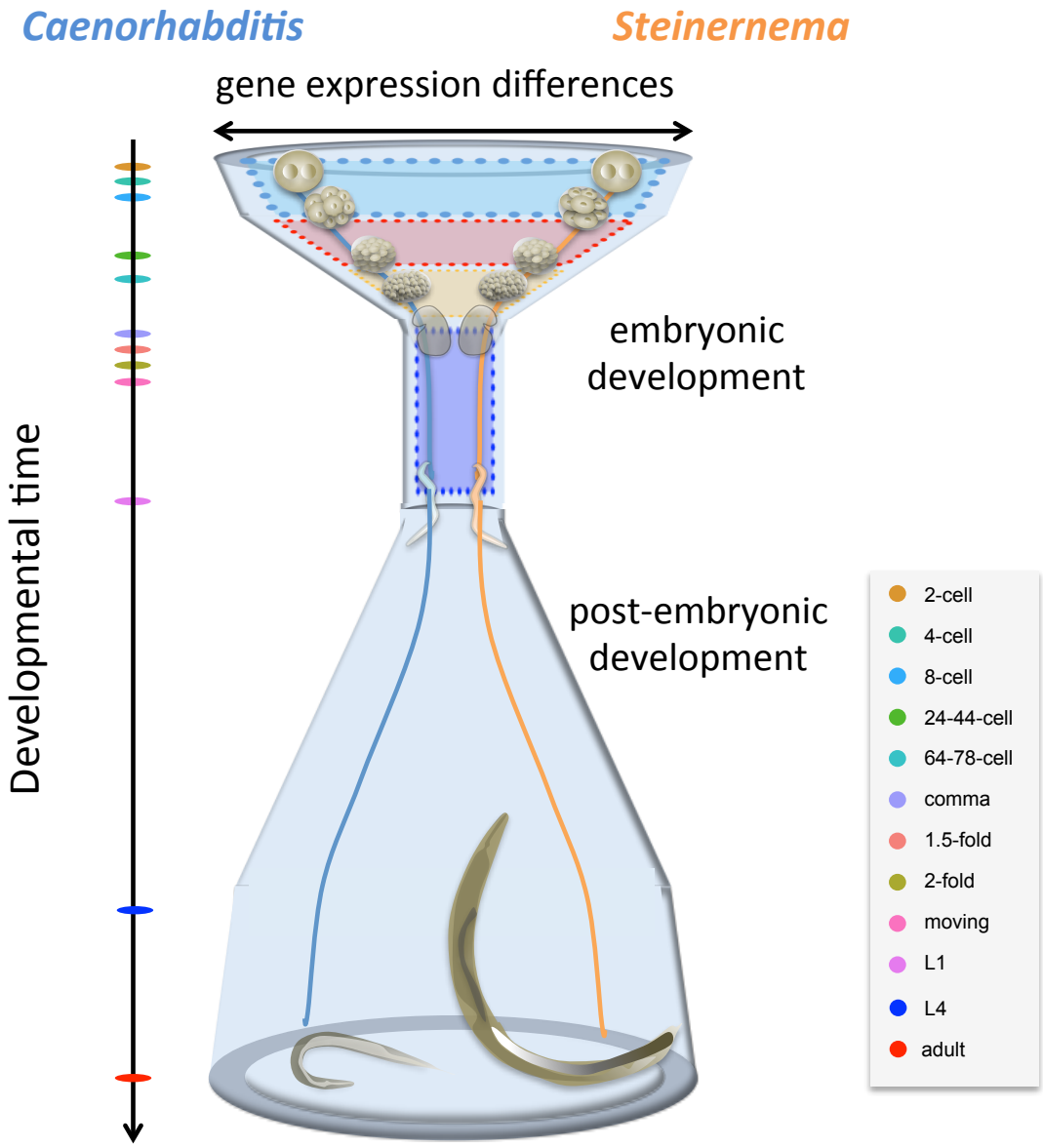


Figure 22. A model of gene expression divergence over embryonic (to scale) and post-embryonic development (not to scale) between distant genera. A funnel model of nematode embryonic development, where 1:1:1:1 ortholog expression variation is high in early stage embryos and low in later stage embryos during embryonic development, and high during post-embryonic development across genera. Embryonic and post-embryonic stages are in the gray figure legend.

Table 1. GO terms for *S. carpocapsae*-specific transcription factors that are not in *Caenorhabditis*

GO term	FDR
Positive regulation of transcription from RNA pol II	3.1 e-20
Negative regulation of RNA pol II promoter	1.1e-5
Positive mesodermal cell fate specification	1.1e-5
Response to retinoic acid	1.9e-5
Dorsal/ventral pattern formation	1.3e-4
N-terminal peptidyl lysine acetylation	4.7e-4
Negative regulation of cell proliferation	9.8e-4
Positive regulation of cell differentiation	1.1e-3
Neuron projection morphogenesis	3.7e-3
Somatic stem cell population maintenance	4.9e-3
Germ line stem cell population maintenance	1.1e-2
BMP signaling pathway	1.8e-2

Table 2. GO terms for *C. elegans* early embryonic maSigPro clusters.

Cluster 1

GO term	FDR
proteasome complex	1.2e-33
cell cycle	3.3e-19
chromosome segregation	9.6e-10
gamete generation	8.7e-10
response to DNA damage	2.5e-7

Cluster 2

GO term	FDR
hermaphrodite genitalia development	2.2e-33
cell cycle	3.1e-26
microtubule cytoskeletal organization	2.9e-15
mRNA processing	2.8e-15
gastrulation with mouth forming first	8.4e-13
nucleoplasm	8.2e-10
epithelium development	3.4e-8
nuclear envelope	2.3e-10
pronuclear migration	3.4e-2

References

1. Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol*, *100*(1), 64-119.
2. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T. et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, *330*(6012), 1775-1787.
3. Araya, C. L., et al. (2014). Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature*, 512, 400-405.
4. Schierenberg, E. (2006). Embryological variation during nematode development. *WormBook*, 1-13.
5. Voronov, D. A., Panchin, Y. V., & Spiridonov, S. E. (1998). Nematode phylogeny and embryology. *Nature*, *395*(6697), 28.
6. Voronov, D. A., & Panchin, Y. V. (1998). Cell lineage in marine nematode Enoplus brevis. *Development*, *125*(1), 143-150.
7. Levin, M., Hashimshony, T., Wagner, F., & Yanai, I. (2012). Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev Cell*, *22*(5), 1101-1108.
8. Dillman, A. R., Macchietto, M., Porter, C. F., Rogers, A., Williams, B., Antoshechkin, I. et al. (2015). Comparative genomics of Steinernema reveals deeply conserved gene regulatory networks. *Genome Biol*, *16*(1), 200.
9. Mortazavi, A., Schwarz, E. M., Williams, B., Schaeffer, L., Antoshechkin, I., Wold, B. J. et al. (2010). Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res*, *20*(12), 1740-1747.
10. Edgar, L. G., Wolf, N., & Wood, W. B. (1994). Early transcription in Caenorhabditis elegans

embryos. *Development*, *120*(2), 443-451.

11. Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, *130*(5), 889-900.
12. Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development*, *136*(18), 3033-3042.
13. Stitzel, M. L., Pellettieri, J., & Seydoux, G. (2006). The C. elegans DYRK Kinase MBK-2 Marks Oocyte Proteins for Degradation in Response to Meiotic Maturation. *Curr Biol*, *16*(1), 56-62.
14. Conesa, A., Nueda, M. J., Ferrer, A., & Talon, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, *22*(9), 1096-1102.
15. Nueda, M. J., Tarazona, S., & Conesa, A. (2014). Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, *30*(18), 2598-2602.
16. Karabey, Y. (2003). Introduction. In *WormAtlas*. http://www.wormatlas.org/ver1/postemblin_1977/intro.html
17. Bosu, D. R., Feng, H., Min, K., Kim, Y., Wallenfang, M. R., & Kipreos, E. T. (2010). C. elegans CAND-1 regulates cullin neddylation, cell proliferation and morphogenesis in specific tissues. *Dev Biol*, *346*(1), 113-126.
18. Enchev, R. I., Schulman, B. A., & Peter, M. (2015). Protein neddylation: beyond cullin-RING ligases. *Nat Rev Mol Cell Biol*, *16*(1), 30-44.
19. Petroski, M. D., & Deshaies, R. J. (2005). Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol*, *6*(1), 9-20.

20. Radford, H. E., Meijer, H. A., & de Moor, C. H. (2008). Translational control by cytoplasmic polyadenylation in Xenopus oocytes. *Biochim Biophys Acta*, *1779*(4), 217-229.

21. Simon, R., Tassan, J. P., & Richter, J. D. (1992). Translational control by poly(A) elongation during Xenopus development: differential repression and enhancement by a novel cytoplasmic polyadenylation element. *Genes Dev*, *6*(12B), 2580-2591.

22. Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*, 135-142.

23. Bateson W. (1894). *Materials for the Study of Variation*. New York, NY: Macmillan.

24. Jason Bryer (2013). timeline: Timelines for a Grammar of Graphics. R package version 0.9. http://CRAN.R-project.org/package=timeline

25. Picelli, S., Faridani, O. R., Bjorklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*, *9*(1), 171-181.

26. Gertz, J., Varley, K. E., Davis, N. S., Baas, B. J., Goryshin, I. Y., Vaidyanathan, R. et al. (2012). Transposase mediated construction of RNA-seq libraries. *Genome Res*, *22*(1), 134-141.

27. Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

28. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25.

29. Li, L., Stoeckert, C. J. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, *13*(9), 2178-2189.

30. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package

for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140.

31. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
32. de Hoon, M. J., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics*, *20*(9), 1453-1454.
33. Saldanha, A. J. (2004). Java Treeview--extensible visualization of microarray data. *Bioinformatics*, *20*(17), 3246-3248.
34. Kandala, S., Kim, I. M., & Su, H. (2014). Neddylation and deneddylation in cardiac biology. *Am J Cardiovasc Dis*, *4*(4), 140-158.

CHAPTER 4

Conclusions

Nearly all previous studies on *Steinernema* have focused on their pathogenicity to insects and their symbiotic relationships with their pathogenic bacteria. No *Steinernema* genomes had been sequenced before our studies, and little was known about which genes are expressed at what times during their development. In chapter 2, we sequenced and assembled the genomes of five *Steinernema* species for comparison purposes in order to uncover genomic features that set the *Steinernema* species apart from other nematodes. We found the expansion of many protein families, such as the fatty- and retinoic acid binding proteins and serine and aspartic acid proteases. We also found that many of these expanded genes were expressed in a stage-specific manner. As a part of this study, we also compared the stage-matched expression of four stages of development (embryonic, L1, IJ, and young adult) across two *Steinernema* species and *C. elegans*. The main question guiding our study is what is the degree of conservation in expression of shared orthologous genes during development between *Steinernema* and *C. elegans*, which have diverged from each other over 200 MYA? We found that about 80% of orthologous genes have conserved expression during development between the *Steinernema* species, but we found that only 61-63% were conserved in expression between *Steinernema* and *C. elegans*. The ortholog expression divergence was less pronounced when we removed the embryonic stage from the developmental comparison. This prompted us to delve further into gene expression during embryonic development, which led us into the study of embryonic development across two *Steinernema* and two *Caenorhabditis* species in chapter 3. We found many features that distinguish development between the two genera. For one, we detected an early upregulation of gene expression during *Steinernema* embryogenesis relative to *Caenorhabditis* embryogenesis. Gene expression increased as early as the 2-cell stage in *S. carpocapsae* and the 4-cell stage in *S. feltiae*, which are two stages and one stage respectively before zygotic transcription begins in *C.*

elegans. This suggests that zygotic transcription initiates at an earlier developmental stage in *Steinernema*. In *C. elegans*, maternally deposited OMA proteins prevent zygotic genome activation by sequestering key transcription factors for zygotic transcription. Coincidentally, we observed a downregulation of *oma-1/2* transcripts in *Steinernema* at the stages where gene upregulation occurred. However, since we have no information about maternal protein degradation (e.g. OMA-1, OMA-2), we still need to determine whether these detectable changes in gene expression are truly from the earlier initiation of zygotic transcription, or whether another mechanism such as differential polyadenylation of maternal mRNAs could be involved. In other organisms such as clams, worms, frogs and mice, it has been found that a subset of maternally deposited transcripts have very short poly(A) tails between 20 and 40 bp ("unpolyadenylated") and are held in translational repression complexes until they are required for translation at the correct times during development (Simon et al., 1992; Radford et al., 2008). When they are needed, the maternal mRNAs are elongated by the addition of 80-250 adenosine residues to make them translationally active. Given that the Smart-seq2 protocol used for the embryonic time course selectively amplifies polyadenylated mRNAs using a 30 bp oligodT primer, unpolyadenylated maternal mRNAs would not be detected (See Figure 4B in chapter 3) (Picelli et al., 2014). Thus, the addition of poly(A) tails to transcripts through this mechanism would be detected as an upregulation of transcription using Smart-seq2. There are several ways in which we can test if this is occurring in *Steinernema*. One method is to conduct single molecule fluorescence *in situ* hybridization (single molecule FISH) on transcripts that are differentially upregulated from zygote to 2-cell. In single molecule FISH, 30 short non-overlapping DNA oligos (20 bp in length) complementary to the transcript of interest are coupled to fluorescent probes. Populations of embryos can be treated with the fluorescent oligos to label

the transcripts. Because multiple oligos tile a transcript with multiple fluorophores, the fluorescent signal becomes strong enough to visualize individual transcripts. If we are able to detect equal levels of these differentially expressed transcripts from the zygote to the 2-cell stage in *S. carpocapsae* using the single molecule FISH method, then this would suggest that differential polyadenylation is responsible for the gene expression upregulation that we are seeing at these stages. If we do not see equal levels, then this could indicate active transcription of the zygotic genome.

Another method to test for zygotic genome activation is to inhibit zygotic transcription in zygote-stage embryos by inhibiting RNA pol II with α-amanitin or actinomycin D (Zeng and Schultz, 2005; Lee et al., 2014). By sequencing the RNA from α-amanitin-treated embryos at the 2-cell, 4-cell, and 8-cell stage, we can test whether transcripts are downregulated relative to the untreated embryos. If gene expression downregulation occurs in the α-amanitin-treated embryos, then those transcripts were the product of zygotic transcription. However, if it does not occur,, it could be indicative of differential transcript polyadenylation. Using either or both of these two proposed methods will allow us to interrogate when transcript expression commences in *Steinernema* relative to *Caenorhabditis*, and whether it is occurring by zygotic genome activation or through differential polyadenylation. These results will help us understand species-specific differences during early embryonic development, and the genes that are important to embryonic development in nematodes in general.

A second result we found from our embryonic analysis showed that ortholog expression was more conserved during late embryonic development (from 64-78-cell to L1 stage) than during early embryonic development (from the zygote to 24-44-cell stage) across the two *Steinernema* and *Caenorhabditis* species. Collectively, our findings on the molecular differences

between nematodes, together with previous studies by others showing the qualitative differences during embryonic development across nematodes, indicates that early embryonic development is rapidly evolvable period across nematodes. At the same time, the convergence of gene expression during later embryonic development suggests that the characteristic roundworm body plan is being set during this time. Thus, we propose a model for nematode embryonic development where gene expression divergence across species is highest during early embryogenesis and lowest during later embryogenesis. This result is different from the results of the Levin et al. study, which showed that *Caenorhabditis* embryonic development follows an hourglass model. The Levin et al. study strictly focused on transcriptome divergence (1-Pearson's correlation coefficient) of orthologous genes between development stages over time. It did not take into consideration the identities of the genes and time was an essential element to their model. In our study of gene expression conservation across species, we investigated specific expression of 1:1:1:1 orthologous genes. Because these are genes that are functionally conserved, they would presumably have similar expression profiles across species. The fact that very few of the orthologous genes expressed during early embryogenesis showed expression conservation across the species indicates that they are more flexible at the early stages of development. The flexibility of the orthologous gene expression decreases after the nematodes reach the 24-44-cell and 64-78-cell stage. All of the developmental stages after these stages appear to express the same orthologous genes at the same morphological stages regardless of the time it takes to reach those stages. Thus, we feel that a funnel-shaped model is the best model to fit our results of how ortholog expression varies during embryogenesis.

While conducting our embryonic analyses, we also found that some of the earliest maternal and well-studied genes/proteins in *C. elegans*, such as *med-1/2* are not found in *Steinernema*. The proteins of these genes are GATA transcription factors that are known to activate other GATA TFs, *end-3* and *end-1,* in the 8-cell and 24-44-cell stage embryo respectively to specify the gut cell lineage (E-cell). Interestingly, gut specification occurs as early at the 8-cell stage in *C. elegans*, and the gene regulatory network involved in specifiying gut cell lineage has already been determined. The lack of *med-1/2* genes in the *Steinernema* gene regulatory network suggests that there is a different maternally deposited regulatory factor or set of regulatory factors upstream of the *end* genes that are activating *end* gene expression at the 8-cell stage in *Steinernema*. It would be interesting to determine which genes are involved in E-cell (gut) specification in *Steinernema* and how gut specification differs from *C. elegans*. One type of analysis we can do to determine the identity of the upstream gut gene regulators through computational means is through binding site detection in the conserved non-coding regions upstream of the *Steinernema end-1/3* gene. Potential candidates could also be found through chromatin immunoprecipitation sequencing (ChIP-seq) data from *C. elegans*. The genes expressed within the E-cell can be further interrogated by mechanically separating 4-cell and 8-cell *Steinernema* embryos into individual cells for single-cell RNA-sequencing. Sequencing the RNA from individual cells of embryos can help us determine how each cell contributes to the expression in the whole embryo. I am particularly interested in seeing what the expression profiles look like for the endoderm cell (E-cell) that gives rise to the gut cell lineage in order to potentially identify those endoderm-specific regulators upstream of the *end-1/3* gene.

Our third major finding is that neddylation is likely to play a bigger role during *Steinernema* development than during *Caenorhabditis* development. In chapter 2, we found that

cullin genes, which act as scaffolds for ubiquitin ligases and target substrates and are activated through neddylation, are expanded in *Steinernema*. We found at least 19 cullins in *Steinernema*, whereas *C. elegans* has 6 cullin proteins (Petroski et al., 2005). In chapter 3, we found that many of these cullin and members of the neddylation pathway are highly upregulated at the 8-cell and 24-44-cell stage during *Steinernema* development. The expansion of cullins coupled with the specific expression upregulation suggests that *Steinernema* has an expanded repertoire of target protein substrates that need to be removed prior to gastrulation. This would suggest that the molecular events occuring prior to gastrulation in *Steinernema* are very different from *Caenorhabditis*. This could make *Steinernema* an interesting model for studying neddylation as well.

In order for *Steinernema* to become a viable satellite model organism, one of the next major steps is to get transfections working. In *C. elegans*, injections of transgenes into the transition region of the hermaphrodite gonad have been very successful because of the syncytial structure of this region. However, efforts to transfect using the injection method in *Steinernema* have not been successful thus far because their gonadal structures appear to be completely cellularized and contain no discernable syncytial region (Zograf et al., 2008). Thus, it is likely that transfections would have to be conducted using an alternative transfection method such as microparticle bombardment with a "gene gun". Transfecting *Steinernema* with a histone-2A GFP transgene would be incredibly useful for studying embryonic development. With this transgene each nuclei within the embryo could be visualized, making it much easier to track the cell count during development and to sort and collect individual cells for single cell RNA-sequencing. Furthermore, the promoter of any lineage-specific gene driving GFP could also be used to separate cells that give rise to the endoderm lineage for RNA-sequencing.

Within the next 5 to 10 years, I could see studies focusing on the genes responsible for nervous system development and jumping behavior of *Steinernema*. Previous neuroanatomical studies of select nematodes found that neuron numbers were fairly conserved across species. However, a recent study on the neuroanatomy of various nematodes found that there are considerable differences in the number of ventral cord (VC) neurons across species, with *Steinernema* IJs having 40% more VC neurons than *C. elegans* (77 *Steinernema* VC neurons versus 52 in *C. elegans*) (Han et al., 2015). Interestingly, this study also found no differences in the number of VC neurons between a jumping IJ (*S. carpocapsae*) and a non-jumping IJ (*S. glaseri*) (Han et al., 2015). However, it is still possible that these additional neurons play a role in the behavior of jumping *Steinernema* species. Laser ablation studies would help to hone in on the neurons responsible for this behavior. Dissecting out and sequencing the RNA of individual neurons and neuron progenitors could help us narrow down genes responsible for this behavior. In conclusion, *Steinernema* nematodes provide an excellent model for studying a variety of topics in development and behavior, beyond classical studies in parasitism and symbiosis. I foresee that these nematodes will be increasingly popular for evo-devo studies.

References

1. Simon, R., Tassan, J. P., & Richter, J. D. (1992). Translational control by poly(A) elongation during Xenopus development: differential repression and enhancement by a novel cytoplasmic polyadenylation element. *Genes Dev*, *6*(12B), 2580-2591.
2. Radford, H. E., Meijer, H. A., & de Moor, C. H. (2008). Translational control by cytoplasmic polyadenylation in Xenopus oocytes. *Biochim Biophys Acta*, *1779*(4), 217-229.
3. Picelli, S., Bjorklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res*, *24*(12), 2033-2040.
4. Lee, M. T., Bonneau, A. R., & Giraldez, A. J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. *Annu Rev Cell Dev Biol*, *30*, 581-613.
5. Zeng, F., & Schultz, R. M. (2005). RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Dev Biol*, *283*(1), 40-57.
6. Petroski, M. D., & Deshaies, R. J. (2005). Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol*, *6*(1), 9-20.
7. Zograf, J. K., Bert, W., Borgonie, G. (2008). The structure of the female reproductive system of nematodes from the genus *Steinernema* (Rhabditida: Steinernematidae). *Nematology*, 10(6), 883-896.
8. Bosu, D. R., Feng, H., Min, K., Kim, Y., Wallenfang, M. R., & Kipreos, E. T. (2010). C. elegans CAND-1 regulates cullin neddylation, cell proliferation and morphogenesis in specific tissues. *Dev Biol*, *346*(1), 113-126.
9. Han, Z., Boas, S., Schroeder, N. E. (2015). Unexpected Variation in Neuroanatomy

among Diverse Nematode Species. *Neuroanat.* 9, 162.