Title

Evolution of a family of developmental transcription factors in Caenorhabditis

Author

Darragh, Antonia C

Publication Date

2022

UNIVERSITY OF CALIFORNIA SAN DIEGO

Evolution of a family of developmental transcription factors in *Caenorhabditis*

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

in

Biology

by

Antonia C. Darragh

Committee in Charge:

Professor Scott Rifkin, Chair
Professor Kimberly Cooper
Professor Justin Meyer
Professor Karen Oegema
Professor Amy Pasquinelli

2022

The dissertation of Antonia C. Darragh is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This work is dedicated to my parents, Belinda and Bob.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

dGRN: developmental gene regulatory network

GATA factors: GATA-type transcription factors

BLAST: basic local alignment search tool

DBD: DNA-binding domain

BLAST: basic local alignment search tool

ELM2: EGL-27 and MTA1 homology 2

SANT: Swi3/Ada2/N-CoR/TFIIIB

ZnF: zinc finger

pHMM: profile hidden Markov model

RERE: arginine-glutamic acid dipeptide repeats

BAH: bromo-adjacent homology

NF: N-terminal zinc finger

CF: C-terminal zinc finger

BR: basic region

NMR: nuclear magnetic resonance

cGATA-1: chicken GATA-1

mGATA-3: mouse GATA-3

hGATA-3: human GATA-3

bps: base pairs

smFISH: single molecule fluorescence *in situ* hybridization

bZIP: basic leucine zipper

NJ: Neighbor Joining

CDS: coding sequence

UTR: untranslated region

DAPI: 4′,6-diamidino-2-phenylindole

SUMO: small ubiquitin-like modifier

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Scott Rifkin, for his perpetual support and enthusiasm for science. Moreover, I appreciate his patience while I learned new techniques including teaching myself how to code. I would also like to thank the members of my doctoral thesis committee, Drs. Kimberly Cooper, Justin Meyer, Karen Oegema, and Amy Pasquinelli for their insightful advice and questions. Thanks to all the past and present members of the Rifkin lab. I am grateful for your suggestions, encouragement, and friendship.

Thanks to the friends I made during my time here, especially Christopher Piggott, Jessica Bloom, Samuel Lin, Hannah Tsunemoto, Jeanae Kaneshiro, Javier How, and Cayce Dorrier. You made my time at UCSD significantly more enjoyable. I also want to thank my many intramural teammates for their comradery.

I want to thank my parents and my sister for their love, for joining me on rejuvenating holidays, and for their interest in my research. I would like to especially thank my mom, Belinda, for inspiring me to be a scientist and for reviewing my writing.

Chapter 1 is a reprint of a manuscript submitted for publication of its contents. It may appear in the journal *Molecular Biology and Evolution*. The manuscript is titled “A GATA factor radiation in *Caenorhabditis* rewired the endoderm developmental gene regulatory network” and is authored by Antonia C. Darragh and Scott A. Rifkin. Experiments and data analysis were led by the dissertation author and S.A.R. The manuscript preparation was led by the dissertation author with assistance from S.A.R. The dissertation author is the primary author on this manuscript.

VITA

Education

University of California San Diego
Ph.D. in Biology 2014 - 2022

University of California Santa Cruz
B.S. in Biochemistry and Molecular Biology 2010 - 2014

Research Experience

University of California San Diego 2015 - 2022
Graduate Student Researcher in the laboratory of Dr. Scott A. Rifkin
Evolution of *Caenorhabditis* GATA-domain-containing proteins

University of California San Diego 2014 - 2015
Graduate School Lab Rotations with:
Dr. Susan Golden – Validated a transposon insertion library and analyzed mutants for antibiotic and herbicide sensitivity in *Synechococcus elongatus*
Dr. Randy Hampton - Investigated differences in the growth of populations of *Saccharomyces cerevisiae* under temperature stress
Dr. Scott Rifkin – Studied variability in the hyperosmotic stress response in wild *Saccharomyces* strains
Dr. Amy Pasquinelli – Investigated the mechanism of argonaute-mediated microRNA autoregulation in *Caenorhabditis elegans*
Dr. Lena Gerwick – Screened the Gerwick marine natural product library for Nrf2-ARE activators and investigated the effectiveness of different natural compounds in the murine macrophage cell line RAW264 as drugs

University of California Santa Cruz 2013 - 2014
Undergraduate Researcher in the laboratory of Dr. Needhi Bhalla
Characterization of chromo-domain-containing proteins in the *C. elegans* meiotic synapsis checkpoint signal transduction pathway

University of California Davis 2013 - 2014
Summer Undergraduate Researcher in the laboratory of Dr. Judy Callis
Assisted graduate student with a project in *Arabidopsis thaliana* examining growth, development, and response to heat shock and conducted BLAST searches on NCBI

and The Arabidopsis Information Resource (TAIR) and updated the lab’s sequence database

Awards and Honors

Marguerite Vogt Award, UC San Diego Biology Department	2019
Graduate Student Association Travel Grant, UC San Diego	2019
NIH-CMG Training Grant, UC San Diego Biology Department (declined)	2015
Honorable Mention, NSF Graduate Research Fellowship Program	2015
Member, Phi Beta Kappa	2014 - present
Undergraduate Research Award in the Sciences, UC Santa Cruz	2013
Grant for Biomedical Research, UC Santa Cruz	2013
Golden Key International Honour Society Chapter Scholarship, UC Santa Cruz	2012
Scholar Athlete, UC Santa Cruz	2011-2014
Finalist, NCAA Post-Graduate Award**,** UC Santa Cruz	2014
Intercollegiate Athlete with Highest G.P.A., UC Santa Cruz	2012 and 2014

Service

Rifkin Lab Area Safety Coordinator	2015 - present
BioEASI Science Class instructor and organizer at county reentry facility	2018 - 2019
Association of Women in Science Outreach Committee Member	2015 - 2018
UCSD STEM Education & Diversity Discussion Group Board Member	2014 - 2018
UCSD Worm Club Seminar Organizer	2017 - 2018
UCSD Biology PhD Program Student Recruitment Leader	2015 - 2016
UCSD Biology PhD Program Peer Mentor	2015 - 2017
UCSD Instructional Assistant: Courses in Recombinant DNA Techniques (laboratory), Evolution (lecture), and Structural Biochemistry (lecture)	2015 - 2016

ABSTRACT OF THE DISSERTATION

Evolution of a family of developmental transcription factors in *Caenorhabditis*

by

Antonia C. Darragh

Doctor of Philosophy in Biology

University of California San Diego, 2022

Professor Scott Rifkin, Chair

Transcription factors regulate gene expression. The coordination of gene expression is fundamental for specifying, determining, and maintaining cell fates. Here I investigate the evolution of a family of developmental transcription factors in the genus of nematodes *Caenorhabditis*. In the model organism *Caenorhabditis elegans*, GATA factors primarily function in either endoderm or hypoderm (epidermal) development. The genomes of *Caenorhabditis* species have diverged extensively however, anatomically these species are barely distinguishable. Which raises the question, have GATA factors evolved while the phenotype of these worms has remained stable? To address this question, Scott Rifkin and I searched for GATA factor orthologs in the proteomes of 58 *Caenorhabditis* species and two outgroup nematode species using the GATA-type zinc finger domain profile from the PROSITE database. Using maximum likelihood approaches, we estimated the evolutionary history of these GATA-domain-containing sequences. We found that GATA factors have radiated extensively within the *Elegans* supergroup. For example, the core of the endodermal cell specification network – *med-1, med-2, end-3, end-1,* and *elt-7* – are unique to the *Elegans* supergroup, suggesting that this expansion rewired *Caenorhabditis* development. Our phylogenetic analyses, gene structure comparisons, and gene expression studies all support an evolutionary scenario in which a duplication of *elt-3* produced the ancestor of *elt-7*, *end-1*, and *end-3* in the *Elegans* supergroup ancestor. If this hypothesis is true, it would be an example of subfunctionalization after gene duplication that resulted in at least three *Elegans* supergroup paralogs that are expressed endoderm-specifically with another, *elt-3*, only expressed in the hypoderm. These gene duplications coincided with an increase in

conserved GATA binding sites in the promoter of another endoderm-specifically expressed GATA factor gene, *elt-2*, which acts downstream of and is a transcriptional target of *elt-7*, *end-1*, and *end-3* during endoderm development in *C. elegans*. One of these sites was found to be necessary for *elt-2* expression. Since ELT-2 also regulates itself, there may be a brief time when all four of these GATA factors are expressed and competing for binding to sites in the *elt-2* promoter. Using bio-layer interferometry I started quantifying the *in vitro* binding kinetics of these GATA factors to a functional GATA site and permutations of this site. My preliminary results show subtle differences in binding kinetics and specificity. These findings give insights into how three closely related transcription factors likely partitioned the ancestral role of a single transcription factor in a conserved developmental process.

Chapter 1:

A GATA factor radiation in *Caenorhabditis* rewired the endoderm developmental gene regulatory network

1.1 Abstract

Studying the evolution of families of transcription factors, especially in the context of developmental gene regulatory networks, is important for uncovering how developmental processes evolve. Members of the GATA family of transcription factors (GATA factors) regulate transcription through one or two GATA-type zinc fingers and one or two poorly characterized activation domains. In the model organism *Caenorhabditis elegans*, GATA factors primarily function in either endoderm or hypoderm (epidermal) development. The recent rise in the number of identified and sequenced *Caenorhabditis* species provides an excellent resource for conducting comparative evolutionary studies to the well-studied *C. elegans*. We searched the genomes of 58 *Caenorhabditis* species and 2 outgroup nematode species (from Wormbase and the Caenorhabditis Genomes Project) for protein-encoding sequences containing matches to the GATA-type zinc finger domain profile from PROSITE. Using maximum likelihood approaches, we estimated the evolutionary history of these GATA-motif-containing sequences. We found that genes in some subclades of our evolutionary tree exhibit a very high turnover rate involving both rapid duplication and frequent deletion. Moreover, GATA factors have radiated extensively within the *Elegans* supergroup. For example, the core of the endodermal cell specification network – *med-1, med-2, end-3, end-1,* and *elt-7* – are unique to the *Elegans* supergroup, suggesting that this expansion rewired *Caenorhabditis* development. Our phylogenetic analyses, gene structure comparisons, and gene expression studies using single-molecule fluorescent *in situ* hybridization in the non-*Elegans* supergroup species *C. angaria* all support an evolutionary scenario in which a duplication of *elt-3*

produced the ancestor of *elt-7*, *end-1*, and *end-3* in the *Elegans* supergroup ancestor. If this hypothesis is true, it would be an example of subfunctionalization after gene duplication that resulted in at least three *Elegans* supergroup paralogs that are expressed endoderm-specifically with another, *elt-3*, only expressed in the hypoderm. These gene duplications coincided with an increase in conserved GATA binding sites in the promoter of another endoderm-specifically expressed GATA factor gene, *elt-2*, which acts downstream of and is a transcriptional target of *elt-7*, *end-1*, and *end-3* during endoderm development in *Elegans* supergroup species. This rewiring of the endoderm network did not accompany any obvious phenotypic changes in endoderm development at the cellular and organismal levels and therefore is likely an illustration of developmental system drift. In sum, we conclude that multiple GATA factor gene duplications in the ancestor species of the *Elegans* supergroup created a GATA factor-centric endoderm gene regulatory network, while maintaining a conserved developmental output.

1.2 Introduction

Genomes are dynamic. Over time, processes like mutation, recombination, and gene conversion change genome sequences and can cause additions, losses, or relocations of genes within those genomes (Brown 2002). Over the last 50 years, many studies have demonstrated that gene duplication has served as a major mechanism through which new genes with novel functions evolve (e.g., Ohno 1970; Gottlieb 1977; Escriva et al. 2006; Assis & Bachtrog 2013; McKeown et al. 2014). For the first model of gene duplication, Ohno hypothesized that one of the duplicated gene paralogs would accumulate mutations while the other would maintain the original gene’s ancestral function since it would be constrained from accumulating mutations through negative selection (Ohno 1970). Because deleterious mutations are more frequent than beneficial ones (e.g., Charlesworth et al. 1994; Lynch & Walsh 1998; Eyre-Walker & Keightley 1999; Denver et al. 2004; Haag-Liautard et al. 2007), this scenario would likely result in the unconstrained paralog becoming non-functional (Nei & Roychoudhury 1973). However, if a beneficial mutation occurred and the new paralog gained a novel function, it could be preserved through positive selection. These two alternative outcomes for gene evolution are referred to as non-functionalization (or pseudogenization) and neofunctionalization, respectively (Ohno 1970).

Empirical evidence gathered since that first model of gene duplication was published (Ohno 1970) suggests that there are more paralogs per genome, but less pseudogenization, than Ohno’s theory predicted, (e.g., Allendorf et al. 1975; Ferris & Whitt 1979; Lundin 1993; Sidow 1996; Nadeau & Sankoff 1997; Postlethwait et al*.* 1998; Wendel 1999; reviewed by Zhang 2003) and so a third outcome of gene duplication was proposed: subfunctionalization (Hughes 1994; Force et al. 1999; Lynch & Force 2000). Subfunctionalization is when duplicated genes each retain some, but not all, of the ancestral gene’s function such that negative selection preserves both genes in the genome. This process occurs through the accumulation of different and complementary deleterious mutations in both genes. If deleterious mutations occur more frequently than beneficial ones, subfunctionalization would be expected to occur more frequently than neofunctionalization and, depending on the specific function of the ancestral gene, at a similar or higher probability than would non-functionalization (Lynch & Force 2000). Another possible outcome of gene duplication is that the resulting gene paralogs may increase the organism's robustness through biochemical redundancy or higher gene expression (Rapoport 1940; Ohno 1970; Nei et al. 2000; Piontkivska et al. 2002; Kondrashov & Kondrashov 2006). In this scenario, both genes, each with the ancestral function, are retained in the genome. Evidence for each of these evolutionary outcomes of gene duplication can be found in nature (e.g., Jozefowicz et al. 2003; He and Zhang 2005b; Gout & Lynch 2015), but it is often difficult to determine the exact evolutionary trajectory since information on extant paralogs is often compatible with several possible histories.

The variability in the numbers of gene family members in different organisms is a testament to the pervasiveness and stochasticity of gene duplications (e.g., Ohno 1970; Jozefowicz et al. 2003; Baker & Woollard 2019). For example, diploid vertebrates have six to eight GATA-type transcription factors (GATA factors), arthropods and lophotrochozoans have four or five, and nematodes harbor from at least one to 36. Known nematode species with more than 14 genes encoding GATA factors are members of the genus *Caenorhabditis* (Lowry & Atchley 2000; He et al. 2007; Gillis et al. 2008; Tang et al. 2014; Eurmsirilerd & Maduro 2020; Maduro 2020; this study). From a starting point of two GATA factors in the ancestor of Bilateria, two whole-genome duplications (Dehal & Boore 2005) and gene loss likely resulted in the six GATA factors currently found in mammals, and a third whole-genome duplication and gene loss likely led to the eight GATA factors found in teleost fish (Gillis et al. 2007; Gillis et al. 2008; Gillis et al. 2009). Evidence suggests that GATA factor evolution in arthropods and lophotrochozoans, which resulted in the four or five GATA factors found in contemporary species, occurred via tandem gene duplications (Gillis et al. 2008). Many tandem duplications of genes also likely occurred in nematodes since no evidence for whole-genome duplications has been found in any *Caenorhabditis* species to date (Semple & Wolfe 1999; Lynch & Conery 2000; Friedman & Hughes 2001; Cavalcanti et al. 2003; Stevens 2020). Changes in selective pressures likely preceeded these duplications. There is evidence for positive selection on some sites in vertebrate GATA factors (Tang et al. 2014), and varying levels of selection on GATA factor DNA-binding domains (DBDs) have been proposed based on how conserved different residues are in these regions (Lowry & Atchley 2000; Maduro 2020). However, to our knowledge GATA factor families have not been tested for relaxed selection, which is expected under the subfunctionalization and pseudogenization gene duplication theories (Force et al. 1999) and which is common for recent paralogs (Lynch & Conery 2000).

Most nematode GATA factors have only been studied in *Caenorhabditis elegans* (Eurmsirilerd & Maduro 2020). Surprisingly, more than half of the GATA factors in *C. elegans* function in the well-characterized gene regulatory network that specifies endoderm identity (reviewed by Maduro 2015;2017). In this network, the maternal transcription factor SKN-1 initiates a feed-forward cascade in which GATA factors that specify, differentiate, and maintain the cell fate of the endoderm are expressed (Bowerman et al. 1992; Blackwell et al. 1994; Maduro & Rothman 2002). SKN-1 is a bZIP/homeodomain-like transcription factor (Bowerman et al. 1992) that activates transcription of the functionally redundant GATA factors *med-1* and *med-2* in the endomesoderm (EMS) cell (Maduro et al. 2001). In parallel with the polarization of the EMS cell by Wnt/MAPK signaling induction of its posterior end from the P_2 cell (Goldstein 1992; Thorpe et al. 1997; Rocheleau et al. 1997, 1999; Meneghini et al. 1999; Shin et al. 1999), MED-1 and MED-2 activate expression of the largely functionally redundant GATA factors *end-3* and then *end-1* during the subsequent two cell divisions in the endoderm lineage (E and 2E stages) (Maduro & Rothman 2002; Baugh et al. 2003; Maduro et al. 2005a; Maduro et al. 2015). Maternally provided POP-1, an HMG box transcription factor, similar to vertebrate TCF and LEF transcription factors (Lin et al. 1995), is converted from a repressor to an activator in the first endoderm cell (1E) due to EMS polarization (Shetty et al. 2005; Phillips & Kimble 2009; Jackson & Eisenmann 2012). Wnt/MAPK induced POP-1 and SYS-1, a beta-catenin cofactor, together likely directly activate *end-1* (Shetty et al. 2005; Phillips et al. 2007). Moreover, *C. elegans* SKN-1 binding sites (Blackwell et al. 1994) are found in both most *Caenorhabditis end-3* and *end-1* promoters suggesting that SKN-1 may also directly activate them (Zhu et al. 1997; Maduro et al. 2005b; Maduro 2020). END-3 and END-1 then activate the expression of the GATA factors *elt-7* and *elt-2* in the 2E and 4E stages, respectively (Zhu et al. 1998; Maduro & Rothman 2002; Sommermann et al. 2010; Boeck et al. 2011). ELT-7 and ELT-2 are partially redundant in the differentiation and maintenance of the endoderm from the 2E to the final twenty intestinal cells that comprise the entire endoderm of these worms (Sulston et al. 1983; Fukushige et al. 1998; McGhee et al. 2007, 2009; Sommermann et al. 2010; Dineen et al. 2018). The expression of *elt-7* and *elt-2* is maintained for the lifetime of the worm through processes of auto- and cross-regulation (Fukushige et al. 1999; Sommermann et al. 2010). *elt-4*, a likely degenerate duplicate of *elt-2,* is also expressed later in the development of the endoderm, but does not have any known function (Fukushige et al. 2003).

The four GATA factors in *C. elegans* that are not involved in gut development all function during hypodermal (ectoderm) cell development (Page et al. 1997; Smith et al. 2005; Woollard 2005). ELT-1 and ELT-3 are both sufficient for differentiation of hypodermal cells (Gilleard & McGhee 2001). However, *elt-1* is expressed earlier, and ELT-1 specifies hypodermal precursors (Page et al. 1997). Without expression of *elt-1,* hypodermal precursors are not specified, a condition lethal to the animal, and so *elt-1* expression is necessary for hypodermal development (Page et al. 1997), whereas no phenotype is observed in worms lacking *elt-3* (Gilleard & McGhee 2001). ELT-3 can be sufficient to specify hypodermal cells in the absence of *elt-1* if *elt-3* is expressed earlier like *elt-1* is in wildtype worms (Gilleard & McGhee 2001). ELT-1 likely directly activates *elt-3* expression in ventral and dorsal hypodermis cells (Gilleard et al. 1999; Gilleard and McGhee 2001; Smith et al. 2005; McGhee 2013). The functions of the other *C. elegans* GATA factors that are not involved in gut development, EGL-18 (also called ELT-5) and ELT-6, are interchangeable in seam (lateral hypodermal) cell differentiation and maintenance, where they likely directly repress *elt-3* expression (Koh and Rothman 2001; Gorrepati et al. 2013). The functions of EGL-18 and ELT-6 also overlap in vulval (Koh et al. 2002) and P cell (ventral hypodermal cell) development (Liu et al. 2014).

GATA factors are potent endodermal regulators throughout bilaterians (Patient & McGhee 2020; Gillis et al. 2007) even when expressed heterologously. For example, if *C. elegans* END-1 is expressed in explant *Xenopus laevis* animal caps (an ectodermal lineage) it activates endoderm development, demonstrating both conservation of endoderm specification capabilities between ecdysozoa and vertebrates and a surprising generalized ability of this GATA factor to function despite a markedly different intracellular context (Shoichet et al. 2000). In *C. elegans*, using the *end-1* promoter to highly express ELT-2 or ELT-7 can compensate for the loss of all four of *end-3*, *end-1*, *elt-7*, and *elt-4* (Wiesenfahrt et al. 2016), suggesting that developmental timing is the primary difference in function among these endoderm-specific GATA factors. However, expression of neither *C. elegans elt-3*, which encodes a hypoderm GATA factor, nor *Mus musculus* GATA-4 expressed using the same *end-1* promoter, were able to rescue loss of both *end-1* and *end-3* in *C. elegans* (Wiesenfahrt et al. 2016), suggesting that GATA factors are not all interchangeable. Identifying attributes responsible for functional redundancy among only some GATA factors has been difficult because these proteins have diverged extensively outside of their DNA-binding domains (Lowry & Atchley 2000; Gillis et al. 2008; Maduro 2020; Eurmsirilerd & Maduro 2020; this study) and (other than the MED orthologs) (Broitman-Maduro et al. 2005; Lowry et al. 2009; Maduro 2020; this study) they are all thought to bind to canonical HGATAR DNA sites (Gerstein et al. 2010; Araya et al. 2014; Narasimhan et al. 2015; Wiesenfahrt et al. 2016; Du et al. 2016; Maduro 2020; this study).

Two studies of the evolution of GATA factors in nematodes have been published recently. A comparison of GATA factor orthologs in 32 species of nematodes suggested that the genome of the ancestor of this phylum contained at least an *elt-1* ortholog and perhaps also an *elt-2* ortholog, and therefore that multiple gene duplications must have occurred since for evolution to have resulted in the 11 GATA factors currently encoded in the *C. elegans* genome (Eurmsirilerd & Maduro 2020). Eurmsirilerd and Maduro also found that the most gene duplications and the most sequence divergence had occurred in the *elt-3* family of orthologs, suggesting that this gene may have evolved faster than the other GATA factors in this phylum. Maduro (2020) found that orthologs of five of the six GATA factors that regulate endoderm development in *C. elegans* - *med-1*, *med-2*, *end-1*, *end-3*, and *elt-7* - are specific to the *Elegans* supergroup, suggesting that these genes arose in the ancestor of this clade. Maduro proposed a model for the origin of the regulatory network specifying endoderm in *C. elegans* based on analysis of a subset of *Caenorhabditis* GATA factors in the genomes of 20 species in the *Elegans* supergroup and four non-*Elegans* supergroup species.

However, dozens of new *Caenorhabditis* species have been discovered sequenced over the last decade (Kiontke et al. 2011; Dey et al. 2012; Félix et al. 2014; Slos et al. 2017; Ferrari et al. 2017; Stevens et al. 2019; Teterina et al. 2020; Stevens 2020). Draft sequences of the genomes of an additional 34 *Caenorhabditis* species (Stevens 2020) are now available for carrying out even more comprehensive comparisons of GATA factors throughout the *Caenorhabditis* genus which should help elucidate the evolutionary history of this family of transcription factors.

Herein we report on our investigations into the evolutionary history of GATA-domain-containing proteins in 58 *Caenorhabditis* species and two outgroup nematode species. We start with a tour of the diversity and evolutionary history of GATA factors in the *Caenorhabditis* genus, highlighting the ongoing radiation in the *Elegans* supergroup, and then we zero in on the effects of this radiation on the developmental genetics of endoderm specification in these developmentally canalized worms. This study illustrates how closely related transcription factors have evolved and how gene duplications fueled the evolution and elaboration of an essential gene regulatory network, all without causing any obvious change in development or morphology.

1.3 Results

***Evolution of GATA-domain-containing proteins in* Caenorhabditis**

Caenorhabditis *GATA-domain-containing proteins form twelve ortholog groups some of which cluster adjacently into larger clades*

To identify potential GATA factors in *Caenorhabditis*, we searched for the characteristic GATA factor DNA-binding domain (DBD) motif defined by the PROSITE profile PS50114 (prosite.expasy.org; see Methods) in all fifty-six *Caenorhabditis* species for which genomic sequence assemblies were available, in *C.* sp. *45* and *C.* sp. *47* for which only transcriptome data were available, and in the genome assemblies of two outgroup *Diploscapter* species (*Caenorhabditis* Genome Project). We identified 890 protein-coding hits and made a preliminary estimation of their evolutionary relationships (data not shown). Because this PROSITE profile method unexpectedly left some species without GATA-domain-containing orthologs of *C. elegans* genes, we also used reciprocal protein basic local alignment search tool (BLASTp) and translated nucleotide BLAST (tBLASTn), to search for missing orthologs by employing genes in their sister species as bait (Altschul et al. 1990;

Camacho et al. 2009; see Methods). This reciprocal BLAST analysis and edits of errors that we identified in multiple gene annotations (see Methods), identified 51 additional proteins which brought our total up to 941 genes (Supp. Table 1). We estimated the evolutionary history of the 884 well-alignable (see Methods) proteins using maximum likelihood approaches (Minh et al. 2020; Kalyaanamoorthy et al. 2017) and a GATA factor from the slime mold *Dictyostelium fasciculatum* to root our phylogenetic analysis (Supp. Fig. 1.1). Some GATA-domain-containing proteins are highly divergent, with long branches that do not group robustly into any clade in our phylogeny (Supp. Fig. 1.1), after excluding these and other probable pseudogenes (Supp. Fig 2; see Methods) left us with 714 GATA-domain-containing protein sequences in which we had high confidence (Supp. Table 1). We used maximum likelihood approaches (Minh et al. 2020; Kalyaanamoorthy et al. 2017) and a GATA factor from *D. fasciculatum* to estimate the evolutionary relationships among these 714 protein sequences (Fig. 1.1).

Our resulting phylogenetic tree contains 12 distinct groups of orthologs, each group named after the *C. elegans* protein(s) within that group (Fig. 1.1A). The clade designations in our tree are similar to those established for *Caenorhabditis* GATA factors previously (Gillis et al. 2008; Maduro 2020; Eurmsirilerd & Maduro 2020). However, because we have included many more *Caenorhabditis* species, all known *Caenorhabditis* GATA factors, and additional *Caenorhabditis* proteins that contain GATA-like domains (e.g., EGL-27, RCOR-1, and SPR-1 orthologs), our tree represents a more complete picture of how *Caenorhabditis* GATA factors evolved. The clades in our phylogenetic tree are robust against ultrafast bootstrapping (Minh et al.

2013; Hoang et al. 2018; see Methods) and have well-supported nodes (Fig. 1.1A). A striking feature of our phylogeny is that GATA-domain-containing proteins have radiated extensively within the *Elegans* supergroup, as exemplified by the END-3, END-1, ELT-7, MED*,* and SPR-1 orthologs, which are unique to the *Elegans* supergroup and by EGL-18 which is also found in the *Guadeloupensis* group (Fig. 1.1A)**.** This pattern for END-3, END-1, ELT-7, and MED orthologs was first identified by Maduro (2020) and then supported by the lack of orthologs for these proteins in other nematodes (Eurmsirilerd and Maduro 2020). We extend this finding to 34 more *Caenorhabditis* species.

Ten of the 12 ortholog groups clustered into four larger clades of adjacent ortholog groups (Fig. 1.1A). We refer to these larger clades by the name of the most ancient ortholog group within the clade or by the names of all the ortholog groups included in the clade when the most ancient group is not apparent. There are four of these larger clades. These clades include the rcor1 clade which contains the RCOR-1 and SPR-1 ortholog groups, the elt6 clade which contains the EGL-18 and ELT-6 ortholog groups, the elt1/2 clade which contains the ELT-1 and ELT-2 ortholog groups, and the elt3 clade which is composed of the ELT-3, ELT-7, END-1, and END-3 ortholog groups (Fig. 1.1A). We present below detailed results for each of these clades and for the MED and EGL-27 ortholog groups.

Protein groups with non-canonical GATA domains and non-GATA domains

rcor1 clade

SPR-1 and RCOR-1 orthologs group adjacently in the rcor1 clade

The SPR-1 and RCOR-1 ortholog groups form a well-supported (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) monophyletic clade in our phylogenetic tree (Fig. 1.1A), which supports their shared homology to the same proteins in other nematodes, CoRest-PI in *Drosophila melanogaster*, and components of the REST corepressor complex in vertebrates (Wormbase.org). In *C. elegans*, SPR-1 and RCOR-1 play non-redundant roles in differentiation of pi cells during gonad and vulval development (Jarriault & Greenwald 2002; Hale et al. 2014; Bender et al. 2006; Vandamme et al. 2015). We only found SPR-1 orthologs in *Elegans* supergroup species, whereas RCOR-1 orthologs were identified in species throughout the *Caenorhabditis* genus as well as in the two *Diploscapter* species included in our analysis (Fig. 1.1). These results (and others described below) support the hypothesis that a *rcor-1* duplication produced *spr-1* in the ancestor of the *Elegans* supergroup.

spr-1 *and* rcor-1 *orthologs have similar gene structures including non-GATA domains*

We compared the structures of extant *spr-1* and *rcor-1* genes (Supp. Fig. 1.3A,B) and used parsimony to predict the *Elegans* supergroup *rcor-1* and *spr-1* ancestral structures (Fig. 1.2A). These predicted structures are very similar, the main difference being that the *spr-1* exon 9 is about three times longer than that of *rcor-1* (Fig. 1.2A). In fact, this exon, or an exon around this position, is more variable in length in extant *spr-1* genes, leading to more variability in the location of the second Myb/SANT (Swi3/Ada2/N-CoR/TFIIIB; Aasland et al. 1996) domain compared to *rcor-1* genes (Supp. Fig. 1.3A,B). Two distinguishing features of most *spr-1* and *rcor-1* genes is that they have a conserved splice site immediately following the first 36 nucleotides that code for their atypical GATA ZnFs (see below) and a conserved splice site after the first 84 nucleotides coding for their first Myb/SANT domain (Fig. 1.2A; Supp. Fig. 1.3A,B). These conserved splice sites are not found in any of the other GATA-domain-containing proteins included in this study (Supp. Fig. 1.3A,B), thus providing further support for the common ancestry between *spr-1* and *rcor-1* genes and for the distant placement of the rcor1 clade away from canonical GATA factor groups in our phylogeny (Fig. 1.1A). Due to extensive homology along the entire lengths of the *rcor-1* and *spr-1* sequences (Supp. Fig. 1.4A), we predict that a full-length duplication of *rcor-1* occurred in the *Elegans* supergroup ancestor, thus generating paralogs with the same gene structure as the predicted *rcor-1 Elegans* supergroup ancestor (Fig. 1.2A). This gene structure consists of 12 exons and has an ELM2 (EGL-27 and MTA1 homology 2; Solari et al. 1999) domain encoded in exon 4, a Myb/SANT domain encoded in exons 5 and 6 with conservation of the intron location between them, a GATA-like zinc finger motif (ZnF) encoded in exons 6 and 7 with conserved location of the intron between them, and a second Myb/SANT domain encoded in exons 10 and 11 separated by an intron (the location of which is not as highly conserved as the location of the introns associated with the first Myb/SANT domain or the ZnF) (Fig. 1.2A). ELM2 and Myb/SANT domains do not share homology with GATA factor domains and are known to be involved in transcriptional repression through chromatin regulation (Boyer et al. 2004; Ding et al. 2003; Wang et al. 2006; Millard et al. 2013). The presence of one ELM2 domain and two Myb/SANT domains is the signature of genes encoding CoREST proteins (Meier & Brehm 2017). But unlike sequences encoding canonical CoREST proteins, nearly all the *Caenorhabditis* SPR-1 and

RCOR-1 orthologs we examined contain a ZnF-like motif ($CX_2CX_{16\text{-}23}CX_2C$; Supp. Fig. 1.5) between the two Myb/SANT domains, which, in most cases, has the same length of a typical animal GATA ZnF (i.e., a $CX_2CX_{17}CX_2C$ motif; Teakle & Gilmartin 1998; Lowry & Atchley 2000). (These ZnF-like motifs are described in more detail below.) Interestingly, most of the ZnF-like motifs in the SPR-1/RCOR-1 homologs of other nematodes and *Drosophila* have shorter loops of only 14 or 15 residues, suggesting that loop length has changed multiple times as these gene families have evolved. Since the four vertebrate SPR-1/RCOR-1 homologs lack any $CX_{2,4}CX_{14\text{-}24}CX_2C$ ZnF-like motifs (Wormbase.org), the ZnF-like motif likely arose in invertebrates. In summary, genes coding for the SPR-1 and RCOR-1 proteins in our phylogenetic tree are similar in their organization, including intron splice locations, positions of non-GATA-related domains and of an atypical ZnF (Fig. 1.2A; Supp. Fig. 1.3A,B), which supports our hypothesis that these ortholog groups share evolutionary history.

spr-1 *and* rcor-1 *orthologs contain a not well-conserved GATA-like domain that likely evolved convergently relative to GATA factors*

We created a hidden Markov model profile (pHMM) (Eddy 2020) of each of the GATA-like domains in RCOR-1 and SPR-1 proteins (see Methods; Fig. 1.3A) and used them to search all the proteins included in our analysis. Other than for SPR-1 orthologs, the SPR-1 pHMM scores were highest for RCOR-1 orthologs and vice versa (Fig. 1.3B), suggesting a high degree of homology among the GATA domains of these proteins. Moreover, the SPR-1 and RCOR-1 pHMMs had low scores for the other proteins included (Fig. 1.3B), suggesting lack of conservation with other

Caenorhabditis GATA domains. This is consistent with the placement of most SPR-1 and all RCOR-1 GATA-like domains closer to each other than to other GATA domains in our GATA domain phylogeny (Supp. Fig. 1.6).

Examination of the SPR-1 and RCOR-1 ZnFs revealed that they all lack an important residue common to GATA factor ZnFs: no proteins in the rcor1 clade have a tryptophan (W) at position 12 in their ZnFs (Fig. 1.3A; Supp. Fig. 1.4A), which is invariant in most GATA factors (the only exception being some plants, where a methionine (M) occupies this position instead; Lowry & Atchley 2000). A tryptophan is found at this position in all other *Caenorhabditis* GATA domains included here (Supp. Fig. 1.4). Instead, the residue at position 12 in the ZnFs of *Caenorhabditis* SPR-1 and RCOR-1 orthologs is generally not well conserved. Most of the SPR-1 orthologs we analyzed have glutamic acid (E) at this position, and most RCOR-1 orthologs have an isoleucine (I) (Fig. 1.3A). A nuclear magnetic resonance (NMR) structure of *Gallus gallus* (chicken) GATA-1 (cGATA-1) C-terminal zinc finger (CF) bound to DNA showed that the tryptophan at ZnF position 12 is important for the structural integrity of the ZnF motif (Omichinski et al. 1993). Additionally, a crystal structure of the human GATA-3 (hGATA-3) CF bound to DNA revealed a hydrophobic interaction between the hGATA-3 CF tryptophan at position 12 and a thymine in the DNA (Chen et al. 2012). The function of this tryptophan is likely related to its large size and hydrophobicity. The residues occupying position 12 the ZnFs of SPR-1 and RCOR-1 proteins – glutamic acid and isoleucine – are both smaller residues than tryptophan. Isoleucine is hydrophobic and glutamic acid is negatively charged. A few other canonical animal GATA factor residues are also missing from rcor1 clade GATA domains, most of which are important for the structure or DNA binding of this domain and are found in most *Caenorhabditis* GATA factors. rcor1 clade proteins lack a threonine at the fifth position after the second cysteine, lack arginines at the 9th and 10th positions after the second cysteine, lack a glycine at the 14th position after the second cysteine, and lack an asparagine after the third cysteine (Fig. 1.3A). These non-canonical residues likely alter the ZnF DNA-binding structure and possibly the binding specificity associated with SPR-1 and RCOR-1 proteins compared to proteins comprising more canonical GATA DNA-binding domains. In fact, there is no evidence that any *Caenorhabditis* SPR-1 or RCOR-1 orthologs bind DNA at all. This domain may instead be used for protein-protein interactions, the structural integrity of the protein, or not at all.

We compared the residue with the highest probability at each position in each pHMM to the residues known to be important for the protein structure and/or DNA binding of animal GATA factors bound to DNA (see Methods; Omichinski et al. 1993; Bates et al. 2008; Lowry et al. 2009; Chen et al. 2012). These important GATA factor residues are shown in Figure 1.3C. The RCOR-1 and SPR-1 pHMMs contain the least number of these residues, including one *C. elegans* MED-1-like residue that is not conserved in either of the MED pHMMs (Fig. 1.3A,C; see below). The RCOR-1 and SPR-1 pHMMs share a similar number of conserved structural and/or binding residues in their GATA-like domains with EGL-27 pHMMs, which is fewer than all other GATA domain pHMMs examined in this study (Fig. 1.3A,C; see below). Moreover, most of the conserved residues in rcor1 domains are different from those in EGL-27 domains (Fig. 1.3A,C; see below), supporting their non-adjacent placement of their GATA-like domains in our GATA domain tree (Supp. Fig. 1.6). If the GATA-like domains in

RCOR-1 and SPR-1 shared a common ancestor with canonical GATA factors, their domains must have diverged extensively. Alternatively, the more likely scenario may be that these motifs evolved convergently.

spr-1 *moved to a new chromosome around the time the ancestral* rcor-1 *duplicated in the ancestor of the* Elegans *supergroup*

We never found *rcor-1* and *spr-1* orthologs on the same piece of assembled genomic DNA (Supp. Fig. 1.7A,B). In each of the six species in this study for which there was chromosome level genome resolution available (as opposed to only scaffolds and contigs), the *rcor-1* and *spr-1* orthologs were found on different chromosomes (1 and 5, respectively; an additional *spr-1* paralog is present on the X chromosome in *C. inopinata*). To expand our analysis of the chromosome locations of GATA-domain containing proteins throughout the genus, we assigned scaffolds or contigs to chromosomes based on the *C. elegans* assembly. For each GATA-domain containing protein on a scaffold or contig in our dataset, we compiled a list of its neighbors, used BLASTp to find their closest homolog in *C. elegans*, and assigned the scaffold or contig to the most common chromosome among these homologs (Fig. 1.4; see Methods). Though there have likely been interchromosomal rearrangements during the evolution of this genus, intrachromosomal rearrangements are more frequent (Stein et al. 2003; Teterina et al. 2020) and this is even the case between *C. elegans* and *C. bovis*, a basal species, which mostly have orthologs on the same chromosomes (Stevens 2020).

In this more detailed dataset we found that while the *Elegans*-supergroup *rcor-1* orthologs were assigned to chromosome 1, non-*Elegans* supergroup *rcor-1* orthologs are assigned to chromosome 5 and so the ancestral gene was probably on chromosome 5 (Fig. 1.4). The most parsimonious explanation for these results is that during or after the duplication of *rcor-1* and before the split of the *Elegans* supergroup species, a *rcor-1* paralog moved from chromosome 5 to chromosome 1 and this paralog stayed *rcor-1*-like. The other paralog, still on chromosome 5, is the ancestor of extant *spr-1* orthologs. Additionally, this data also suggests that a *C. inopinata*-specific *spr-1* duplication may have occurred interchromosomally or that movement between chromosomes occurred after a tandem duplication.

EGL-27 ortholog group

EGL-27 orthologs form a monophyletic clade that is more likely to share a more recent common ancestor with orthologs of the rcor1 clade than with canonical GATA factors

C. elegans EGL-27 functions in embryonic patterning by controlling cell polarity, migration, and fusion (Herman et al. 1999; Solari et al. 1999; Ch'ng & Kenyon 1999). We identified at least one EGL-27 ortholog in 59 of 60 species included in our phylogenetic analysis (Supp. Fig. 1.1; Supp. Table 2), suggesting that *egl-27* originated before split of *Caenorhabditis* and *Diploscapter* and may have been lost in *C. virilis*. EGL-27 orthologs form a well-supported (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) group in our phylogeny (Fig. 1.1A). These orthologs form a tight, distinct cluster closer to the rcor1 clade than to any other clade or ortholog group (Fig. 1.1A). Like the rcor1 clade genes, they contain ELM2 and

Myb/SANT domains (Fig. 1.2B) as well as an atypical GATA ZnF (Fig. 1.3A-C). And like the rcor1 clade orthologs, EGL-27 orthologs share homology with non-GATA factor proteins in other organisms (Wormbase.org), such as the human arginine-glutamic acid dipeptide repeats protein encoded by the RERE gene (reviewed in Wang & Tsai 2008), which has the same domains and similar organization as *Caenorhabditis* EGL-27s but with an additional Atrophin-1 domain (reviewed in Wang & Tsai 2008). Atrophin proteins repress transcription through recruitment of histone modifiers (Wang et al. 2006; 2008) but can be transcriptional activators in other contexts (Shen et al. 2007; Vilhais-Neto et al. 2010). Taken together, these data suggest that the EGL-27 orthologs likely share a more recent common ancestor with genes of the rcor1 clade than with any of the other genes we examined in this study.

egl-27 *genes are longer and are divergent from those in other ortholog groups*

We compared extant *egl-27* gene structures and made predictions about their ancestral gene(s) (Supp. Fig. 1.3C). For example, based on the large numbers of exons in the extant *egl-27* genes (more than in any other gene family included in this study) we predict that the ancestral *Guadeloupensis* group and *Elegans* supergroup *egl-27* had 14 exons (Fig. 1.2B). Moreover, EGL-27 protein sequences are substantially longer than all other protein sequences in our study (Fig. 1.2; Supp. Fig. 1.3). Only the EGL-27 orthologs contain a bromo-adjacent homology (BAH; Nicolas & Goodwin 1996) domain (Fig. 1.2; Supp. Fig. 1.3). BAH domains, which are often associated with gene silencing through protein-protein interactions (Callebaut et al. 1999), are large - in the case of EGL-27 orthologs the domain usually spans 4 exons

(Fig. 1.2B; Supp. Fig. 1.3C). Additionally, EGL-27 orthologs only have a single Myb/SANT domain whereas most rcor1 clade orthologs have two (Fig. 1.2A,B; Supp. Fig. 1.3A-C). *egl-27* orthologs are also distinguished by a conserved splice site immediately following the first 21 nucleotides coding for their ELM2 domain, a somewhat conserved splice site immediately following the first 24 or 102 nucleotides coding for their Myb/SANT domain, and a conserved splice site immediately following the first 22 nucleotides coding for their atypical GATA ZnFs (see below) (Fig. 1.2B; Supp. Fig. 1.3C). These conserved splice sites are not found in any of the other GATA-domain-containing proteins included in this study (Supp. Fig. 1.3), which further supports the divergent evolutionary history of *egl-27* orthologs. Overall, the EGL-27 orthologs are encoded by large genes, with conserved splice sites at different locations from those conserved in the other ortholog groups in this study, and have three non-GATA factor domains, two of which are also found in rcor1 clade genes. These results support their placement close to, but not within, the rcor1 clade in our phylogeny (Fig. 1.1A).

EGL-27 orthologs contain atypical GATA DBDs that are likely on a divergent evolution trajectory in relation to those of other Caenorhabditis *GATA-domain containing proteins*

We created an pHMM (Eddy 2020) of the atypical GATA domain in EGL-27 orthologs (Fig. 1.3A) and used it to search against all the proteins in our analysis (see Methods). The EGL-27 profile only scored EGL-27 orthologs well (Fig. 1.3B), which indicates the uniqueness of this GATA-like domain in *Caenorhabditis*. There are two primary reasons EGL-27 ZnFs are classified as GATA factors. One is due to the conserved tryptophan (W) at position 12 in their ZnFs and the other is their similarly spaced cysteines, i.e., $CX_2CX_{16}CX_2C$ (Fig. 1.3A; Supp. Fig. 1.4B; Supp. Fig. 1.5; prosite.expasy.org; Lowry & Atchley 2000). All animal GATA factors have a tryptophan at position 12; however, most of their cysteine pairs are 17 residues apart, as opposed to the 16 in EGL-27 orthologs (Teakle & Gilmartin 1998; Lowry & Atchley 2000; Gillis et al. 2008; Maduro 2020; this study). This shorter ZnF loop in *Caenorhabditis* EGL-27 orthologs is why their ZnFs are classified as atypical (prosite.expasy.org; Lowry & Atchley 2000). Only in three *Japonica* group sister species, *C. nouraguensis*, *C. becei*, and *C. yunquensis*, do EGL-27s have the typical 17 residues in their ZnF loops (Supp. Fig. 1.4B), which suggests that a 3 base pair (bp) insertion/addition to the *egl-27* ZnF coding sequence or an alternative splice site likely occurred in their most recent common ancestor. Interestingly, *Diploscapter* EGL-27 paralogs, EGL-27 orthologs from two Clade II nematodes, EGL-27 orthologs from a Clade III nematode, and the mouse and human RERE proteins all have ZnF loops that are 18 residues long, whereas the EGL-27 ortholog from the Clade V *Pristionchus pacificus* nematode has a ZnF loop with 19 residues. *Drosophila* species appear to lack an EGL-27 ortholog altogether (Wormbase; Haag et al. 2018). Overall, the loop size of EGL-27/RERE orthologs has likely changed multiple times in different lineages, and the orthologous gene may have been lost from the ancestor of the *Drosophila* lineage. Although there are a few exceptions (see details in the elt1/2 clade and MED ortholog group sections below), since most other *Caenorhabditis* GATA factor orthologs have ZnFs with the typical animal loop length of 17 residues (Supp. Fig. 1.5), the atypical length of most

EGL-27 ortholog ZnF loops is more consistent with a divergent evolutionary history thus supporting this clade's placement in our phylogeny more distant from the clades with more canonical GATA factors (Fig. 1.1A).

There are 21 residues in EGL-27 ZnFs and BRs that distinguish them from typical *Caenorhabditis* GATA domains, and 10 of them are also found in the human and mouse RERE proteins, which have identical GATA DBDs (Supp. Fig. 1.8), suggesting that these 10 have been highly conserved since the vertebrates and invertebrate ancestors diverged. The 21 EGL-27-specific ZnF and BR residues are: histidines (H) at positions three and four, a tyrosine (Y) or a phenylalanine (F) at position six, aspartic acid (D) at position 12, histidines at positions 14 and 15, a threonine (T) at position 25, an aspartic acid or a glutamic acid (E) at position 26, an arginine (R) at position 28, usually a leucine (L) at position 37, usually a non-proline (non-P) at position 39, usually an alanine (A) or a valine (V) at position 41, usually an asparagine (N) at position 42, a proline at position 44, usually a threonine at position 45, usually a glutamine (Q) at position 46, a proline at position 48, usually an alanine at position 49, a cysteine (C) at position 50, a leucine at position 51, and a phenylalanine at position 52 (Supp. Fig. 1.8). Similar residues (Henikoff & Henikoff 1992; see Methods) are rarely found in these positions in canonical *Caenorhabditis* GATA domains (Fig. 1.3A,D,E; Supp. Fig. 1.4D-F). The few GATA factors that do have some of these residues do not have more than three of them (Supp. Fig. 1.4D-F). These minor similarities between canonical GATA ZnFs and EGL-27 ZnFs are likely due to convergent evolution. 11 of these 21 EGL-27-specific ZnF residues are known to be important for the structure and/or binding function of the cGATA-1 DBD or the

DNA binding of mouse and human GATA-3s (Omichinski et al. 1993; Bates et al. 2008; Chen et al. 2012) and so may affect the DNA-binding kinetics or specificity of EGL-27 orthologs. For example, the *C. elegans* EGL-27 ortholog binds not only to a GATA site (GATAAG) but also to the non-GATA sites GAGAAG and RGRMGRWG (Xu & Kim 2012).

The number of binding or structural residues that are identical or similar between canonical animal GATA domains and those in the EGL-27 and rcor1 clade orthologs are approximately the same (Fig. 1.3C; see above), and there are fewer similarities compared to the more typical GATA factor ortholog groups in this study (Fig. 1.3C). Additionally, the similarities in EGL-27 orthologs are different from the similarities in rcor1 clade orthologs (Fig. 1.3C), suggesting that their GATA-like domains likely do not share a recent common ancestor. The shared ELM2 and Myb/SANT domains rather than the GATA domain may be the shared features between EGL-27 and rcor1 clade orthologs that place them on a closer branch to each other than to GATA factors in our phylogeny (Fig. 1.1A), especially since these GATA domains do not group adjacently in our GATA domain phylogeny (Supp. Fig. 1.6). The EGL-27 GATA domains are highly conserved across the genus (Fig. 1.3A; Supp. Fig. 1.6), suggesting they have experienced strong negative selection. Moreover, many of the non-canonical features of the EGL-27 GATA-like motifs are conserved among vertebrate RERE proteins (Supp. Fig. 1.8) suggesting that EGL-27 ZnF and BRs were diverged from canonical GATA factor DBDs at least since the split between vertebrate and invertebrates.

Most egl-*27 orthologs are not syntenic with any other GATA-domain-containing proteins*

We only found two *Caenorhabditis* species with their *egl-27* ortholog syntenic with any GATA factors or rcor1 clade orthologs included in this study (Supp. Fig. 1.5C; data not shown). The first is in *C*. sp. *44*, where we found the *egl-27* ortholog is 12 Mb from the *rcor-1* ortholog (data not shown). The second is in *C. tropicalis* where the *egl-27* ortholog is on the same scaffold as the *elt-2*, *elt-3*, and *egl-18* orthologs, a truncated *elt-6* ortholog, and three truncated *med* orthologs. This is unexpected, since most of these orthologs are located on three different chromosomes, whereas *med* orthologs are found on all chromosomes (Fig. 1.5). In the few species with available chromosome-level assemblies (which does not include *C. tropicalis*) *egl-27* orthologs are found on chromosome 2, *elt-2* and *elt-3* orthologs are located on the X chromosome, and *egl-18* and *elt-6* orthologs are found on chromosome 4. Interestingly, we found that *C. tropicalis egl*-27, *egl-18*, *elt*-2, and *elt-3* neighbors within 70 kb BLASTp top hit in *C. elegans* are mainly on chromosome 1, 4, X, and X, respectively, which are the same chromosomes these orthologs are found on in species with chromosome-level assemblies (Fig. 1.5). This suggests that interchromosomal rearrangements of at least kilobase long multi-gene sequences from at least three chromosomes may have occurred in *C. tropicalis***.** It will be interesting to re-examine synteny in *C. tropicalis* once its genome has been more fully assembled. Until then, our data suggest that *C. tropicalis* has undergone substantial chromosome rearrangements as compared to most of the other species included in this study.

Atypical Caenorhabditis *GATA factors*

MED ortholog group

MED sequences in different species are highly variable likely due to quick turnover

The MED ortholog group is the largest and contains many of the most divergent GATA factors (Fig. 1.1). The topology of the MED ortholog group is only somewhat consistent with the phylogenetic tree of *Caenorhabditis* species (Fig. 1.1; Stevens 2020), which was also found by Maduro (2020). MEDs in sister species are mostly clustered closer to each other than to MEDs in non-sister species and most *Elegans* group MEDs group closer to each other than to *Japonica* group MEDs, as expected (Fig. 1.1). However, all *C. brenneri* and all but one *C.* sp. *48* MED orthologs group within the *Japonica* group MEDs (Fig. 1.1). Moreover, the branches leading to sister species MED ancestors are shorter than the branches of most species-specific paralog MED ancestors (Fig. 1.1). For example, *C. elegans* MED-1 and MED-2 group adjacently, on short (0.02 substitutions per site) branches from their most recent common ancestor node, while a long branch (1.1) connects the *C. elegans* MED ancestor node to the nearest node of the most recent common ancestor of other species MED orthologs (Fig. 1.1). This suggests that either *C. elegans* MED-1 or MED-2 arose from a recent duplication of the other, or that gene conversion has made these sequences more similar to each other than to their more recent common ancestors in other species. This pattern of MED paralogs within a species grouping together holds true for most of the MED orthologs included in our study (Fig. 1.1) and was also noted by Coroian et al. (2006) in their examination of MED orthologs in *C.*

elegans, *C. briggsae*, and *C. remanei* and Maduro's (2020) phylogeny of 20 *Caenorhabditis* species MED ortholog. Most of the branches leading to species-specific MED paralog ancestral nodes are longer than most branches in other ortholog groups (Fig. 1.1A), a testament to the protein sequence variability in this group. This variability suggests that these proteins have evolved, and may still be evolving, at faster rates than other GATA-domain-containing proteins. Additionally, most *Caenorhabditis* species possess variable numbers of species-specific *med* gene paralogs (Fig. 1.1; Maduro 2020). For example, we only found a single *med* ortholog in *C*. sp. *46* but we found at least 14 *med* paralogs in *C. doughertyi* (Fig. 1.1; Supp. Table 2). This supports the hypothesis that the extant *med* genes are the product of rapid turnover, a conclusion also reached by Maduro (2020).

Most MED orthologs have a single GATA domain, which is the case for all *Caenorhabditis* GATA-domain containing proteins outside the elt1/2 clade (Maduro 2020; Eurmsirilerd & Maduro 2020; this study). However, we found a few MED orthologs with multiple GATA domains (Supp. Table 2), some of which were found previously by Maduro (2020). As Maduro (2020) proposed, the additional GATA domains likely evolved through interstitial ZnF duplications since the dual ZnFs of these proteins are more like each other than the ZnFs of paralogous *meds* (data not shown). This supports the hypothesis of quick evolutionary dynamics of these proteins. We also found examples of what are likely small insertions and deletions in some *med* genes (data not shown), perhaps indicative of pseudogenization. Collectively, this evidence suggests that at least some *med* paralogs have likely experienced relaxed selection. The fast rate of evolution of *med* genes have allowed *med* orthologs to explore more evolutionary space than any of the other GATA-domain-containing ortholog groups (Fig. 1.1).

Many MEDs are likely orthologous, despite their sequence variability

Although MED protein sequences are highly variable between species (Fig. 1.1; see above), the MED ortholog group is robust (100%) to ultrafast bootstrapping (Minh et al. 2013; Hoang et al. 2018; Fig. 1.1A) and previous phylogenies, one of *C. elegans* and *C. briggsae* GATA factors (Gillis et al. 2008) and another of MED, END-1, and END-3 orthologs in 20 *Elegans* supergroup species (Maduro 2020), have also grouped MEDs into monophyletic clades. This hypothesis is also supported experimentally by Coroian et al. (2006) who found that two *C. briggsae* and five *C. remanei med* genes were able to compensate for the otherwise lethal knock-out of both *med* genes in *C. elegans*, suggesting that even though *med* sequences have diverged, their functionality has been conserved, at least within the *Elegans* group (Coroian et al. 2006). This point is especially significant considering that the *C. elegans*, *C. briggsae*, and *C. remanei* MEDs do not cluster with each other (Fig. 1.1; Maduro 2020) and that none of these are sister species (Stevens 2020). We conclude that the almost identical *med* genes in *C. elegans*, are likely orthologous to at least the *Elegans* group *med* genes included in our phylogeny. Additionally, we predict that *Japonica* group *meds* are likely orthologous at least to each other and probably to the *Elegans* group *meds* as well.

The most recent common med *ancestor was likely in the ancestor of the* Elegans *supergroup species*

MED orthologs are almost completely restricted to the *Elegans* supergroup, and we found at least one in all 35 *Elegans* supergroup species (Fig. 1.1). The most parsimonious explanation for this is that the *med* ancestor originated in the ancestor of the *Elegans* supergroup. This hypothesis was originally espoused by Maduro (2020). The 18 additional non-*Elegans* supergroup species included in our study, all of which lack a canonical *med* ortholog, provide further support for this theory. The only potential non-*Elegans*-supergroup MED is a gene in *C.* sp. *45* (Fig. 1.1), which codes for an *end-1*-like ZnF motif (i.e., $CX_4CX_7WX_9CX_2C$), that may have evolved convergently since otherwise the DBD of this protein is more similar those of MED orthologs than to those of END-1 orthologs (Supp. Fig. 1.6; data not shown). This *C.* sp. *45* protein is on a long (4 substitutions per site), relatively isolated branch and groups most basally to the MED ortholog group with another possibly misplaced protein from *C.* sp. *51* (Fig. 1.1). The *C.* sp. *51* gene also has non-*med* features including its gene structure (see below) and is also on a long branch (2.8 substitutions per site). The long branches of these non-canonical *C.* sp. *45* and *C.* sp. *51* proteins (Fig. 1.1) suggests that their evolutionary history is unclear. Maduro (2020) predicted that the *med* genes originated from an *end-1* or *end-3* duplication. However, our phylogeny does not provide support for this, or for any hypothesized origin, since the MED ortholog group does not cluster close to any other ortholog group, including the END-1 and END-3 groups (Fig. 1.1). Our additional evidence for whether an elt3 clade gene duplication could have produced the ancestral *med* gene is relatively weak (see below). The origin of the *med* gene remains unresolved.

Most Elegans *group* med *genes have lost all introns while the gene structures of most* Japonica *group* med *genes resemble the gene structures of the* end-1 *and* end-3 *clades*

The *med* genes in *Elegans* and *Japonica* group species both have small coding sequences, but their gene structures are divergent from each other (Supp. Fig. 1.3D,E). We predict that the ancestor of the *Elegans* group *med* genes contained a single large exon, while the ancestor of the *Japonica* group *med* genes contained three exons (Fig. 1.2C). Both ancestral *med* genes coded for the shortest GATA-domain-containing proteins associated with this study (with lengths of 187 and 193 residues, respectively) (Fig. 1.2; Supp. Fig. 1.3). All but one *Japonica* group *med* orthologs have the same splice site location in their ZnF as is found in the orthologs in the elt3 clade and the ELT-2 ortholog group (Supp. Fig. 1.3E,I-M). This suggests a shared ancestry since this splice site is not found in any of the other GATA-domain-containing proteins included in this study and is unlikely to have arisen twice independently. The short introns (data not shown) and few exons of the *med* genes in the *Japonica* group most closely resemble the gene structures of *end-1* and *end-3* orthologs (Supp. Fig. 1.3E,L,M), which supports the hypothesis of Maduro (2020) that an *end-3* duplication produced the *med* ancestral gene. The *med* genes in *Elegans* group have more variable gene structures (Supp. Fig. 1.3D) which give few clues as to how they arose. Relative to *Japonica* group *meds* (Supp. Fig. 1.3E) and considering the median *C. elegans* exon size (123 bps (Wormbook; Spieth et al. 2018)) we hypothesize that the *Elegans* group *med* ancestor lost at least two introns to become a large single-exon gene (Fig. 1.2C). While most *med* genes in *Elegans* group species have no introns, one divergent *C.* sp. *51 med*-like paralog (see above) contains a splice site at the same location in its basic region as in the elt6 clade and the *elt-1* orthologs (Supp. Fig. 1.3F-H; see below). This gene may be a pseudogenizing elt6 clade or *elt-1* paralog, or the similarity could be the result of convergent evolution.

MED orthologs share non-canonical residues and have more variability in their GATA domains compared to other Caenorhabditis *GATA factors*

We created pHMMs (Eddy 2020) of the *Japonica* group and *Elegans* group MED DBDs, respectivly (Fig. 1.3A) and used them to query all the proteins in our analysis (Fig. 1.3B; see Methods). Except in ZnF loop length, *Japonica* group and *Elegans* group MED DBDs are more like each other than to any of the other GATA domains in this study (Fig. 1.3A) and so, as expected, their pHMMs scored each of the MEDs from the other group second highest, after the MEDs from their own species group (Fig. 1.3B). Both MED pHMMs specifically score MEDs higher than other GATA domain containing proteins (Fig. 1.3B), supporting the MED ortholog group in our phylogeny (Fig. 1.1A). Some primarily MED-specific residues are conserved within most MED DBDs, supporting predictions of their more recent common ancestry and divergent binding specificity, while other poorly conserved residues suggest these genes are experiencing rapid evolution and/or relaxed selection. For example, one of the similar residues arginine (R), lysine (K), or glutamine (Q) (Henikoff & Henikoff

1992) is usually found at position nine in MED ZnFs whereas similar residues are not found at this position in any of the other GATA domains (Fig. 1.3A). The arginine at this position in *C. elegans* MED-1 forms hydrogen bonds with the 5' region of the non-canonical MED DNA binding site (Lowry et al. 2009), thus part of this non-canonical binding may be conserved in most, if not all, MEDs. On the other hand, a tyrosine (Y) found at position 19 in the BRs of most *Elegans* group MEDs is not conserved in *Japonica* group MEDs (Fig. 1.3A; Supp. Fig. 1.4C) and in *C. elegans* MED-1 this tyrosine interacts with the 3' region of the non-canonical MED DNA binding site (Lowry et al. 2009). Since this position was not found to be involved in the binding of other animal GATA factors to DNA, *Japonica* group MEDs may have a more canonical DNA binding site than *Elegans* group MED orthologs. Some poorly conserved residues in MED DBDs include position nine in the ZnF, which is a highly conserved threonine (T) in canonical GATA factors but a threonine, serine (S), asparagine (N), or cysteine (C) in MED orthologs (Fig. 1.3A; Supp. Fig. 1.4C) and four residues before the third zinc coordinating cysteine which is a glycine (G) in most canonical GATA factors, but an asparagine, aspartic acid (D), glutamic acid (E), threonine, lysine, arginine, or serine in the MEDs.

Most (86 of 94) *Elegans* group MED ZnFs have loops 18 residues long (Fig. 1.3A; Supp. Fig. 1.5; Supp. Fig. 1.4C). This is atypical for animal GATA factors and for the other genes in this study, although canonical plant and fungi GATA factors have loops of this length (Teakle & Gilmartin 1998) (Fig. 1.3A; Supp. Fig. 1.5). The eight other *Elegans* group MEDs all have ZnF loops that are 17 residues long (Supp. Fig. 1.5; Supp. Fig. 1.4C), like canonical animal GATA factors (Teakle & Gilmartin 1998).

These MEDs are found in three non-basal species, two of which are sister species. Due to the placement of the two sister species MEDs in our phylogeny (Fig. 1.1) the most parsimonious explanation is that a three nucleotide deletion, or an alternative splice site, in the ZnF loop coding sequence of one ancestral MED occurred in the most recent common ancestor of *C*. sp. *48* and *C. brenneri*. A *C*. sp. *51* MED-like protein also only has a 17 residue long loop, but this may be due to this protein having a more recent common ancestor with a non-MED GATA-domain containing protein (see above). We think that the *Elegans* group ancestral MED had a ZnF with 18 residues in its loop. The ZnFs loops of most (23 of 36) *Japonica* group MEDs are also an atypical length - they are a residue shorter than canonical GATA factors (Supp. Fig. 1.5; Supp. Fig. 1.4C; Lowry & Atchley 2000) but the same length as most *Caenorhabditis* EGL-27 ZnF loops (Fig. 1.3A; Supp. Fig. 1.5; Supp. Fig. 1.4B). However, due to the lack of homology between *Japonica* group MEDs and EGL-27 orthologs, this likely reflects convergent evolution. The other 13 *Japonica* group MEDs have ZnF loops of length 17 (Supp. Fig. 1.5; Supp. Fig. 1.4C), like canonical animal GATA factors (Teakle & Gilmartin 1998), but since they are not found in basal *Japonica* group species, we propose that the ancestral *Japonica* group MED had a 16 residue long ZnF loop. Due to their placement in our phylogeny (Fig. 1.1) and since GATA ZnF insertions are relatively rare compared to substitutions, at least eight of these 13 *Japonica* group MEDs likely shared the same three nucleotide insertion, or alternative splice site, in their most recent common ancestor.

We compared the residue with the highest probability at each position in the *Japonica* group and *Elegans* group pHMMs (Fig. 1.3A) to the residues known to be important structurally or for DNA binding in animal GATA factors bound to DNA (Fig. 1.3C; see Methods; Omichinski et al. 1993; Bates et al. 2008; Lowry et al. 2009; Chen et al. 2012). Both of these MED pHMMs have nine of the 24 important DNA interaction residues, and four residues that are similar. Of the 18 structurally important residues, the *Japonica* group and *Elegans* group pHMMs have eight and 10 identical and three and two similar residues, respectively (Fig. 1.3C). Compared to the EGL-27 and rcor1 clades, MED DBDs contain a couple more residues important for GATA factor DNA binding or structural integrity in common with canonical animal GATA domains (Fig. 1.3C). Moreover, MED GATA domains more closely resemble canonical *Caenorhabditis* GATA domains than they do the atypical EGL-27 or rcor1 clade GATA domains (Fig. 1.3A), supporting the hypothesis that the MEDs arose from one of the canonical *Caenorhabditis* GATA factors instead of from a different GATA-domain-containing protein.

Many species have closely related syntenic med *paralogs, yet sister species often have* meds *on different chromosomes*

We searched for genes that are coded on the same chromosome, scaffold, or contig (see Methods) and found that confident *med* paralogs in 19 species are syntenic (24 species if we include non-confident *meds*), and, with the exception of some confident *C. brenneri meds* and a pair of confident *C. latens meds,* they are relatively close to each other (range of 319 bp to 13 kb, median of 3.1 kb) (Supp. Fig. 1.7D). Most of these close confident syntenic *med* paralogs also have high sequence conservation (ranging from 48% to 100% identical, median of 89%) (Supp. Fig. 1.9A-

C). Together this suggests that these paralogs likely arose through recent tandem duplications. The range of percent identity for non-syntenic *med* paralogs (37%-98%), is similar to syntenic *meds* however, the median is much lower (53%) (Supp. Fig. 1.9D). The syntenic *meds* may simply be more recent duplicates with less time to diverge, which would imply that *med* paralogs tend to move farther apart. In our chromosome assignment analysis, we found that at least one *med* ortholog is found on every chromosome and that most sister species have *med* paralogs on at least two different chromosomes (Fig. 1.4). Only the sister species *C. oiwi* and *C. kamaaina* have the same total number of *meds* and these *meds* are predicted to be on the same chromosomes (Supp. Table 2; Fig. 1.4). Overall, the synteny and percent identity between *med* paralogs lead us to conclude that the *med* genes turn over with rapid duplications and deletions along species-specific lineages and that they repeatedly move around the genome during evolution.

Canonical Caenorhabditis *GATA factors*

elt6 clade

egl-18 *originated through a duplication of* elt-6 *that occurred in the common ancestor of the* Elegans *supergroup and* Guadeloupensis *group species*

Our phylogeny places the ELT-6 and EGL-18 (also known as ELT-5) ortholog groups in a well-supported (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) monophyletic clade (Fig. 1.1A). A previous phylogenetic study of *C. elegans* and *C. briggsae* also placed ELT-6 and EGL-18 orthologs in a well-supported monophyletic clade (Gillis et al. 2008). In *C. elegans*, ELT-6 and EGL-18 function redundantly in seam cell specification (Koh and Rothman, 2001), larval seam cell maintenance (Gorrepati et al. 2013), male tail morphogenesis (Nelson et al. 2011), and inhibition of cell fusion during vulval development (Koh et al. 2002). However, ELT-6 and EGL-18 do not necessarily contribute equally to these phenotypes (Gorrepati et al. 2013; Nelson et al. 2011). While the functions of ELT-6 and EGL-18 orthologs in other *Caenorhabditis* species have not been studied, conservation in protein sequence among the ELT-6 and EGL-18 orthologs in this study (Supp. Fig. 1.4D) make it likely that their functions are also conserved. Because EGL-18 orthologs are only found in *Elegans* supergroup and *Guadeloupensis* species, whereas ELT-6 orthologs are present throughout the *Caenorhabditis* genus as well as in the *Diploscapter* species included in our study (Fig. 1.1), we conclude that an *elt-6* duplication produced *egl-18* in the ancestor of the *Elegans* supergroup and *Guadeloupensis* group.

The Elegans *supergroup ancestral* egl-18 *likely lost three introns while the* Guadeloupensis *group ancestral* elt-6 *likely lost a single exon*

Most of the singleton *elt-6* orthologs (30 of 40) have seven exons and code for a single GATA ZnF within their third to last exon (Supp. Fig. 1.3F). This gene structure appears to have been conserved among *egl-18* genes in *Guadeloupensis* group species as well, at least in the two species for which genomic sequences are available (Supp. Fig. 1.3G), which suggests that a full-length *elt-6* duplication produced the *egl-18* ancestor in the *Guadeloupensis* and *Elegans* supergroup ancestor. Interestingly, all 30 singleton *egl-18* orthologs in the *Elegans* supergroup have only four exons, and they also code for a single GATA ZnF in their second to last exon (Supp. Fig. 1.3G). Yet despite these differences in their gene structures, the protein sequences of ELT-6 and EGL-18 orthologs are of similar lengths (medians of 389 and 420 residues, respectively) and align along their entire lengths (Supp. Fig. 1.4D). These data suggest that the *Elegans* supergroup ancestral *egl-18* lost three introns (Fig. 1.2D). The locations of the ZnF-coding exons indicate that the *Elegans* supergroup *egl-18* gene lost intron 6, the final intron in its *elt-6* ancestor (Fig. 1.2D). Additionally, *Elegans* supergroup *egl-18s* have large second exons (median 636 nts), which are larger than the combined lengths of exons 2, 3, and 4 of most *elt-6* genes (median total 595.5 nts) (Supp. Fig. 1.4D). This suggests that the *Elegans* supergroup *egl-18* ancestor also lost introns 2 and 3, relative to its *elt-6* ancestor (Fig. 1.2D). *Guadeloupensis* group *elt-6s* likely also lost intron 2 as compared to most other *elt-6* orthologs in this study (Fig. 1.2D). *Guadeloupensis* group *elt-6* orthologs also have larger exon 2s (median 383.5 nts) which approximate the combined lengths of exons 2 and 3 (median 371.5 nts) in most *elt-6* genes (Fig. 1.2D). Our analyses of phylogeny (Fig. 1.1) and synteny (see below) suggest that the loss of intron 2 in *Elegans* supergroup *egl-18* genes and *Guadeloupensis* group *elt-6* genes occurred convergently, although another possibility is that the *Elegans* supergroup ancestral *egl-18* is more closely related to the *Guadeloupensis* group ancestral *elt-6* than to the *Guadeloupensis* ancestral *egl-18*.

<u>GATA domains of elt6 clade orthologs contain a conserved intron in their basic regions</u>

The *elt-6* and *egl-18* orthologs have a conserved intron 24 nucleotides into the basic region (BR), which is directly C-terminal to the ZnF motif (Supp. Fig. 1.3F,G).

This intron location is conserved for all singleton *elt-6* and *egl-18* genes in this study as well as in all singleton *elt-1* CFs apart from the *C. sinica elt-1* CF (Supp. Fig. 1.3F-H). Eurmsirilerd & Maduro (2020) found this conserved intron location in many, mostly non-*Caenorhabditis*, nematode *egl-18*/*elt-6*, *elt-1*, and Clade I *elt-2* CFs and we extend this finding to *elt-6*, *egl-18*, and *elt-1* orthologs in 53 more *Caenorhabditis* species. This splice site location is also found in vertebrate and some arthropod GATA factors and was likely the splicing location in the bilaterian ancestral GATA factor CF (Gillies et al. 2008). This conserved intron location supports our phylogenetic placement of the ELT-6 and EGL-18 ortholog groups into a monophyletic clade. It also hints at a relationship between the elt6 clade genes and the CFs of genes in the elt1/2 clade.

EGL-18 and ELT-6 orthologs have similar GATA DNA-binding domains

We created pHMMs (Eddy 2020) of the ELT-6 and EGL-18 DBDs (Fig. 1.3A) and used them to search against all the proteins included in our analysis (Fig. 1.3B; see Methods). We found the ELT-6 ortholog group pHMM scores higher against EGL-18 orthologs than any others except ELT-6 orthologs and vice versa (Fig. 1.3B), indicating that the DBDs of these ortholog groups are most like each other. The ELT-6 and EGL-18 orthologs uniquely and exclusively have an alanine (A) at position seven in their ZnF loops (Fig. 1.3A). This alanine is also conserved in all the ZnFs of ELT-6/EGL-18 orthologs identified across the Nematode phylum (Eurmsirilerd & Maduro 2020). An NMR structure of cGATA-1 DBD bound to DNA showed that the leucine (L) at position seven in the cGATA-1 ZnF loop interacted with DNA in the major groove (Omichinski et al. 1993); and most other *Caenorhabditis* GATA factors have a leucine conserved at this position as well (Fig. 1.3). The *Aspergillus nidulans* AreA GATA factor also has a leucine at this position, as do most animal and some other fungi GATA factors (Teakle & Gilmartin 1998). Mutations of this leucine to a valine (V) significantly alter the binding specificity and affinity of AreA *in vivo* (Ravagnani et al. 1997) and *in vitro*, such that it binds better to TGATA better than to CGATA DNA sites (Starich et al. 1998b). A mutation in AreA at this same site to a Methionine (M) has weaker but opposite effect (Arst & Scazzocchio 1975; Gorton 1983). Because both valine and alanine are smaller than leucine, whereas methionine is larger, the alanine residue found in ELT-6 and EGL-18 orthologs may also increase specificity for TGATA over CGATA DNA sites and likely alters the ZnF interactions with DNA in some way that could change the DNA-binding kinetics of these orthologs.

We also compared the residues of highest probability at each position in both the ELT-6 and EGL-18 pHMMs to the residues found to be important for structure or DNA binding in animal GATA factors bound to DNA (Fig. 1.3C; see Methods; Omichinski et al. 1993; Bates et al. 2008; Lowry et al. 2009; Chen et al. 2012). The ELT-6 pHMM has 19 residues identical to and one residue similar to the 24 residues found to interact with DNA. Additionally, the ELT-6 pHMM has 13 residues identical to and two residues similar to the 18 residues found to be important for the structural integrity of the cGATA-1 DBD (Fig. 1.3C). The EGL-18 pHMM is very similar to the ELT-6 pHMM with 18 residues identical to and two residues similar to the 24 residues important for DNA binding and 13 residues identical to and one residue similar to the 18 residues important for the DBD structure (Fig. 1.3C). Of the twelve ortholog groups in our phylogeny, the ELT-6 and EGL-18 pHMMs contain the second and third most conserved functional residues, respectively, suggesting that their DBDs are likely under similar levels of stabilizing selection.

egl-18 *likely originated from a tandem duplication of* elt-6*, and dicistron transcription of these genes may be conserved throughout the* Elegan*s supergroup and* Guadeloupensis *group species*

elt-6 and *egl-18* are adjacent to each other on the same chromosome/scaffold/contig in 30 of the 31 species where we could confidently identify orthologs of both genes (Supp. Fig. 1.7E,F). Moreover, in all these species *egl-18* orthologs are found upstream of *elt-6* orthologs and on the same strand and thus transcribed in the same direction (Supp. Fig. 1.7E,F). In *C. elegans, egl-18* and *elt-6* are sometimes transcribed together as a dicistron (Koh & Rothman 2001), and since this operon-like structure is conserved in many of the species in this study, these additional species may also express these genes via a dicistron. The tight synteny between *elt-6* and *egl-18* orthologs supports our hypothesis that a duplication of *elt-6* produced the *egl-18* ancestor, and this ancient duplication was probably a tandem duplication.

ELT-2 ortholog group

The ELT-2 ortholog group includes both the *C. elegans* ELT-2 and ELT-4 proteins; ELT-4 is on a branch near *C. elegans* ELT-2, but it does not group adjacent to it (Supp. Fig. 1.1). *C. elegans* ELT-4 has no known function and is thought to be a pseudogene that resulted from a *C. elegans*-specific *elt-2* duplication (Fukushige et al.

2003). We did not identify any *elt-4* orthologs or even other *elt-2* duplications in any of the 57 other *Caenorhabditis* species, supporting the hypothesis that a duplication of *elt-2* within *C. elegans* produced *elt-4* and that its position away from *C. elegans elt-2* in our tree reflects a loss of selective constraint on its sequence.

elt1/2 clade

<u>*ELT-1 and ELT-2 orthologs group adjacently forming an elt1/2 clade with ancient origins*</u>

The elt1/2 clade consists of the ELT-1 and ELT-2 ortholog groups, which cluster adjacently forming a well-supported monophyletic clade (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) (Fig. 1.1A). We found at least one *elt-1* ortholog and one *elt-2* ortholog in every *Caenorhabditis* and *Diploscapter* species included in this study (Fig. 1.1), and *elt-1* and *elt-2* orthologs have also been found in all extant nematodes with fully sequenced genomes (Eurmsirilerd & Maduro 2020). This suggests the ancestors of these genes were present at the beginning of the nematode phylum. In *C. elegans,* ELT-1 and ELT-2 orthologs function in different germ layers: ELT-1 is involved in the specification and differentiation of hypoderm precursors (Page et al. 1997; Gilleard & McGhee 2001), while ELT-2 is involved in differentiation and maintenance of endoderm cells (Fukushige et al. 1998). Additionally, the gene structures of *elt-1* and *elt-2* orthologs are different (Supp. Fig. 1.3H,I; Eurmsirilerd & Maduro 2020; see below) and we found no cases in which *elt-2* and *elt-1* orthologs were syntenic (Supp. Fig. 1.7).

ELT-1 and ELT-2 are unusual in *Caenorhabditis* because they both have two zinc fingers (Fig. 1.2E; Supp. Fig. 1.3H,I; Supp. Fig. 1.4E). Their C-terminal ZnFs (CF) are similar (Fig. 1.3A; Supp. Fig. 1.4E; see below), but their NF sequences have diverged (Fig. 1.3D,E; Supp. Fig. 1.4E; see below). In fact, it appears that the ELT-2 NF has experienced relaxed selection to the point where most ELT-2 NFs are barely recognizable other than their two cysteine pairs (Fig. 1.3E; Supp. Fig. 1.4E; see below). Vertebrate GATA factors have two canonical GATA ZnF motifs whereas invertebrate, plant, and fungi GATAs have either one or two motifs (Reyes et al. 2004). We therefore hypothesize that the shared ancestry of the elt1/2 clade goes back further than for any of the other well-supported clades in our phylogeny and that the sequences and likely the function(s) of the NFs of these GATA factors have diverged extensively since their most recent common ancestor.

<u>elt-1 *and* elt-2 *ortholog GATA domains are encoded differently*</u>

Even though many *elt-1* and *elt-2* orthologs have nine exons and code for proteins of similar length, there are conserved gene structure features that distinguish the genes of these two ortholog groups from each other (Supp. Fig. 1.3H,I). Our predicted ancestral elt1/2 clade gene structures provide a visual summary of the differences (Fig. 1.2E; Supp. Fig. 1.3H,I). Most of the ZnF motifs are coded in different exons and have different splice sites (Supp. Fig. 1.3H,I). For example, the NF in the *Elegans* supergroup ancestral *elt-2* is coded in exon 4 whereas it is coded in exon 5 in the *Elegans* supergroup and *Guadeloupensis* (and the *Elegans* supergroup only) ancestral *elt-1* (Fig. 1.2E). Moreover, the introns in extant *elt-1* orthologs are nearly always longer than the introns in *elt-2* orthologs, and the gene structures of the latter are relatively compact (data not shown). *elt-2* orthologs all have a conserved intron located just upstream of the last seven nucleotides comprising their CFs (Supp. Fig. 1.3I; Fig. 1.2E). This feature is shared with the elt3 clade (see below) and *Japonica* group *meds* (see above), and is found in most nematode *elt-2* and *elt-3* orthologs (Eurmsirilerd & Maduro 2020) and most previously examined *Caenorhabditis end-1*, *end-3*, *elt-2*, and *Japonica* group *med* orthologs (Maduro 2020). *elt-2* NFs have a less well conserved intron position located upstream of the last 19 nucleotides comprising that motif that is not found in any of the other genes in this study (Supp. Fig. 1.3I; Fig. 1.2E). *elt-1* orthologs have a conserved intron position located just after the first 24 nucleotides comprising their CF basic regions, which is also found in elt6 clade genes (see above), and their NFs have a conserved intron position located 60 nucleotides after the end of the ZnF sequence (Supp. Fig. 1.3H; Fig. 1.2E). These two splice site locations are likely conserved from the bilaterian ancestral GATA factor (Gillis et al. 2008) and were previously found in some nematode *elt-1* and *elt-6* orthologs (Eurmsirilerd & Maduro 2020). This NF intron location is not found in any of the other genes in this study (Supp. Fig. 1.3). Additionally, the spacing between the NF and CF in *elt-1* orthologs is highly conserved. All singleton *elt-1* orthologs outside the *Elegans* supergroup have 29 residues between their NF and CF motifs, while in most singleton *Elegans* supergroup *elt-1* genes that spacing is 30 residues (Supp. Fig. 1.3H). Twenty-nine residues between the N- and CFs is also the predicted state of the bilaterian ancestral GATA factor (Gillis et al. 2008), suggesting that this spacing represents the ancestral state and that the ancestor *elt-1* of the *Elegans* supergroup acquired another residue between its ZnFs**.** On the other hand, the numbers of residues between the NFs and CFs of singleton *elt-2* orthologs is more variable, ranging from 36 to 89 residues (Supp. Fig. 1.3I). Overall, the divergent organization of these ZnF domains in *elt-1* and *elt-2* orthologs indicate divergent evolutionary paths since they last shared a common ancestor.

ELT-1 and ELT-2 orthologs have C-terminal GATA domains that are similar to each other and to the GATA factors of arthropod and vertebrate species

The CF GATA domains of ELT-1 and ELT-2 orthologs share long stretches of common residues (Fig. 1.3A; Supp. Fig. 1.4E), and this pattern holds across the nematodes (Eurmsirilerd & Maduro 2020). Furthermore, the sequences of their CFs are more similar to those of arthropod and vertebrate GATA CFs than to the single ZnF domains of other *Caenorhabditis* GATA-motif-containing proteins (Fig. 1.3C; data not shown). The similarities between ELT-1 and ELT-2 CFs are captured by our pHMMs in which each of these proteins' pHMMs scores the other ortholog's CF sequences second highest after that of their own ortholog's CFs (Fig. 1.3B). Moreover, the ELT-1 and ELT-2 CFs group adjacently in our GATA domain tree (Supp. Fig. 1.6). The highly conserved PVCNACGLY[FY]KLH sequence, located at positions 20-25 of the CF and followed by the first seven residues of the BR (Fig. 1.3A), illustrates the similarities between the ELT-1 and ELT-2 CF domains. Structures of vertebrate GATA factors showed that this sequence encodes for the second anti-parallel beta-sheet and all but the last residue of the alpha helix of some canonical GATA factors (Omichinski et al. 1993; Clore et al. 1997; Bates et al. 2008; Chen et al. 2012; Wilkinson-White et al. 2012). This sequence is found in all singleton ELT-1 and ELT-2 orthologs in this study, with the single exception of the *C. monodelphis* ELT-2, which has a threonine (T) instead of the proline (P) at the start (Supp. Fig. 1.4E). This particular sequence is also highly conserved in the CFs of most arthropod and vertebrate GATA1/2/3 factors and, with only one residue different, in the CFs of vertebrate GATA4/5/6 factors, which have a methionine (M) instead of a phenylalanine (F) or tyrosine (Y) (Teakle & Gilmartin 1998). The CFs of most (107 of 109) singleton ELT-1 and ELT-2 orthologs contain another highly conserved sequence, TTLWRRN, in positions 11-17 (Fig. 1.3A, which is also highly conserved in arthropod and vertebrate GATA factors CFs (Teakle & Gilmartin 1998) and in ELT-3 orthologs (Fig. 1.3A). The TTLWRRN sequence is also present in two non-sister species END-1s, but this is more likely due to convergent evolution since most END-1s have the sequence TTLWRRT at this location (Fig. 1.3A; Supp. Fig. 1.4E). No other *Caenorhabditis* GATA-domain-containing proteins have TTLWRRN in their ZnF sequences (Supp. Fig. 1.4; Supp. Table 1). The mostly invariant sequence conservation between the CF domains of ELT-1 and ELT-2 orthologs is evidence of their more recent shared ancestry and suggests strong functional constraint of these domains through negative selection.

We also compared the more probable residue at each position in both the ELT-1 and ELT-2 CF pHMMs to the residues known to be important for the structure or DNA binding in animal GATA factors (Fig. 1.3C; see Methods; Omichinski et al. 1993; Bates et al. 2008; Lowry et al. 2009; Chen et al. 2012). The ELT-1 CF pHMM contains 22 of the 24 residues important for DNA interactions plus one similar residue (Fig. 1.3C). Additionally, this pHMM contains 13 of the 18 structurally important residues plus two similar residues (Fig. 1.3C; see Methods). The ELT-2 pHMM has 16 identical and three similar residues to the 24 residues important for DNA binding and 11 identical and two similar residues to the 18 structurally important residues (Fig. 1.3C). Of the twelve ortholog groups in our phylogeny, the ELT-1 CF pHMM contains the most, while the ELT-2 CF pHMM is tied with the ELT-7 pHMM for the fifth most, of these highly conserved structural and/or DNA-binding residues, suggesting that ELT-1 orthologs have likely experienced stronger negative selection pressures since the divergence of ELT-2 and ELT-1 paralogs.

The N-terminal zinc fingers of ELT-1 and ELT-2 orthologs have diverged

The NFs of *Caenorhabditis* ELT-2 orthologs branch off from a very long internal branch within the ELT-1 NF clade in our GATA domain phylogeny (Supp. Fig. 1.6). This topology is similar to their nematode-wide pattern (Eurmsirilerd & Maduro 2020), suggesting that the NFs of these two ortholog groups are on different evolutionary trajectories. In fact, the ELT-2 NF is highly diverged from the canonical animal GATA factor NF sequence, containing both fewer conserved residues and a variable ZnF loop length (Fig. 1.3E; Supp. Fig. 1.4E). The ELT-1 NF pHMM contains 13 of the 24 residues important for CF DNA interactions, 11 of 11 residues found involved in NF DNA binding (two of which are not included in the 13 CF residues), and eight of the 18 structurally important CF residues (Fig. 1.3C,D). The ELT-2 pHMM contains only one of the 24 residues important for CF DNA binding, zero of 11 residues found involved in NF DNA binding, and only four of the 18 structurally important residues (Fig. 1.3C,E). Because it cannot independently bind DNA, the NF of the *C. elegans* ELT-2 ortholog has been suggested to be non-functional (Hawkins and McGhee 1995) and degenerate due to its lack of many of the conserved residues in canonical GATA factors and the overall lack of conservation of this domain (Eurmsirilerd & Maduro 2020). However, some canonical and highly conserved vertebrate NFs do not bind to DNA independently (Martin & Orkin 1990; Yang & Evans 1992), and some NFs are known to participate in protein-protein interactions (Tsang et al. 1997; Lu et al. 1999; Ono et al. 1998). Therefore, even though ELT-2 NFs have likely experienced relaxed selection, they may still serve a function in these proteins, especially since all but one of the *Caenorhabditis* species included in our study (the basal species, *C. plicata*) contain a singleton ELT-2 ortholog with an NF with sequence $CX_2CX_{15-17}CX_2C$. Although most residues are non-canonical and not that well conserved, a few canonical GATA factor residues are conserved. These include a glycine (G) four residues upstream of the third ZnF-coordinating cysteine (found in 52 of 57 orthologs) (Fig. 1.3E; Supp. Fig. 1.4E), which is highly conserved in non-*med Caenorhabditis* orthologs (Fig. 1.3A; Supp. Fig. 1.4D-F) and canonical GATA factors (Teakle & Gilmartin 1998), and an acidic lysine (K) or arginine (R) at NF position 13 (found in 55 of 57 orthologs) (Fig. 1.3E; Supp. Fig. 1.4E), which is a structurally important arginine in the canonical NF sequence that also directly interacts with bound DNA (Omichinski et al. 1993). Both the NFs of ELT-1 orthologs and their downstream basic regions are highly conserved when compared to canonical NFs of vertebrates (Fig. 1.3D; Teakle & Gilmartin 1998; Lowry & Atchley 2000). In summary, the NFs of ELT-2 orthologs have diverged extensively from their ELT-1 counterparts and from each other, but why they have not been lost from more than a single species remains a mystery. Both the *C.*

elegans ELT-1 and ELT-2 orthologs bind to canonical WGATAR DNA sites (Shim et al. 1995; Araya et al. 2014; McGhee et al. 2009, 2007; Du et al. 2016; Wiesenfahrt et al. 2016). ELT-1 also binds non-canonical GATC DNA sites (Shim et al. 1995) and GATR followed by AGAT, 3 bps apart on the opposite strand (Araya et al. 2014) while ELT-2 has only been found to bind to single WGATAA sites (McGhee et al. 2009, 2007; Du et al. 2016; Wiesenfahrt et al. 2016; Lancaster & McGhee 2020). This difference in binding preference between ELT-1 and ELT-2 orthologs could be due to the ELT-1 NF if it binds DNA and expands the ELT-1 binding repertoire.

elt-1 *and* elt-2 *orthologs were not found on the same scaffold and are likely on different chromosomes in most Caenorhabditis species*

We never found *elt-1* and *elt-2* orthologs on the same piece of assembled genomic DNA (Supp. Fig. 1.7G,H). In each of the few species in this study for which there was chromosome level genome resolution available, the *elt-1* and *elt-2* orthologs were found on different chromosomes (4 and X, respectively). Through our synteny analysis (Fig. 1.4, see Methods) we predict that, with few exceptions, *elt*-2 orthologs lie on the X chromosome throughout the clade and *elt*-1 orthologs are found on chromosome 4.

elt3 clade

An elt-3 *duplication likely produced the ancestor of* end-1/end-3/elt-7 *in the ancestor of the* Elegans *supergroup*

The ELT-7, END-1, END-3, and ELT-3 ortholog groups cluster together in a well-supported (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) monophyletic clade (Fig. 1.1A), suggesting that these proteins share a more recent common ancestor with each other than with any of the other GATA-motif-containing proteins included in this study. ELT-7, END-1, and END-3 orthologs are only found in *Elegans* supergroup species, whereas ELT-3 orthologs are found throughout the *Caenorhabditis* genus as well as in the *Diploscapter* species included in this study (Fig. 1.1A). This is consistent with previous studies with fewer *Caenorhabditis* species and other nematodes (Maduro 2020; Eurmsirilerd & Maduro 2020). The topology of this clade supports a more recent common ancestor between END-3 and END-1 orthologs, which is consistent with previous results (Maduro et al. 2005a; Gillis et al. 2008; Boeck et al. 2011; Maduro 2020), but the evolutionary relationship between ELT-3, ELT-7, and the END ancestor is less clear (Fig. 1.1A). The ELT-7 and ELT-3 ortholog group ancestral nodes are 1.58 and 0.74 substitutions per site away from the END clade ancestral node respectively (Fig. 1.1). This topology suggests that an *elt-3* duplication produced the *elt-7* ancestor, and a second *elt-3* duplication produced the *end* gene ancestor. However, if we consider the distance to the ancestral node of most singleton and representative paralog (Supp. Fig. 1.10; see Methods) ELT-3s, or ELT-7s, or the *Elegans* supergroup ELT-3 ancestral node the distance is 1.88, 1.94, or 2.63 substitutions per site respectively to the END clade ancestral node (Fig. 1.1). This comparison makes it unclear whether the *end* genes arose from an ancestral *elt-3* or an ancestral *elt-7* gene. Phylogenetic analysis may not be able to distinguish between these two scenarios because the duplications that produced the ancestral *elt-7* and *end* genes happened in a short period of evolutionary time (Fig. 1.1A). Therefore, we propose that one or two *elt-3* duplications in the ancestor of the *Elegans* supergroup produced the ancestral gene(s) that evolved into *end-1*, *end-3*, and *elt-7*, and we refer to the clade as the elt3 clade.

Evidence for a partial elt-3 *duplication or sequence loss in the* Elegans *supergroup* end-1*/*end-3*/*elt-7 *ancestor*

We compared the gene structures of extant *end-1*, *end-3*, *elt-7*, and *elt-3* orthologs respectively (Supp. Fig. 1.3J-M) and predicted their *Elegans* supergroup ancestral gene structures (Fig. 1.2F). The gene structures of elt3 clade orthologs contain similar features (Supp. Fig. 1.3J-M; Fig. 1.2) that suggest two alternative hypotheses for how their gene structures evolved (Supp. Fig. 1.11A,B). Most singleton and representative paralog (Supp. Fig. 1.10; see Methods) elt3 clade homologs have an intron located at the same position in their ZnF coding sequence, which is also found in most *elt-2* and *Japonica* group *med* orthologs (Supp. Fig. 1.3E,I-M; Fig. 1.2C,E,F; see above; Maduro 2020; Eurmsirilerd & Maduro 2020). This conserved intron supports the monophyly of the elt3 clade (Fig. 1.1A). A single *end-3* representative paralog in *C*. sp. *51* has likely lost this intron (Supp. Fig. 1.3M). Most of the singleton *end-1*, *end-3*, and *elt-7* orthologs have four exons of which the last two contain the code for their single GATA ZnF (Supp. Fig. 1.3K-M; Fig. 1.2F). Most confident singleton *elt-3* orthologs also code for their single GATA ZnF in the last two exons, but the entire gene is coded by eight exons (Supp. Fig. 1.3J; Fig. 1.2F). The median protein length of ELT-3 orthologs used in this study (322 residues) is substantially longer than the median lengths of ELT-7, END-1, and END-3 proteins, which are more similar to each other (202, 226, and 240 residues, respectively; Supp. Fig. 1.3J-M). These data suggest that if an *elt-3* duplication produced the ancestral *end-1/end-3/elt-7* gene, it was either (a) a partial duplication of *elt-3* that possibly involved additional sequence loss as well or (b) a full gene duplication followed by subsequent sequence loss. This evidence would appear to favor an already shortened ancestral *elt-7* duplication producing the ancestral *end* gene, instead of two shortening events of an ancestral *elt-3* duplication. However, ELT-7 and the ENDs do not share other obvious conservation of sequence (Supp. Fig. 1.4F) leaving the origin of the *end* genes and their lineal relationship with *elt-7* ambiguous (Supp. Fig. 1.11A,B).

elt3 clade GATA DBDs are more similar to each other than to those in the other GATA-domain-containing proteins

Representations of the consensus GATA DBDs for each elt3 clade ortholog group, created from pHMMs (Eddy 2020) of each, are shown at the bottom of Figure 1.3A. Each of the elt3 clade pHMMs generally score the other clade proteins higher than other GATA proteins (Fig. 1.3B). Conserved residues within the elt3 clade GATA DBDs further support common ancestry among the ELT-7, END-1, END-3, and ELT-3 orthologs. For example, there is a glutamic acid (E) at position 17 in the ZnF loops of all confident representative (Supp. Fig. 1.10; see Methods) ELT-7, END-1, END-3, and ELT-3 homologs, and this residue is not conserved in any other canonical GATA factor groups (Fig. 1.3A; Supp. Fig. 1.4D-F). (This residue is conserved in all singleton RCOR-1 ZnFs and in all but one singleton SPR-1 ZnFs, even though rcor1 clade ZnFs are quite divergent from canonical GATA factor ZnF motifs overall (Fig. 1.3A; Supp. Fig. 1.4A), suggesting convergent evolution). An NMR structure of the DBD of cGATA-1 bound to DNA showed that the valine (V) at position 17 in the cGATA-1 ZnF loop is important for the structural integrity of the ZnF motif (Omichinski et al. 1993), and most of the canonical *Caenorhabditis* GATA factors in our study also contain a valine at this position (Fig. 1.3A; Supp. Fig. 1.4D,E). Another residue uniquely found in all representative ELT-7 and ELT-3 homologs, and all but one END-3 ortholog, is an asparagine (N) serving as the first residue of the basic region (Fig. 1.3A; Supp. Fig. 1.4F). This asparagine with its larger polar side chain is quite different from the small flexible non-polar glycine (G) residue conserved at this position in most canonical GATA factors (Fig. 1.3A; Supp. Fig. 1.4D,E). Most END-1 orthologs have a serine (S), which is polar and smaller than asparagine but larger than glycine, at this position (Fig. 1.3A; Supp. Fig. 1.4F). Two additional, adjacent residues are conserved in all representative ELT-3 homolog ZnFs and in most representative END-3 and END-1 homolog ZnFs: an arginine (R) followed by a glutamic acid (E) at positions three and four in the ZnF loop (Fig. 1.3A; Supp. Fig. 1.4F). Arginine is also found in the same position in some rcor1 clade orthologs, and glutamic acid is also found in the same position in many MED orthologs and in some EGL-27 orthologs (Fig. 1.3A; Supp. Fig. 1.4A-C). However, this combination of a negatively charged residue adjacent to a positively charged residue in these positions is not found in the ZnF motifs of any other proteins included in this study (Fig. 1.3A; Supp. Fig. 1.4), nor is it typical for other GATA factors (Teakle & Gilmartin 1998). These non-canonical residues that are conserved among the elt3 clade GATA DBDs support the monophyly of this clade.

ELT-3 and ELT-7 orthologs have more canonical animal GATA factor residues in their GATA DBDs than END-1 and END-3 orthologs

Figure 1.3C shows that the ELT-3 pHMM has 16 residues identical to and three residues similar to the 24 residues important for DNA interactions and 12 of the 18 structurally important residues plus two similar residues for animal GATA factors (Omichinski et al. 1993; Bates et al. 2008; Lowry et al. 2009; Chen et al. 2012; see Methods). The ELT-7 pHMM has 14 of the 24 residues important for DNA interactions plus four similar residues and 13 residues identical to the 18 structurally important residues (Fig. 1.3C). The END-1 pHMM has 13 residues identical to and three residues similar to the 24 DNA-interacting residues and 10 of the 18 structurally important residues plus two similar residues (Fig. 1.3C). The END-3 pHMM has 11 of the 24 DNA-interacting residues plus three similar residues and 11 residues identical to and one residue similar to the 18 structurally important residues (Fig. 1.3C). Overall, the ELT-3 and ELT-7 pHMMs have about the same number of conserved important residues (Fig. 1.3C), which may contribute to the similar binding preference of the *C. elegans* orthologs for TGATAA DNA sites (Gerstein et al. 2010; Narasimhan et al. 2015). The END-1 and END-3 pHMMs also have a similar number of conserved functional residues. However, they have fewer than the ELT-3 and ELT-7 pHMMs (Fig. 1.3C), which may be why *C. elegans* END-1 and END-3 have a similar binding preference of GATA DNA sites, but no specificity for flanking sequences (Narasimhan et al. 2015). The conservation of residues important for DBD binding and structure suggests that END-1 and END-3 orthologs have lost some of the residues that we presume they initially acquired from an ancestral *elt-3* or *elt-7* duplication. If this is the case, the *end* ancestor probably experienced weaker selection compared to its paralog and/or was selected for a broader range of binding.

end-1, end-3*, and* elt-7 *orthologs are often found on the same scaffold whereas* elt-3 *orthologs are likely on a different chromosome in most Caenorhabditis species*

We found that *elt-7*, *end-1*, and *end-3* orthologs were syntenic in eight species and we always found *end-1* and *end-3* orthologs closer to each other than to *elt-7* orthologs (Supp. Fig. 1.7J-L). *elt-7* orthologs range from 1.1 to 9.2 Mb away from *end-1* orthologs, with a median of 2.0 Mb (Supp. Fig. 1.7J). Due to the large distance between *elt-7* and *end* orthologs and the short length of some species scaffolds/contigs we except to find *elt-7* orthologs syntenic with more *end* orthologs once more *Caenorhabditis* genomes are fully assembled. We found that *end-1* and *end-3* were syntenic in 23 species, ranging from 19.8 to 528 kb apart with a median of 35.3 kb (Supp. Fig. 1.7K,L). Our results are consistent with a previous study that found microsynteny between *end-1* and *end-3* orthologs in 15 *Elegans* supergroup species, ranging from 20 to 528 kb with an average distance of about 37 kb between them (Maduro et al. 2005a; Maduro 2020), and we extend this finding to eight additional species. Moreover, consistent with Maduro (2020) we find that the orientation of syntenic *end-1* and *end-3* orthologs is variable, suggesting that intrachromosomal rearrangements are common, like previously identified in *C. elegans* (Coghlan & Wolfe 2002).

Even though we predict that an *elt-3* paralog in the *Elegans* supergroup ancestor led to the origin of *elt-7*, *end-1*, and *end-3*, we found no instances in which an *elt-3* ortholog occurred on the same piece of assembled genomic DNA as an *elt-7*, *end-1*, or *end-3* ortholog (Supp. Fig. 1.7I). In fact, in the six species with chromosome-level assemblies, *elt-3* orthologs are found on the X chromosome, and this location was consistent in our synteny analysis (Fig. 1.4). This suggests that an *elt-3* paralog moved to chromosome 5 during or after the duplication of *elt-3* and before the radiation of the *Elegans* supergroup. Once fully assembled genomes of all the *Caenorhabditis* species in our study are available, we can more precisely compare the chromosomal locations of elt3 clade orthologs and make more informed predictions about how these genes evolved.

Relaxed selection on one paralog relative to the other

To test whether the intensity of selection changed after the duplications in the elt3 clade, we used the RELAX hypotheses testing framework (Wertheim et al. 2015). RELAX compares two sets of branches in a tree and evaluates whether the data is better fit by a single distribution of a few dN/dS rate categories among all branches or by different distributions for each set where the rate categories in one are related to the rate categories in the other by an exponentiation factor (k). We performed four tests: (1) *Elegans* supergroup *elt-3s* vs. the *ends*; (2) *Elegans* supergroup *elt-3s* vs. *elt-7s;* (3) *end-1s* vs *end-3s*; and (4) non-*Elegans* supergroup *elt-3s* vs. *Elegans* supergroup *elt-3s.*

In the *Elegans* supergroup both the *elt-7s* (p<0.0001; k=0.76) and the *ends* (p<0.0001; k=0.73) have experienced less intense selection than the *elt-3*s (Supp. Fig. 1.12). In turn, selection intensity relaxed on the *end-3* ortholog group after duplication compared to the *end-1*s (p<0.00001; k=0.63), and this relaxation of selection intensity may be related to the greater variability in *end-3* homolog numbers compared to *end-1* (Supp. Table 2). However, the *Elegans* supergroup *elt-3*s have not experienced a change in selection intensity compared to the non-*Elegans* supergroup *elt-3*s (p=0.095; k=0.95), even though our data suggests that there was a functional change in the *Elegans* supergroup *elt-3*s. All of these patterns of selection intensity are concordant with the differences in branch lengths in these groups that are readily apparent in the tree (Fig. 1.1).

Re-wiring of the* Elegans *supergroup endoderm developmental gene regulatory network

Six zygotic transcription factors comprise the *C. elegans* endoderm dGRN, and they are all GATA factors. These GATA factors function, with varying levels of redundancy, in a feed-forward cascade to specify endoderm cells and to activate *elt-7* and *elt-2* expression. These last two GATA factors guide differentiation of the endoderm cells into the final twenty intestinal cells present in adult worms and the maintenance of intestinal function throughout a worm's life (reviewed in Maduro 2015). Interestingly, if *C. elegans elt-2* is overexpressed under the *end-1* promoter it can rescue animals null for *end-3*, *end-1*, and *elt-7*, suggesting that *elt-2* can specify endoderm despite its limited sequence homology outside of the DBD to these other

GATA factors (Wiesenfahrt et al. 2016). In *C. elegans, end-3, end-1, elt-7,* and *elt-2* are exclusively expressed in the endoderm (Zhu et al. 1997; Fukushige et al. 1998; Maduro & Rothman 2002; Robertson et al. 2004; Maduro et al. 2005a; McGhee et al. 2007; Sommermann et al. 2010; Raj et al. 2010) while *med-1* and *med-2* are expressed in the EMS cells which give rise to both endoderm and mesoderm (Maduro et al. 2001). Neither our searches nor those conducted by other researchers (Maduro 2020; Eurmsirilerd & Maduro 2020) identified any *med-1*, *med-2*, *end-3*, *end-1*, or *elt-7* orthologs in any *Caenorhabditis* species outside of the *Elegans* supergroup, suggesting that this GATA factor radiation and fixation happened during a short period of evolutionary time in the ancestor of this supergroup. A corollary of this finding is that of all the GATA factors comprising the *C. elegans* endoderm dGRN, only orthologs of *elt-2* are found among non-*Elegans* supergroup species (Fig. 1.1A). This begs the question: could *elt-2* constitute the entire endoderm dGRN on its own in non-*Elegans* supergroup species, or are other transcription factors involved? Although *elt-3* is expressed specifically in hypoderm in *C. elegans* (Gilleard et al. 1999; Gilleard & McGhee 2001), our phylogenetic analysis placed *elt-3* orthologs as the closest relatives of *end-3*, *end-1*, and *elt-7* (Fig. 1.1A) and so we hypothesized that *elt-3* might play a role in endoderm specification outside the *Elegans* supergroup. To test this hypothesis, we used single molecule fluorescence *in situ* hybridization (smFISH; Raj et el. 2008) to visualize *elt-3* and *elt-2* mRNA expression in *C. angaria*, a non-*Elegans* supergroup species.

C. angaria elt-3 *mRNA is expressed in early endoderm cells and later in hypoderm cells*

We initially detected *elt-3* mRNA in the two endoderm (2E) cells of 14-cell *C. angaria* embryos (beginning of 2E cell stage) prior to gastrulation (Fig. 1.5B,C). Expression of *elt-3* in *C. angaria* was observed in all endoderm cells throughout early gastrulation, including the beginning of the 4E cell stage (when embryos are comprised of 45 to at least 58 cells, Fig. 1.5B,C). We did not detect *elt-3* expression in any endoderm cells after that point (Fig. 1.5B,C). This expression pattern resembles that of *C. elegans end-1* (Zhu et al. 1997; Raj et al. 2010) and is consistent with a role for ELT-3 in endoderm specification in *C. angaria*. Single-embryo RNA-sequencing found slightly earlier and higher levels of expression of *elt-3* in *C. angaria* than in *C. elegans* (Macchietto et al. 2017; Supp. Fig. 1.13A). This pattern of expression is consistent with our smFISH results Fig. 1.5B,C). *C. angaria elt-3* mRNA was also expressed in hypoderm cells by the bean stage of development (Fig. 1.5B,C) and continuing through later stages (e.g., comma stage, Fig. 1.5B,C). This expression pattern resembles *elt-3* in *C. elegans* (Gilleard et al. 1999; Supp. Fig. 1.13A) and suggests that the role ELT-3 plays in *Caenorhabditis* hypoderm development is likely conserved throughout the genus.

Conserved expression of the C. angaria elt-2 *ortholog during endoderm embryogenesis*

Our smFISH analysis detected *C. angaria elt-2* mRNA initially in the four endoderm (4E) cells of 45-cell embryos (Fig. 1.5A,C), like the expression pattern observed for *C. elegans elt*-2 (Fukushige et al. 1998; Raj et al. 2010; Macchietto et al. 2017; Supp. Fig. 1.13B). Throughout all later stages of embryo development examined, e.g., the comma and bean stages (Fig. 1.5A,C), *C. angaria elt-2* expression remained endoderm-specific. This expression pattern suggests that *C. angaria* ELT-2 functions in endoderm differentiation like it does in *C. elegans*, supporting our hypothesis that the function of ELT-2 orthologs has been conserved throughout the *Caenorhabditis* genus.

elt-2 *orthologs in non-*Elegans *supergroup species may be regulated by a GATA-binding transcription factor(s)*

To investigate how *elt-2* was regulated before the *Elegans* supergroup ancestral GATA factor expansion (see above), we searched for conserved transcription factor binding sites in the *elt-2* promoters of non-*Elegans* supergroup species (see Methods). We found more canonical GATA factor binding sites (HGATAR; Ravagnani et al. 1997) in the *elt-2* orthologs than expected by chance (Fig. 1.6B), suggesting that these orthologs may autoregulate themselves, as the *C. elegans* ELT-2 ortholog does (Fukushige et al. 1999; Wiesenfahrt et al. 2016), or that these sites could be bound by some other HGATAR-binding transcription factor. Our finding that *C. angaria elt-3* is expressed in endoderm cells prior to initiation of *elt-2* expression (Fig. 1.5) suggests that ELT-3 may bind to *elt-2* HGATAR sites in non-*Elegans* supergroup species.

The organization of the elt-2 *promoter differs markedly between* Caenorhabditis *subclades and suggests coordinated evolution of* cis- *and* trans-*regulation in the endoderm*

We also examined *Elegans* supergroup *elt-2* promoters and found a striking conservation of HGATAR sites (Fig. 1.6B). There are six highly conserved HGATAR sites in most *Elegans* supergroup *elt-2* promoters (Fig. 1.6B). Three of these sites are TGATAA in all *Elegans* supergroup species, the only exception is in *C. elegans* which does not have a HGATAR site that aligns with the most 3' of these sites (Fig. 1.6B). TGATAA sites are important for *C. elegans elt-2* expression (Wiesenfahrt et al. 2016; Du et al. 2016). They are the most overrepresented DNA site in *C. elegans elt-2* target genes (McGhee et al. 2007, 2009), and TGATAA sites have been found to be the preferred DNA-binding site of *C. elegans* ELT-7, ELT-6, and ELT-3 GATA factors (Narasimham et al. 2015). Moreover, *C. elegans* ELT-2, ELT-7. END-3, and END-1 bind to TGATAA sites *in vitro* (Wiesenfahrt et al. 2016; Du et al. 2016). Two of the other conserved HGATAR sites are AGATAG and CGATAA, which are found in all *Elegans* supergroup *elt-2* promoters (Fig. 1.6B). The sixth HGATAR site is the least well conserved; most (23 of 35) species having CGATAG, but four species have AGATAA, three species have TGATAG, two species have AGATAG, and three species do not have a conserved HGATAR site that aligns at this position (Fig. 1.6B). Some of these HGATAR sites even align in a some non-*Elegans* supergroup species; however, no non-*Elegans* supergroup species has more than one of these six sites (Fig. 1.6B). Overall, HGATAR sites are less abundant and less spatially conserved in the promoters of *elt-2* orthologs in non-*Elegans* supergroup species as compared to *elt-2* promoters in *Elegans* supergroup species (Fig. 1.6B). The organization of the *Elegans* supergroup *elt-2* promoter evolved in parallel with the expansion of GATA factors involved in the endoderm dGRN (see above) and has remained highly conserved since.

*non-*Elegans *supergroup* elt-3 *orthologs may be regulated by a Sp1 family transcription factor,* SPTF-3

To look for clues as to how *elt-3* was regulated before the expansion of GATA factors in the ancestor of the *Elegans* supergroup (see above), we searched for conserved transcription factor binding sites in the promoters of *elt-3* orthologs from non-*Elegans* supergroup species (see Methods). We found significant numbers of Sp1-like binding sites (CYCCRCCY; Saito et al. 2013) and/or SPTF-3 (a *C. elegans* Sp1 family homolog; Ulm et al. 2011) binding sites (MCGCCCMY; Narasimham et al. 2015) in 13 of 18 non-*Elegans* supergroup and non-*Guadeloupensis* group species *elt-3* promoters whereas Sp1-like sites are only significant in two and zero *Elegans* supergroup and *Guadeloupensis* group species *elt-3* promoters, respectivly (Fig. 1.6A; data not shown). Sp1-like sites are found in many *C. elegans* promoters (Grishkevich et al. 2011) and knock-down of expression of *C. elegans sptf-3* via RNA interference (RNAi) results in fewer embryos with correctly specified endoderms and reduced *end-3* and *end-1* reporter expression (Sullivan-Brown et al. 2016). This suggests that SPTF-3 positively regulates *C. elegans* endoderm development, likely at least somewhat through *end-3* and *end-1* activation. Sp1-like binding sites were also found previously in the promoters of most *med, end-1*, and *end-3* orthologs (Maduro 2020), suggesting that *end-3*, *end-1*, *med*, and non-*Elegans* supergroup *elt-3* orthologs may be regulated similarly. If that is the case, these Sp1-like binding-site sequences in the promoters may have duplicated along with the GATA factor coding sequences in the multiple gene duplications we think occurred during the evolution of this gene family (see the elt3 clade and MED ortholog group sections above and the Discussion). This would support the hypothesis that a full-length *elt-3* duplication occurred in the ancestor of the *Elegans* supergroup.

Angaria *group* elt-3 *orthologs may be regulated by SKN-1 orthologs*

Other than Sp1-like binding sites, we did not identify any additional strongly conserved transcription factor binding sites in the promoters of all non-*Elegans* supergroup *elt-3* orthologs, but we did identify a conserved site in a non-*Elegans* supergroup subclade. We found an invariant motif, TACTATATATAGTGCATGCGCAA, in all seven promoters of *Angaria* group *elt-3* orthologs (Fig. 1.6A). We then searched the JAPSPAR 2018 core non-redundant database (jaspar.genereg.net) for similar motifs. *Arabidopsis thaliana* FUS3, a B3 DBD protein was the top hit, presumably because it binds to GCATGC; however, B3 DBDs are known to be plant-specific (Yang et al. 2021). The next best match to this invariant *Angaria* group *elt-3* motif was the *Homo sapiens* Nrf1 site: GCGCNTGCGC (jaspar.genereg.net). A BLASTp search (e-value cutoff of 0.01) did not reveal any highly conserved Nrf1 orthologs in any of the *Caenorhabditis* species included in this study (data not shown). However, Nrf1 contains a basic leucine zipper (bZIP) DBD, and the *C. elegans* transcription factor SKN-1 also contains the basic region of a bZIP domain. Moreover, the invariant *Angaria* motif starts with a TATA-rich region, and the *C. elegans* SKN-1 DBD also contains part of a homeo-like domain which binds T/A-rich sequences (Blackwell et al. 1994; Carroll et al. 1997; Pal et al. 1997; Lo et al. 1998). Even though the *C. elegans* SKN-1 bZIP-like domain binds RTCAT sequences with high affinity (1nM; Blackwell et al. 1994), and this exact sequence is not found in the invariant *Angaria* group motif, the specificity of *C. elegans* SKN-1 may be diverged from other *Caenorhabditis* species or this could be a secondary binding site for SKN-1 orthologs. The SKN-1 orthologs in *C. elegans and C. briggsae* contribute extensively to initiating the endoderm dGRN, primarily by activating *med-1* and *med-2* and possibly by directly activating *end-3* (Lin et al. 2009; Bowerman et al. 1992; Maduro et al. 2001; Maduro et al. 2007; Maduro et al. 2005b). Additionally, SKN-1 binding sites are enriched in the promoters of most *med* and many *end* orthologs (Zhu et al. 1997; Maduro 2020). Given the resemblance of the invariant motif in *Angaria* group *elt-3* promoters to a possible SKN-1 binding site and the involvement of SKN-1 in the endoderm dGRN—at least in *Elegans* supergroup species, we hypothesize that *Angaria* group SKN-1 orthologs bind to this invariant sequence to activate *elt-3* expression in early endoderm cells. If true and if ELT-3 is indeed part of the endoderm dGRN in non-*Elegans* supergroup species, then regulation involving SKN-1 would be conserved in the initial stages of endoderm specification, despite a change in the SKN-1 binding site.

Conservation of TGATAA DNA sites in the promoters of intestine-specific/enriched Caenorhabditis *orthologs*

McGhee and colleagues (2007, 2009) identified 197 intestine-specific or enriched genes in *C. elegans* by comparing expression of genes in the intestine to total somatic gene expression at different developmental stages. The putative promoters of these genes are enriched with TGATAA sites (McGhee et al. 2007, 2009). This is the preferred binding site of many *C. elegans* GATA factors (Narasimhan et al. 2015) including, most importantly, ELT-2 and ELT-7 which are expressed in the developing and adult intestine (Sommermann et al. 2010; Hawkins & McGhee 1995; Wiesenfahrt et al. 2016; Du et al. 2016). ELT-2 has been shown to regulate one of these intestine genes through two TGATAA sites *in vivo* (Lancaster & McGhee 2020). Using reciprocal BLASTp, we identified orthologs of the *C. elegans* intestine-specific or enriched genes in the 57 other sequenced *Caenorhabditis* species (see Methods). We found three or more TGATAA sites in many of these putative promoters (Fig. 1.7A), whereas we mostly find fewer than three sites in the promoters of orthologs of muscle (Fig. 1.7B), hypoderm, and neural genes (Supp. Fig. 1.14C,D). This supports the idea that GATA-factor (probably ELT-2) regulation of intestinal differentiation is conserved throughout the genus.

Figure 1.1. Inferred evolutionary history of *Caenorhabditis* GATA-domain-containing proteins. **(A)** Maximum likelihood phylogeny of 714 “confident” GATA-domain-containing proteins (Supp. Fig. 2; see Methods) in 58 *Caenorhabditis* and two outgroup nematode species. A GATA factor from the slime mold *Dictyostelium fasciculatum* was used to root the phylogenic tree (located between the ELT-1 and EGL-27 ortholog groups). The colors in the ring encircling the tree correspond to the species in which the protein was identified (the key to color-species correspondence is given in B below). The names of the 12 ortholog groups the 714 proteins were categorized into (see Results) are indicated in the lighter of the two outer gray rings (with white gaps between groups). Clades of multiple ortholog groups are highlighted by the darker gray outer ring (with white gaps between clades). The intensity of shading of each branch of the tree is indicative of its degree of bootstrap support, darker shading indicates stronger support (see Methods). The key for translating branch length into evolutionary distance (in units of amino acid substitutions per site) is shown to the right of the tree. **(B)** Phylogenetic relationships among the 60 species used in this study (based on Stevens 2020). Each species is designated by a different color shade; the same color-species designations are used in A above.

Figure 1

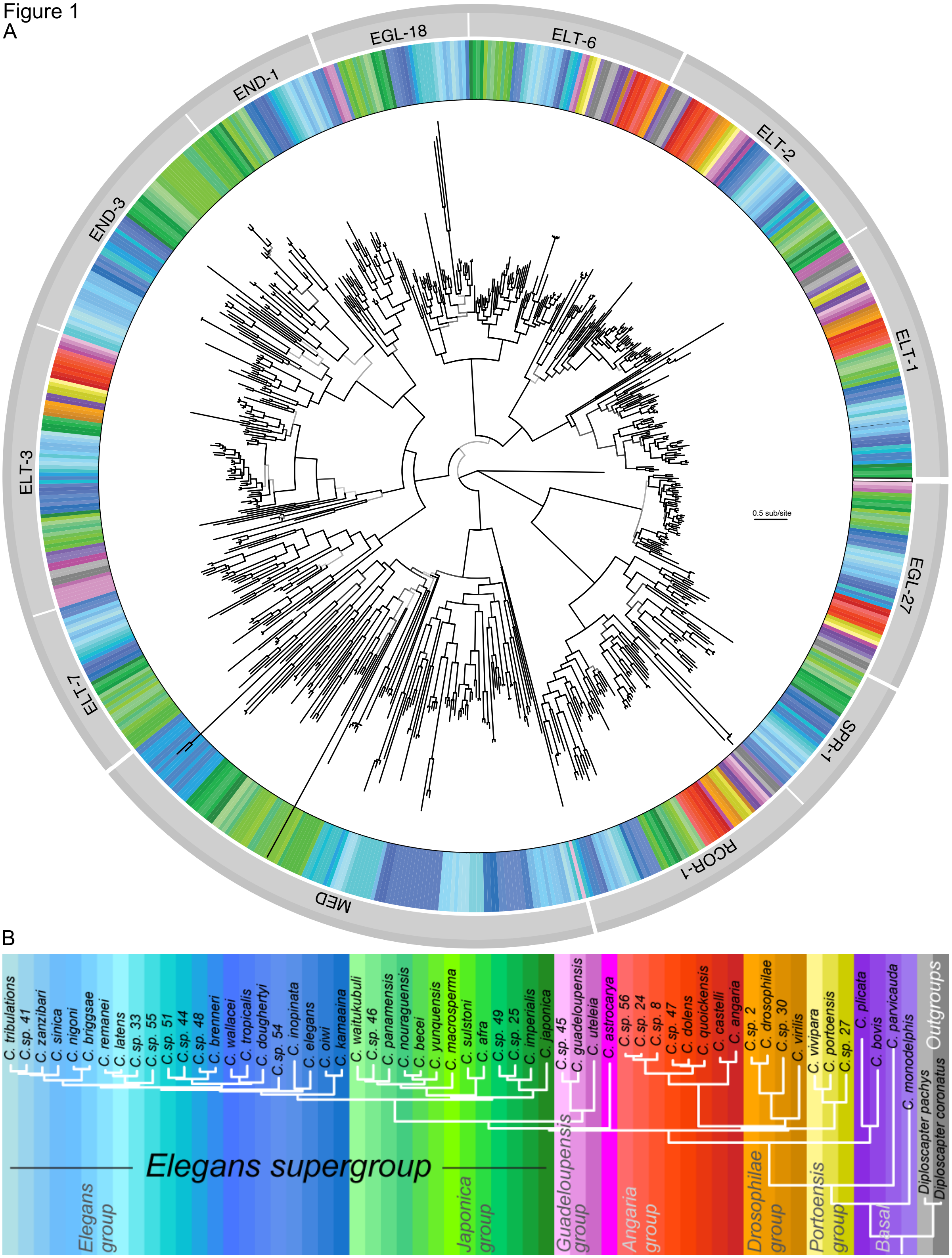

A
EGL-18
ELT-6
ELT-2
ELT-1
EGL-27
SPR-1
RCOR-1
MED
ELT-7
ELT-3
END-3
END-1
0.5 sub/site
B
C. tribulations
C. sp. 41
C. zanzibari
C. sinica
C. nigoni
C. briggsae
C. remanei
C. latens
C. sp. 33
C. sp. 55
C. sp. 51
C. sp. 44
C. sp. 48
C. brenneri
C. wallacei
C. tropicalis
C. doughertyi
C. sp. 54
C. inopinata
C. elegans
C. oiwi
C. kamaaina
C. waitukubuli
C. sp. 46
C. panamensis
C. nouraguensis
C. becei
C. yunquensis
C. macrosperma
C. sulstoni
C. afra
C. sp. 49
C. sp. 25
C. imperialis
C. japonica
C. sp. 45
C. guadeloupensis
C. uteleia
C. astrocarya
C. sp. 56
C. sp. 24
C. sp. 8
C. sp. 47
C. dolens
C. quoickensis
C. castelli
C. angaria
C. sp. 2
C. drosophilae
C. sp. 30
C. virilis
C. vivipara
C. portoensis
C. sp. 27
C. plicata
C. bovis
C. parvicauda
C. monodelphis
Diploscapter pachys
Diploscapter coronatus
Outgroups
Elegans supergroup
Elegans group
Japonica group
Guadeloupensis group
Angaria group
Drosophilae group
Portoensis group
Basal

Figure 2

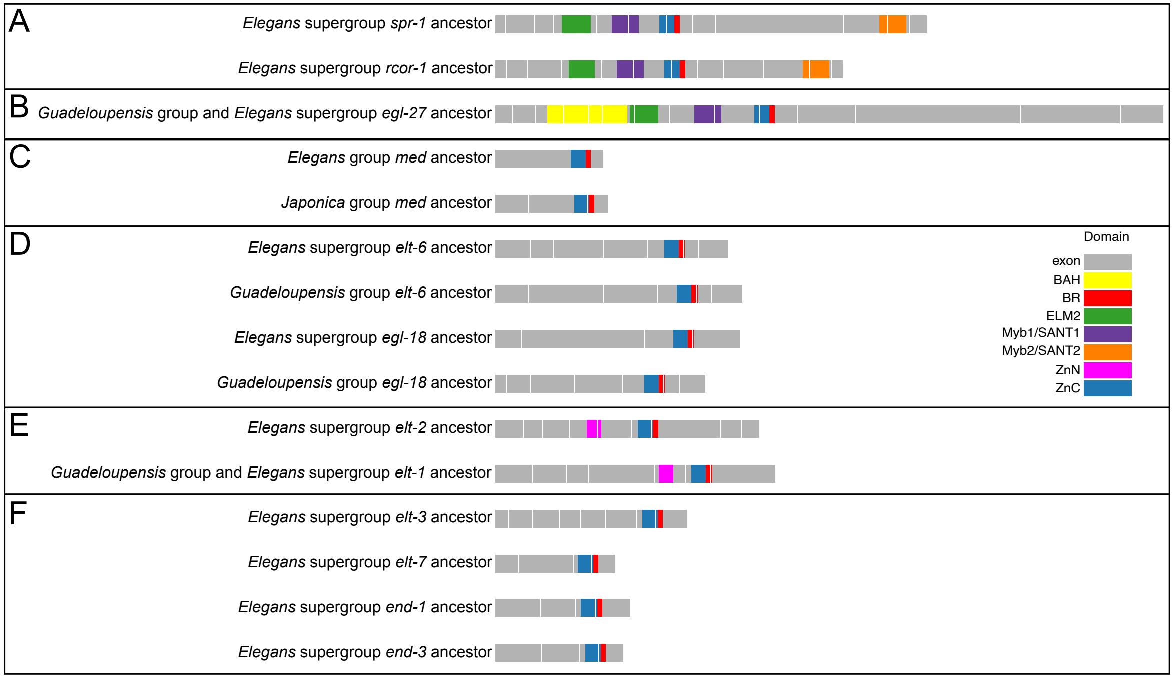
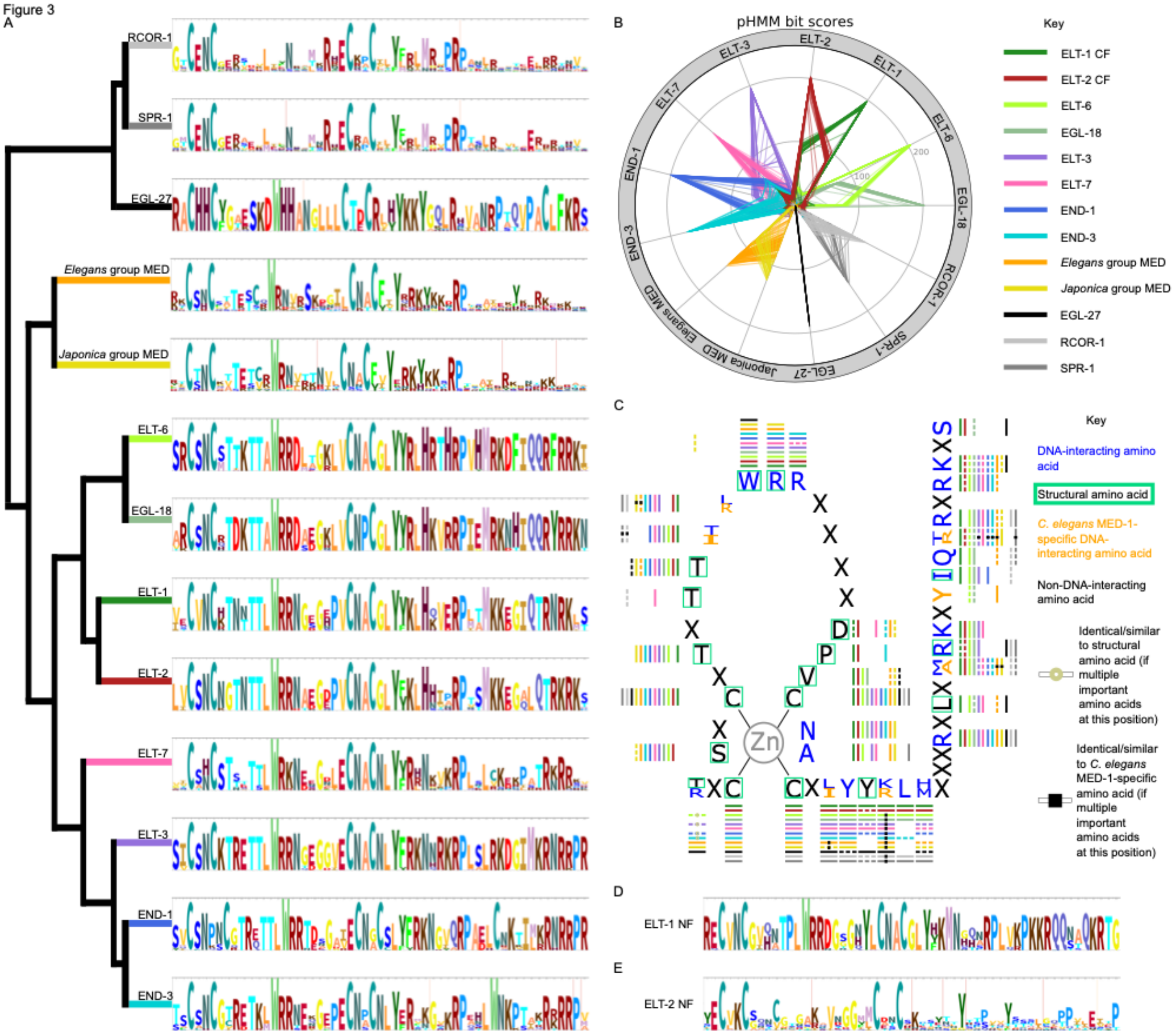

A
Elegans supergroup *spr-1* ancestor
Elegans supergroup *rcor-1* ancestor
B
Guadeloupensis group and *Elegans* supergroup *egl-27* ancestor
C
Elegans group *med* ancestor
Japonica group *med* ancestor
D
Elegans supergroup *elt-6* ancestor
Guadeloupensis group *elt-6* ancestor
Elegans supergroup *egl-18* ancestor
Guadeloupensis group *egl-18* ancestor
E
Elegans supergroup *elt-2* ancestor
Guadeloupensis group and *Elegans* supergroup *elt-1* ancestor
F
Elegans supergroup *elt-3* ancestor
Elegans supergroup *elt-7* ancestor
Elegans supergroup *end-1* ancestor
Elegans supergroup *end-3* ancestor
Domain
exon
BAH
BR
ELM2
Myb1/SANT1
Myb2/SANT2
ZnN
ZnC

Figure 1.2. Comparisons of predicted ancestral *Caenorhabditis* GATA-domain-containing gene structures. (A) *rcor-1* and *spr-1* predicted *Elegans* supergroup ancestral gene structures. **(B)** Predicted gene structure of the *Guadeloupensis* group and *Elegans* supergroup *egl-27* ancestor. **(C)** Predicted *Elegans* and *Japonica* group ancestral *med* gene structures, respectively. **(D)** *elt-6* and *egl-18* predicted *Elegans* supergroup ancestral gene structures, respectively and *elt-6* and *egl-18* predicted *Guadeloupensis* group ancestral gene structures, respectively. **(E)** Predicted *Elegans* supergroup ancestral *elt-2* gene structure and *Guadeloupensis* group and *Elegans* supergroup *elt-1* ancestral gene structure (which is also representative of the *elt-1 Elegans* supergroup ancestral gene structure). **(F)** *elt-3*, *elt-7*, *end-1*, and *end-3* predicted *Elegans* supergroup ancestral gene structures, respectively. The key to the color coding of the protein domains encoded in the gene structures is shown on the right: exons are shown in gray (with intron positions indicated by white vertical lines); BAH (bromo adjacent homology) domains are shown in yellow; the basic regions of GATA domains (BR) are in red; ELM2 (EGL-27 and MTA1 homology 2) domains are shown in green; the Myb/SANT (Swi3/Ada2/N-CoR/TFIIIB) domains nearer the 5’ end of a gene (Myb1/SANT1) are in purple; the Myb/SANT domains nearer the 3’ end of a gene (Myb/SANT2) are in orange; N-terminal GATA(-like) zinc fingers (ZnN) are in pink; and C-terminal GATA(-like) zinc fingers (ZnC) are in blue.

Figure 1.3. Comparisons of *Caenorhabditis* GATA(-like) DNA-binding domains. (A) Profile Hidden Markov model (pHMM) amino acid sequence logos, of the C-terminal zinc finger domain (CF/ZnF) for each ortholog group in the Figure 1A phylogeny (however, the MED ortholog group pHMM is divided into *Elegans* and *Japonica* group MED pHMMs because most of the ZnFs in these two groups have different lengths). The total height of the stack of amino acid(s) at each position represents the total information content at that position (Wheeler et al. 2014; see Methods). Amino acid(s) with above background frequency scores are shown as subdivisions of the total stack height depending on the probability of that amino acid at that position (Wheeler et al. 2014). The relationships among the ortholog groups in the phylogeny (Fig. 1A) are indicated through the cladogram on the left, the branches of which display the name of the particular ortholog group. **(B)** Radar plot depicting the pHMM bit scores for each of the GATA-domain-containing proteins scored on the 13 pHMMs shown in A. Key to the color-coding by ortholog group is given on the right side of the panel. The scale for the bit scores is depicted by the concentric circles on the figure (see the ELT-6 pHMM radius for numerical values). **(C)** Conservation of important animal GATA factor DNA-binding domain residues in *Caenorhabditis* GATA(-like) DNA-binding domains. Whether or not the residue with the highest probability in an ortholog groups pHMM (Fig. 3A) is the same or is similar (see Supp. Methods) to the indicated residues is denoted by a colored bar to the exterior of the figure. (These bars are color-coded by ortholog group as indicated in the key to the right of B.) Non-dashed colored bars mean that the residues with the highest probability at that position in that ortholog group's pHMM is identical to the shown residue. If the colored bar is dashed that means that the residues with the highest probability at that position in that ortholog groups pHMM has a similar residue (see Supp. Methods). Chicken GATA-1 (Omichinski et al. 1993), mouse GATA-3 (Bates et al. 2008), human GATA-3 (Chen et al. 2012), and/or *C. elegans* MED-1 (Lowry et al. 2009) residues that interact with DNA are shown in blue. Residues involved in the structural integrity of the chicken GATA-1 bound to DNA (Omichinski et al. 1993) are shown with an aquamarine box around them. *C. elegans* MED-1-specific DNA-interacting residues are shown in orange. Residues not found to interact with DNA in any of these animal GATA factor-DNA structures are shown in black. Some positions have different residues found to be important for the structural integrity and/or DNA-binding in different animal GATA factor-DNA structures (e.g., at the first position a threonine (T) was found to be structurally important in the cGATA-1-DNA structure while an arginine (R) was found to interact with DNA at this position in the mGATA-3-DNA structure). In this case an off-white ring in the middle of the ortholog group colored bar indicates that T has the highest probability of being at this position in this ortholog groups pHMM. Whereas a colored bar without an off-white ring at this position indicates that an R has the highest probability of being at this position in this ortholog groups pHMM. A similar classification, but with a black square, is used for positions where a different residue was found to interact with DNA in the *C. elegans* MED-1-DNA structure compared to at least one of the vertebrate GATA factor-DNA structures. **(D** and **E)** Amino acid sequence logos of the N-terminal zinc finger (NF) pHMMs, of ELT-1 (D) and ELT-2 (E). (Compare these logos based on NFs to those **in** A based on CF/ZnFs.)

Figure 3


A
RCOR-1
SPR-1
EGL-27
Elegans group MED
Japonica group MED
ELT-6
EGL-18
ELT-1
ELT-2
ELT-7
ELT-3
END-1
END-3
B
pHMM bit scores
ELT-2
ELT-1
ELT-6
EGL-18
RCOR-1
SPR-1
EGL-27
Japonica MED
Elegans MED
END-3
END-1
ELT-7
ELT-3
100
200
Key
ELT-1 CF
ELT-2 CF
ELT-6
EGL-18
ELT-3
ELT-7
END-1
END-3
Elegans group MED
Japonica group MED
EGL-27
RCOR-1
SPR-1
C
Zn
Key
DNA-interacting amino acid
Structural amino acid
C. elegans MED-1-specific DNA-interacting amino acid
Non-DNA-interacting amino acid
Identical/similar to structural amino acid (if multiple important amino acids at this position)
Identical/similar to C. elegans MED-1-specific amino acid (if multiple important amino acids at this position)
D
ELT-1 NF
E
ELT-2 NF

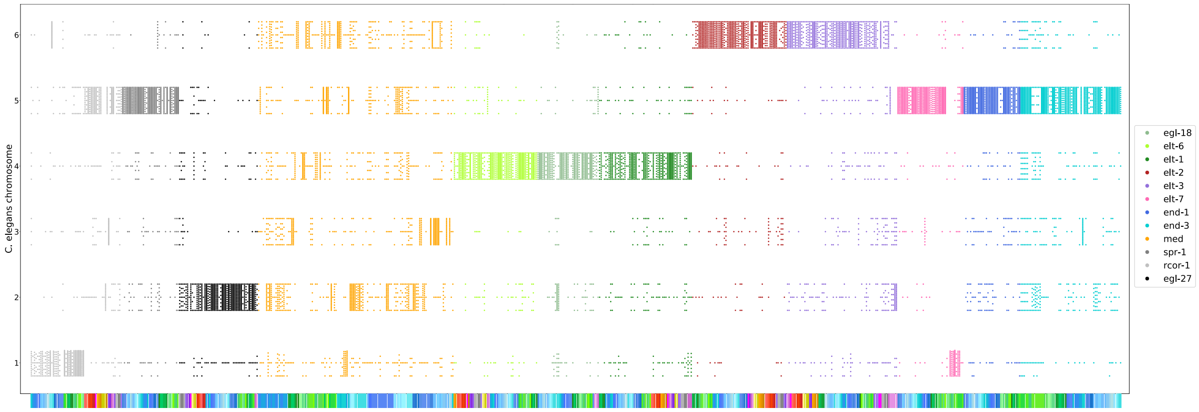


Figure 1.4. Chromosome assignment for GATA-domain-containing genes on scaffolds or contigs. To expand our analysis of the chromosome locations of GATA-domain containing proteins throughout the genus, we assigned scaffolds or contigs to chromosomes based on the *C. elegans* assembly. For each GATA-domain containing protein on a scaffold or contig in our dataset, we compiled a list of its neighbors, used BLASTp to find their closest homolog in *C. elegans*, and assigned the scaffold or contig to the most common chromosome among these homologs (see Methods). Each dot corresponds to a GATA-domain containing gene (on the x-axis) neighbor. The dot location on along the y-axis shows the chromosomal location of the *C. elegans* homolog of that neighbor. The color of the dot indicates the ortholog group of the GATA-domain containing gene. A key for these colors is shown on the right (which assigns the same ortholog group colors as those used in Figure 3). The genes are ordered by the Figure 1A phylogeny, and their species colors are shown below the plot (as in Figure 1B). The numbers on the left refer the numbers of the six *C. elegans* chromosomes; chromosome 6 refers to the *C. elegans* X chromosome.

Figure 5

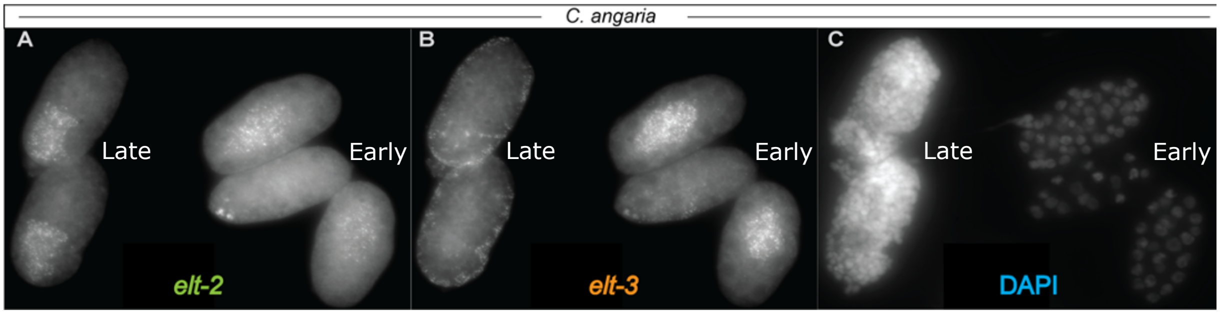


Figure 1.5. Expression of *elt-3* and *elt-2* mRNA in *C. angaria*, a non-*Elegans* supergroup species. Image of five embryos, each at a different developmental stage, illustrating the patterns of *elt-3* and *elt-2* mRNA expression observed in *C. angaria* using single molecule fluorescent *in situ* hybridization (smFISH, see Methods). The embryo depicted at the top left is at the comma stage (approximately) and contains more than 100 cells; the embryo at the bottom left is at the bean stage (approximately) and contains more than 100 cells; the embryo at the top right contains 54 cells; the embryo in the middle on the right contains 16 cells; and the embryo at the bottom right contains 25 cells. **(A)** Visualization of *C. angaria elt-2* mRNA after hybridization with a smFISH probe specific for *C. angaria elt-2*. **(B)** Visualization of *C. angaria elt-3* mRNA after hybridization with a smFISH probe specific for *C. angaria elt-3*. **(C)** DAPI-stained nuclei of *C. angaria* embryos (proxy for developmental stage).

Figure 1.6. Comparison of transcription factor binding sites in *Caenorhabditis elt-3* and *elt-2* promoters. Transcription factor binding sites of interest, including those found significantly more than expected by chance, are indicated in the predicted proximal promoters of the *elt-3* (A) and *elt-2* (B) orthologs from the *Caenorhabditis* species included in this study. Aligned promoter sequences are represented by gray boxes, whereas gray horizontal lines between the boxes represent gaps in the alignment. Each entry represents a *elt-3* (A) or *elt-2* (B) promoter sequence and they are listed in the same order (top to bottom) as the *Caenorhabditis* species in the phylogeny shown in Figure 1B (left to right). Species colors (from Figure 1B) are shown to the left of each gene name. The black boxes delineate the different species clades. The keys to the different transcription factor binding site motifs (depicted using triangles of different colors), and the highly conserved HGATAR sites (depicted using circles of different colors), are shown between panels A and B. **(A)** *elt-3* promoter sequences. Note the highly conserved HGATAR site in the *Elegans* group species (indicated above the panel). **(B)** *elt-2* promoter sequences. Note the highly conserved HGATAR sites in the *Elegans* supergroup species (indicated above the panel).

Figure 6

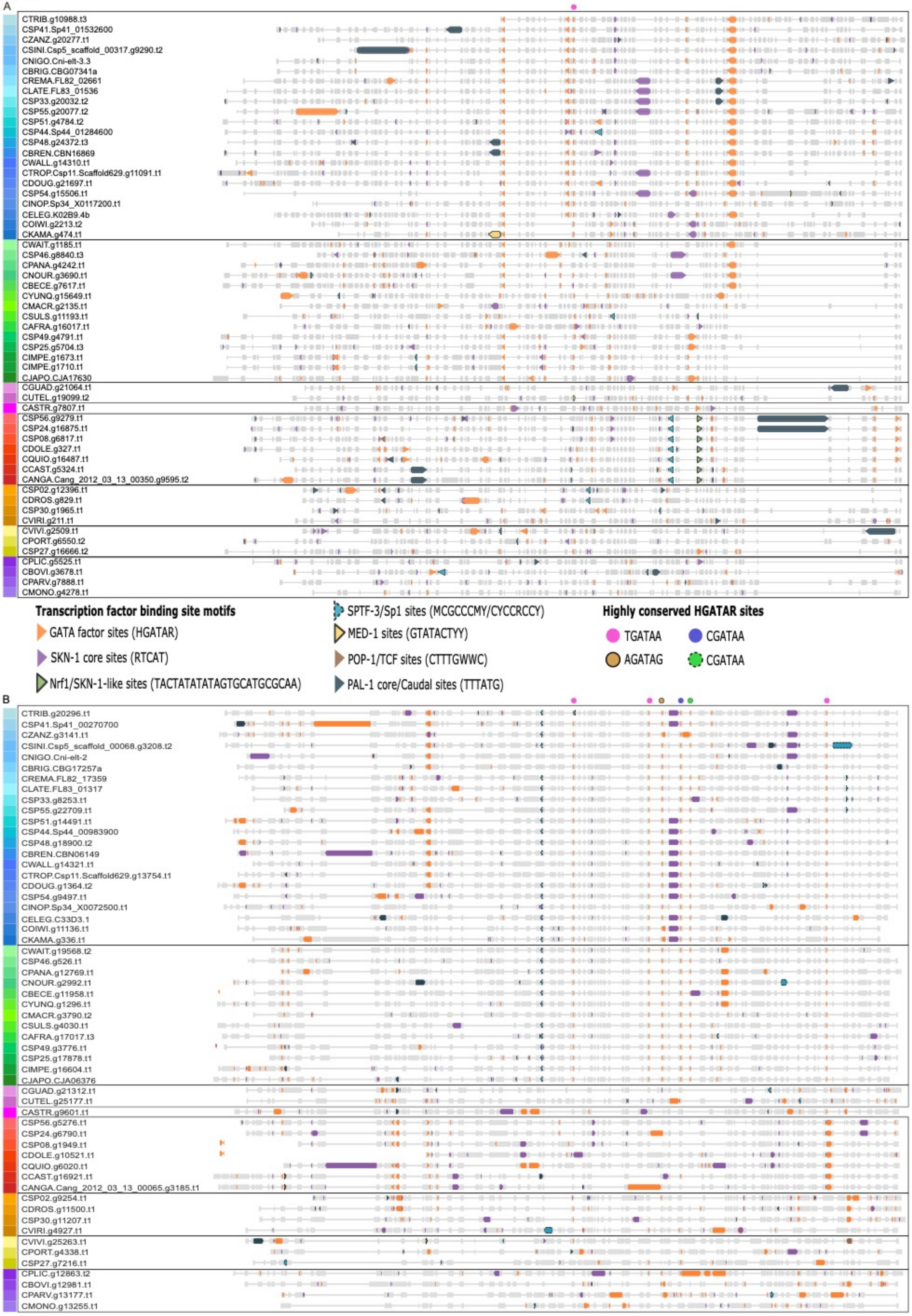

A
CTRIB.g10988.t3
CSP41.Sp41_01532600
CZANZ.g20277.t1
CSINI.Csp5_scaffold_00317.g9290.t2
CNIGO.Cni-elt-3.3
CBRIG.CBG07341a
CREMA.FL82_02661
CLATE.FL83_01536
CSP33.g20032.t2
CSP55.g20077.t2
CSP51.g4784.t2
CSP44.Sp44_01284600
CSP48.g24372.t3
CBREN.CBN16869
CWALL.g14310.t1
CTROP.Csp11.Scaffold629.g11091.t1
CDOUG.g21697.t1
CSP54.g15506.t1
CINOP.Sp34_X0117200.t1
CELEG.K02B9.4b
COIWI.g2213.t2
CKAMA.g474.t1
CWAIT.g1185.t1
CSP46.g8840.t3
CPANA.g4242.t1
CNOUR.g3690.t1
CBECE.g7617.t1
CYUNQ.g15649.t1
CMACR.g2135.t1
CSULS.g11193.t1
CAFRA.g16017.t1
CSP49.g4791.t1
CSP25.g5704.t3
CIMPE.g1673.t1
CIMPE.g1710.t1
CJAPO.CJA17630
CGUAD.g21064.t1
CUTEL.g19099.t2
CASTR.g7807.t1
CSP56.g9279.t1
CSP24.g16875.t1
CSP08.g6817.t1
CDOLE.g327.t1
CQUIO.g16487.t1
CCAST.g5324.t1
CANGA.Cang_2012_03_13_00350.g9595.t2
CSP02.g12396.t1
CDROS.g829.t1
CSP30.g1965.t1
CVIRI.g211.t1
CVIVI.g2509.t1
CPORT.g6550.t2
CSP27.g16666.t2
CPLIC.g5525.t1
CBOVI.g3678.t1
CPARV.g7888.t1
CMONO.g4278.t1
Transcription factor binding site motifs
GATA factor sites (HGATAR)
SKN-1 core sites (RTCAT)
Nrf1/SKN-1-like sites (TACTATATATAGTGCATGCGCAA)
SPTF-3/Sp1 sites (MCGCCCMY/CYCCRCCY)
MED-1 sites (GTATACTYY)
POP-1/TCF sites (CTTTGWWC)
PAL-1 core/Caudal sites (TTTATG)
Highly conserved HGATAR sites
TGATAA
CGATAA
AGATAG
CGATAA
B
CTRIB.g20296.t1
CSP41.Sp41_00270700
CZANZ.g3141.t1
CSINI.Csp5_scaffold_00068.g3208.t2
CNIGO.Cni-elt-2
CBRIG.CBG17257a
CREMA.FL82_17359
CLATE.FL83_01317
CSP33.g8253.t1
CSP55.g22709.t1
CSP51.g14491.t1
CSP44.Sp44_00983900
CSP48.g18900.t2
CBREN.CBN06149
CWALL.g14321.t1
CTROP.Csp11.Scaffold629.g13754.t1
CDOUG.g1364.t2
CSP54.g9497.t1
CINOP.Sp34_X0072500.t1
CELEG.C33D3.1
COIWI.g11136.t1
CKAMA.g336.t1
CWAIT.g19568.t2
CSP46.g526.t1
CPANA.g12769.t1
CNOUR.g2992.t1
CBECE.g11958.t1
CYUNQ.g1296.t1
CMACR.g3790.t2
CSULS.g4030.t1
CAFRA.g17017.t3
CSP49.g3776.t1
CSP25.g17878.t1
CIMPE.g16604.t1
CJAPO.CJA06376
CGUAD.g21312.t1
CUTEL.g25177.t1
CASTR.g9601.t1
CSP56.g5276.t1
CSP24.g6790.t1
CSP08.g1949.t1
CDOLE.g10521.t1
CQUIO.g6020.t1
CCAST.g16921.t1
CANGA.Cang_2012_03_13_00065.g3185.t1
CSP02.g9254.t1
CDROS.g11500.t1
CSP30.g11207.t1
CVIRI.g4927.t1
CVIVI.g25263.t1
CPORT.g4338.t1
CSP27.g7216.t1
CPLIC.g12863.t2
CBOVI.g12981.t1
CPARV.g13177.t1
CMONO.g13255.t1

Figure 1.7. Conservation of TGATAA sites in putative promoters of gut- and muscle-specific/-enriched orthologs. Heatmaps of the number of TGATAA sites in the promoter regions of gut-specific/-enriched orthologs (A) versus muscle-specific/-enriched orthologs (B) in the 59 non-*C. elegans* species included in this study. The columns comprising the x-axis represent each species, in the same order (left to right) as the listing of species in the phylogeny shown in Figure 1B. Each row on the y-axis represents the promoter region of a *C. elegans* gene specifically expressed, or enriched in expression, in gut (A) or muscle (B) tissue, ordered using hierarchical clustering with Euclidean distance metric. The color key is shown to the right of each heatmap plot. To make the color scaling more informative, the few promoter regions that had more than 10 TGATAA sequences are shown as having only 10 TGATAA sites within their promoters**.** White space in heatmaps shows which species we did not find orthologs for. **(A)** *C. elegans* gut-specific/enriched ortholog promoters. **(B)** *C. elegans* muscle-specific/enriched ortholog promoters.

Figure 7

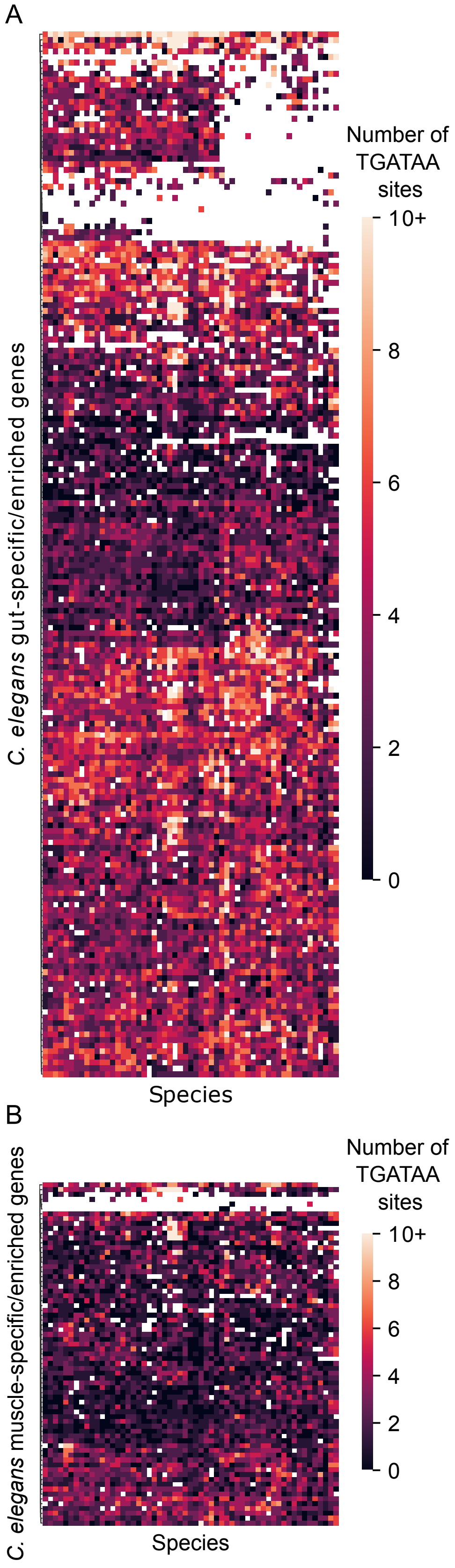

A
C. elegans gut-specific/enriched genes
Species
Number of TGATAA sites
10+
8
6
4
2
0
B
C. elegans muscle-specific/enriched genes
Species
Number of TGATAA sites
10+
8
6
4
2
0

Figure 8

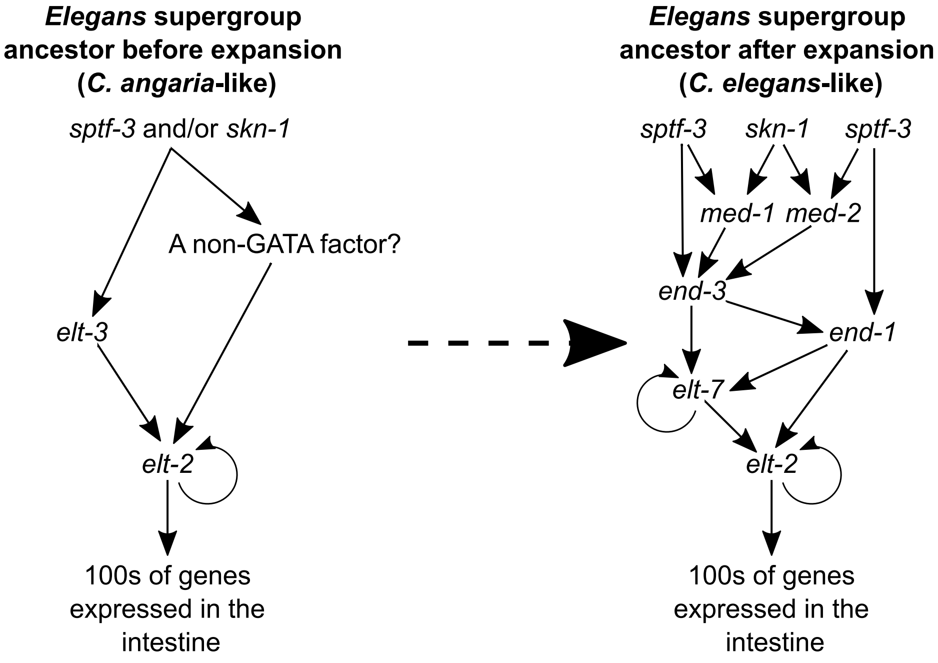
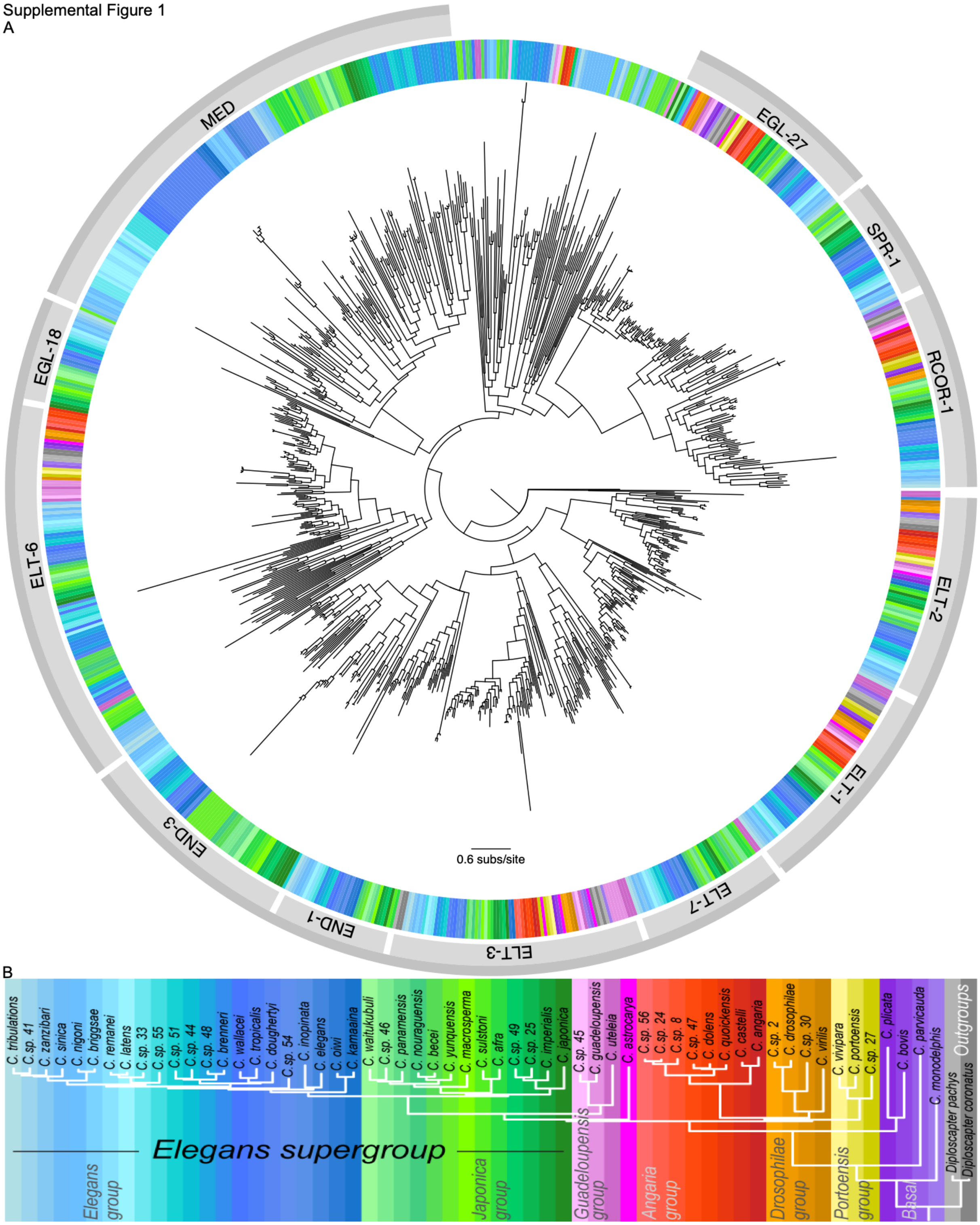

Elegans supergroup ancestor before expansion (*C. angaria*-like)
sptf-3 and/or *skn-1*
A non-GATA factor?
elt-3
elt-2
100s of genes expressed in the intestine
Elegans supergroup ancestor after expansion (*C. elegans*-like)
sptf-3
skn-1
sptf-3
med-1
med-2
end-3
end-1
elt-7
elt-2
100s of genes expressed in the intestine

Figure 1.8. Prediction of how the endoderm developmental gene regulatory network (dGRN) evolved in the *Elegans* supergroup ancestor. Data from this study are consistent with this evolutionary model in which, prior to our proposed expansion of the *elt-3* gene in the *Elegans* supergroup ancestor, the endoderm dGRN was initiated by *sptf-3* and/or *skn-1*. These transcription factors activate *elt-3* and possibly another transcription factor that is expressed earlier (depicted as “A non-GATA factor?” in the Figure). *elt-3* and possibly other transcription factors then activate *elt-2*. *elt-2* then likely regulates hundreds of intestine-specific/-enriched genes and may auto-regulate itself. This network is shown on the left and is expected to be similar to the endoderm dGRNs found in non-*Elegans* supergroup species, like *C. angoria*. We predict that an expansion of *elt-3* added two or three GATA factors between the endoderm dGRNs initiation with *sptf-3* and *skn-1* and it’s differentiation by *elt-2* leading to the network shown on the right. During the expansion *elt-3* paralogs subfunctionalized into an *elt-3*-like gene expressed only in the hypoderm (not shown), and endoderm-specific *elt-7* and ancestor of the *end* genes. (See Supplemental Figure 11 for molecular scenarios of how this might have happened). Data from this study also support the hypothesis (previously proposed by Maduro et al. (2005a) and Coroian et al. (2006)) that an additional *end* gene duplication produced the ancestors of *end-1* and *end-3*, and that another *end* gene duplication likely produced the ancestor *med* gene (previously proposed by Maduro (2020)). Together these data suggest how the endoderm dGRN in the ancestor of the *Elegans* supergroup may have operated before (panel on left) and after (panel on right) the hypothesized *elt-3* radiation that re-wired the developmental network took place.

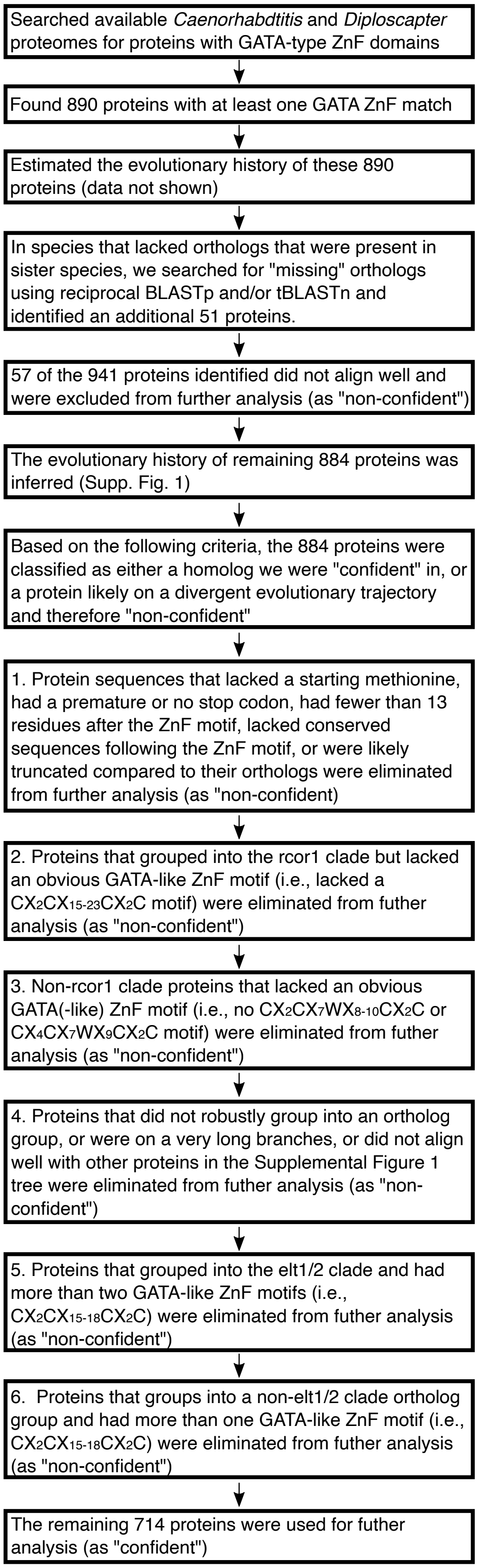
Supplemental Figure 1.1. Preliminary inferred evolutionary history of *Caenorhabditis* GATA-domain-containing proteins. (A) Maximum likelihood phylogeny of 884 alignable GATA-domain-containing proteins in 58 *Caenorhabditis* and two outgroup nematode species, created using a GATA factor from the slime mold *Dictyostelium fasciculatum* to root the tree. The colors in the ring encircling the tree correspond to the species in which the proteins in the tree were identified; the key to color-species correspondence is given in B below. The names of the 12 ortholog groups the 884 proteins were categorized into (see Results) are indicated in the lighter of the two outer gray rings (with white gaps between groups). Groupings of ortholog groups that share adjacent clades in the tree are highlighted by a darker gray line on the outermost edge of the figure (with white gaps between groups). The key for translating branch length into evolutionary distance (in units of amino acid substitutions per site) is shown near the bottom of the phylogenetic tree. **(B)** Phylogenetic relationships among the 60 species used in this study (based on Stevens 2020). Each species is designated by a different color shade; the same color-species designations are used in A above.

Supplemental Figure 1



Supplemental Figure 1.2. Method of selecting protein sequences for further analyses. The process carried out to select among putative GATA-domain-containing protein sequences those for further analyses are depicted in this figure. Protein sequences were classified as “confident” for use in further analyses versus “not_confident” using this decision tree and its selection criteria. (Supplemental Table 1 comprises a list of the resulting classifications for each protein sequence.) ZnF stands for zinc finger.

Suppplemental Figure 2



Searched available *Caenorhabdtitis* and *Diploscapter* proteomes for proteins with GATA-type ZnF domains

Found 890 proteins with at least one GATA ZnF match

Estimated the evolutionary history of these 890 proteins (data not shown)

In species that lacked orthologs that were present in sister species, we searched for "missing" orthologs using reciprocal BLASTp and/or tBLASTn and identified an additional 51 proteins.

57 of the 941 proteins identified did not align well and were excluded from further analysis (as "non-confident")

The evolutionary history of remaining 884 proteins was inferred (Supp. Fig. 1)

Based on the following criteria, the 884 proteins were classified as either a homolog we were "confident" in, or a protein likely on a divergent evolutionary trajectory and therefore "non-confident"

1. Protein sequences that lacked a starting methionine, had a premature or no stop codon, had fewer than 13 residues after the ZnF motif, lacked conserved sequences following the ZnF motif, or were likely truncated compared to their orthologs were eliminated from further analysis (as "non-confident)

2. Proteins that grouped into the rcor1 clade but lacked an obvious GATA-like ZnF motif (i.e., lacked a $CX_2CX_{15-23}CX_2C$ motif) were eliminated from futher analysis (as "non-confident")

3. Non-rcor1 clade proteins that lacked an obvious GATA(-like) ZnF motif (i.e., no $CX_2CX_7WX_{8-10}CX_2C$ or $CX_4CX_7WX_9CX_2C$ motif) were eliminated from futher analysis (as "non-confident")

4. Proteins that did not robustly group into an ortholog group, or were on a very long branches, or did not align well with other proteins in the Supplemental Figure 1 tree were eliminated from futher analysis (as "non-confident")

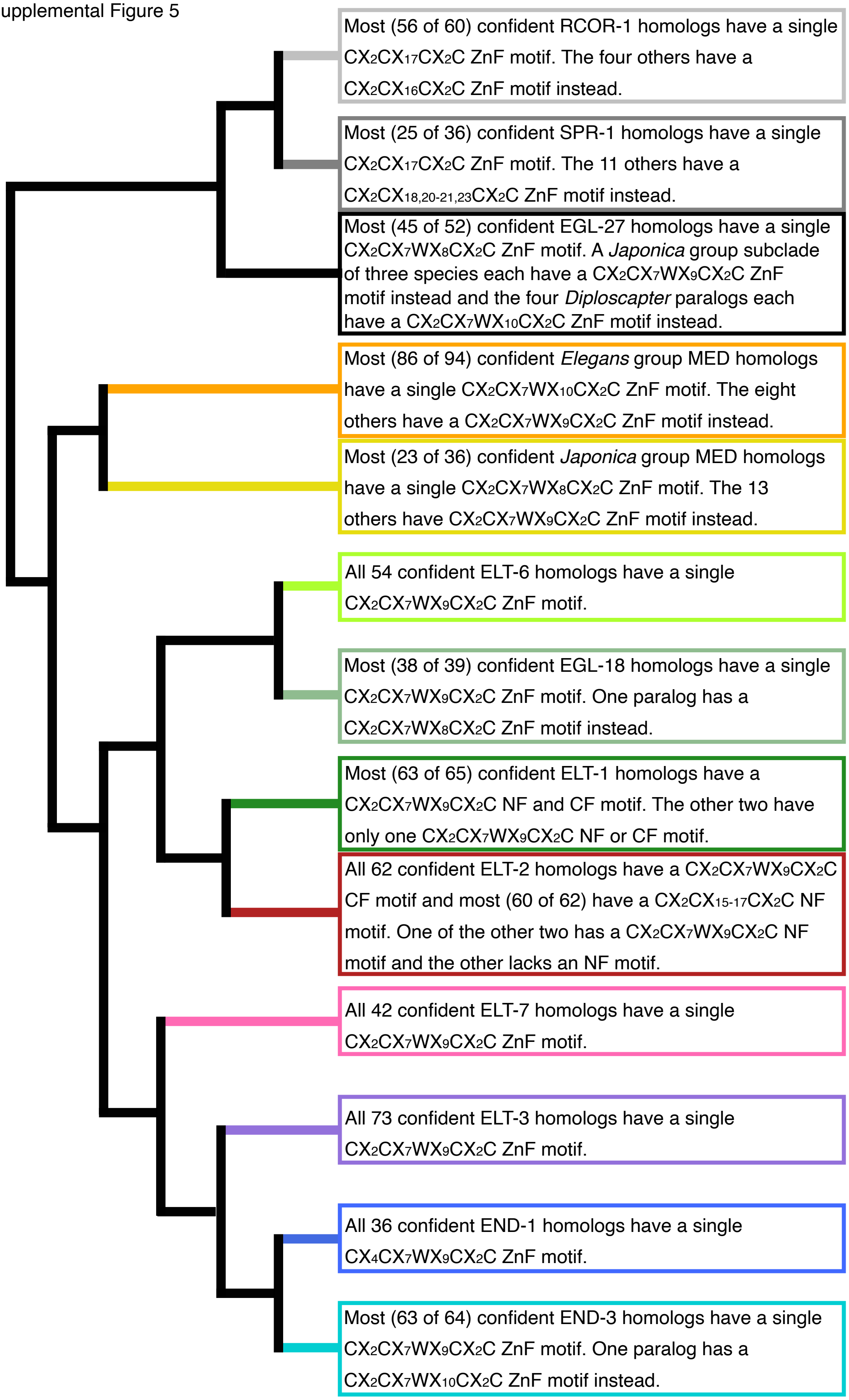
5. Proteins that grouped into the elt1/2 clade and had more than two GATA-like ZnF motifs (i.e., $CX_2CX_{15-18}CX_2C$) were eliminated from futher analysis (as "non-confident")

6. Proteins that groups into a non-elt1/2 clade ortholog group and had more than one GATA-like ZnF motif (i.e., $CX_2CX_{15-18}CX_2C$) were eliminated from futher analysis (as "non-confident")

The remaining 714 proteins were used for futher analysis (as "confident")

Supplemental Figure 1.5. Comparison of zinc finger motifs found in each ortholog group. The loop size(s) of each GATA factor zinc finger (ZnF), and the presence or absence of the highly conserved tryptophan (W) at position eight in the ZnF loop, is shown for all genes deemed confident in each of the 12 ortholog groups.

Supplemental Figure 5



Most (56 of 60) confident RCOR-1 homologs have a single $CX_2CX_{17}CX_2C$ ZnF motif. The four others have a $CX_2CX_{16}CX_2C$ ZnF motif instead.

Most (25 of 36) confident SPR-1 homologs have a single $CX_2CX_{17}CX_2C$ ZnF motif. The 11 others have a $CX_2CX_{18,20\text{-}21,23}CX_2C$ ZnF motif instead.

Most (45 of 52) confident EGL-27 homologs have a single $CX_2CX_7WX_8CX_2C$ ZnF motif. A *Japonica* group subclade of three species each have a $CX_2CX_7WX_9CX_2C$ ZnF motif instead and the four *Diploscapter* paralogs each have a $CX_2CX_7WX_{10}CX_2C$ ZnF motif instead.

Most (86 of 94) confident *Elegans* group MED homologs have a single $CX_2CX_7WX_{10}CX_2C$ ZnF motif. The eight others have a $CX_2CX_7WX_9CX_2C$ ZnF motif instead.

Most (23 of 36) confident *Japonica* group MED homologs have a single $CX_2CX_7WX_8CX_2C$ ZnF motif. The 13 others have $CX_2CX_7WX_9CX_2C$ ZnF motif instead.

All 54 confident ELT-6 homologs have a single $CX_2CX_7WX_9CX_2C$ ZnF motif.

Most (38 of 39) confident EGL-18 homologs have a single $CX_2CX_7WX_9CX_2C$ ZnF motif. One paralog has a $CX_2CX_7WX_8CX_2C$ ZnF motif instead.

Most (63 of 65) confident ELT-1 homologs have a $CX_2CX_7WX_9CX_2C$ NF and CF motif. The other two have only one $CX_2CX_7WX_9CX_2C$ NF or CF motif.

All 62 confident ELT-2 homologs have a $CX_2CX_7WX_9CX_2C$ CF motif and most (60 of 62) have a $CX_2CX_{15\text{-}17}CX_2C$ NF motif. One of the other two has a $CX_2CX_7WX_9CX_2C$ NF motif and the other lacks an NF motif.

All 42 confident ELT-7 homologs have a single $CX_2CX_7WX_9CX_2C$ ZnF motif.

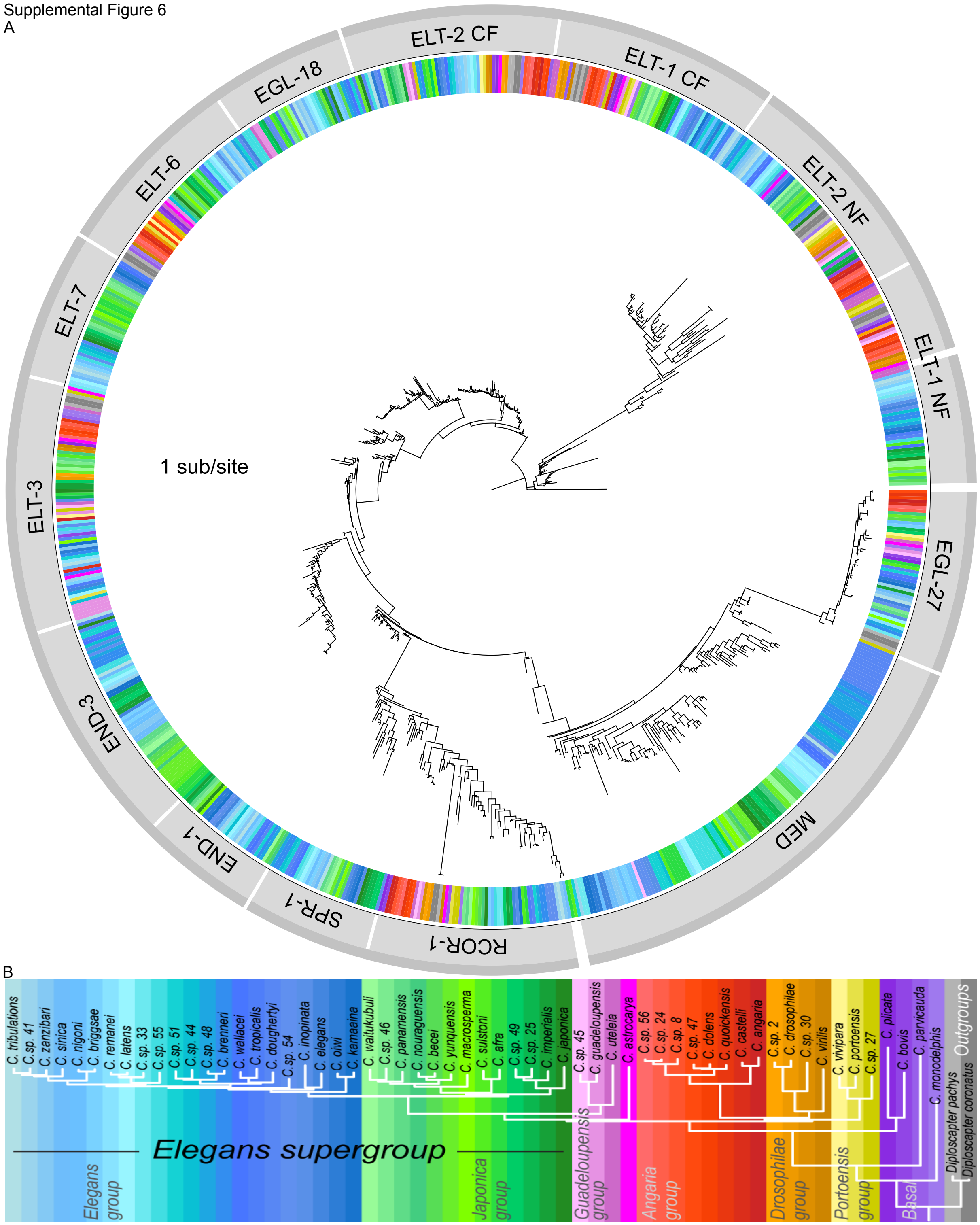
All 73 confident ELT-3 homologs have a single $CX_2CX_7WX_9CX_2C$ ZnF motif.

All 36 confident END-1 homologs have a single $CX_4CX_7WX_9CX_2C$ ZnF motif.

Most (63 of 64) confident END-3 homologs have a single $CX_2CX_7WX_9CX_2C$ ZnF motif. One paralog has a $CX_2CX_7WX_{10}CX_2C$ ZnF motif instead.

Supplemental Figure 1.6. Phylogenetic analysis of GATA (or GATA-like) domains. Maximum likelihood phylogeny of confident GATA domains in 58 *Caenorhabditis* and two outgroup nematode species. The GATA domain from a *D. fasciculatum* (slime mold) GATA factor was used to root the tree. The colors in the ring encircling the tree correspond to the species in which the GATA domain is from (the key to color-species correspondence is the same as in Supplemental Figure 1B). The 14 different groups that the GATA domain cluster into are labeled in light gray bars on the outside of the species color ring. Groupings of ortholog groups that share adjacent clades in the tree are highlighted by a darker gray line on the outermost edge of the figure. The key for translating branch length into evolutionary distance (in units of amino acid substitutions per site) is shown near the bottom of the phylogenetic tree.

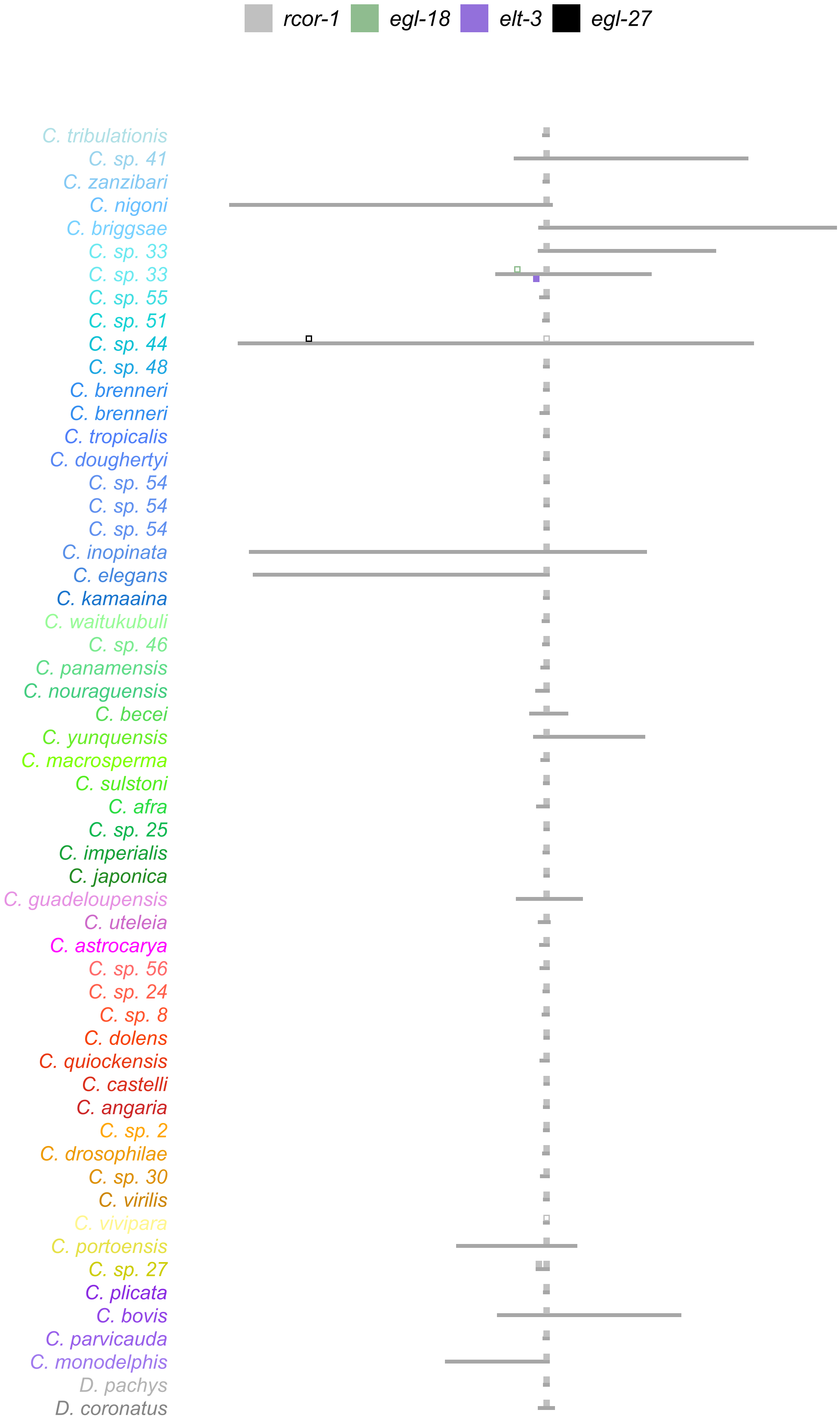
Supplemental Figure 6



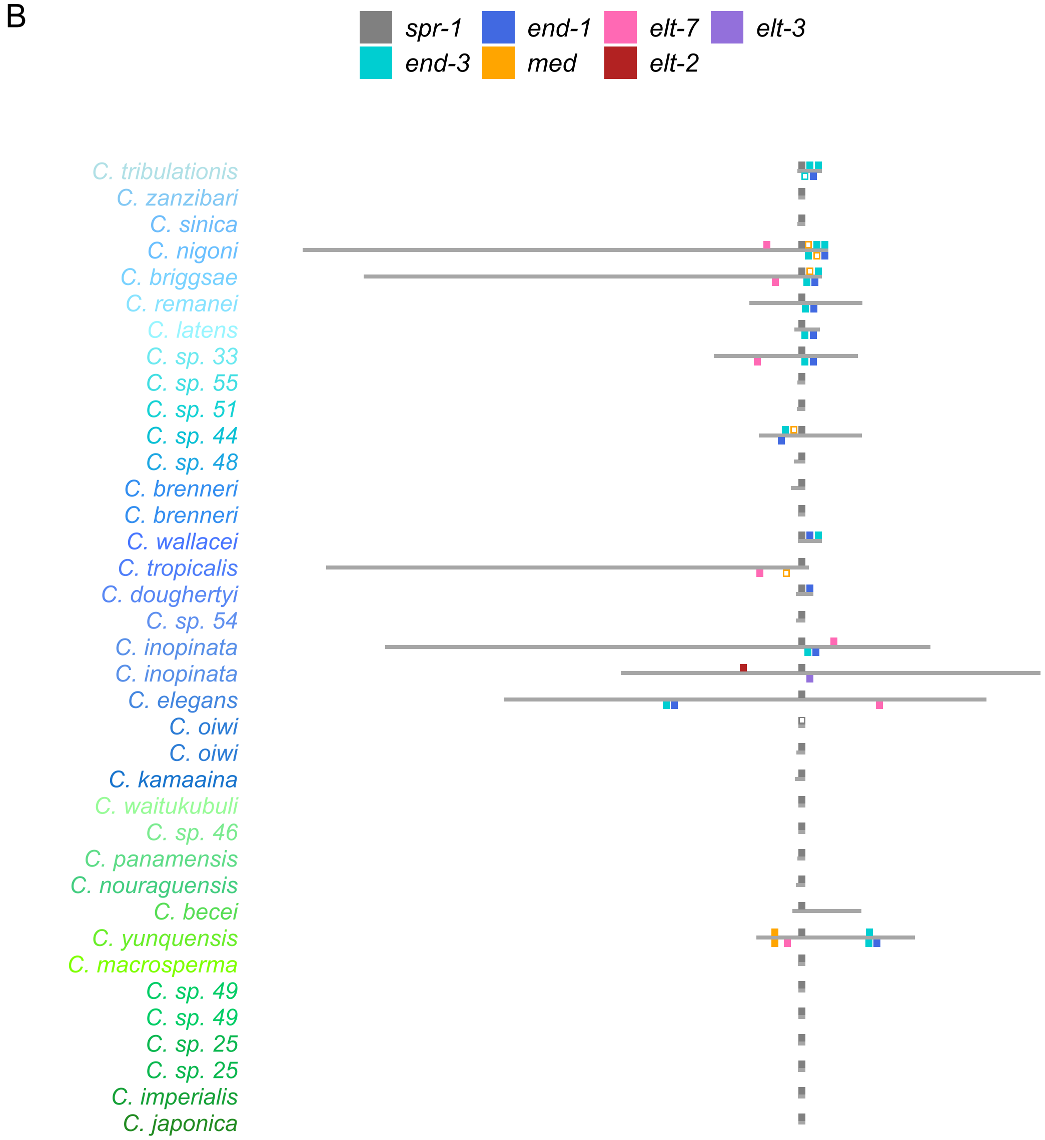
Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins. Contigs/scaffolds/chromosomes (depicted as gray horizontal rectangles) are anchored on a respective GATA-domain-containing gene (depicted as colored squares). The relative locations of any other GATA-domain-containing genes (depicted as other color squares) on the same scaffold/chromosome (i.e., syntenic GATA-domain-containing genes) are shown above or below a given contig/scaffold/chromosome, indicating their orientation on the same or opposite strand, respectively, as the anchored gene. Genes deemed confident and non-confident (see Methods) are depicted as filled in or outlined colored squares, respectively. Genes from each ortholog group are designated using the same color, as noted in the key at the top of each plot. The species from which each respective contig/scaffold/chromosome was sequenced is indicated on its left. The species names are in the order of the species phylogeny (Stevens 2020) and color-coded as in Supplemental Figure 1B. (For visual clarity, the sizes and exact relative locations of the colored squares representing GATA-domain-containing genes have been adjusted slightly in some cases, and large contigs/scaffolds/chromosomes were scaled down (based on their actual length per plot) while the smallest contigs/scaffolds were lengthened.) The gene serving as the anchor in each panel is as follows: **(A)** *rcor-1*; **(B)** *spr-1*; **(C)** *egl-27*; **(D)** *med*; **(E)** *elt-6*; **(F)** *egl-18*; **(G)** *elt-1*; **(H)** *elt-2*; **(I)** *elt-3*; **(J)** *egl-18*; **(K)** *end-1*; and **(L)** *end-3*.

Supplemental Figure 7

A

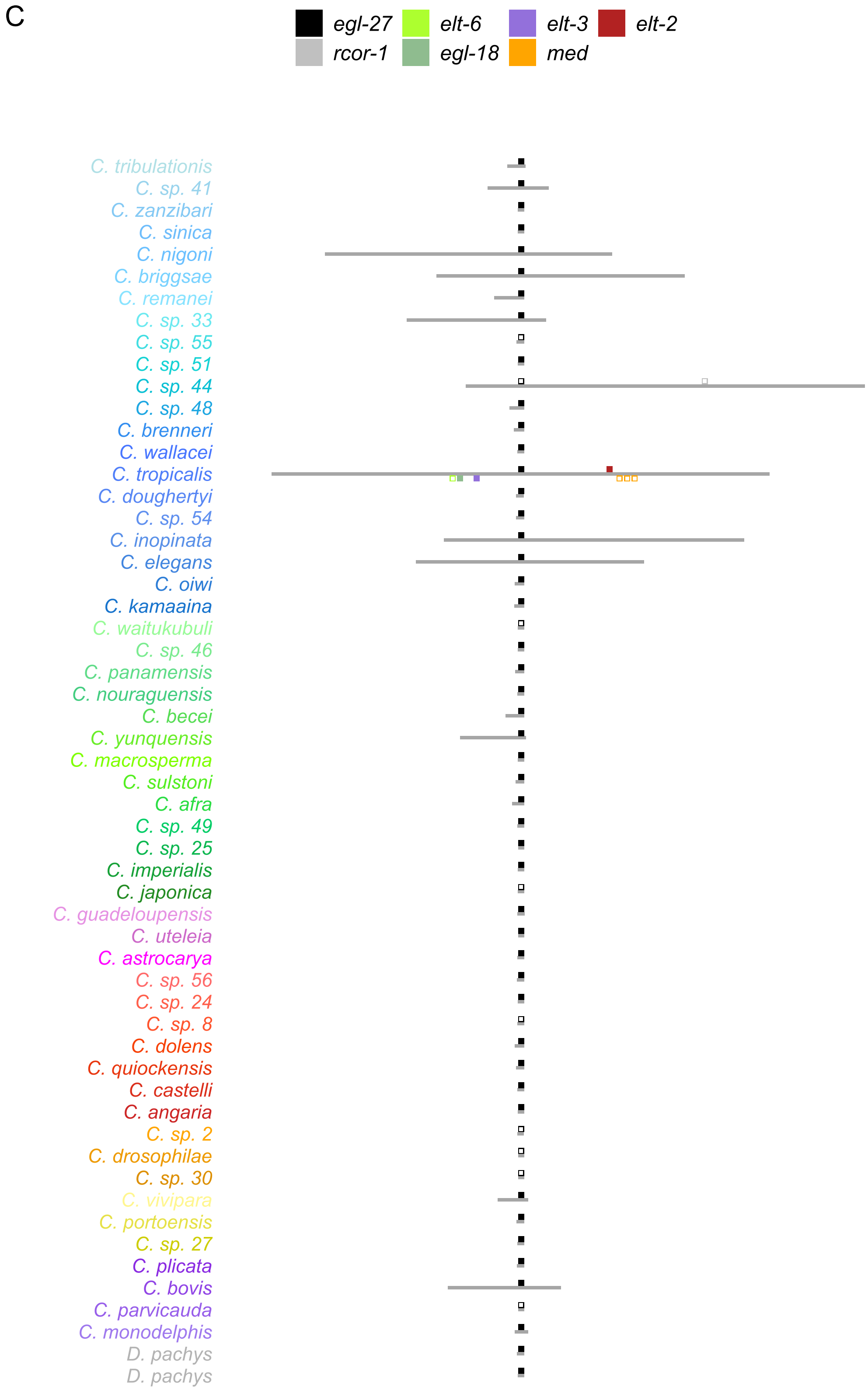


Supplemental Figure 7



Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

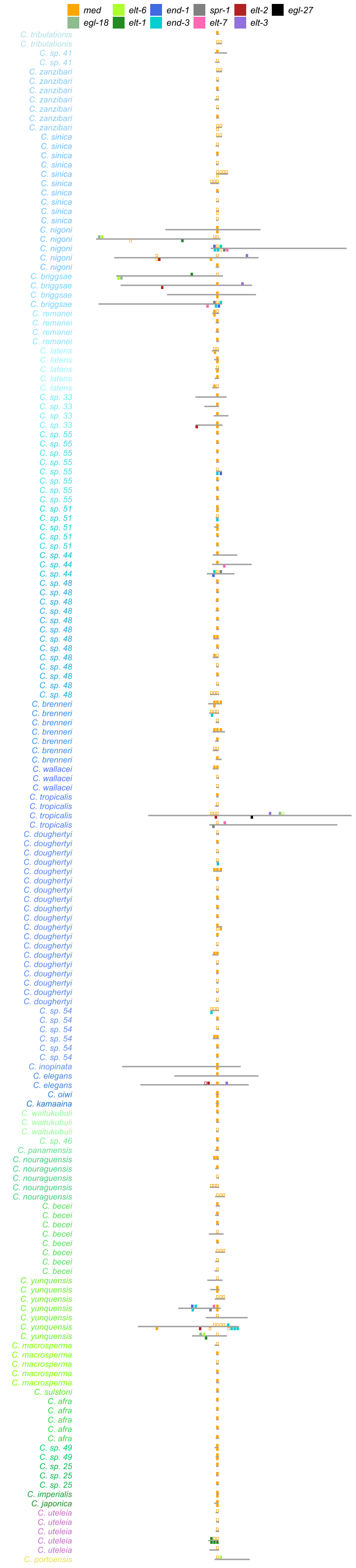
Supplemental Figure 7



Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

Supplemental Figure 7

D

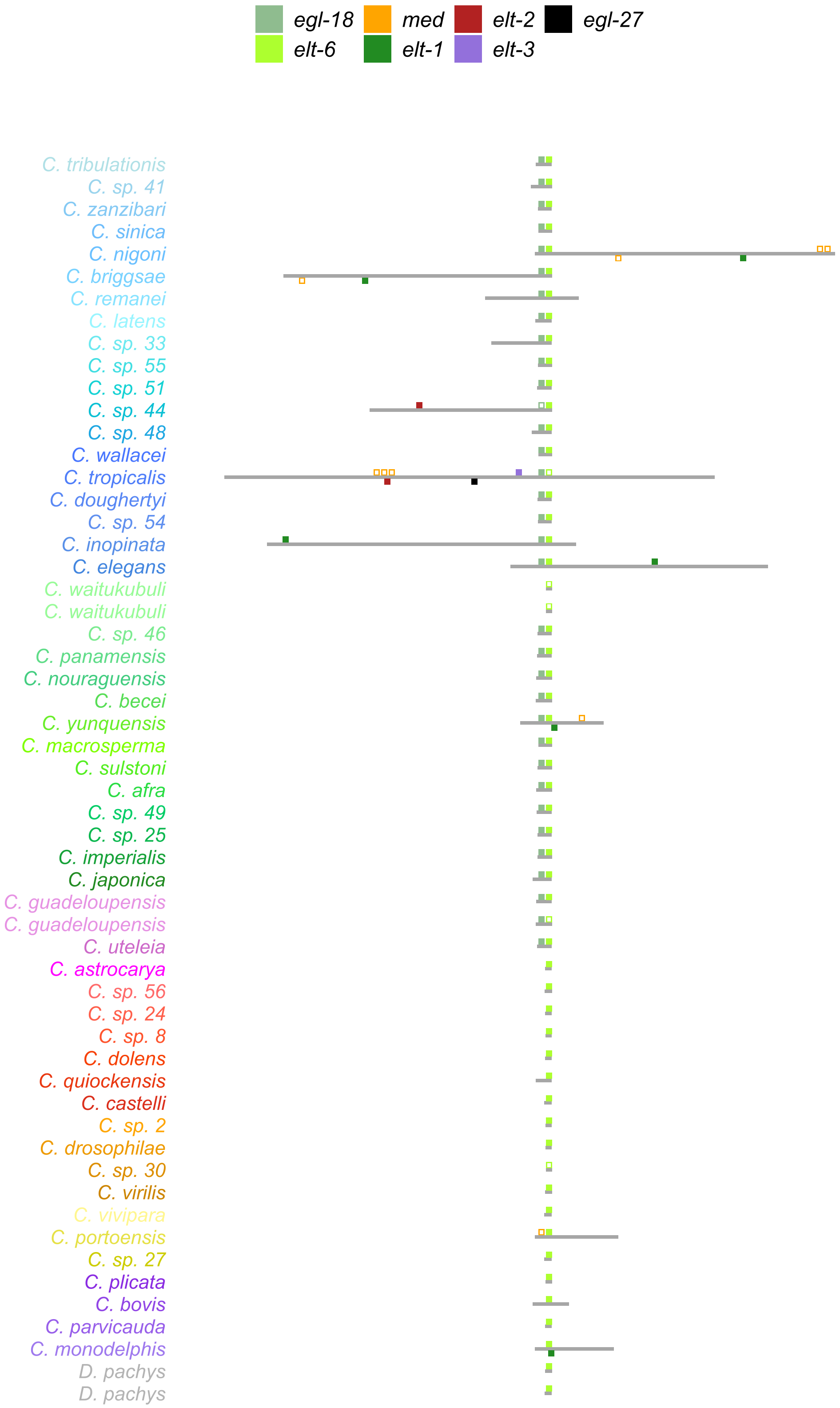


Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

)

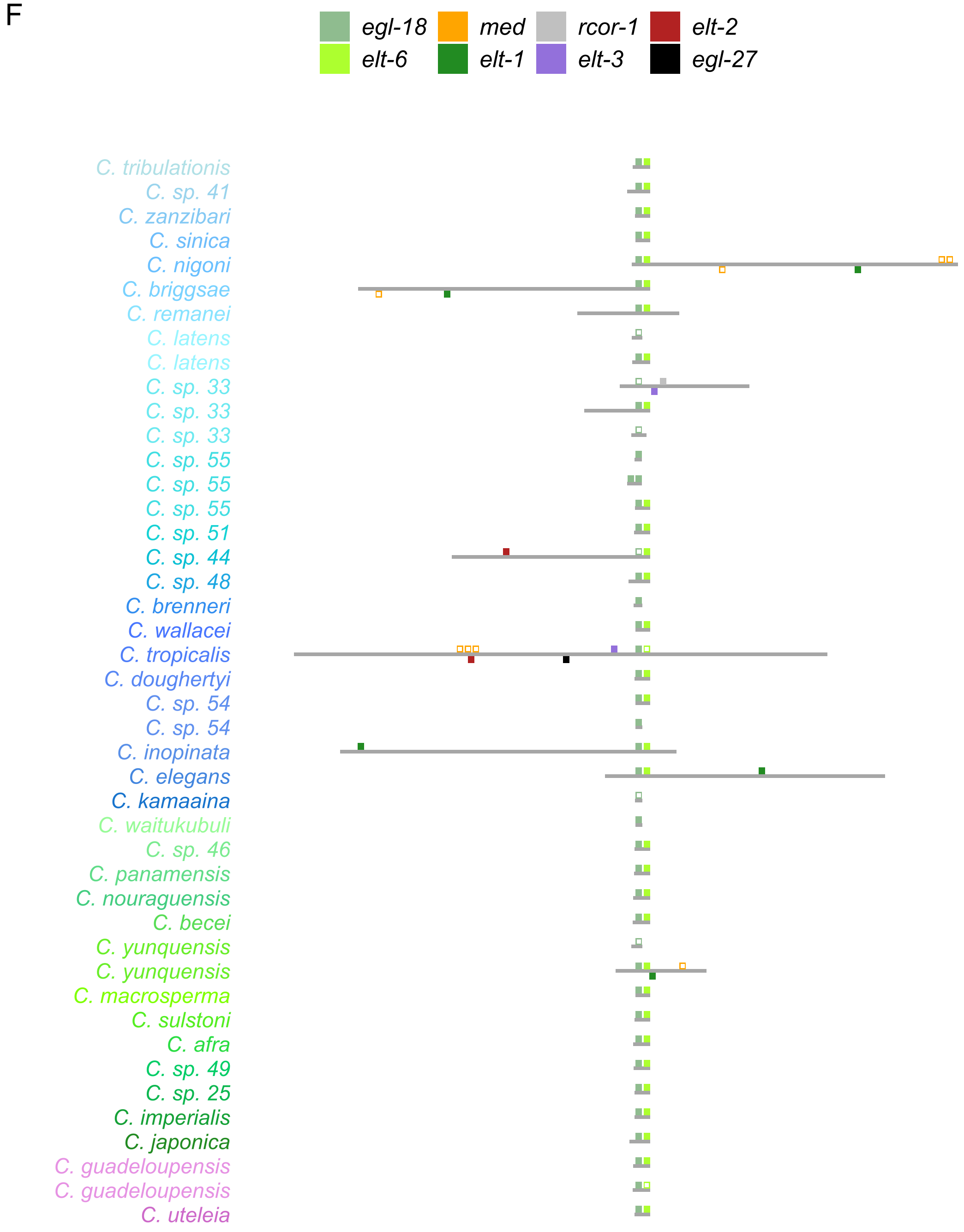
Supplemental Figure 7

E



Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

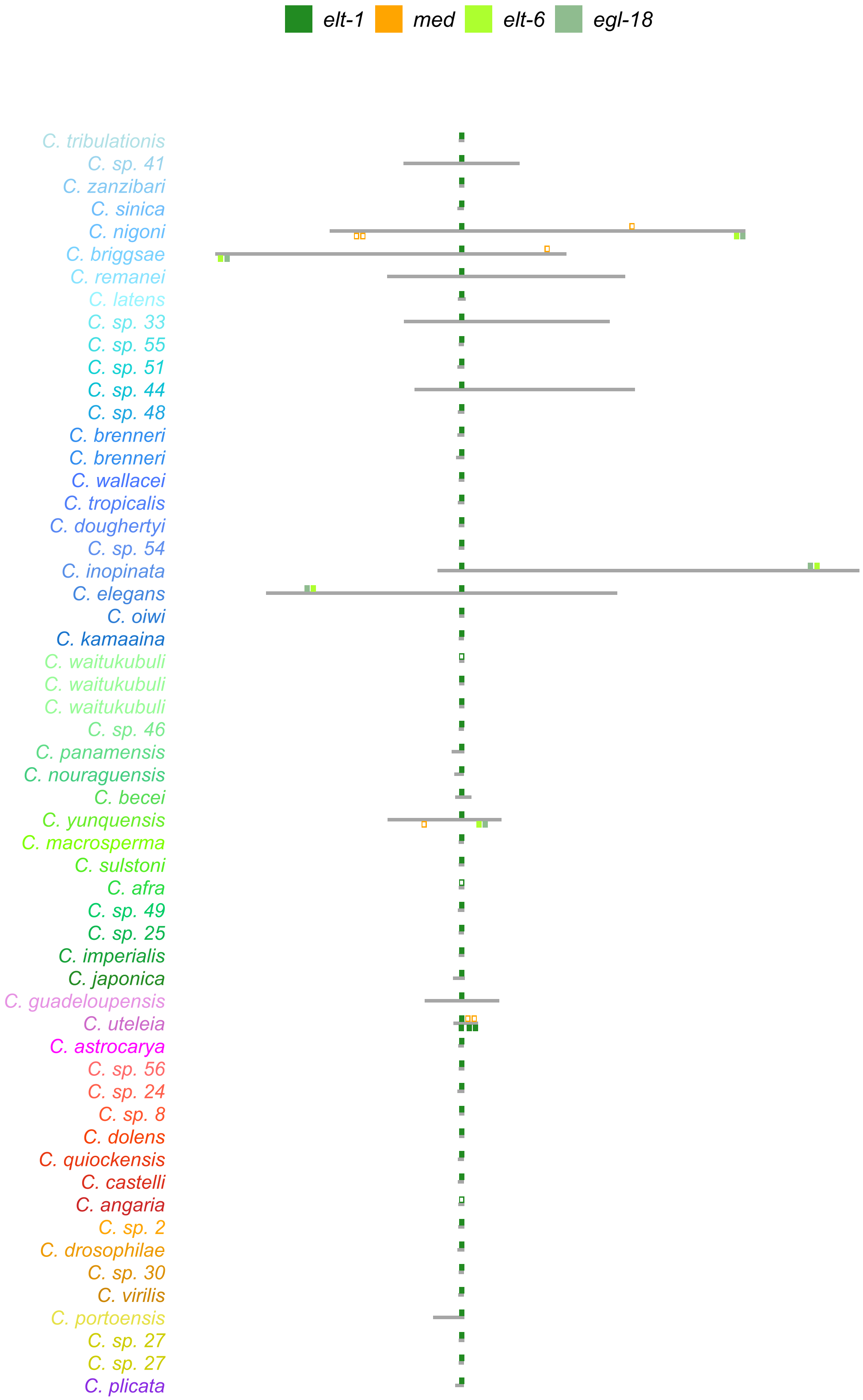
Supplemental Figure 7



Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

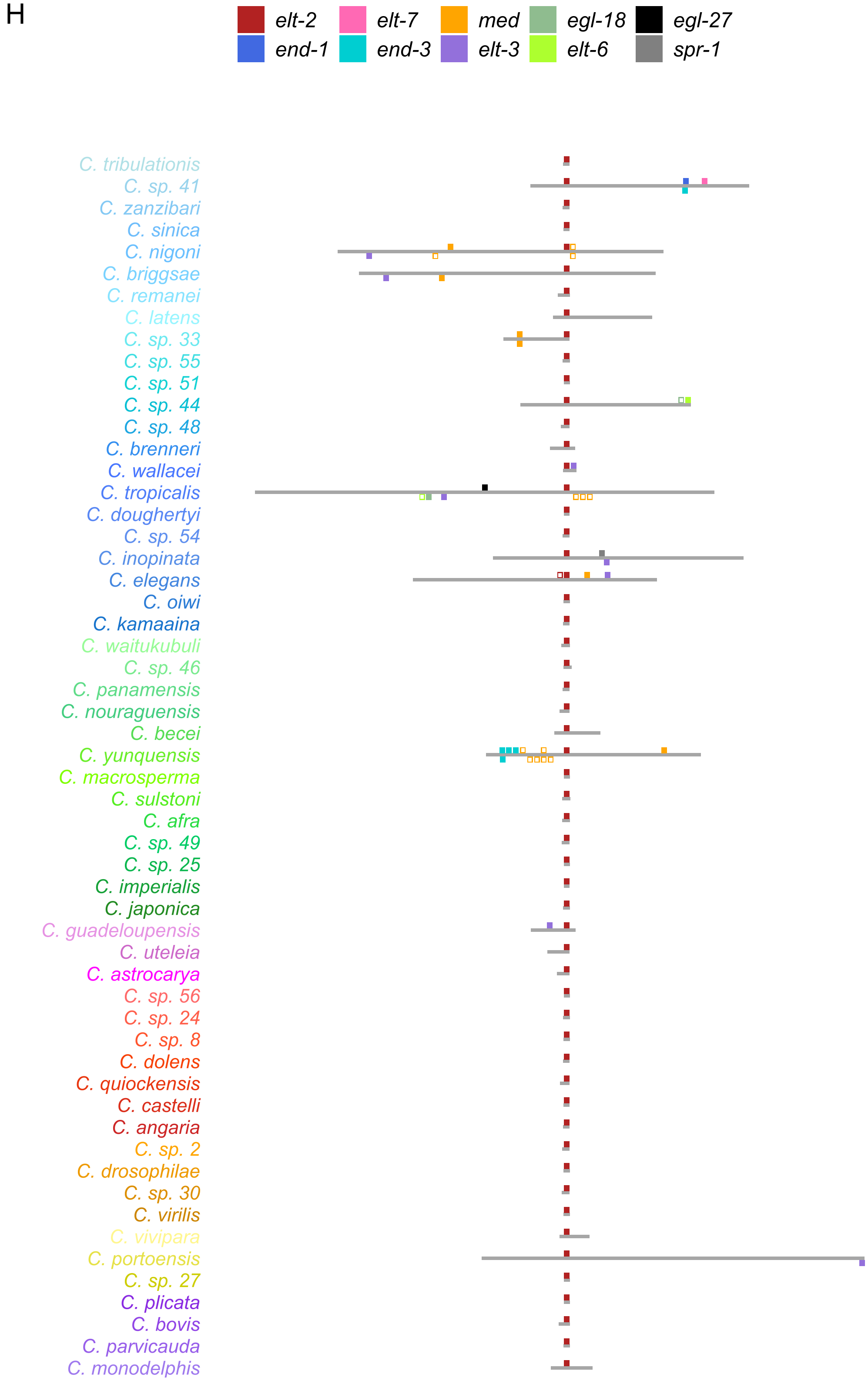
Supplemental Figure 7

G



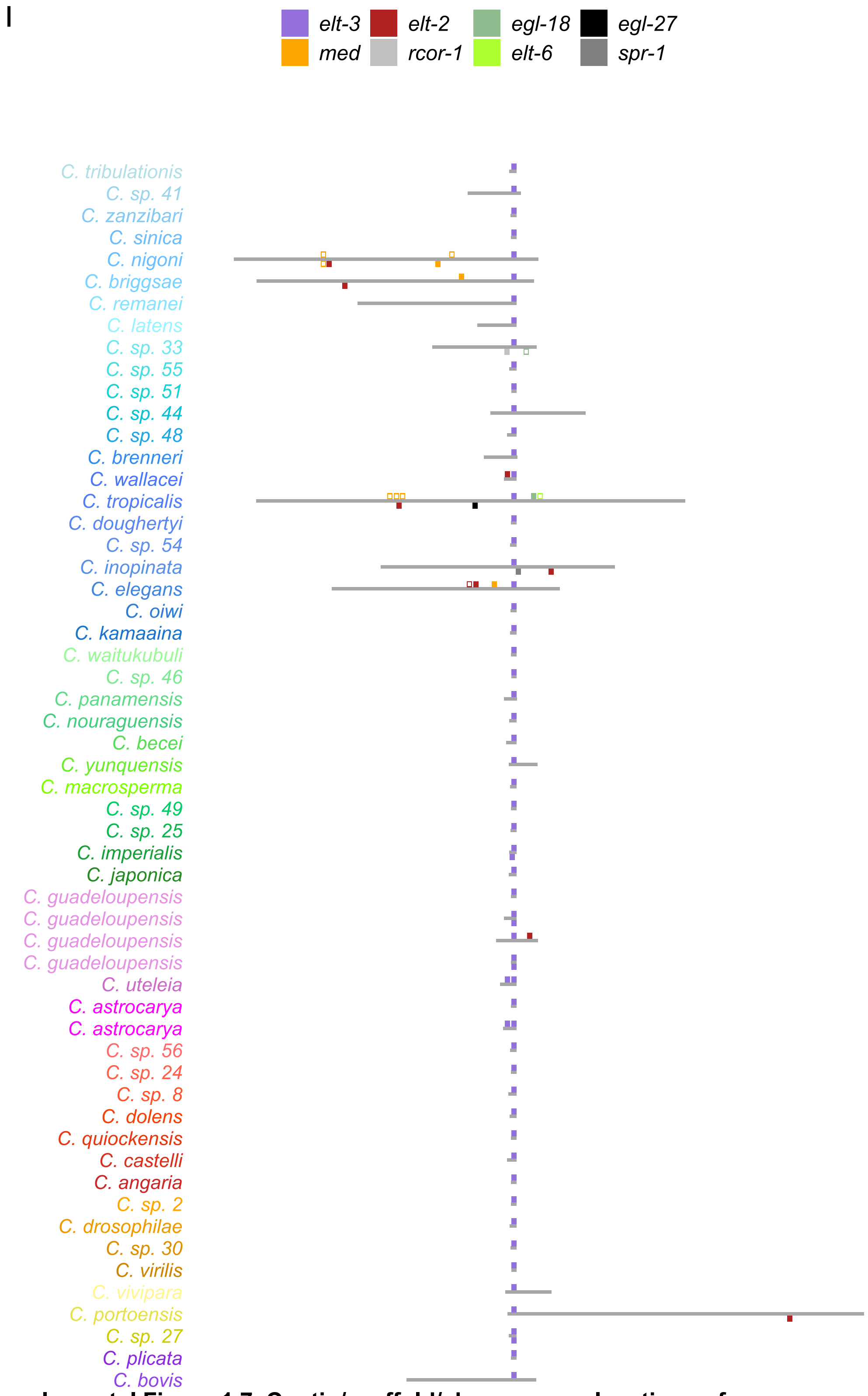
Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

Supplemental Figure 7



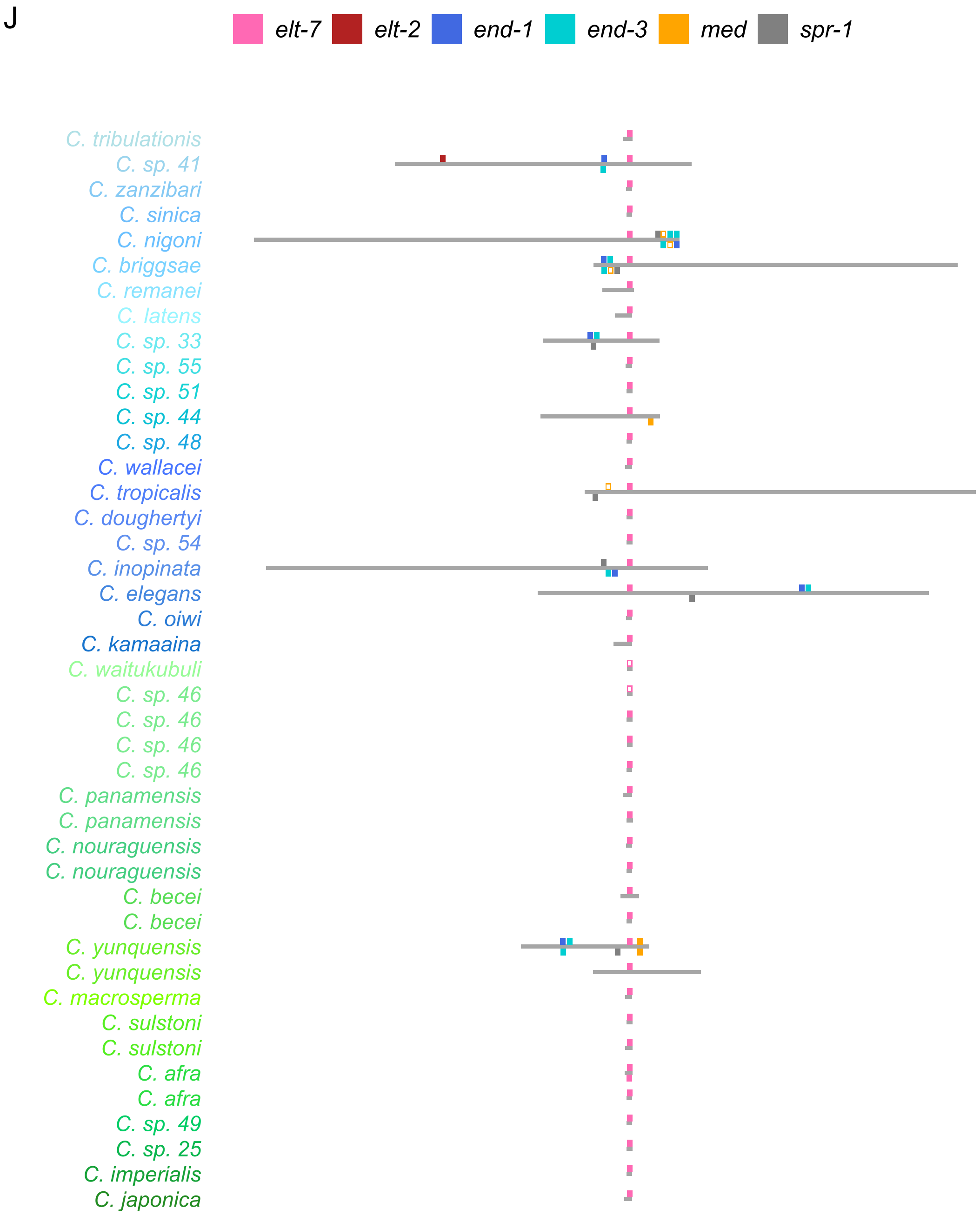
Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

Supplemental Figure 7



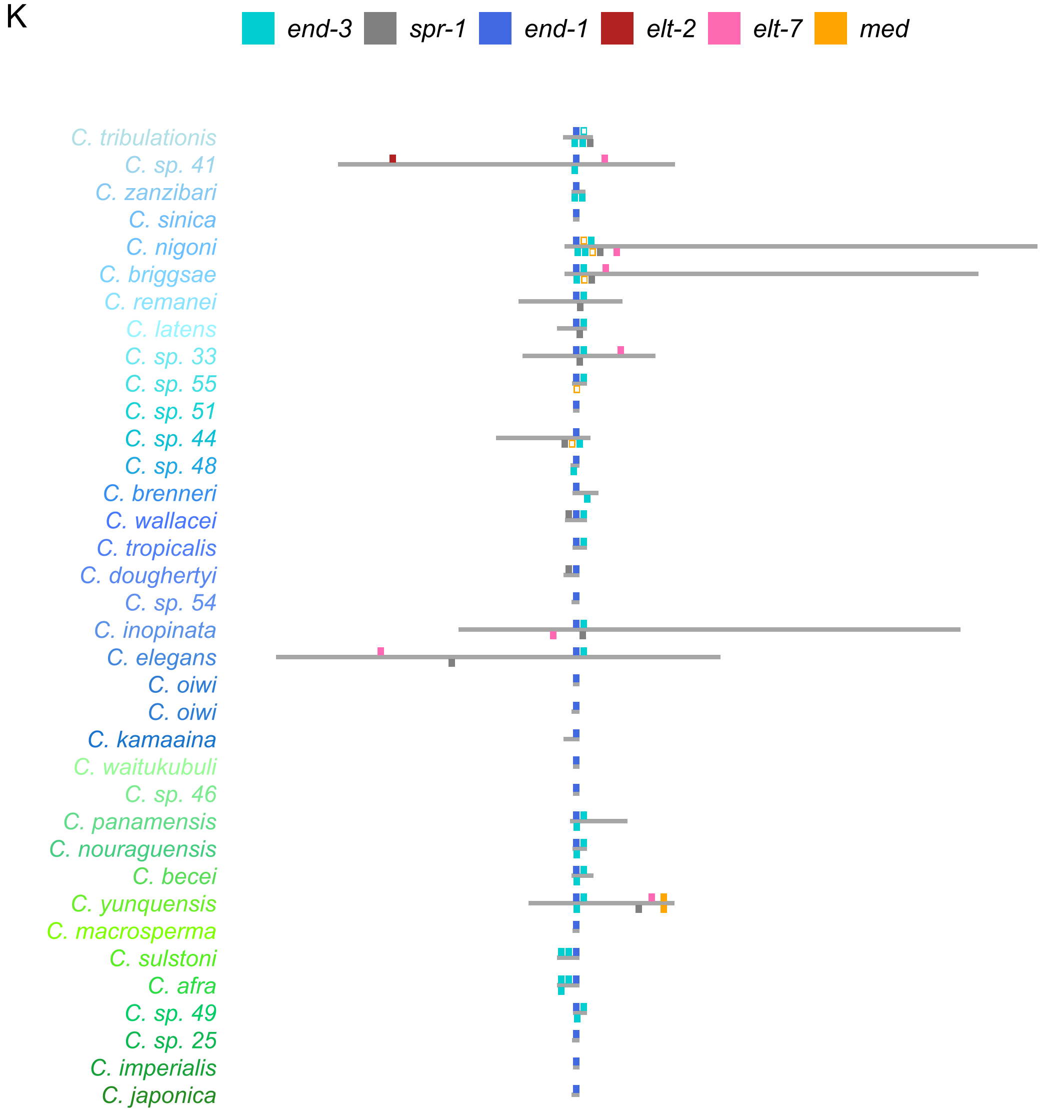
Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

Supplemental Figure 7

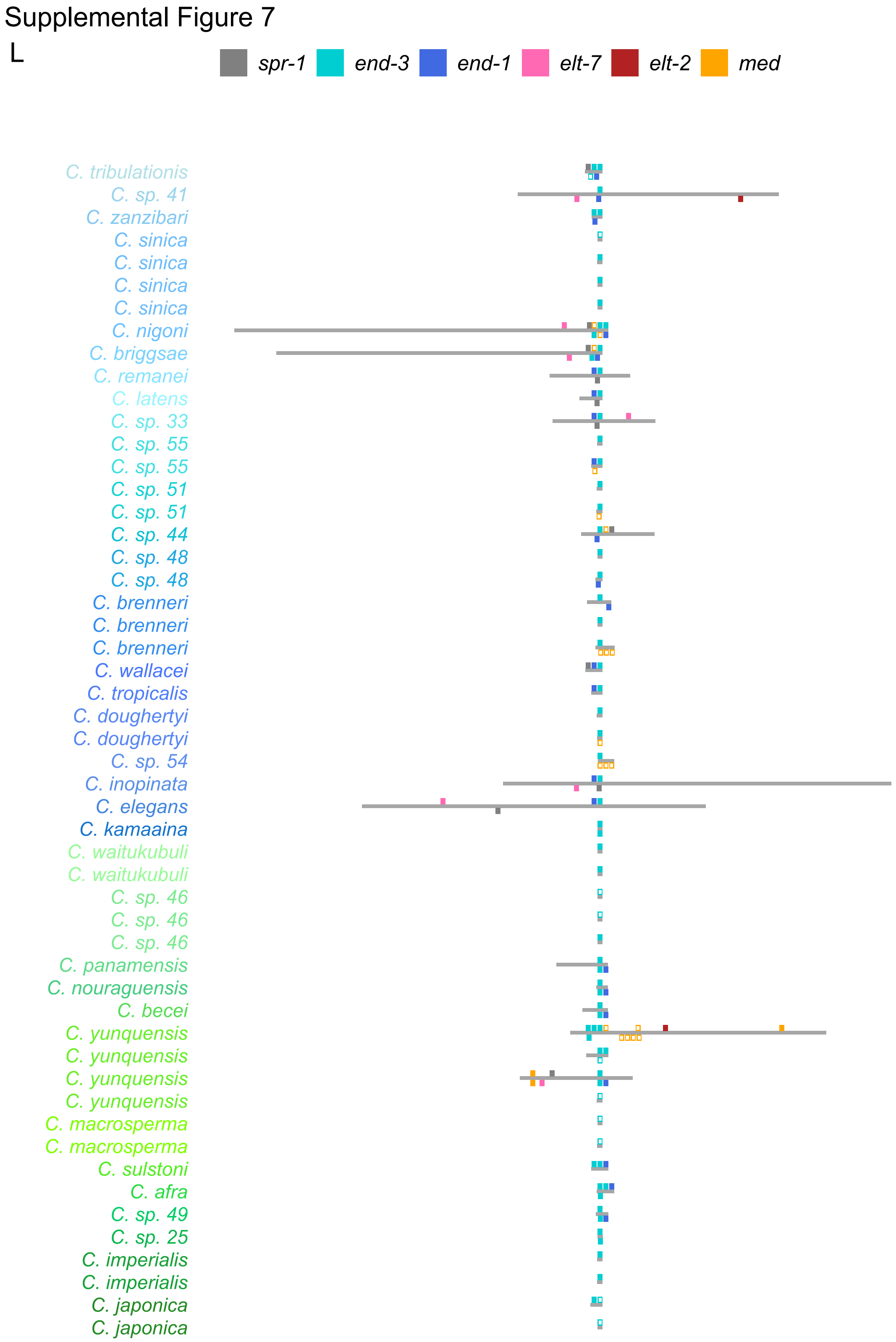


Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

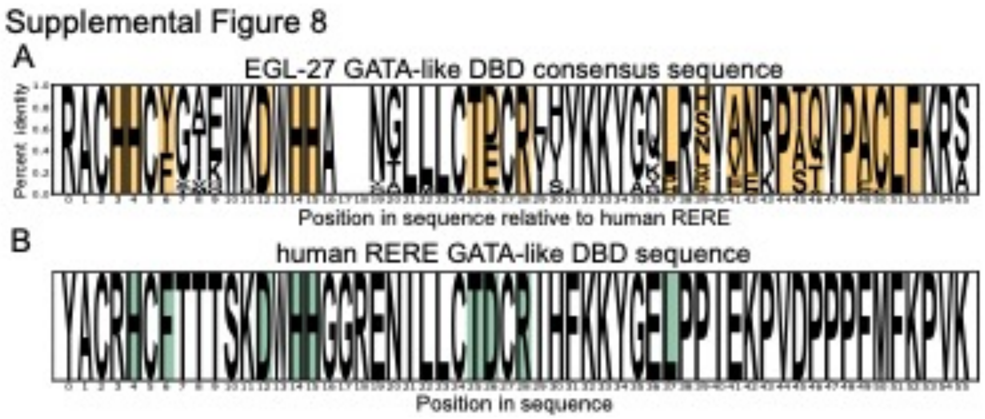
Supplemental Figure 7



Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.

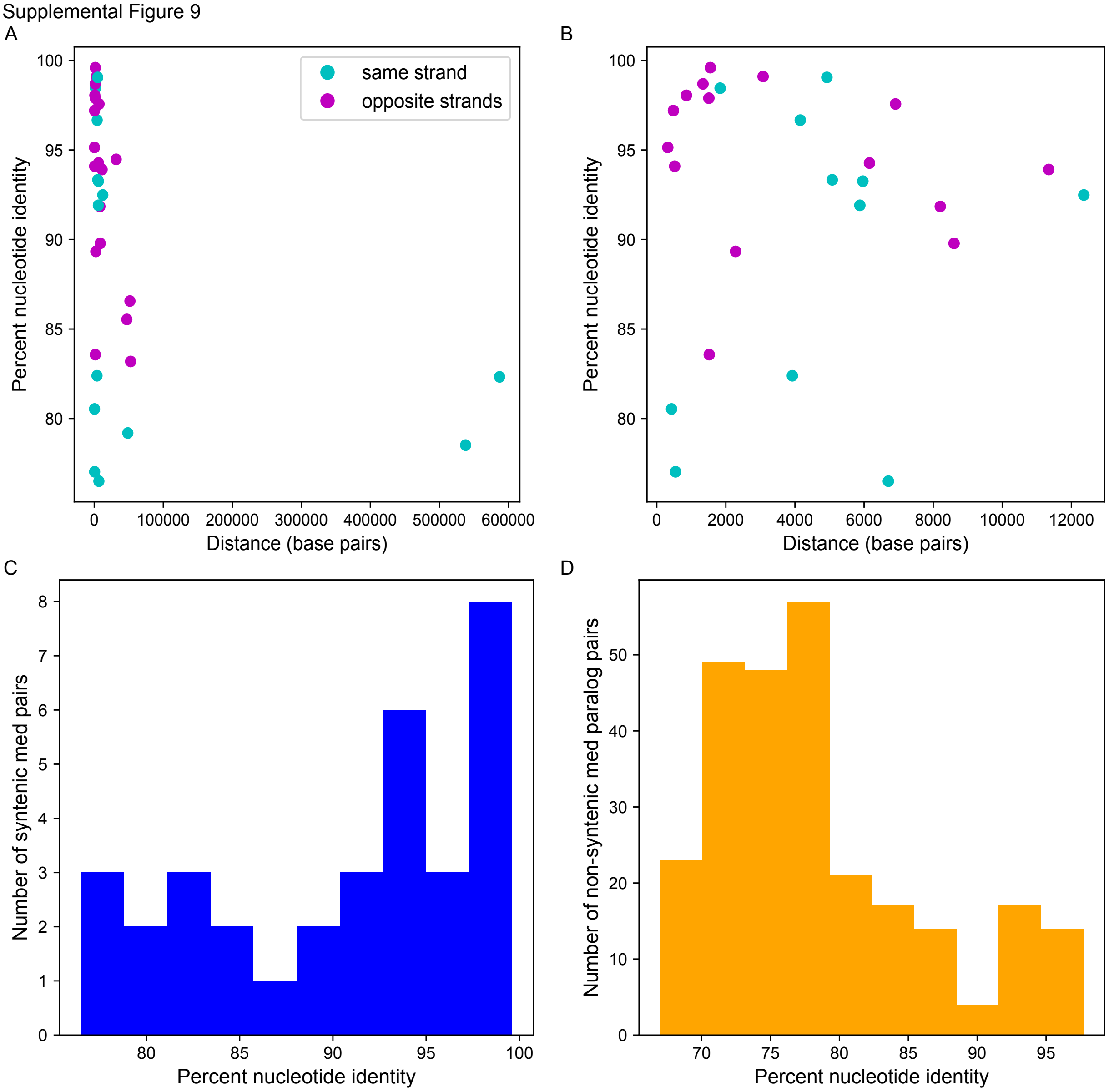


Supplemental Figure 1.7. Contig/scaffold/chromosome locations of genes encoding *Caenorhabditis* GATA-domain-containing proteins, continued.


Supplemental Figure 8
A
EGL-27 GATA-like DBD consensus sequence
Percent identity
Position in sequence relative to human RERE
B
human RERE GATA-like DBD sequence
YACRHCFTTTSKDNHHGGRENLLLCTDCRIHFKKYGELPPIEKPVDPPPFMFKPVK
Position in sequence

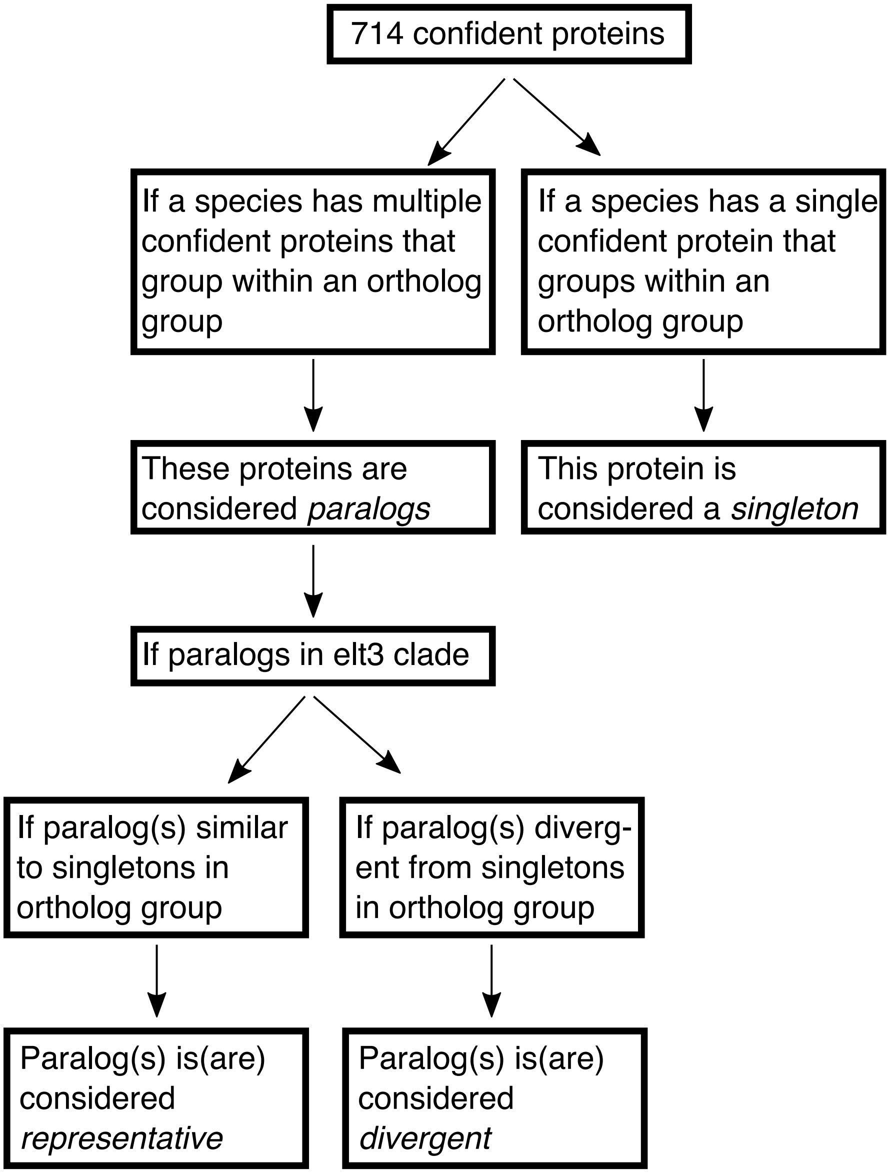
Supplemental Figure 1.8. Comparison of GATA DBD sequences in *Caenorhabditis* EGL-27 and human RERE. (A) The consensus GATA DNA-binding domain (DBD) for *Caenorhabditis* EGL-27. Ten residues not found in any other GATA DBDs in *Caenorhabditis* species are highlighted in light orange. (Gaps were included at positions 17 and 18 to provide alignment with the human RERE protein.) **(B)** The GATA DBD sequence of human RERE. The nine (of the 10) residues specific to the consensus *Caenorhabditis* EGL-27 DBD are highlighted in light blue.

Supplemental Figure 9



Supplemental Figure 1.9. Relatedness of syntenic and non-syntenic *med* paralogs. (A) Plot depicting the degree of identify (percent identity) between all pairs of syntenic *med* paralogs versus the chromosomal distance, in base pairs (bp), between them. Paralogs with the same orientation (on the same strand) are depicted with cyan-colored dots and those on opposite strands are depicted with magenta-colored dots (as noted in the key in the top right). **(B)** Plot depicting the degree of identity (percent identity) between pairs of syntenic *med* paralogs in close proximity to each other (less than 13 kb) versus the distance (in bp) between them. Six *C. brenneri med* paralog pairs and one from *C. latens* were excluded so as to promote better visualization of the distribution of *med* paralogs located closer to each other. Color-coding the same as in A. **(C)** Histogram illustrating the numbers of syntenic *med* pairs (y-axis) versus their relatedness to each other (percent identity, x-axis). **(D)** Histogram illustrating the number of non-syntenic *med* pairs versus their relatedness to each other (percent identity, x-axis).

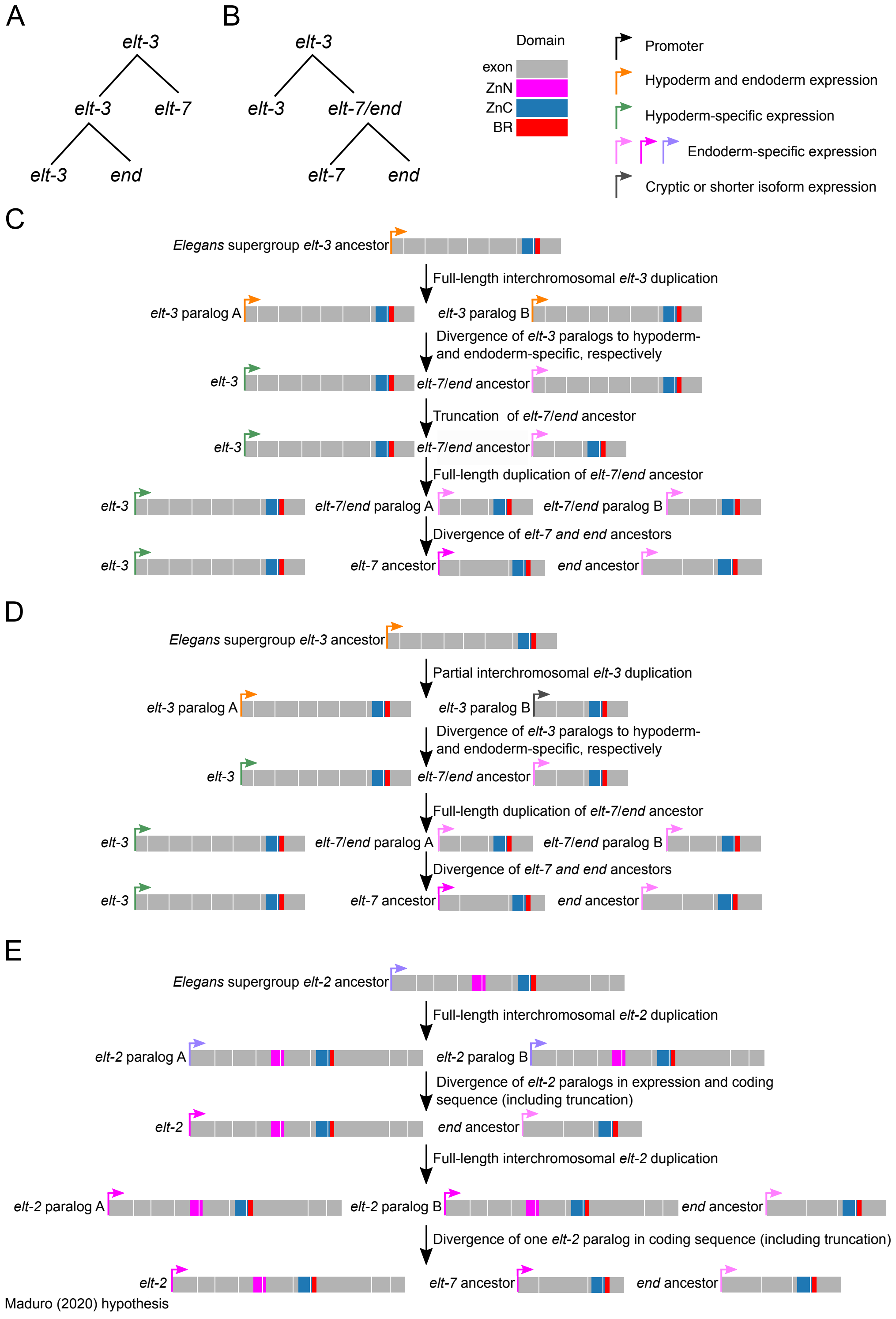
Supplemental Figure 10


714 confident proteins
If a species has multiple confident proteins that group within an ortholog group
If a species has a single confident protein that groups within an ortholog group
These proteins are considered *paralogs*
This protein is considered a *singleton*
If paralogs in elt3 clade
If paralog(s) similar to singletons in ortholog group
If paralog(s) diverg-ent from singletons in ortholog group
Paralog(s) is(are) considered *representative*
Paralog(s) is(are) considered *divergent*

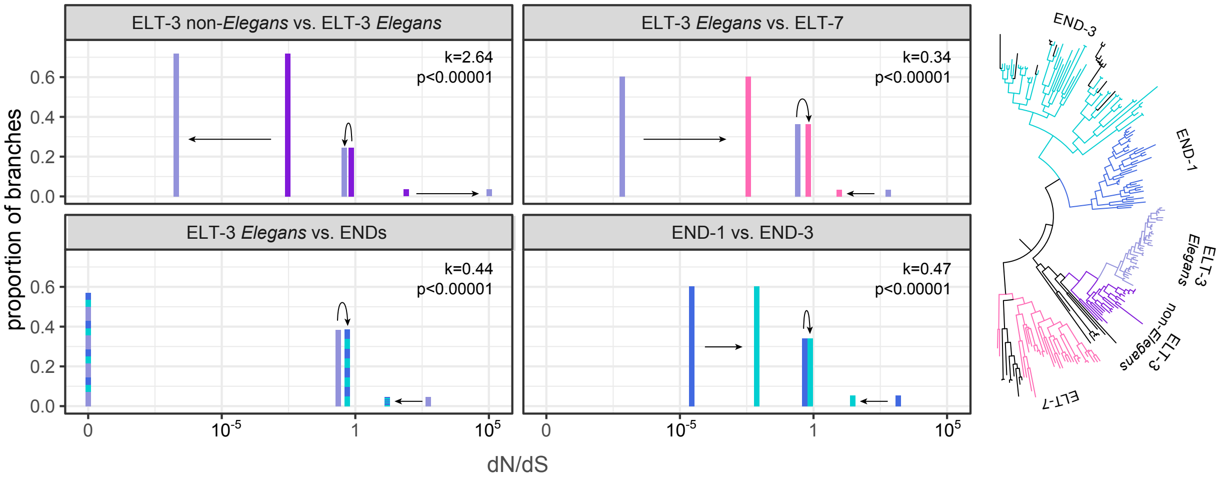
Supplemental Figure 1.10. Classification of paralogs in the Elt3 clade. Decision tree used for classifying paralogs in the elt3 clade as representative or divergent.

Supplemental Figure 1.11. Different gene duplication scenarios that could have contributed to the expansion of GATA factors in the *Elegans* supergroup ancestor. Comparison of possible gene duplication scenarios, one supported by our results (A-D), another proposed by Maduro (2020) (E). **(A)** Scenario involving two duplications of *elt-3,* one which produced the ancestor of *elt-7* and another which produced the ancestor *end* gene. The phylogenic topology findings of this study (see Fig. 1A) support this scenario. **(B)** Scenario involving a single *elt-3* duplication, in which one duplication of *elt-3* produced the ancestor *elt-7/end* gene and then a subsequent duplication of the *elt-7/end* ancestral gene produced the ancestors of the *elt-7* and *end* genes. Data on gene structures (Fig. 2F; Supp. Fig. 3J-M), chromosome locations of elt3 clade genes (Supp. Fig. 7I-L; Fig. 4), expression patterns of *elt-3* and *elt-2* in *C. angaria* (Fig. 5), as well as previously published additional data (see Results and Discussion) support this scenario. Details of how this scenario might have played out are provided in (D) and (E). **(C)** Proposed details of a scenario involving a single duplication of a full-length *elt-3*. Alternatively, if instead of one, two full-length *elt-3* duplications occurred, then the first three steps of this scenario could occur twice to produce the *elt-7* and *end* ancestor genes. **(D)** Proposed details of a scenario involving a single, partial duplication of *elt-3*. Alternatively, if instead of one, two partial-length *elt-3* duplications occurred, the first two steps of this scenario could occur twice to produce the *elt-7* and *end* ancestor genes. **(E)** Molecular representation of a previously published hypothesis for how two *elt-2* duplications could have produced the *elt-7* and *end* ancestor genes (based on Maduro 2020). The key to color-coding of gene domains and expression patterns is located in the upper right corner of the figure.

Supplemental Figure 11


A
elt-3
elt-3
elt-7
elt-3
end
B
elt-3
elt-3
elt-7/end
elt-7
end
Domain
exon
ZnN
ZnC
BR
Promoter
Hypoderm and endoderm expression
Hypoderm-specific expression
Endoderm-specific expression
Cryptic or shorter isoform expression
C
Elegans supergroup elt-3 ancestor
Full-length interchromosomal elt-3 duplication
elt-3 paralog A
elt-3 paralog B
Divergence of elt-3 paralogs to hypoderm- and endoderm-specific, respectively
elt-3
elt-7/end ancestor
Truncation of elt-7/end ancestor
elt-3
elt-7/end ancestor
Full-length duplication of elt-7/end ancestor
elt-3
elt-7/end paralog A
elt-7/end paralog B
Divergence of elt-7 and end ancestors
elt-3
elt-7 ancestor
end ancestor
D
Elegans supergroup elt-3 ancestor
Partial interchromosomal elt-3 duplication
elt-3 paralog A
elt-3 paralog B
Divergence of elt-3 paralogs to hypoderm- and endoderm-specific, respectively
elt-3
elt-7/end ancestor
Full-length duplication of elt-7/end ancestor
elt-3
elt-7/end paralog A
elt-7/end paralog B
Divergence of elt-7 and end ancestors
elt-3
elt-7 ancestor
end ancestor
E
Elegans supergroup elt-2 ancestor
Full-length interchromosomal elt-2 duplication
elt-2 paralog A
elt-2 paralog B
Divergence of elt-2 paralogs in expression and coding sequence (including truncation)
elt-2
end ancestor
Full-length interchromosomal elt-2 duplication
elt-2 paralog A
elt-2 paralog B
end ancestor
Divergence of one elt-2 paralog in coding sequence (including truncation)
elt-2
elt-7 ancestor
end ancestor
Maduro (2020) hypothesis

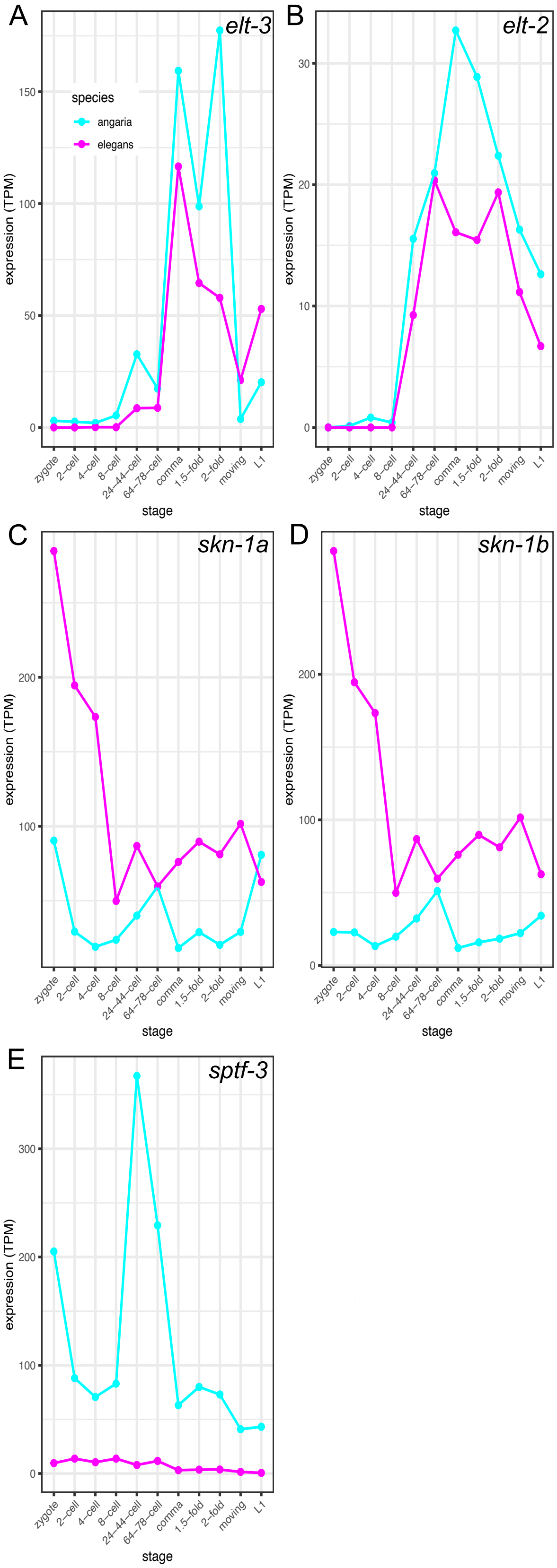
Supplemental Figure 12



Supplemental Figure 1.12. Comparison of selection intensity along elt3 clade branches after *elt-3* expansion in *Elegans* supergroup ancestor. The RELAX test (see Methods) was used to compare the intensity of selection imposed on elt3 clade genes. Phylogenetic tree branches used for comparisons are color-coded as per the phylogenetic tree depicted on the right side of the figure (i.e., END-3 branches are in teal, END-1 branches are in blue, *Elegans* supergroup ELT-3 branches are in light purple, non-*Elegans* supergroup ELT-3 branches are in dark purple, and ELT-7 branches are in pink). Branches of divergent paralogs (see Supp. Fig. 10) were not included. The ratio of the number of non-synonymous substitutions per site to the number of synonymous substitutions per site (dN/dS) is depicted on the x-axis (the scale of which is the same for all four panels). The proportion of branches in each of the three dN/dS rate categories per test is depicted on the y-axis (the scale of which is the same for all four panels). The top left panel depicts the RELAX test results comparing selection on non-*Elegans* supergroup ELT-3 branches (dark purple) to that on *Elegans* supergroup ELT-3 branches (light purple). The top right panel shows the RELAX results comparing selection on *Elegans* supergroup ELT-3 branches (light purple) to that on ELT-7 branches (pink). The bottom left panel depicts the RELAX results comparing selection on *Elegans* supergroup ELT-3 branches (light purple) to that on both the END-1 and END-3 (END) branches (alternating teal and blue). The bottom right panel shows the RELAX results comparing selection on END-1 branches (blue) to that on END-3 branches (teal). The exponentiation factors (k) and p-values for differences in dN/dS rate category distributions for each test is shown in the top right corner of each panel. Arrows in the panels indicate the direction of selection pressure; arrows pointing towards a dN/dS ratio of one indicate relaxed selection, those pointing toward values less than one indicate increasing negative selection, and those pointing to values greater than one indicate increasing positive selection.

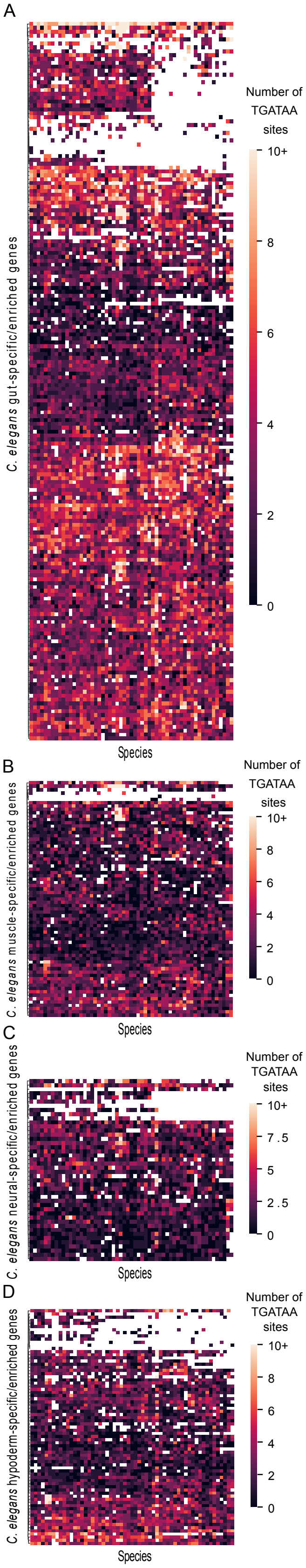
Supplemental Figure 1.13. RNA-seq analysis of *C. angaria* and *C. elegans* genes of interest. RNA-seq data (Macchietto et al. 2017) from 10 different stages of embryo development and the first larval stage (L1), in *C. angaria* and *C. elegans*, were used as RNA-seq inputs. The developmental stages that were sampled are listed (in chronological order) on the x-axis. The number of transcripts corresponding to each gene, normalized as transcripts per million (TPM), is shown on the y-axis. In all panels, *C. angaria* data is shown in cyan and *C. elegans* data is shown in magenta. **(A)** *elt-3* mRNA expression. **(B)** *elt-2* mRNA expression. **(C)** *skn-1* (isoform A in *C. angaria*) mRNA expression. **(D)** *skn-1* (isoform B in *C. angaria*) mRNA expression. **(E)** *sptf-3* mRNA expression.

Supplemental Figure 13



Supplemental Figure 1.14. Conservation of TGATAA sites in putative promoters of gut-, muscle-, neural-, and hypoderm-specific/enriched orthologs. Heatmaps of the number of TGATAA sites in the putative gene promoter regions of orthologs of *C. elegans* specifically expressed, or enriched for expression, in various tissues. The columns comprising the x-axis represent each species, in the same order (left to right) as the listing of species in the phylogeny shown in Figure 1B. Each row on the y-axis represents the putative promoter region of a *C. elegans* gene specifically expressed, or enriched in expression, in gut **(A)**, muscle **(B)**, neural **(C)**, or hypoderm **(D)** tissue, ordered using hierarchical clustering with Euclidean distance metric. The color key is shown to the right of each heatmap plot. To make the color scaling more informative, the few promoter regions that had more than 10 TGATAA sequences are depicted as having only 10 TGATAA sites within their promoters**.** White space in heatmaps shows which species we did not find orthologs for.

Supplemental Figure 14



1.4 Discussion

Radiation of GATA-domain-containing proteins in the Elegans *supergroup ancestor*

We found additional evidence for Maduro's (2020) proposed expansion of GATA factors – *end-1*, *end-3*, *elt-7*, and *med* – in the ancestor of the *Elegans* supergroup and predict that a duplication of the *Elegans* supergroup ancestor *rcor-1* produced the *rcor-1* and *spr-1* orthologs found in contemporary species (Fig. 1.1; see Results). Moreover, we found that a tandem *elt-6* duplication likely produced *egl-18* in the ancestor of the *Guadeloupensis* group and the *Elegans* supergroup (Fig. 1.1; see Results). It is intriguing that at least five duplications of GATA-domain-containing regulatory proteins were likely fixed in the *Elegans* supergroup ancestor, and it will be interesting to look genome wide to see whether other protein families also expanded at this time, particularly those involved in development. There is a lot of variation in protein-coding gene count between the 48 likely non-heterozygous *Caenorhabditis* draft genomes analyzed here (Stevens 2020), suggesting that protein-coding genes duplication and losses are frequent across the genus. Moreover, the finding that the spontaneous gene duplication rate in *C. elegans* was three orders of magnitude larger (Konrad et al. 2018) than the point mutation per nucleotide site rate (Denver et al. 2009), suggests that gene duplication would be a quicker way to increase the gene expression of a locus than nucleotide substitution (Lipinski et al. 2011). In fact, evidence in multiple natural populations support this idea (e.g., Nair et al. 2007; Perry et al. 2007). Not only did this radiation re-wire at least one dGRN (Maduro 2020; see below) but it is often associated with an increase in copy number variations of these *Elegans* supergroup-specific paralogs relative to the *Caenorhabditis*-wide paralog

(Supp. Table 2). This radiation is not associated with any known change in the environment or morphology of these animals and so the evolutionary forces driving this change remain unknown. Overall, this GATA-domain containing protein radiation likely re-wired endoderm development in the *Elegans* supergroup ancestor, creating a dGRN with at least five transcription factors from the same family (Maduro 2020; see below). This concentration of a single type of transcription factor in a GRN – especially one as temporally and spatially restricted as the endoderm network – is extraordinarily rare and creates the potential for extensive developmental system drift.

Expansion of GATA factors in the Guadeloupensis *group*

We found many (22 and 16 respectively) GATA domain-containing proteins in the draft genomes of *C. uteleia* and *C. guadeloupensis*, which is similar to the median number that we found in *Elegans* supergroup species of 19 (Supp. Table 2). Many of these proteins are paralogs and some of the *C. uteleia* proteins are highly divergent, with long branches that do not robustly group into any clade in our phylogeny (Supp. Fig. 1.1). These proteins are either species-specific duplications or have diverged extensively since the *Guadeloupensis* group branched off from the *Elegans* supergroup or since *C. uteleia* and *C. guadeloupensis* diverged (Stevens 2020; Fig. 1.1B). Interestingly, three *C. uteleia* and five *C. guadeloupensis* GATA factors group basally to the ELT-3 ortholog group within the elt3 clade and two *C. uteleia* GATA factors group basally to the ELT-7 ortholog group within the elt3 clade (Supp. Fig. 1.1). Once there is a high quality genome for the third *Guadeloupensis* group species, *C.* sp. *45* and potential heterozygosity for the non-inbred *C. guadeloupensis* genome has been removed (or an inbred strain has been sequenced), we will have more clarity as to when these duplications occurred. It will be especially interesting to discover where these proteins are expressed and their genomic locations, particularly if these *elt-3* paralogs independently subfunctionalized their roles in hypoderm and endoderm development.

elt-3 *duplications in* C. astrocarya

We found two divergent *elt-3* paralogs in *C. astrocarya* that are placed on long sister branches basally within the ELT-3 ortholog group (Fig. 1.1). One paralog is the adjacent gene to the representative *elt-3* paralog (Supp. Fig. 1.7I), suggesting a relatively recent tandem duplication produced this gene. However, the other divergent paralog is not syntenic with these genes (Supp. Fig. 1.7I), suggesting it is likely a more ancient duplication. There is limited phylogenetic signal for the *C. astrocarya* species branch placement in the *Caenorhabditis* species phylogeny - the most support being as sister to the *Guadeloupensis* group and *Elegans* supergroup (Stevens 2020). Therefore, it is also possible that an *elt-3* duplication occurred in the common ancestor of *C. astrocarya*, *Guadeloupensis* group, and *Elegans* supergroup species, but that these paralogs have been on different evolutionary trajectories since then.

GATA domain intron/exon structure supports the clades in our phylogeny but is not sufficient to resolve the complete Caenorhabditis *GATA factor birth order*

Recent non-retrotransposed gene duplicates will often share gene structure characteristics, such as locations of introns and domains, which can help untangle the evolutionary history of a gene family (Gillis et al. 2008; Maduro 2020; Eurmsirilerd & Maduro 2020). Maduro (2020) found conservation of an intron in the coding sequence for the ZnF of all *end-1*, *end-3*, and *Japonica* group *med* homologs he examined and loss of this intron in *Elegans* group *med* homologs, which mostly lack introns completely. The same intron location, immediately before the last seven nucleotides that code for the ZnF, was also found in *C. elegans elt-2* CF (Fukushige et al. 1998) and *elt-7* (Sommermann et al. 2010) and some additional (but not specified) *Elegans* supergroup species *elt-2* and *elt-7* orthologs (Maduro 2020). In addition, this splice site location was found in most nematode *elt-2* and *elt-3* orthologs examined, including five *Caenorhabditis* species (Eurmsirilerd & Maduro 2020). However, this differs from most nematode Clade I *elt-2* orthologs which instead have a CF intron at the same location as most nematode *elt-1* CFs, the *Drosophila melanogaster* (fruit fly) GATA factor *Grain* CF, and the *Gallus gallus* (chicken) GATA-1 CF (Eurmsirilerd & Maduro 2020), supporting the elt1/2 clade found in our phylogeny (Fig. 1.1A). We found an intron immediately before the last seven base pairs that code for the ZnF in most of the additional *end-1, end-3*, *elt-2*, *elt-3*, and *Japonica* group *med* orthologs included in this study (Fig. 1.2), corroborating the findings of both Maduro (2020) and Eurmsirilerd & Maduro (2020). Only one *C*. sp. *51 end-3* paralog, two *C. macrosperma med* paralogs, and three *elt-2* orthologs have lost this intron (Fig. 1.2). Moreover, we found an intron at this location in all *elt-7*s (Fig. 1.2). This shared intron between *end-1*, *end-3*, *elt-2*, *elt-3*, *elt-7*, and *Japonica* group *med* orthologs suggests common ancestry. More recent shared ancestry between *end-1*, *end-3*, *elt-3*, and *elt-7* is captured by our phylogenetic analysis (Fig. 1.1), our GATA domain pHMMs (Fig. 1.3A), and the pHMM scores (Fig. 1.3B), but none of these methods provide evidence of additional homology for *elt-2* and *Japonica* group *med* orthologs, suggesting that extensive sequence divergence has obscured any evolutionary relationship.

We did not find this conserved splicing location in any of the other GATA domain-containing proteins in our study. However, we did find other conserved intron locations within and between clades. In corroboration with previous results (Eurmsirilerd & Maduro 2020) we found conservation of the *elt-1* CF intron location in the basic regions of all singleton confident *elt-6s*, *egl-18*s, and *elt-1* CFs (Supp. Fig. 1.2). This shared intron supports the more recent common ancestry of *elt-6* and *egl-18* orthologs in our phylogeny (Fig. 1.1). However, more distant evolutionary history is still not clear. This is because the ancestral elt6 clade GATA factor likely originated at the base of Chromadoria, and the only two GATA factor orthologs in Clade I nematodes, *elt-1* and *elt-2* orthologs, both had an intron conserved at the same position in their CF BRs at this time (Eurmsirilerd & Maduro 2020). Thus, it is unclear whether an *elt-1* or an *elt-2* duplication produced the elt6 clade ancestor (Eurmsirilerd & Maduro 2020).

Caenorhabditis *GATA domain hidden Markov model profiles can be used to help identify orthologs in newly sequenced* Caenorhabditis *and other nematodes*

We created hidden Markov model profiles (pHMMs) (Eddy 2020) representing the GATA DNA-binding domains of the 12 ortholog groups in our phylogeny (Fig. 1.3A). We searched for alignments of these pHMMs against all proteins used to create these profiles and found that even these relatively short profiles could clearly distinguish between the 12 ortholog group domains, since an ortholog group profile scored said orthologs from the group highest, as expected (Fig. 1.3B). Moreover, the profiles of ortholog groups in larger clades in our tree, scored these orthologs next highest (e.g., EGL-18 profile scored ELT-6 profile second highest) (Fig. 1.3B), suggesting that at least some of the phylogenetic signal that these ortholog groups are related is captured in the 55 residues of these profiles. Additionally, we used these pHMMs to score paralogs that were not used to create the profiles. Similar to singletons, paralogs scored highest on their ortholog groups pHMM (data not shown). The inclusion of up to 58 *Caenorhabditis* species GATA domain sequences to create these pHMMs, the comparable speeds of HMMER searches to those of BLAST, and the specificity of these profiles suggests that they can be used to identify orthologs and even divergent paralogs in newly sequenced *Caenorhabditis* species and likely other nematode species as well. Our search for GATA factor orthologs used the PROSITE GATA-type zinc finger domain profile, which was designed to detect distantly related GATA domain-containing proteins, which is why in addition to canonical GATA factor domains from fungi, plants, and animals this profile included *C. elegans* EGL-27 and SPR-1 atypical GATA domains (prosite.expasy.org). In addition to the *Caenorhabditis* GATA factor ortholog group specific profiles (Fig. 1.3A) we created a more general *Caenorhabditis* GATA factor profile that included ELT-6, EGL-18, ELT-1 CF, ELT-2 CF, ELT-3, ELT-7, END-1, END-3, and MED GATA domain sequences in its creation (data not shown). We hope that these *Caenorhabditis* GATA factor pHMMs will be a valuable resource to the *Caenorhabditis* and greater worm community.

Prediction of the Elegans *supergroup ancestral endoderm developmental gene regulatory network*

Maduro (2020) hypothesized that two *elt-2* duplications in the *Elegans* supergroup ancestor produced an ancestral *end/med* and an ancestral *elt-7* gene (Supp. Fig. 1.11C). This hypothesis was supported by the fact that *elt-2*, *elt-7*, *end-1*, *end-3*, and *med* orthologs all function in the *C. elegans* endoderm dGRN (Zhu et al. 1997; Maduro et al. 2005a; McGhee et al. 2007, 2009; Sommermann et al. 2010; Dineen et al. 2018), that *elt-2* orthologs are conserved throughout the *Caenorhabditis* genus, and *elt-2* orthologs share a conserved intron location in their ZnFs with *end-1*, *end-3*, *elt-7*, and *Japonica* group *meds*. However, *elt-3* orthologs are also conserved in all sequenced *Caenorhabditis* species and have the same conserved intron in their ZnFs (Supp. Fig. 1.2). Moreover, our phylogenetic inference places ELT-3 orthologs as sharing a more recent common ancestor with ELT-7, END-1, and END-3 orthologs compared to all other *Caenorhabditis* GATA-domain-containing groups (Fig. 1.1) and our smFISH analysis found *C. angaria elt-3* mRNA expressed in the early endoderm (Fig. 1.5B,C). This evidence suggests that one or two *elt-3* duplications in the *Elegans* supergroup ancestor produced the ancestors of the *end* and *elt-7* genes (see below). We found no additional evidence for the *med* ancestor originating from an *elt-2* duplication. In fact, we think there is more evidence that the *med* ancestor originated from an elt3 clade duplication in the *Elegans* supergroup ancestor (see below).

Our promoter and expression analyses suggest that endoderm specification in the *Elegans* supergroup ancestor, before the *elt-3* radiation, started with the Sp1 family transcription factor SPTF-3 and/or SKN-1 or another bZIP transcription factor activating *elt-3* expression in the first two endoderm cells in a 13 cell embryo (Fig. 1.6A; Fig. 1.5B,C; Fig. 1.8). SKN-1 is known to help initiate the *C. elegans* (Bowerman et al. 1992; Maduro et al. 2001) and the *C. briggsae* (Lin et al. 2009) endoderm dGRNs, while SPTF-3 and SKN-1 are hypothesized to function in the endoderm dGRNs of most, if not all, *Elegans* supergroup species (Maduro 2020). There is also evidence of *C. angaria skn-1* and *sptf-3* mRNA in early embryogenesis from whole embryo single-cell RNA-seq (Macchietto et al. 2017; Supp. Fig. 1.13C-E). We predict that ELT-3 then activates *elt-2* expression by binding to GATA sites in the *elt-2* promoter (Fig. 1.8). *elt-3* endoderm expression stops before the bean stage (Fig. 1.5B,C), while *elt-2* expression remains endoderm-specific (Fig. 1.5A,C), likely for the lifetime of the worm like it does in *C. elegans* (McGhee et al. 2007, 2009). We found more TGATAA sites in the promoters of *C. elegans* gut-specific or enriched *Caenorhabditis* and *Diploscapter* orthologs than in the promoters of orthologs for other tissues (Fig. 1.7; Supp. Fig. 1.14), suggesting that *elt-2s* role of regulating hundreds of genes in the *C. elegans* intestine is likely conserved throughout the genus (Fig. 1.8).

It is unclear whether one or two elt-3 *duplications in the* Elegans *supergroup ancestor produced the* elt-7 *and* end *gene ancestors*

We found strong evidence that ELT-3 orthologs share a more recent common ancestor with ELT-3, END-1, and END-3 orthologs in the *Elegans* supergroup ancestor compared to other GATA-domain containing proteins (Fig. 1.1; Fig. 1.3A,B). Therefore, we posited that one or two *elt-3* duplication in the *Elegans* supergroup ancestor produced the *elt-7* and *end* gene ancestors (Supp. Fig. 1.11A,B). However, it is less clear whether the *end* genes arose from a duplication of *elt-7* or *elt-3*. The elt3 clade topology has ELT-7 orthologs branching off of ELT-3 orthologs before the END ancestor branches off of ELT-3 orthologs (Fig. 1.1A), suggesting that the END orthologs are more closely related to ELT-3 orthologs than to ELT-7. This supports a scenario where two *elt-3* duplications produced the *elt-7* and *end* ancestors instead of a single *elt-3* duplication producing the ancestor of both the *elt-7* and *end* genes followed by a second duplication of this *elt-7*/*end* gene producing the *elt-7* and *end* ancestors. However, the gene structures and chromosome locations of elt3 clade genes support the later scenario. The gene structures of *end-1*, *end-3*, and *elt-7* orthologs are more similar to each other than to those of *elt-3* orthologs because they are predominantly encoded in only four exons and code for shorter proteins compared to *elt-3* orthologs which are mainly coded by eight exons (Fig. 1.2F; Supp. Fig. 1.3J-M). The most parsimonious explanation for this would be that there was a single shortening event giving rise to the *elt-7/end* ancestor. Whether this shortening occurred by a partial duplication of *elt-3* or after a full length *elt-3* duplication is also not clear. Moreover, the five species with chromosome-level assemblies all have *end-1*, *end-3*, and *elt-7* orthologs on the same chromosome, whereas *elt-3* orthologs are found on a different chromosome (Supp. Fig. 1.7I-L). Based on our synteny analysis, we predict that these chromosome locations are conserved in most *Caenorhabditis* species that have these orthologs (Fig. 1.4).

The profile scores also supported a closer relationship between the END genes and ELT-7 since the END pHMMs scored ELT-7 orthologs slightly higher than ELT-3 orthologs (Fig. 1.3B). But we did not find any additional sequence homology to support this scenario. In fact, we found a similar N-terminus poly-serine (poly-S) region in ELT-3 orthologs (data not shown) as previously found in END-1 and END-3 orthologs (Maduro et al. 2005a; Maduro 2020). Eurmsirilerd & Maduro (2020) defined a poly-S region as at least six serines in ten adjacent residues and we found such a motif in 30 of 35 *Elegans* supergroup ELT-3 orthologs (data not shown). The five *Elegans* supergroup ELT-3 orthologs without a poly-S motif have a poly-S/T motif (defined as at least six serines or threonines in ten adjacent residues), therefore it is likely that the ancestral ELT-3 in the *Elegans* supergroup ancestor had a similar motif. Similar to Maduro (2020), we find an N-terminal poly-S motif in most (31 of 35 species) END-1 and (26 of 35 species) END-3 orthologs (data not shown). If the ELT-3 and END poly-S motifs are homologous, it supports a scenario where *elt-3* fully duplicated in the *Elegans* supergroup ancestor to produce the *end* ancestor, followed by sequence loss to four exons. We do not find a similar poly-S motif at the beginning of ELT-7 orthologs (data not shown). If *elt-7* and the *ends* arose from the same *elt-3* duplication, then their ancestor quickly duplicated again since they experienced different trajectories of mutation and deletion. However, if the *elt-7* and *end* ancestors resulted from two different *elt-3* duplications, then the lack of a N-terminal poly-S motif in ELT-7 orthologs could be the result of a partial *elt-3* duplication producing the ancestral *elt-7* gene. The ambiguity in the precise birth order of *elt-7* and *end* gene ancestors may reflect the fact that this small radiation happened in an evolutionarily short period of time such that both *elt-7* and *end* orthologs are about equally diverged from *elt-3* orthologs but in different ways.

Elegans supergroup ancestral elt-3 paralogs likely subfunctionalized into hypoderm- and endoderm-specific GATA factors

Given the expression pattern of *C. angaria elt-3* mRNA in early endoderm and hypoderm cells (Fig. 1.5B,C), the hypoderm-specific function and expression of the *C. elegans elt-3* ortholog (Gilleard & McGhee 2001; Gilleard et al. 1999), and the sequence conservation of *elt-3* orthologs (Fig. 1.1; Fig. 1.3A; Supp. Fig. 1.6F), we hypothesize that the ancestral *elt-3* gene in the ancestor of the *Elegans* supergroup species functioned in both hypoderm and endoderm development. If true, then the subsequent duplications led to subfunctionalization such that one *elt-3* paralog become hypoderm-specific and the *elt-7*/*end-1*/*end-3* ancestral paralogs became only expressed in endoderm cells. Because the *C. elegans elt-3* ortholog is not essential for hypoderm development (Gilleard & McGhee 2001), we expect that the *Elegans* supergroup ancestral *elt-3* was also probably not essential for hypoderm development and therefore an *elt-3* duplication may have been tolerated better than the duplication of an essential gene could have been. Moreover, if the *Elegans* supergroup ancestral *elt-3* had modular subfunctions (i.e. in hypoderm and endoderm development) it may have been easier to subfunctionalize into endoderm and hypoderm-specific paralogs. Examining the expression and function of more non-*Elegans* supergroup *elt-3* orthologs should shed more light on the process that led from one putatively multifunctional ancestor to four specialized descendants.

Any endoderm function of non-Elegans supergroup elt-3 orthologs depends on the protein sequence and not just expression

As mentioned previously, many *C. elegans* GATA factors have redundant functions and can at least partially complement each other (e.g, Maduro et al. 2001; Gilleard & McGhee 2001; Koh & Rothman 2001; Koh et al. 2002; Maduro & Rothman 2002; Maduro et al. 2005a; Wiesenfahrt et al. 2016). Moreover, these genes can work across species: some *C. briggsae* and *C. remanei* GATA factors are able to rescue loss of their *C. elegans* orthologs (Maduro & Rothman 2002; Coroian et al. 2006) and *C. elegans end-1* is able to convert *Xenopus laevis* ectoderm cells into endoderm cells (Shoichet et al. 2000). Some of this functional redundancy is due to a recent common ancestor between the respective GATA factors, whereas other cases demonstrate at least some functional conservation over hundreds of millions of years. This deep conservation is likely primarily due to the conservation of the GATA DBD's ability to bind to canonical HGATAR sites in DNA, since all non-MED GATA factors have been found to bind to these sites (e.g., Merika & Orkin 1993; Omichinski et al. 1993; Shim et al. 1995; Hawkins & McGhee 1995; Starich et al. 1998; Bates et al. 2008; Chen et al. 2012; Narasimhan et al. 2015; Wiesenfahrt et al. 2016). However, most GATA factors have diverged extensively outside of the DBD (Lowery & Atchely 2000; Gillis et al. 2008; Maduro 2020; Eurmsirilerd & Maduro 2020), and any effector domains in these regions are far more poorly characterized. This is typical for transcription factor families: they are defined by their similar DBDs but have effector domains that generally evolve more quickly (Peter & Davidson 2015).

If the *C. elegans elt-2* ortholog is overexpressed under the control of the *C. elegans end-1* promoter, it can compensate for loss of *end-3*, *end-1*, *elt-7*, and *elt-4* (Wiesenfahrt et al. 2016). *C. elegans elt-3* cannot do this. This is especially surprising considering that ELT-3 orthologs are more closely related to END-3, END-1, and ELT-7 orthologs than *elt-2* orthologs (Fig. 1.1). This suggests that the *C. elegans* ELT-3 (and likely all *Elegans* supergroup ELT-3 orthologs) has lost the ability to specify endoderm even when ectopically expressed there and even though it can bind TGATAA sites (Narasimhan et al. 2015). However, we found that the *C. angaria elt-3* ortholog mRNA is expressed in early endoderm cells (Fig. 1.5B,C) in a pattern reminiscent of *end-1*, which suggests that it probably also functions there. Not only did the *Elegans* supergroup ancestral *elt-3* likely subfunctionalize its expression pattern (see above), something else must of changed about the coding region of its descendants such that one branch preserved its capacity to function in the endoderm while the other branch lost this. To investigate this we compared the protein sequences of *Elegans* supergroup versus non-*Elegans* supergroup *elt-3* orthologs. We did not find any obvious differences in the *Elegans* supergroup versus non-*Elegans* supergroup *elt-3* DBDs (Supp. Fig. 1.6F), however we did find a highly conserved and possible SUMOylation site (Chang et al. 2018) towards the N-terminus of most non-*Elegans* supergroup *elt-3* orthologs, and this [VIA]KE[ED] motif has been lost from all *Elegans* supergroup *elt-3* orthologs (data not shown). We speculate this post-translation modification could be involved in an endoderm-specific protein-protein interaction in non-*Elegans* supergroup *elt-3* orthologs.

Did the ancestral med *originate from an elt3 clade gene duplication?*

As mentioned above, Maduro (2020) speculated that the *med* ancestor arose from an *end-1* or *end-3* ancestral duplication in the *Elegans* supergroup ancestor.

Support for this hypothesis only considered GATA factors and not all GATA-domain containing proteins like this study did. However, our results also support a GATA factor origin of the MEDs since the MED ortholog groups ancestral node is 0.76 substitutions per site away from the ancestral canonical GATA factor node but 2.14 substitutions per site away from the split between the EGL-27 ortholog group and the rcor1 clade (Fig. 1.1). Moreover, the node that the MED ortholog group branches off from its most recent common ancestor with canonical GATA factors is well-supported (100% ultrafast bootstrap support (Minh et al. 2013; Hoang et al. 2018)) (Fig. 1.1). We also found that MED DBDs have more canonical GATA factor residues conserved than EGL-27, RCOR-1, and SPR-1 GATA-like DBDs (Fig. 1.3C).

The long branches and many species-specific paralogs found in the MED ortholog group (Fig. 1.1) suggest that the *meds* turnover quickly, and this was also proposed previously (Maduro 2020). This quick turnover has likely erased any strong evidence for the homologous relationship between MED orthologs and other GATA factors. This is supported by the topology of our tree where the MED ortholog group does not cluster close to any other GATA-domain-containing protein ortholog groups (Fig. 1.1). Moreover, we did not find evidence for the MED ZnF being more similar to any other GATA domain (Fig. 1.3A,B; Supp. Fig. 1.5). The strongest evidence for the *med* origin (and which is what supports Maduro's (2020) hypothesis) is the location of an intron in the ZnF of most *Japonica* group *meds* (that has been lost from *Elegans* group *meds*), which is found at the same location in the ZnFs of *end-1* and *end-3* orthologs. But an intron at this location is also found in *elt-2*, *elt-3*, and *elt-7* orthologs (Fig. 1.2E,F; Supp. Fig. 1.3I-K). The gene structures of *Japonica* group *meds* are most similar to those of *end-3* orthologs (Fig. 1.2; Supp. Fig. 1.3). Because most *Japonica* group species have at least one *end-3* ortholog/paralog with four exons and because intron loss is more frequent than intron gain in a few *Elegans* group species (Roy & Penny 2006), we expect that the *Elegans* supergroup ancestral *end-3* gene had four exons (Fig. 1.2). However, most *end-3* orthologs/paralogs have only three exons (Supp. Fig. 1.3M), and we think that the *Japonica* group *med* ancestor had three exons (Fig. 1.2). An *Elegans* supergroup ancestral *end-3* with only three exons would be consistent with a full *end-3* duplication producing the *med* ancestor. However, a partial duplication of the *Elegans* supergroup ancestral *end-1*, *elt-2*, *elt-3*, or *elt-7* orthologs is also possible as is a full duplication of any of these ancestral genes followed by coding sequence loss. Since we have shown that three elt3 clade duplications occurred and were fixed in the *Elegans* supergroup ancestor (see above), it is plausible that at least one more elt3 clade duplication could have occurred and be fixed, producing the *med* ancestor**.**

Neofunctionalization of at least the Elegans *group* med *ancestor for binding to a non-GATA DNA site and intercalation into the mesoderm dGRN*

The *C. elegans meds* and at least some *C. briggsae* and *C. remanei meds*, function early in both mesoderm and endoderm dGRNs (Maduro et al. 2001; Coroian et al. 2006). The *C. elegans meds* bind to a non-GATA DNA site (Broitman-Maduro et al. 2005; Lowry et al. 2009). Since some *C. briggsae* and *C. remanei meds* are able to compensate for lack of both the *C. elegans meds*, it is likely that they similarly bind GTATACTYYY instead of canonical HGATAR sites. This non-canonical binding and function in the mesoderm is unique among *C. elegans* GATA factors (Hawkins & McGhee 1995; Shim et al. 1995; Zhu et al. 1997; Gilleard et al. 1999; Koh & Rothman 2001; Gilleard & McGhee 2001; Maduro & Rothman 2002; Fukushige et al. 2003; Narasimhan et al. 2015; Wiesenfahrt et al. 2016; Du et al. 2016). Thus, these novel features likely arose in an ancestral *med*. To narrow down the timeline, it would be fruitful to examine whether *Japonica* group *med* expression, binding, and function mirror that of *Elegans* group *meds* or that of canonical GATA factors. If these novel properties are conserved, it suggests that the ancestral *med* neofunctionalization occurred in the *Elegans* supergroup ancestor.

Another example of developmental system drift in Caenorhabditis

Other than male tails, *Caenorhabditis* species are very similar anatomically (Kiontke et al. 2011; Slos et al. 2017) and are thought to have similar cell lineages as *C. elegans*, since the lineages of *C. briggsae* (Zhao et al. 2008) and *Pristionchus pacificus* (Vangestel et al. 2008) are almost identical to that of *C. elegans* (Sulston et al. 1983). Therefore, we think that the GATA-domain-containing duplications, subfunctionalizations, and neofunctionalizations that we have described here and that rewired dGRNs have had little overt effect on the developmental output of these animals. This phenomenon has been termed developmental system drift (True & Haag 2001) and has been previously documented in *Caenorhabditis* (Félix 2007; Ellis and Lin 2014; Verster et al. 2014; Haag et al. 2018; Maduro 2020). In the case of the *Caenorhabditis* endoderm specification network, we hypothesize that *elt-2* plays a similar role throughout the genus and that *elt-3* functions above it in *non-Elegans* supergroup species and ancestrally. However, there is still a gap in developmental time between maternally deposited factors and the start of *elt-3* expression in *C. angaria*. In *C. elegans*, this gap is filled in by the expression of the *meds* and *end-3* as if the duplications expanded to fill the available developmental time. Either early regulators (maternally deposited or zygotically expressed) in non-*Elegans* supergroup species transmit their positional information directly to *elt-3* or there are analogous intervening genes upstream of *elt-3* that were displaced by the radiating GATA-factors, all without any apparent phenotypic change (Fig. 1.8). We expect that additional examples of developmental system drift in *Caenorhabditis* will continue to be uncovered as development in more species and strains is characterized in greater detail.

1.5 Methods

Identifying GATA-type zinc finger domain-containing proteins and putative homologs of GATA factors

We downloaded the proteome files for 56 *Caenorhabditis* species and two *Diploscapter* species, and the transcriptome files for *C.* sp. *45* and *C*. sp. *47* from Caenorhabditis.org (the *Caenorhabditis* Genome Project) in early 2020. We searched for GATA DNA-binding domains in the proteome and transcriptome files that matched the PROSITE GATA-type zinc finger (ZnF) domain profile PS50114 (prosite.expasy.org) using the pftools3 (https://github.com/sib-swiss/pftools3) pfsearchV3 tool (Schuepbach et al. 2013). We identified 890 proteins with at least one match of score 8 or more and used them (and a few more, see below) for this study.

For any case in which we expected to identify an ortholog of a *C. elegans* GATA-domain-containing protein but did not, we performed reciprocal protein-protein BLAST (BLASTp) searches using the Biopython NcbiblastpCommandline wrapper and/or protein query-translated subject BLAST (tBLASTn) searches using the NcbitblastnCommandline wrapper (Altschul et al. 1990; Camacho et al. 2009; Cock et al. 2009) against the *C. elegans* ortholog, and/or sister species ortholog(s), to identify any GATA domain-containing proteins we may have missed. tBLASTn was performed on downloaded scaffold files for species of interest from Caenorhabditis.org (the *Caenorhabditis* Genome Project) in early 2020. We used an e-value cutoff of 0.001 to non-*C. elegans* species and an e-value cutoff of 0.1 back to *C. elegans* for these BLAST searches. We identified an additional 23 EGL-27 homologs, 13 RCOR-1 homologs, and 12 SPR-1 homologs, and one ELT-7 homolog (in *C. becei*), one MED homolog (in *C. panamensis*, referred to as CPANA.med in Supp. Table 1), another MED homolog (in *C. macrosperma*, referred to as CMACR.med in Supp. Table 1), using this method. This brought the total number of proteins for our phylogenetic analysis to 941.

We suspect that many of the putative homologs of *C. elegans* GATA-domain-containing proteins we identified through BLAST searches were not identified using the PROSITE profile because their ZnF motifs are too divergent. For example, most of the BLAST-identified SPR-1 homologs have a $CX_2CX_{18,20\text{-}21,23}CX_2C$ ZnF motif as opposed to the canonical animal GATA ZnF motif ($CX_2CX_{17}CX_2C$). The ZnFs in these proteins with atypical GATA-like domains, including putative homologs of EGL-27, RCOR-1, and SPR-1, may be diverging due to directional selection or because they are degrading. The ELT-7 and MEDs GATA factor homologs we identified using BLAST but not using PROSITE, on the other hand, contained annotation errors; the annotation of the *C. becei* genome ELT-7 contained a premature stop codon in its ZnF and in the genome annotations of *C. panamensis* and *C. macrosperma* did not have the two putative MED homologs as coding regions.

Although we may have missed additional proteins with GATA-like DNA binding domains in our searches, the genes encoding them would likely be undergoing pseudogenization or on some dramatically different evolutionary trajectory than the sequences that we included in our analysis; it is also possible we might have missed a few other proteins due to additional annotation errors. Additionally, some of the paralogs we included may represent artifacts related to possible heterozygosity in the sequenced strain and the quality and coverage of the genome assemblies for some of the species (Barriere et al. 2009; Haag & Thomas 2015; Stevens 2020), this possibility being most likely for paralogs in the MED ortholog group which often have many highly similar sequences (Fig. 1.1; see Results).

Phylogenetic analysis

Phylogenetic analysis is dependent on alignment of multiple sequences, but sequence alignments can be noisy and dependent on arbitrary factors like sequence direction and the order of sequences in a list. Furthermore, when aligning many sequences, some are going to be alignable over a greater length than others. This is particularly true in our case where all the sequences share a common alignable core GATA ZnF but can be otherwise so different that they fall into different subfamilies.

Within a subfamily, sequences will be alignable over greater distances than between subfamilies, and this is evident in our alignments. A single alignment of multiple sequences, however, provides no way to estimate how variable the alignment between two sequences is at any given residue and therefore no way to weight the information at each aligned position by sequence pair combination for downstream applications like phylogenetic inference or population genetic estimations. To capture the variation in alignability, we followed an approach like Penn and colleagues and generated a heterogeneous set of alignments (Penn et al. 2010). The rationale is that alignable regions should be fairly impervious to variations in the guide tree used to build the alignment or to the sequence direction, while unalignable regions will be sensitive to these manipulations. Across a collection of multiple alignments, the consistent signal from alignable residues will contribute more than the conflicting noise from poorly aligned regions. Our multiple alignments were constructed following 5 steps: (1) divide the sequences into separate groups (see below for justification); (2) build an initial alignment; (3) bootstrap this alignment; (4) generate neighbor joining (NJ) trees from these bootstrapped alignments; (5) serially align the sequences using these NJ trees as guide trees.

Only 884 of the 941 proteins have a well-aligning GATA-domain (the 57 without are marked as "poorly_aligning" in the "fullTreeOrthologGroup" column in Supp. Table 1). We aligned the longest isoforms of these 884 protein sequences using the default options of Clustal Omega (Sievers et al. 2011). By construction these sequences include a common core domain that varies slightly in residues and length; beyond that domain the proteins are much less conserved. Because Clustal Omega uses profile

Hidden Markov model (pHMM) alignment (Sievers & Higgins 2018), it is ideal for aligning sequences like these and generated multiple alignments with excellent alignment of the GATA ZnFs that was much more compact than, for example, an alignment generated using MAFFT (Katoh et al. 2002). However, when we aligned all sequences at once, a few invariably were nearly completely offset from the rest. To overcome this, we serially aligned the sequences in batches. After the first batch of sequences was aligned, an pHMM of the alignment was used to align them with the next batch of sequences, and this was repeated until all sequences were aligned. We also constrained the alignments to always align with the first cysteine in the C-terminal ZnF (CF). These procedures resulted in alignments of the conserved residues in the CF. END-1 orthologs have an extra two residues between the first pair of cysteines in their ZnFs (see Fig. 1.3A); for a consistent alignment, we therefore set the sequence of non-END-1 proteins to CX--XC.

After this initial alignment of 884 sequences, we bootstrapped (Felsenstein 1985) these alignments and used them to make new NJ guide trees (Saitou & Nei 1987). We then used each new guide tree to make a new Clustal Omega (Sievers et al. 2011) alignment using the serial alignment approach (described above) for sequences in both the forward and reverse direction. We repeated this process 10 times and then randomly chose nine of these alignments (five forward; four reverse) and concatenated them to the original one, reversing the orientation of the reversed ones so that all sequences were in the same direction. This procedure effectively weights each position x sequence pair combination by how often they are consistently aligned. We removed any uninformative positions (i.e., invariant or present in no more than 2 sequences) from the alignment. Using IQ-TREE 2 version 2.1.2 ModelFinder (Kalyaanamoorthy et al. 2017) on the Cyberinfrastructure for Phylogenetic Research (CIPRES) Science Gateway V. 3.3 (phylo.org) we identified the VT+F+I+G4 model (variable time (Muller & Vingron 2000), empirical base frequencies from supplied alignment, allowing invariable sites, and discrete Gamma model (Yang 1994) with four rate categories, respectively) as the best model of evolution for this alignment. We then used the IQ-TREE 2 version 2.1.2 tree inference (Minh et al. 2020) to estimate the evolutionary history of these sequences. The resulting maximum likelihood phylogeny is shown in Supplemental Figure 1.1. We used this same procedure, besides the addition of 3000 ultrafast bootstraps (Minh et al. 2013; Hoang et al. 2018), for phylogenetic inference of the 714 protein sequences we deemed “confident” (see below for the protocol used to determine “confident” proteins); the resulting maximum likelihood tree is shown in Figure 1.1.

Clade and ortholog group terminology

We refer to the 12 ortholog groups revealed in our phylogenies (Fig. 1.1; Supp. Fig. 1.1) by the name of the *C. elegans* protein(s) within that group. For example, the group containing the *C. elegans* ELT-3 ortholog, and all the proteins from the other 59 species that we classified as ELT-3 orthologs, is referred to as the ELT-3 ortholog group (Fig. 1.1; Supp. Fig. 1.1). We gave the ortholog groups that grouped adjacent to another ortholog group(s), thus forming larger clades, the name of the most ancient ortholog group within that larger clade, formatted in lower case, without a hyphen. For example, we refer to the larger clade containing the ELT-3, ELT-7, END-1, and END-3 ortholog groups as the elt3 clade (Fig. 1.1; Supp. Fig. 1.1). However, one clade contains two ortholog groups that both contain orthologs from all species included in this study, so their evolutionary history is beyond the scope of our analysis, and we refer to this clade as elt1/2 to represent both the ancient ELT-1 and ELT-2 ortholog groups within it (Fig. 1.1; Supp. Fig. 1.1).

Process for editing gene annotations

Upon visual inspection and comparison of ortholog group sequences it became clear that some of the gene annotations in the files used for this study were likely incorrect because the coding sequences were highly divergent from the majority of their orthologs and less divergent coding sequences were possible with minor adjustments. We therefore selectively edited some annotations to make the sequences more homologous. For example, the initial annotations of a few genes had premature stop codons, but by simply adding an alternative splice site these coding sequences became full-length sequences with much better sequence homology to their orthologs. Additionally, there were examples of multiple genes that had been annotated as a single gene, and parts of genes annotated as multiple genes, and addition of alternative splice sites to these genes resulted in improved homology among their orthologs as well. In total we edited the annotations of 226 genes. We annotated coding sequences instead of exons because the data for a few of the 60 species include untranslated regions (UTRs) in their exon sequences; although most of the data files did not contain this information, we wanted consistent sequences that started from the first coding ATG. Notes on the types of edits made are included under the “editingNotes” column in Supp. Table 1. These edits increased the number of genes we considered “confident” (see below) in using for further studies and the robustness of our phylogeny (compare Fig. 1.1 to Supp. Fig. 1.1).

Identifying confident sequences for further analyses

For all additional analyses of the protein sequences in our phylogeny we focused on a subset of the 884 that based on their sequence features and how robustly they grouped into a clade in our preliminary phylogeny (Supp. Fig. 1.1), best met important criteria for this study. For most gene families, highly divergent protein sequences are likely on different evolutionary trajectories compared to their conserved relatives and may even be pseudogenizing; we therefore had less confidence in those protein sequences and did not include them in further analysis (Supp. Fig. 1.2). For example, we did not include protein sequences that branched off on long branches at the base of an ortholog group, comprised exceptionally long branches, and/or had weak branch support values. The features of protein sequences deemed “not_confident” (see the “confidence” column in Supp. Table 1) can be found in either the “geneQuality” column, the “notes” column, and/or the “fullTreeOrthologGroup**”** column of Supp. Table 1. For example, protein sequences comprising their own very long branches, which did not align well with other protein sequences, and/or did not group robustly in the confident phylogeny (Fig. 1.1), are marked as “strange” in the “geneQuality” column (Supp. Table 1). Proteins that grouped basally in groups or those that are on long branches without a *C. elegans* ortholog (unlabeled proteins in Supp. Fig. 1.1) in the full tree (Supp. Fig. 1.1) are labeled “basal_group_name” (e.g., basal_med) or “divergent”, respectively, in the “fullTreeOrthologGroup**”** column (Supp. Table 1).

To assess gene annotation quality, in early 2020 we downloaded from Caenorhabditis.org (the *Caenorhabditis* Genome Project) the annotation and scaffold files for the 56 *Caenorhabditis* species and two *Diploscapter* species for which we had obtained proteome files for this study and examined the protein-coding and neighboring sequences for each protein. Using a custom Python script, we extracted the coding sequence for each protein from that species' annotation file, starting with the first ATG, if present (see “exonSeq” column in Supp. Table 1). We then marked whether the gene had features listed under the “geneQuality” column in Supp. Table 1 that reduced our confidence in it. These features include: premature stop codons (“prematureStop”); lack of an obvious GATA ZnF (“noZnF”); fewer than 13 amino acid residues coded for after their ZnFs, suggesting an incomplete basic region (“shortBR”); lack of conservation/alignment in the sequences following the ZnF (“noBR”); absence of a start codon (“noMet”); absence of a stop codon (“noStop”); or truncation as compared to its orthologs (“truncatedStart” or “truncatedEnd”). There were also three protein sequences marked as either “nonHomologousBR”, “NsInGene”, or “NsUpsteamOfTruncation” which, as the labels imply, had either a non-homologous basic region, an incomplete sequence (i.e. unknown amino acids, AKA N’s) in the coding sequence, or N’s upstream of a putative gene truncation. If a gene’s annotation passed all our criteria, it was labeled “good” in the “geneQuality” column (Supp. Table 1).

We considered $CX_{2,4}CX_7WX_9CX_2C$ as the canonical GATA ZnF motif for this study because the pattern, of similarly spaced four cysteines (C's) and a tryptophan (W) at position 8 in the ZnF loop, is found in the CFs of all canonical animal GATA factor domains (Teakle & Gilmartin 1998; Lowry & Atchley 2000; Supp. Fig. 1.5). In addition, to ensure that we did not miss any possibly homologous ZnFs, we included the following divergent ZnF motifs in our analysis (see Results): $CX_2CX_{15\text{-}17}CX_2C$ for ELT-2 ortholog N-terminal ZnFs (NFs), $CX_2CX_7WX_8CX_2C$ for some *Japonica* group MEDs and for most EGL-27 orthologs, $CX_2CX_{16\text{-}17}CX_2C$ for all RCOR-1 orthologs, and $CX_2CX_{17\text{-}18,20\text{-}21,23}CX_2C$ for all SPR-1 orthologs (Supp. Fig. 1.5). Protein sequences that lacked the expected GATA-like ZnF motif of the ortholog group that the protein grouped into are marked "not_confident" (see Supp. Table 1; Supp. Fig. 1.2). Of the 31 protein sequences that were added to our analysis based on reciprocal BLASTp searches against orthologs in other species because no significant matches to the PROSITE GATA-type ZnF domain had been found for them among the proteome files (see above), 16 did not have an obvious GATA ZnF ("noZnF") as a result of this analysis either, as expected. Most of the remaining 15 "noZnF" protein sequences appeared to have degrading/non-functional ZnFs since they contained only two or three of the usual four zinc-coordinating cysteines; those that had four ZnF cysteines had fewer than 15 residues between the two cysteine pairs, a pattern which is not found in any canonical GATA factor ZnFs. MED sequences that had multiple ZnFs were also considered "not_confident" (see Supp. Table 1) because most MEDs (and all *Caenorhabditis* GATA factors in the non-elt1/2 clade) have only a single ZnF (this study; Lowry & Atchely 2000; Maduro 2020; Eurmsirilerd & Maduro 2020).

Overall, proteins that grouped robustly into a clade in our phylogeny, were classified as “good” in the “geneQuality” column in Supplemental Table 1 and contained an expected GATA(-like) ZnF motif for the ortholog group that the protein clustered into were deemed “confident”. A total of 714 of the 884 proteins with well-aligning GATA-domains fit this description and were used for further analyses.

Process for classifying proteins as singletons, paralogs, representative paralogs, or divergent paralogs

We classified the 714 proteins we were most confident in (see above) as *singleton* or *paralog*, depending on how many proteins from each species clustered into the 12 ortholog groups in our phylogeny (Fig. 1.1). For example, a species with only a single ELT-3-like sequence grouping in the ELT-3 ortholog group was considered a singleton ELT-3 ortholog whereas a species with multiple ELT-3-like sequences that robustly grouped into the ELT-3 ortholog group were deemed paralog ELT-3s. Other than in the MED ortholog group, which was comprised primarily of paralogs, the elt3 clade in our phylogeny contained the most confident paralogs (Supp. Table 2), and we were especially interested in how the elt3 clade genes had evolved (see Results: Re-wiring of the *Elegans supergroup* endoderm developmental gene regulatory network)**.** We therefore further categorized elt3 clade paralogs as either *representative* or *divergent* depending on how conserved the paralog's sequence was in relation to the sequences of singletons in the same ortholog group (Supp. Fig. 1.10; Supp. Table 1); if an individual paralog within a species had noticeably higher levels of conservation to singleton orthologs than did other paralogs, we choose that paralog as the representative one and labeled the others divergent. If multiple paralogs within a species had approximately equal levels of conservation to singleton orthologs, they were all considered representative paralogs. For most analyses of the elt3 clade, we used both representative paralogs and singletons, for most analyses of the MED ortholog group, we used single zinc finger paralogs and singletons, and for the other clades we used only singletons for most analyses.

Gene structure comparisons and predictions of ancestral gene structures

Using a custom Python script, the exon sequences listed in the Supp. Table 1 column “exonSeq” (see above), and the respective scaffold sequence we determined intron lengths in our selected genes (data not shown). Using the ELM2 (PF01448), BAH (PF01426), and Myb (PF00249) Pfam domain seed alignments (pfam.xfam.org) and HMMER v3.3 tools (hmmer.org), we made pHMMs of each of these domains. Then we used HMMER v3.3 tools to search for pHMM domain matches (with no significance cutoff, default settings) in each protein sequence and found its corresponding location in each protein’s gene structure. Using a custom Python script we found the locations of the GATA ZnF domains that we identified in each confident protein (Supp. Fig. 1.5; see above). Using the exon lengths and the domain location information, we created representations of the gene structures of all the confident genes in this study for which genomic data was available (Supp. Fig. 1.3) using a custom R script. (Note: the *C*. sp. *45* and *C*. sp. *47* genes were excluded because only transcriptome data were available for these species.)

We visually compared the gene structures of 714 confident genes (Supp. Fig. 1.3) and, using the principle of parsimony (and when parsimony was not sufficient to distinguish between two alternatives also treating intron loss as more frequent than intron gain (Roy & Penny 2006)), then predicted some ancestral gene structures (exon number and domain location(s)) for each ortholog group (Supp. Fig. 1.3; Fig. 1.2). To estimate the lengths of the exons and introns in the ancestral genes, we calculated and used the median lengths of the exons and introns of the orthologs that had the same gene structure as the predicted ancestor (Supp. Fig. 1.3).

Construction, comparison, and use of hidden Markov model profiles for ortholog group

We aligned singleton protein sequences (see above) within each of the ELT-1, ELT-2, ELT-6, EGL-18, RCOR-1, SPR-1, and EGL-27 ortholog groups, the singleton and representative paralog (see above; Supp. Fig. 1.10) protein sequences within each of the ELT-3, ELT-7, END-1, and END-3 ortholog groups, and all *Elegans* group and *Japonica* group protein sequences within the MED ortholog group respectively, using MUSCLE (Edgar 2004), MAFFT FFT-NS-2 (Katoh et al. 2002), and Clustal Omega (Sievers et al. 2011) default settings. Additionally, we aligned all protein sequences that we considered GATA factors (i.e. those in the elt1/2, elt6, and elt3 clades and in the MEDs ortholog group). Overall, all three alignment algorithms aligned the ZnFs similarly, although the surrounding regions contained more differences. Upon visual inspection, we concluded that the MUSCLE alignments did the best job of aligning the most ZnF-neighboring residues (i.e. introduced more gaps than Clustal Omega, but fewer than MAFFT, such that likely conserved residues were aligned between orthologs) and, therefore, MUSCLE was used for all alignments with the exceptions of the ELT-2 NFs and the all-GATA-factor alignments for which MAFFT and Clustal Omega, respectively, were used instead.

We trimmed the MUSCLE alignments to three different sizes. The small alignment includes the GATA ZnF and part of the adjacent basic region (BR) up to the well conserved arginine (R), proline (P) pair which is usually located 12 to 13 residues after the end of the ZnF (Lowry & Atchely 2000; Maduro 2020; Eurmsirilerd & Maduro 2020; this study). The medium-sized alignment includes the two residues before the start of the GATA ZnF and most, often all, of the adjacent BR (i.e. 28 residues after the ZnF), which contains all the residues involved in the structure and DNA-binding of the cGATA-1 CF (Omichinski et al. 1993) and *C. elegans* MED-1 (Lowry et al. 2009). The large alignments comprise all reasonable-looking alignment positions surrounding the ZnF on either side, but not including any positions from non-GATA DNA-binding domains, as determined by visual inspection. Each category of truncated alignment was used to make two pHMMs using the HMMER version 3.3 hmmbuild tool (hmmer.org), one with default and the other with enone settings. (Enone uses the actual number of aligned sequences for the effective number of sequences, which maximizes the information content per position.) We then used hmmscan with no bit score cutoff (Mistry et al. 2013) to identify and score profile matches among all the protein sequences included in this study. Enone medium pHMMs are shown and used in Figure 1.3. We used a custom Python script to create radar plots depicting the bit scores for the protein sequences (see Fig. 1.3B). We used Skylign and its “information content above background amino acid frequencies” option (Wheeler et al. 2014) to make logos of the pHMMs (Fig. 1.3A,D,E). This option displays the total information content per position as the total height of the stack of amino acid(s). Only amino acid(s) with frequencies at that position above the background frequency of that amino acid in the BLOSUM62 substitution matrix (Henikoff & Henikoff 1992) are included in the stack.

pHMM comparison to residues known to be important for animal GATA factors bound to DNA

We compared the residue with the highest probability at each position in each pHMM to the residues known to be important for the protein structure and/or DNA binding of animal GATA factors bound to DNA. We considered residues similar if they have a BLOSUM62 substitution score of 1 or higher (Henikoff & Henikoff 1992). The one exception was substitutions of lysines (K) with glutamic acid (E), in the basic region of the GATA domain. This exception was only implemented once for the MEDs which have a glutamic acid instead of a lysine as position 17 in their basic regions.

Identifying syntenic GATA-domain-containing genes

Using a custom Python script, we analyzed the annotation files for each species to identify syntenic genes. We thereby established the scaffold/chromosome containing the sequence coding for each of the proteins in this study, and then determined whether any of those sequences were on the same scaffolds. The syntenic GATA-domain-containing genes, their shared scaffold, and the distance between these syntenic genes are listed in Supp. Table 3. Since most species genome assemblies lack chromosome-level resolution we also used a custom Python script, to find all annotated genes within 70 kb upstream and downstream of each confident gene (“neighbor genes”). We then used BLASTp (evalue cutoff 0.1) to search for the neighbor genes longest isoform tophit in the *C. elegans* proteome. We then found what chromosome the *C. elegans* tophit was coded on. These neighbor gene *C. elegans* tophit chromosome is what is plotted for each confident GATA-domain containing gene in Figure 1.4.

Worm maintenance

C. angaria strain PS1010 was grown at room temperature (RT, approximately 21-22□C) on Nematode Growth Medium Lite (NGM Lite, 34.22 mM NaCl, 4g/L Bactotryptone, 22.04 mM KH_2PO_4, 2.87 mM K_2HPO_4, 20.69 uM Cholesterol, 59.47 mM Agar, in distilled (DI) water) in Petri dishes containing a lawn of *Escherichia coli* strain OP50 as a food source, in a manner similar to that standardly used for culturing and maintaining *C. elegans* (Brenner 1974; Stiernagle 2006).

Single molecule fluorescence in situ *hybridization*

C. angaria strain PS1010 embryos were isolated from gravid adults using worm-bleaching solution (250mM NaOH, 1% NaOCl, in distilled water) and then, following standard *C. elegans* protocols for synchronizing them (Stiernagle 2006), grown in liquid M9 (22 mM KH_2PO_4, 42 mM Na_2HPO_4, 85.6 mM NaCl, 1 mM MgSO4, in DI water) for 1 day until they hatched. Synchronized, larval stage 1 (L1s) worms were then pipetted onto NGM Lite agar plates with *E. coli* lawns and grown, using standard procedures, at RT for 4 days until the by then adult worms started laying eggs. To enrich for early embryos (i.e., those still inside the worm), the *C. angaria* on the plates were washed off with DI water and into a 40 um filter set-up which retained adults and let already laid eggs pass through to be discarded. The adult worms were then bleached (with same solution as above) to extract early embryos (Raj et al. 2008). Embryos were then fixed with 4% formaldehyde in PBS (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, 2 mM KH_2PO_4 in nuclease free water), freeze-thawed using liquid nitrogen, washed with 1x phosphate-buffered saline (PBS), placed in 70% ethanol in nuclease-free water (Ambion), and stored at 4C for at least overnight and up to one week. Embryos were then washed in a solution (wash buffer) comprised of 10% formamide, 2x SSC (300mM NaCl 30 mM $Na_3C_6H_5O_7$) prepared using nuclease-free water. Hybridizations were carried out as previously described (Raj et al., 2010, 2008) in 100uL hybridization buffer (10% formamide, 2x SSC, 0.1g/mL dextran sulfate in nuclease-free water) to which 1 ul of each of two smFISH probes, one designed to hybridize to *elt-2* mRNA (Atto 647::*elt-2*, Biosearch) and the other to *elt-3* mRNA (Quasar 570::*elt-3*, Biosearch), had been added; embryos were incubated in the hyrbridization solution for 16 hours at 30C. Embryos were then washed with wash buffer and their nuclei stained with 5ug/mL DAPI (4′,6-diamidino-2-phenylindole (Roche)) prepared in wash buffer for 10 minutes at 30C.

For imaging, embryos were suspended in RT glox buffer comprised of 20mM Tris Cl pH 7.5, 2x SSC, 0.4% glucose in nuclease-free water, 37 ug/mL glucose oxidase, and 1ul catalase (Sigma-Aldrich; Raj et al. 2010). Embryos were imaged in Z-stacks with 0.3 um spacing at 100x magnification on a Nikon epifluorescent compound microscope. smFISH signals were quantified using a machine-learning spot-classification tool, AroSpotFinding Suite (Rifkin 2011; Wu & Rifkin 2015), and visually confirmed. Nuclei were counted manually with the help of a custom MATLAB script (available upon request). An embryo's nuclei count was used as a proxy for its developmental stage.

Identifying conserved transcription factor binding sites in elt-3 *and* elt-2 *promoters*

Using a custom Python script, we extracted the sequences comprising 1200 base pairs (bps) upstream of the start codons (i.e., proximal promoters) of *elt-2* singletons and *elt-3* singletons and representative paralogs from the scaffold files of each species (see above and note that there were no confident *elt-2* paralogs). If another annotated coding sequence occurred within the upstream 1200 bps of a gene, we shortened the proximal promoter so as to eliminate the annotated coding sequences. To look for possible *cis*-regulatory motifs within these sequences we used meme-5.2.0 command line tools (downloaded from meme-suite.org/tools/meme) to identify any enriched motifs in the *elt-2* and *elt-3* promoters respectively. To look for clade-specific motifs we also compared *elt-2* promoters and *elt-3* promoters between the *Elegans* supergroup and non-*Elegans* supergroup species. The parameters used for our MEME analysis included consideration of both strands (revcomp), motif widths between 5 to 12, expected site distribution, and any number of repetitions (anr) (which often found more repetitive A/T-rich motifs in *elt-3* promoters) or zero or one occurrence (zoops); the program usually stopped finding motifs after a significant motif with an e-value greater than 0.001 had been identified. For evaluations of *Elegans* supergroup *elt-2* promoters using the zoops option, the analysis reached the maximum number of motifs to be identified of 20.

Additionally, we looked for conserved binding sites for specific transcription factors. We aligned *elt-2* and *elt-3* promoters, respectively, using MAFFT FFT-NS-2 (Katoh et al. 2002) and searched for the sites identified using MEME, canonical and non-canonical GATA factor binding sites, and binding sites for *C. elegans* endoderm transcription factors (i.e. SKN-1, POP-1, and PAL-1) (Fig. 1.6). Using a custom Python script, we determined whether these sites occurred in individual promoters more often than expected by chance, assuming a Poisson distribution and the sequence composition of the given promoter, in a manner similar to prior analyses carried out on previously available *end-1*, *end-3*, and *med* promoters (Maduro et al. 2015; Maduro 2020).

Identifying TGATAA sites in orthologous gut-, muscle-, neural-, and hypoderm-specific/enriched promoters

Using a custom Python script, we found the longest isoform of each of the 197 intestine-specific/enriched, 71 muscle, 47 neural, and 79 hypoderm *C. elegans* genes from McGhee et al. (2007; 2009). We then used reciprocal BLASTp (evalue cutoff 0.01) to search for putative orthologs of these genes in the 57 available non-*C. elegans Caenorhabditis* and two outgroup nematode proteomes. We only included orthologs with an ATG at the start of their coding sequence and genes with at least two orthologs. Next, we got the putative proximal promoters (i.e., 2 kb upstream of coding sequence start) of each ortholog. We did not include promoters less than 5 bps long. We then found the number of TGATAA sites in each promoter. Next, we used hierarchical clustering with Euclidean distance metric to organize the genes by number of TGATAA sites in their promoters (and whether species had a putative ortholog). This is what is plotted in Figure 1.7 and Supplemental Figure 1.14.

Data availability

Custom Python, R, and MATLAB scripts are available upon request.

Acknowledgements

Chapter 1 is a reprint of a manuscript submitted for publication of its contents. It may appear in the journal *Molecular Biology and Evolution*. The manuscript is titled “A GATA factor radiation in *Caenorhabditis* rewired the endoderm developmental gene regulatory network” and is authored by Antonia C. Darragh and Scott A. Rifkin. Experiments and data analysis were led by the dissertation author and S.A.R. The manuscript preparation was led by the dissertation author with assistance from S.A.R. The dissertation author is the primary author on this manuscript.

Works Cited

Aasland R, Stewart AF, Gibson T. 1996. The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. *Trends Biochem Sci* 21:87–88.

Allendorf F, Utter F. 1973. Gene Duplication within the Family Salmonidae: Disomic Inheritance of Two Loci Reported to Be Tetrasomic in Rainbow Trout. *Genetics*.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.

Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D, Niu W, Boyle AP, Xie D, Ma L, Murray JI, Reinke V, Waterston RH, Snyder M. 2014. Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* 512:400–405.

Arst HN, Scazzocchio C. 1975. Initiator constitutive mutation with an “up-promoter” effect in Aspergillus nidulans. *Nature* 254:31–34.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. *Proc Natl Acad Sci U S A* 110:17409–17414.

Baker EA, Woollard A. 2019. How Weird is The Worm? Evolution of the Developmental Gene Toolkit in Caenorhabditis elegans. *J Dev Biol* 7:E19.

Barrière A, Yang S-P, Pekarek E, Thomas CG, Haag ES, Ruvinsky I. 2009. Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Res* 19:470–480.

Bates DL, Chen Y, Kim G, Guo L, Chen L. 2008. Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. *J Mol Biol* 381:1292–1306.

Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. 2003. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development* 130:889–900.

Bender AM, Kirienko NV, Olson SK, Esko JD, Fay DS. 2007. lin-35/Rb and the CoREST ortholog spr-1 coordinately regulate vulval morphogenesis and gonad development in C. elegans. *Dev Biol* 302:448–462.

Blackwell TK, Bowerman B, Priess JR, Weintraub H. 1994. Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. *Science* 266:621–628.

Boeck ME, Boyle T, Bao Z, Murray J, Mericle B, Waterston R. 2011. Specific roles for the GATA transcription factors end-1 and end-3 during C. elegans E-lineage development. *Dev Biol* 358:345–355.

Borgonie G, Bert W, Houthoofd W, Vangestel S. 2008. The early embryonic development of the satellite organism Pristionchus pacificus: differences and similarities with Caenorhabditis elegans. *Nematol* 10:301–312.

Bowerman B, Eaton BA, Priess JR. 1992. skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell* 68:1061–1075.

Boyer LA, Latek RR, Peterson CL. 2004. The SANT domain: a unique histone-tail-binding module? *Nat Rev Mol Cell Biol* 5:158–163.

Brenner S. 1974. The genetics of Caenorhabditis elegans. *Genetics* 77:71–94.

Broitman-Maduro G, Maduro MF, Rothman JH. 2005. The noncanonical binding site of the MED-1 GATA factor defines differentially regulated target genes in the C. elegans mesendoderm. *Dev Cell* 8:427–433.

Brown TA. 2002. Genomes. 2nd ed. Oxford: Wiley-Liss Available from: http://www.ncbi.nlm.nih.gov/books/NBK21128/

Callebaut I, Courvalin JC, Mornon JP. 1999. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *FEBS Lett* 446:189–193.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Carroll AS, Gilbert DE, Liu X, Cheung JW, Michnowicz JE, Wagner G, Ellenberger TE, Blackwell TK. 1997. SKN-1 domain folding and basic region monomer stabilization upon DNA binding. *Genes Dev* 11:2227–2238.

Cavalcanti ARO, Ferreira R, Gu Z, Li W-H. 2003. Patterns of gene duplication in Saccharomyces cerevisiae and Caenorhabditis elegans. *J Mol Evol* 56:28–37.

Chang C-C, Tung C-H, Chen C-W, Tu C-H, Chu Y-W. 2018. SUMOgo: Prediction of sumoylation sites on lysines by motif screening models and the effects of various post-translational modifications. *Sci Rep* 8:15512.

Charlesworth D, Lyons EE, Litchfield LB. 1994. Inbreeding depression in two highly inbreeding populations of Leavenworthia. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 258:209–214.

Chen Y, Bates DL, Dey R, Chen P-H, Machado ACD, Laird-Offringa IA, Rohs R, Chen L. 2012. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep* 2:1197–1206.

Ch’ng Q, Kenyon C. 1999. egl-27 generates anteroposterior patterns of cell fusion in C. elegans by regulating Hox gene expression and Hox protein function. *Development* 126:3303–3312.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Coghlan A, Wolfe KH. 2002. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res* 12:857–867.

Coroian C, Broitman-Maduro G, Maduro MF. 2006. Med-type GATA factors and the evolution of mesendoderm specification in nematodes. *Dev Biol* 289:444–455.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3:e314.

Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledó JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M, Baer CF. 2009. A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. *Proc Natl Acad Sci U S A* 106:16310–16314.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature* 430:679–682.

Dey A, Jeon Y, Wang G-X, Cutter AD. 2012. Global population genetic structure of Caenorhabditis remanei reveals incipient speciation. *Genetics* 191:1257–1269.

Dineen A, Osborne Nishimura E, Goszczynski B, Rothman JH, McGhee JD. 2018. Quantitating transcription factor redundancy: The relative roles of the ELT-2 and ELT-7 GATA factors in the C. elegans endoderm. *Dev Biol* 435:150–161.

Ding Z, Gillespie LL, Paterno GD. 2003. Human MI-ER1 alpha and beta function as transcriptional repressors by recruitment of histone deacetylase 1 to their conserved ELM2 domain. *Mol Cell Biol* 23:250–258.

Du L, Tracy S, Rifkin SA. 2016. Mutagenesis of GATA motifs controlling the endoderm regulator elt-2 reveals distinct dominant and secondary cis-regulatory elements. *Dev Biol* 412:160–170.

Eddy SR. 2020. Hmmer. http://hmmer.org/.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Ellis RE, Lin S-Y. 2014. The evolutionary origins and consequences of self-fertility in nematodes. *F1000Prime Rep* 6:62.

Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet* 2:e102.

Eurmsirilerd E, Maduro MF. 2020. Evolution of Developmental GATA Factors in Nematodes. *J Dev Biol* 8:E27.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. *Nature* 397:344–347.

Félix M-A. 2007. Cryptic quantitative evolution of the vulva intercellular signaling network in Caenorhabditis. *Curr Biol* 17:103–114.

Félix M-A, Braendle C, Cutter AD. 2014. A streamlined system for species diagnosis in Caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. *PLoS One* 9:e94723.

Felsenstein J. 1985. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* 39:783–791.

Ferrari C, Salle R, Callemeyn-Torre N, Jovelin R, Cutter AD, Braendle C. 2017. Ephemeral-habitat colonization and neotropical species richness of Caenorhabditis nematodes. *BMC Ecol* 17:43.

Ferris SD, Whitt GS. 1977. Loss of duplicate gene expression after polyploidization. *Nature*. 265:258–260.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Friedman R, Hughes AL. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res* 11:373–381.

Fukushige T, Goszczynski B, Tian H, McGhee JD. 2003. The evolutionary duplication and probable demise of an endodermal GATA factor in Caenorhabditis elegans. *Genetics* 165:575–588.

Fukushige T, Hawkins MG, McGhee JD. 1998. The GATA-factor elt-2 is essential for formation of the Caenorhabditis elegans intestine. *Dev Biol* 198:286–302.

Fukushige T, Hendzel MJ, Bazett-Jones DP, McGhee JD. 1999. Direct visualization of the elt-2 gut-specific GATA factor binding to a target promoter inside the living Caenorhabditis elegans embryo. *Proc Natl Acad Sci U S A* 96:11883–11888.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung M-S, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dosé AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, Muroyama A, Murray JI, Ooi S-L, Pham H, Phippen T, Preston EA, Rajewsky N, Rätsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan K-K, Zeller G, Zha Z, Zhong M, Zhou X, modENCODE Consortium, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330:1775–1787.

Gilleard JS, McGhee JD. 2001. Activation of hypodermal differentiation in the Caenorhabditis elegans embryo by GATA transcription factors ELT-1 and ELT-3. *Mol Cell Biol* 21:2533–2544.

Gilleard JS, Shafi Y, Barry JD, McGhee JD. 1999. ELT-3: A Caenorhabditis elegans GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev Biol* 208:265–280.

Gillis WJ, Bowerman B, Schneider SQ. 2007. Ectoderm- and endomesoderm-specific GATA transcription factors in the marine annelid Platynereis dumerilli. *Evol Dev* 9:39–50.

Gillis WQ, Bowerman BA, Schneider SQ. 2008. The evolution of protostome GATA factors: molecular phylogenetics, synteny, and intron/exon structure reveal orthologous relationships. *BMC Evol Biol* 8:112.

Gillis WQ, St John J, Bowerman B, Schneider SQ. 2009. Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. *BMC Evol Biol* 9:207.

Goldstein B. 1992. Induction of gut in Caenorhabditis elegans embryos. *Nature* 357:255–257.

Gorrepati L, Thompson KW, Eisenmann DM. 2013. C. elegans GATA factors EGL-18 and ELT-6 function downstream of Wnt signaling to maintain the progenitor fate during larval asymmetric divisions of the seam cells. *Development* 140:2093–2102.

Gorton DJ. 1983. Genetical and biochemical studies of the uptake of purines and their degradation products in *Aspergillus nidulans*, University of Essex, UK. PhD thesis.

Gottlieb LD. 1977. Evidence for duplication and divergence of the structural gene for phosphoglucoisomerase in diploid species of clarkia. *Genetics* 86:289–307.

Gout J-F, Lynch M. 2015. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. *Mol Biol Evol* 32:2141–2148.

Grishkevich V, Hashimshony T, Yanai I. 2011. Core promoter T-blocks correlate with gene expression levels in C. elegans. *Genome Res* 21:707–717.

Haag ES, Fitch DHA, Delattre M. 2018a. From “the Worm” to “the Worms” and Back Again: The Evolutionary Developmental Biology of Nematodes. *Genetics* 210:397–433.

Haag ES, Fitch DHA, Delattre M. 2018b. From “the Worm” to “the Worms” and Back Again: The Evolutionary Developmental Biology of Nematodes. *Genetics* 210:397–433.

Haag ES, Thomas CG. 2015. Fundamentals of Comparative Genome Analysis in Caenorhabditis Nematodes. *Methods Mol Biol* 1327:11–21.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature* 445:82–85.

Hawkins MG, McGhee JD. 1995. elt-2, a second GATA factor from the nematode Caenorhabditis elegans. *J Biol Chem* 270:14666–14671.

He C, Cheng H, Zhou R. 2007. GATA family of transcription factors of vertebrates: phylogenetics and chromosomal synteny. *J Biosci* 32:1273–1280.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89:10915–10919.

Herman MA, Ch'ng Q, Hettenbach SM, Ratliff TM, Kenyon C, Herman RK. 1999. EGL-27 is similar to a metastasis-associated factor and controls cell polarity and cell migration in C. elegans. *Development* 126:1055–1064.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* 35:518–522.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.

Jackson BM, Eisenmann DM. 2012. β-catenin-dependent Wnt signaling in C. elegans: teaching an old dog a new trick. *Cold Spring Harb Perspect Biol* 4:a007948.

Jarriault S, Greenwald I. 2002. Suppressors of the egg-laying defective phenotype of sel-12 presenilin mutants implicate the CoREST corepressor complex in LIN-12/Notch signaling in C. elegans. *Genes Dev* 16:2713–2728.

Jozefowicz C, McClintock J, Prince V. 2003. The fates of zebrafish Hox gene duplicates. *J Struct Funct Genomics* 3:185–194.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066.

Kiontke KC, Félix M-A, Ailion M, Rockman MV, Braendle C, Pénigault J-B, Fitch DHA. 2011. A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evol Biol* 11:339.

Koh K, Peyrot SM, Wood CG, Wagmaister JA, Maduro MF, Eisenmann DM, Rothman JH. 2002. Cell fates and fusion in the C. elegans vulval primordium are regulated by the EGL-18 and ELT-6 GATA factors -- apparent direct targets of the LIN-39 Hox protein. *Development* 129:5171–5180.

Koh K, Rothman JH. 2001. ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in C. elegans. *Development* 128:2867–2880.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. *J Theor Biol* 239:141–151.

Konrad A, Flibotte S, Taylor J, Waterston RH, Moerman DG, Bergthorsson U, Katju V. 2018. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 115:7386–7391.

Lancaster BR, McGhee JD. 2020. How affinity of the ELT-2 GATA factor binding to cis-acting regulatory sites controls Caenorhabditis elegans intestinal gene transcription. *Development* 147:dev190330.

Lin KT-H, Broitman-Maduro G, Hung WWK, Cervantes S, Maduro MF. 2009. Knockdown of SKN-1 and the Wnt effector TCF/POP-1 reveals differences in endomesoderm specification in C. briggsae as compared with C. elegans. *Dev Biol* 325:296–306.

Lin R, Thompson S, Priess JR. 1995. pop-1 encodes an HMG box protein required for the specification of a mesoderm precursor in early C. elegans embryos. *Cell* 83:599–609.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in Caenorhabditis elegans. *Curr Biol* 21:306–310.

Liu W-J, Reece-Hoyes JS, Walhout AJM, Eisenmann DM. 2014. Multiple transcription factors directly regulate Hox gene lin-39 expression in ventral hypodermal cells of the C. elegans embryo and larva, including the hypodermal fate regulators LIN-26 and ELT-6. *BMC Dev Biol* 14:17.

Lo MC, Ha S, Pelczer I, Pal S, Walker S. 1998. The solution structure of the DNA-binding domain of Skn-1. *Proc Natl Acad Sci U S A* 95:8455–8460.

Lowry JA, Atchley WR. 2000. Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain. *J Mol Evol* 50:103–115.

Lowry JA, Gamsjaeger R, Thong SY, Hung W, Kwan AH, Broitman-Maduro G, Matthews JM, Maduro M, Mackay JP. 2009. Structural analysis of MED-1 reveals unexpected diversity in the mechanism of DNA recognition by GATA-type zinc finger domains. *J Biol Chem* 284:5827–5835.

Lu JR, McKinsey TA, Xu H, Wang DZ, Richardson JA, Olson EN. 1999. FOG-2, a heart- and brain-enriched cofactor for GATA transcription factors. *Mol Cell Biol* 19:4495–4502.

Lundin LG. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16:1–19.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.

Lynch M, Walsh B. 1998. Genetics and Analysis of Quantitative Traits. 1st edition. Sunderland, Mass: Sinauer Associates is an imprint of Oxford University Press.

Macchietto M, Angdembey D, Heidarpour N, Serra L, Rodriguez B, El-Ali N, Mortazavi A. 2017. Comparative Transcriptomics of Steinernema and Caenorhabditis Single Embryos Reveals Orthologous Gene Expression Convergence during Late Embryogenesis. *Genome Biol Evol* 9:2681–2696.

Maduro MF. 2015. Developmental robustness in the Caenorhabditis elegans embryo. *Mol Reprod Dev* 82:918–931.

Maduro MF. 2017. Gut development in C. elegans. *Semin Cell Dev Biol* 66:3–11.

Maduro MF. 2020. Evolutionary Dynamics of the SKN-1 → MED → END-1,3 Regulatory Gene Cascade in Caenorhabditis Endoderm Specification. *G3 (Bethesda)* 10:333–356.

Maduro MF, Broitman-Maduro G, Choi H, Carranza F, Wu AC-Y, Rifkin SA. 2015. MED GATA factors promote robust development of the C. elegans endoderm. *Dev Biol* 404:66–79.

Maduro MF, Broitman-Maduro G, Mengarelli I, Rothman JH. 2007. Maternal deployment of the embryonic SKN-1-->MED-1,2 cell specification pathway in C. elegans. *Dev Biol* 301:590–601.

Maduro MF, Hill RJ, Heid PJ, Newman-Smith ED, Zhu J, Priess JR, Rothman JH. 2005. Genetic redundancy in endoderm specification within the genus Caenorhabditis. *Dev Biol* 284:509–522.

Maduro MF, Kasmir JJ, Zhu J, Rothman JH. 2005. The Wnt effector POP-1 and the PAL-1/Caudal homeoprotein collaborate with SKN-1 to activate C. elegans endoderm development. *Dev Biol* 285:510–523.

Maduro MF, Lin R, Rothman JH. 2002. Dynamics of a developmental switch: recursive intracellular and intranuclear redistribution of Caenorhabditis elegans POP-1 parallels Wnt-inhibited transcriptional repression. *Dev Biol* 248:128–142.

Maduro MF, Meneghini MD, Bowerman B, Broitman-Maduro G, Rothman JH. 2001. Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in C. elegans. *Mol Cell* 7:475–485.

Maduro MF, Rothman JH. 2002. Making worm guts: the gene regulatory network of the Caenorhabditis elegans endoderm. *Dev Biol* 246:68–85.

Martin DI, Orkin SH. 1990. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev* 4:1886–1898.

McGhee JD. 2013. The Caenorhabditis elegans intestine. *Wiley Interdiscip Rev Dev Biol* 2:347–367.

McGhee JD, Fukushige T, Krause MW, Minnema SE, Goszczynski B, Gaudet J, Kohara Y, Bossinger O, Zhao Y, Khattra J, Hirst M, Jones SJM, Marra MA, Ruzanov P, Warner A, Zapf R, Moerman DG, Kalb JM. 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the C. elegans intestine, from embryo to adult. *Dev Biol* 327:551–565.

McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA, Baillie DL, Kohara Y, Marra MA, Jones SJM, Moerman DG, Robertson AG. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev Biol* 302:627–645.

McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. 2014. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* 159:58–68.

Meier K, Brehm A. 2014. Chromatin regulation: how complex does it get? *Epigenetics* 9:1485–1495.

Meneghini MD, Ishitani T, Carter JC, Hisamoto N, Ninomiya-Tsuji J, Thorpe CJ, Hamill DR, Matsumoto K, Bowerman B. 1999. MAP kinase and Wnt pathways converge to downregulate an HMG-domain repressor in Caenorhabditis elegans. *Nature* 399:793–797.

Merika M, Orkin SH. 1993. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol* 13:3999–4010.

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* 30:1188–1195.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37:1530–1534.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 41:e121.

Müller T, Vingron M. 2000. Modeling Amino Acid Replacement. *Journal of Computational Biology* 7:761–776.

Nadeau JH, Sankoff D. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147:1259–1266.

Nair S, Nash D, Sudimack D, Jaidee A, Barends M, Uhlemann A-C, Krishna S, Nosten F, Anderson TJC. 2007. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Mol Biol Evol* 24:562–573.

Narasimhan K, Lambert SA, Yang AWH, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi HS, Reece-Hoyes JS, Fuxman Bass JI, Walhout AJM, Weirauch MT, Hughes TR. 2015. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. *Elife* 4.

Nei M, Roychoudhury AK. 1973. Probability of Fixation of Nonfunctional Genes at Duplicate Loci. *The American Naturalist* 107:362–372.

Nelson MD, Zhou E, Kiontke K, Fradin H, Maldonado G, Martin D, Shah K, Fitch DHA. 2011. A bow-tie genetic architecture for morphogenesis suggested by a genome-wide RNAi screen in Caenorhabditis elegans. *PLoS Genet* 7:e1002010.

Nicolas RH, Goodwin GH. 1996. Molecular cloning of polybromo, a nuclear protein containing multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain. *Gene* 175:233–240.

Ohno S. 1970. Evolution by Gene Duplication. Springer-Verlag.

Omichinski JG, Clore GM, Schaad O, Felsenfeld G, Trainor C, Appella E, Stahl SJ, Gronenborn AM. 1993. NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science* 261:438–446.

Ono Y, Fukuhara N, Yoshie O. 1998. TAL1 and LIM-only proteins synergistically induce retinaldehyde dehydrogenase 2 expression in T-cell acute lymphoblastic leukemia by acting as cofactors for GATA3. *Mol Cell Biol* 18:6939–6950.

Page BD, Zhang W, Steward K, Blumenthal T, Priess JR. 1997. ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in Caenorhabditis elegans embryos. *Genes Dev* 11:1651–1661.

Pal S, Lo MC, Schmidt D, Pelczer I, Thurber S, Walker S. 1997. Skn-1: evidence for a bipartite recognition helix in DNA binding. *Proc Natl Acad Sci U S A* 94:5556–5561.

Patient RK, McGhee JD. 2002. The GATA family (vertebrates and invertebrates). *Current Opinion in Genetics & Development* 12:416–422.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res* 38:W23-28.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256–1260.

Peter I, Davidson EH. 2015. Genomic Control Process: Development and Evolution. Academic Press.

Phillips BT, Kidd AR, King R, Hardin J, Kimble J. 2007. Reciprocal asymmetry of SYS-1/beta-catenin and POP-1/TCF controls asymmetric divisions in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 104:3231–3236.

Phillips BT, Kimble J. 2009. A new look at TCF and beta-catenin through the lens of a divergent C. elegans Wnt pathway. *Dev Cell* 17:27–34.

Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, Brownlie A, Donovan A, Egan ES, Force A, Gong Z, Goutel C, Fritz A, Kelsh R, Knapik E, Liao E, Paw B, Ransom D, Singer A, Thomson M, Abduljabbar TS, Yelick P, Beier D, Joly JS, Larhammar D, Rosa F, Westerfield M, Zon LI, Johnson SL, Talbot WS. 1998. Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* 18:345–349.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5:877–879.

Raj A, Rifkin SA, Andersen E, van Oudenaarden A. 2010. Variability in gene expression underlies incomplete penetrance. *Nature* 463:913–918.

Rapoport, IA. 1940. Mnogokratnye linejnye povtoreniya uchastkov khromosom i ikh evolyucionnoe znachenie. [Multiple linear repeats of chromosome segments and their evolutionary significance]. *Zhurnal Obshchej Biologii*. 1:235–270. [In Russian]

Ravagnani A, Gorfinkiel L, Langdon T, Diallinas G, Adjadj E, Demais S, Gorton D, Arst HN, Scazzocchio C. 1997. Subtle hydrophobic interactions between the seventh residue of the zinc finger loop and the first base of an HGATAR sequence determine promoter-specific recognition by the Aspergillus nidulans GATA factor AreA. *EMBO J* 16:3974–3986.

Reyes JC, Muro-Pastor MI, Florencio FJ. 2004. The GATA family of transcription factors in Arabidopsis and rice. *Plant Physiol* 134:1718–1732.

Rifkin SA. 2011. Identifying fluorescently labeled single molecules in image stacks using machine learning. *Methods Mol Biol* 772:329–348.

Robertson SM, Shetty P, Lin R. 2004. Identification of lineage-specific zygotic transcripts in early Caenorhabditis elegans embryos. *Dev Biol* 276:493–507.

Rocheleau CE, Downs WD, Lin R, Wittmann C, Bei Y, Cha YH, Ali M, Priess JR, Mello CC. 1997. Wnt signaling and an APC-related gene specify endoderm in early C. elegans embryos. *Cell* 90:707–716.

Rocheleau CE, Yasuda J, Shin TH, Lin R, Sawa H, Okano H, Priess JR, Davis RJ, Mello CC. 1999. WRM-1 activates the LIT-1 protein kinase to transduce anterior/posterior polarity signals in C. elegans. *Cell* 97:717–726.

Roy SW, Penny D. 2006. Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. *Mol Biol Evol* 23:2259–2262.

Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, Morishita S. 2013. The transcription start site landscape of C. elegans. *Genome Res* 23:1348–1361.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.

Schuepbach T, Pagni M, Bridge A, Bougueleret L, Xenarios I, Cerutti L. 2013. pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. *Bioinformatics* 29:1215–1217.

Semple C, Wolfe KH. 1999. Gene duplication and gene conversion in the Caenorhabditis elegans genome. *J Mol Evol* 48:555–564.

Shen Y, Lee G, Choe Y, Zoltewicz JS, Peterson AS. 2007. Functional architecture of atrophins. *J Biol Chem* 282:5037–5044.

Shetty P, Lo M-C, Robertson SM, Lin R. 2005. C. elegans TCF protein, POP-1, converts from repressor to activator as a result of Wnt-induced lowering of nuclear levels. *Dev Biol* 285:584–592.

Shim YH, Bonner JJ, Blumenthal T. 1995. Activity of a C. elegans GATA transcription factor, ELT-1, expressed in yeast. *J Mol Biol* 253:665–676.

Shin TH, Yasuda J, Rocheleau CE, Lin R, Soto M, Bei Y, Davis RJ, Mello CC. 1999. MOM-4, a MAP kinase kinase kinase-related protein, activates WRM-1/LIT-1 kinase to transduce anterior/posterior polarity signals in C. elegans. *Mol Cell* 4:275–280.

Shoichet SA, Malik TH, Rothman JH, Shivdasani RA. 2000. Action of the Caenorhabditis elegans GATA factor END-1 in Xenopus suggests that similar mechanisms initiate endoderm development in ecdysozoa and vertebrates. *Proc Natl Acad Sci U S A* 97:4076–4081.

Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev* 6:715–722.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* 27:135–145.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

Slos D, Sudhaus W, Stevens L, Bert W, Blaxter M. 2017*. Caenorhabditis monodelphis* sp. n.: defining the stem morphology and genomics of the genus *Caenorhabditis*. *BMC Zool*. 2:4.

Smith JA, McGarr P, Gilleard JS. 2005. The Caenorhabditis elegans GATA factor elt-1 is essential for differentiation and maintenance of hypodermal seam cells and for normal locomotion. *J Cell Sci* 118:5709–5719.

Solari F, Bateman A, Ahringer J. 1999. The Caenorhabditis elegans genes egl-27 and egr-1 are similar to MTA1, a member of a chromatin regulatory complex, and are redundantly required for embryonic patterning. *Development* 126:2483–2494.

Sommermann EM, Strohmaier KR, Maduro MF, Rothman JH. 2010. Endoderm development in Caenorhabditis elegans: the synergistic action of ELT-2 and -7 mediates the specification→differentiation transition. *Dev Biol* 347:154–166.

Spieth J, Lawson D, Davis P, Williams G, Howe K. 2014. Overview of gene structure in C. elegans. *WormBook*:1–18.

Starich MR, Wikström M, Arst HN, Clore GM, Gronenborn AM. 1998. The solution structure of a fungal AREA protein-DNA complex: an alternative binding mode for the basic carboxyl tail of GATA factors. *J Mol Biol* 277:605–620.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DHA, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol* 1:E45.

Stevens L. 2020. Genome evolution in the genus Caenorhabditis. [Doctoral Dissertation, University of Edinburgh]. Edinburgh Research Archive.

Stevens L, Félix M-A, Beltran T, Braendle C, Caurcel C, Fausett S, Fitch D, Frézal L, Gosse C, Kaur T, Kiontke K, Newton MD, Noble LM, Richaud A, Rockman MV, Sudhaus W, Blaxter M. 2019. Comparative genomics of 10 new Caenorhabditis species. *Evol Lett* 3:217–236.

Stiernagle T. 2006. Maintenance of C. elegans. *WormBook*:1–11.

Sullivan-Brown JL, Tandon P, Bird KE, Dickinson DJ, Tintori SC, Heppert JK, Meserve JH, Trogden KP, Orlowski SK, Conlon FL, Goldstein B. 2016. Identifying Regulators of Morphogenesis Common to Vertebrate Neural Tube Closure and Caenorhabditis elegans Gastrulation. *Genetics* 202:123–139.

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

Tang Y, Wei Y, He W, Wang Y, Zhong J, Qin C. 2014. GATA transcription factors in vertebrates: evolutionary, structural and functional interplay. *Mol Genet Genomics* 289:203–214.

Teakle GR, Gilmartin PM. 1998. Two forms of type IV zinc-finger motif and their kingdom-specific distribution between the flora, fauna and fungi. *Trends Biochem Sci* 23:100–102.

Teterina AA, Willis JH, Phillips PC. 2020. Chromosome-Level Assembly of the Caenorhabditis remanei Genome Reveals Conserved Patterns of Nematode Genome Organization. *Genetics* 214:769–780.

True JR, Haag ES. 2001. Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev* 3:109–119.

Tsang AP, Visvader JE, Turner CA, Fujiwara Y, Yu C, Weiss MJ, Crossley M, Orkin SH. 1997. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* 90:109–119.

Ulm EA, Sleiman SF, Chamberlin HM. 2011. Developmental functions for the Caenorhabditis elegans Sp protein SPTF-3. *Mech Dev* 128:428–441.

Vandamme J, Sidoli S, Mariani L, Friis C, Christensen J, Helin K, Jensen ON, Salcini AE. 2015. H3K23me2 is a new heterochromatic mark in Caenorhabditis elegans. *Nucleic Acids Res* 43:9694–9710.

Verster AJ, Ramani AK, McKay SJ, Fraser AG. 2014. Comparative RNAi screens in C. elegans and C. briggsae reveal the impact of developmental system drift on gene function. *PLoS Genet* 10:e1004077.

Vilhais-Neto GC, Maruhashi M, Smith KT, Vasseur-Cognet M, Peterson AS, Workman JL, Pourquié O. 2010. Rere controls retinoic acid signalling and somite bilateral symmetry. *Nature* 463:953–957.

Wagner GP. 2007. The developmental genetics of homology. *Nat Rev Genet* 8:473–479.

Wang L, Charroux B, Kerridge S, Tsai C-C. 2008. Atrophin recruits HDAC1/2 and G9a to modify histone H3K9 and to determine cell fates. *EMBO Rep* 9:555–562.

Wang L, Rajan H, Pitman JL, McKeown M, Tsai C-C. 2006. Histone deacetylase-associating Atrophin proteins are nuclear receptor corepressors. *Genes Dev* 20:525–530.

Wang L, Tsai C-C. 2008. Atrophin proteins: an overview of a new class of nuclear receptor corepressors. *Nucl Recept Signal* 6:e009.

Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol Biol* 42:225–249.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol* 32:820–832.

Wheeler TJ, Clements J, Finn RD. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* 15:7.

Wiesenfahrt T, Berg JY, Osborne Nishimura E, Robinson AG, Goszczynski B, Lieb JD, McGhee JD. 2016. The function and regulation of the GATA factor ELT-2 in the C. elegans endoderm. *Development* 143:483–491.

Wu AC-Y, Rifkin SA. 2015. Aro: a machine learning approach to identifying single molecules and estimating classification error in fluorescence microscopy images. *BMC Bioinformatics* 16:102.

Xu X, Kim SK. 2012. The GATA transcription factor egl-27 delays aging by promoting stress resistance in Caenorhabditis elegans. *PLoS Genet* 8:e1003108.

Yang HY, Evans T. 1992. Distinct roles for the two cGATA-1 finger domains. *Mol Cell Biol* 12:4562–4570.

Yang T, Guo L, Ji C, Wang H, Wang J, Zheng X, Xiao Q, Wu Y. 2021. The B3 domain-containing transcription factor ZmABI19 coordinates expression of key factors required for maize seed development and grain filling. *Plant Cell* 33:104–128.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39:306–314.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18:292–298.

Zhao Z, Boyle TJ, Bao Z, Murray JI, Mericle B, Waterston RH. 2008. Comparative analysis of embryonic cell lineage between Caenorhabditis briggsae and Caenorhabditis elegans. *Dev Biol* 314:93–99.

Zhu J, Fukushige T, McGhee JD, Rothman JH. 1998. Reprogramming of early embryonic blastomeres into endodermal progenitors by a Caenorhabditis elegans GATA factor. *Genes Dev* 12:3809–3814.

Zhu J, Hill RJ, Heid PJ, Fukuyama M, Sugimoto A, Priess JR, Rothman JH. 1997. end-1 encodes an apparent GATA factor that specifies the endoderm precursor in Caenorhabditis elegans embryos. *Genes Dev* 11:2883–2896.

Chapter 2:

Studies of the binding affinities and binding specificities of closely related *C. elegans* GATA factors to a key binding site for endoderm development: an effort toward building a biochemical model of a developmental process and uncovering mechanistic details of how paralog binding diverges within a developmental gene regulatory network

2.1 Introduction

The coordination of gene expression is fundamental for establishing different cell types during development. Sequence-specific transcription factors regulate gene expression by binding to DNA sequences in the regulatory region(s) of genes and recruiting or inhibiting transcriptional machinery (Ptashne 1967; Carey et al. 2009). The binding affinity (strength of interaction) of a transcription factor for, and its specificity (accuracy of binding to a specific sequence) to, its binding site(s) are characteristics through which gene expression can be fine-tuned. These binding attributes can also serve as targets for evolutionary processes (e.g., Gao & Stock 2015; Crocker et al. 2016).

Within an organism the members of a family of transcription factors generally carry out different functions yet bind to the same or similar DNA sequences (Kribelbauer et al. 2019; e.g., Weirauch et al. 2014; Narasimha et al. 2015). This apparent paradox occurs for various reasons. For example, homologous transcription factors can be expressed at different times or in different places (Ferris & Whitt 1979; Wagner 2000; Gu et al. 2002) and thereby have access to different binding sites due to different chromatin accessibility (Klemm et al. 2019). Different family members might recognize and only bind to the common binding sequence when the DNA is in one conformation versus another (Rohs et al. 2009; 2010). Homologous transcription factors may form distinct protein-protein interactions which have different gene regulatory outcomes (e.g., Bartlett 2017). Additionally, there could be subtle differences in preferences of DNA sequence to bind to among the closely related members of a family of transcription factors (e.g., Narasimha et al. 2015).

Evolutionary forces contribute to this apparent paradox. Although a full-length duplication of a gene encoding a transcription factor would be expected to create paralogs with the same coding sequence and expression pattern initially, such paralogs are likely to diverge over time unless redundancy is evolutionarily selected for (Ohno 1970). But few mechanistic examples exist of how paralogous transcription factors diverge from each other within a conserved developmental gene regulatory network.

An intriguing and tractable system in which to study such evolution of transcription factors is the endoderm developmental gene regulatory network in the nematode *Caenorhabditis elegans*. This network is well-characterized genetically and is primarily composed of six related GATA-type transcription factors (GATA factors) (reviewed by McGhee 2013 and Maduro 2015; 2017). Five of these GATA factors act in a feedforward loop to activate expression of *elt-2*; *elt-2* encodes the final GATA factor in the network and is required for proper development of a functional intestine in *C. elegans* (McGhee 2013). Recently, the genes encoding three of the upstream GATA factors in the network, *end-3*, *end-1*, and *elt-7*, were found to be closely related to each other and likely all paralogs (Darragh AD & Rifkin SA, unpublished data). The spatial and temporal expression patterns for these three GATA factors also overlap, and evidence suggests they likely all participate directly in regulating *elt-2* (McGhee 2013). Du *et al*. (2016) found that a single GATA-binding site in the *elt-2* promoter was necessary for *elt-2* expression, and the sequence flanking this site matches well with the binding site preferred by END-3, END-1, ELT-7 (Narasimha et al. 2015; Tracy S &

Rifkin SA personal communication). Moreover, ELT-2 regulates its own expression (Fukushige et al. 1999) and is known to bind to other similar sites (Hawkins & McGhee 1995; Wiesenfahrt et al. 2016; Du et al. 2016; Lancaster & McGhee 2020). Based on these data, we hypothesize that END-3, END-1, ELT-7, and ELT-2 all compete for binding to this particular site in the *elt-2* promoter *in vivo* (Fig. 2.1).

As an initial test of this hypothesis, I carried out measurements of the *in vitro* binding kinetics of END-3, END-1, ELT-7, and ELT-2 to this key binding site in the *elt-2* promoter. The relative binding affinity of ELT-2 to this site as compared to its binding to the same site but for a single base pair substitution (thymine (T) to adenine (A) in the position preceding the guanine (G) of the GATA core) was recently reported to be about 56% stronger (Lancaster & McGhee 2020). However, the relative binding kinetics and affinities of the other three GATA factors, and of ELT-2 to other binding sites, have not been quantified. Most binding studies of animal GATA factors indicate that the *in vitro* affinities of these proteins for binding to DNA sequences containing a GAT core are at nanomolar levels (e.g., Merika & Orkin 1993; Figueira et al. 2010; Chen et al. 2012), whereas other GATA factors are known to bind sequences that lack a GAT core with affinities that are three orders of magnitude higher (and thus weaker) (Omichinski et al. 1993). Pedone et al. (1997) found similar differences in binding affinity between the N-terminal zinc fingers (NFs) of *Gallus gallus* (chicken) GATA-2 and GATA-3 as compared to the chicken GATA-1 NF in binding to their preferred GATC sequence. Experiments testing the effects of modifying the core sequence and/or the sequence flanking the binding site core have been conducted. For example, Merika & Orkin (1993) found that GATA factors did not bind to a GATA DNA sequence in which a G preceded the G (GGATA), or followed the T (GATG), in the core sequence. The binding kinetics of GATA factor-DNA interactions have only been measured in a few studies; the few association and dissociation rates, k_{on} and k_{off} (see paragraph below), reported have been on the order of 10^{6}-10^{7} $M^{-1}s^{-1}$ and 10^{-3}-$10^{0}s^{-1}$, respectively (Ko & Engel 1993; Chen et al. 2012).

A key metric used to describe the effectiveness of binding between molecules is the equilibrium disassociation constant K_D, a measure of binding affinity (or strength). For a one-to-one molecular interaction, K_D is the concentration of molecule A at which half of the molecules of B are bound by molecule A at equilibrium. A method to quantify the K_D is to measure the association (k_{on}) and dissociation (k_{off}) rates (i.e., the binding kinetics) of a molecular interaction, and then calculate the K_D using the equation: $K_D = k_{off}/k_{on}$. This chapter describes the experiments I carried out, using various methods, to measure the *in vitro* binding kinetics of END-3, END-1, ELT-7, and ELT-2 to the key binding site in the *elt-2* promoter and establish the K_D for each of the binding interactions. The results, while preliminary, contribute to our understanding of the binding dynamics and the relative strength of the binding of each of these GATA factors to that key binding site. They pave the way for future studies that will elucidate the mechanistic details of the molecular interactions that result in endoderm development in *C. elegans*.

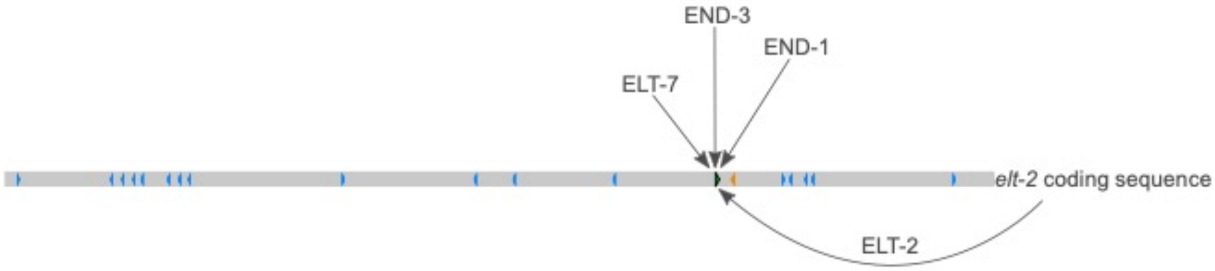


Figure 2.1. Depiction of predicted competition over an important GATA binding site in the *C. elegans elt-2* promoter. The grey bar represents the *C. elegans elt-2* promoter and comprises 1879 bp upstream of the *elt-2* start codon. HGATAR motifs in this promoter sequence are indicated by blue triangles. Binding kinetics of the END-1, END-3, ELT-7 and ELT-2 DBDs were measured to two sites in the *elt-2* promoter, the ACTGATAAGA motif at which we predict these GATA factors compete for binding (depicted by a green triangle outlined in black) located 527 bp upstream of the start codon, and an AGCGATTATG motif 490 bp upstream of the start codon (depicted by the orange triangle). The direction of the triangles indicates which DNA strand each site is on.

2.2 Preliminary Results

Differences in the binding kinetics of four C. elegans *gut-specific GATA factors to an important GATA factor-binding site in the* elt-2 *promoter are subtle*

I measured the binding kinetics of the DNA-binding domains (DBDs) of ELT-2, END-1, END-3, and ELT-7 to the important GATA factor-binding site in the *elt-2* promoter (Du et al. 2016) using bio-layer interferometry (BLI) (Concepcion et al. 2009; see Methods). I found that the ELT-2 DBD containing both its NF and C-terminal zinc finger (CF), the ELT-2 DBD with only its CF, and the END-3 DBD all bound to the *elt-2* promoter binding site with higher affinity (lower K_D) than did the END-1 and ELT-7 DBDs (Table 2.1). The higher affinities of the ELT-2 and END-3 DBDs appeared due to their longer residence time (lower k_{off}) on this DNA sequence because the association rates (k_{on}) of the ELT-7 and END-1 DBDs, while higher than those of the ELT-2 and END-3 DBDs, did not contribute to the differences in K_D as much (Table 2.1). Overall, the measured differences in binding affinity to the *elt-2* binding site among these gut-specific GATA factors were low; all the K_Ds were within one order of magnitude of each other in the nanomolar range (Table 2.1). Considerable variability was also observed between experiments (e.g., compare Table 2.1 rows three and four), and therefore more repetitions are required for accurate comparisons.

Changing the nucleotide preceding GATA from a T to an A had little effect on the binding affinities of the DBDs of ELT-2, END-1, END-3, and ELT-7

Narasimhan et al. (2015), using protein binding microarrays, found that some *C. elegans* GATA factors, including ELT-7, prefer binding to GATA sites with a thymine

(T) in the position immediately preceding the guanine (G) while others, including END-1 and END-3, are less picky about the nucleotide at this position. Other evidence also suggests that END-1 and END-3 do not exhibit as strong a preference for TGATA sites as ELT-7 does, but that TGATA is still slightly preferred over sites comprising AGATA. An adenine (A) preceding GATA was also the second most preferred nucleotide at this position for ELT-7 but it was relatively much less compared to the END genes (Tracy S & Rifkin SA personal communication, high-throughput SELEX data). Lancaster and McGhee (2020) found that full-length ELT-2 binds to DNA of the sequence ACTGATAAGA (which contains the same TGATA sequence used for this study) with a 56% stronger affinity compared to its binding to the same sequence but for the first T being substituted for an A. To further investigate binding preferences and affinity differences, I used BLI to measure the binding kinetics of the DBDs of ELT-2, END-1, END-3, and ELT-7 to a DNA sequence the same as the key *elt-2* promoter binding site but with a T to A substitution immediately preceding the GATA in the sequence (Table 2.2; see Methods). I found that binding affinities for the END-3, END-1, and ELT-7 DBDs to the modified binding site sequence were slightly lower (higher K_D) than to the original sequence, whereas binding affinities for the ELT-2 DBD to both sequences were about the same (compare results in Table 2.2 versus Table 2.1). Differences in the underlying kinetics of the binding were also observed; the ELT-2 and END-3 DBDs had shorter residence times (higher k_{off}) while those for the END-1 and ELT-7 DBDs were about the same, and the latter had slower association rates (lower k_{on}) whereas those for the ELT-2 and END-3 DBDs were faster (compare results in Table 2.2 versus Table 2.1).

DBDs of ELT-2, END-1, END-3, and ELT-7 bind with only slightly weaker affinity to a GATT-containing site in the C. elegans elt-2 *promoter as compared to a GATA-containing site*

Previous work had indicated that the third most preferred binding site sequence for END-1 is TGATT (Tracy S & Rifkin SA personal communication). Additionally, mouse GATA-3 CF bound to DNA sequences comprising either a GATT or GATA site similarly as determined by X-ray crystallography (Bates et al. 2008). I therefore hypothesized that END-1 would also have a high binding affinity to TGATT and similar sequences. There is a CGATT sequence close to the key TGATA site in the *elt-2* promoter that could be an important binding site for END-1 *in vivo* (Fig. 2.1 orange triangle). I therefore measured the biding kinetics of the DBDs of END-1, END-3, ELT-7, and ELT-2 to DNA sequences containing this CGATT sequence using BLI (Table 2.3; see Methods). Results indicated that the binding affinities of the DBDs of END-1, END-3, and ELT-2 were slightly weaker (higher K_D) for CGATT than for TGATA or AGATA, and that the ELT-7 DBD likely does not bind this sequence specifically, at least under the conditions used (compare Table 2.3 to Tables 2.1 and 2.2; see Methods). The ELT-2 DBD had the shortest residence time (higher k_{off}) and an intermediate association rate (k_{on}) for this atypical CGATT-containing binding site, as compared to the GATA-containing sites (compare Table 2.3 to Tables 2.1 and 2.2). The END-1 DBD also had its shortest residence time (higher k_{off}) but fastest association rate (k_{on}) for the atypical CGATT-containing binding site as compared to the GATA-containing sites (compare Table 2.3 to Tables 2.1 and 2.2). The dissociation rate (k_{off}) for the END-3 DBD was about the same as for that observed with the AGATA sequence whereas its association rate was intermediate but closer to the slower rate of binding that had been observed for the END-3 DBD to the TGATA site (compare Table 2.3 to Tables 2.1 and 2.2).

GATA factors have similar binding affinities to GATA- and TGAC-containing DNA binding sites under both physiologically relevant conditions and conditions of higher ionic strength

In vitro measurements of protein-to-DNA binding are limited by the conditions under which they are measured, such as ionic strength, temperature, pH, and the assay itself (e.g., migration of a binding reaction on a gel or tagged molecules). Members of the Rifkin lab had previously used 10 mM HEPES, 200 mM $NH_4CH_3CO_2$, 30 mM NaCl, 1.5 mM $MgCl_2$, 0.5 uM $Zn(CH_3CO_2)_2$, 0.2 mg/ml BSA, 1 mM DTT, 8% glycerol, 20 mM imidazole, pH 7 at room temperature as binding reaction conditions (Du et al. 2016). These conditions have an approximate ionic strength of 0.244 (244 mM), as calculated using the Lewis and Randall (1921) equation and defined by the International Union of Pure and Applied Chemistry (IUPAC 1997). To simplify the number of ingredients, minimize the number of hydrogens in the buffer (in anticipation of conducting 1D-NMR), use a more stable reducing agent (i.e., TCEP instead of DTT), and create more “physiologically relevant” binding conditions, I changed my binding buffer to 10 mM HEPES, 150 mM NaCl, 100 uM $ZnSO_4$, 0.2 mg/ml BSA, 1mM TCEP, 10% glycerol, 20 mM imidazole, adjusted to pH 7 with 2.4 mM NaOH. This buffer has an approximate ionic strength of 0.164 (164 mM), 33% lower than the previous conditions. To put these ionic strengths in perspective, the ionic strength of 1X PBS (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, 2 mM KH_2PO_4, pH 7.4), a common “physiologically relevant” mammalian buffer, is approximately 0.169 (169 mM). Most protein-DNA binding buffers contain salt concentrations in the range of 50-210 mM (Halford 2009). Because interactions between DNA and proteins can often be electrostatic, the ionic strength of the binding conditions can affect these interactions (e.g., Sevall et al. 1975; Esadze et al. 2014).

To check that END-1, END-3, ELT-7, and ELT-2 DBDs were binding to the various GAT-containing sites (mentioned above) specifically under the initial binding conditions, I used BLI to measure the binding kinetics of these DBDs to a DNA sequence I thought would serve as a control for non-specific binding. For this purpose, I used a DNA sequence that contained TGAC instead of GATA but was otherwise identical to the key GATA factor-binding site in the *elt-2* promoter (see Fig. 2.1). Surprisingly, I found that the DBDs of ELT-2, ELT-7, and END-3 all showed specific one-to-one binding, and comparable affinity, to the modified site (see Tables 2.4-2.6) as compared to the binding I had observed using the TGATA-type binding sites under the same conditions (compare Table 2.4,2.5,2.1,2.2). (I did not test END-1 in this experiment because I had recently found that its construct contained a non-synonymous point mutation in the basic region of the DBD. However, this mutation was fixed in the reported END-1 experiments.) I also confirmed that both the ELT-2 DBD containing only the CF and the ELT-2 DBD containing both the NF and CF bound to the TGAC site and to the otherwise identical GATA site similarly using electromobility shift assays (EMSAs) (Fig. 2.2,2.3). Moreover, at least for the ELT-2

DBD, binding to the TGAC-containing sequence was comparable to its binding to the typical TGATA-containing sequence under three different conditions: physiologically relevant ionic strength (164 mM) without addition of random competitor DNA, physiologically relevant ionic strength (164 mM) with addition of random competitor DNA, and conditions of higher ionic strength (244 mM) with addition of random competitor DNA (Table 2.4-2.6). The ELT-2 DBD displayed the strongest affinity (lowest K_D) under the condition of lower ionic strength without competitor DNA (Table 4), an intermediate affinity under the condition of lower ionic strength with competitor DNA (Table 2.5), and the lowest affinity (highest K_D) under the condition of higher ionic strength with competitor DNA (Table 2.6).

To start to determine how specific the binding to TGAC was, I carried out additional experiments using a sequence that Du et al. (2016) had found did not bind to the END-1, END-3, ELT-7, and ELT-2 DBDs (Du et al. 2016). This non-binding DNA sequence contains CTGT instead of TGAC. I found that the ELT-2 DBD containing both the NF and CF bound to this CTGT-containing sequence similarly to its binding to the otherwise identical GATA-containing sequence under the conditions of "physiologically relevant" ionic strength that I had been using (Fig. 2.4). However, under the conditions of higher ionic strength used by Du et al. (2016), I found that the ELT-2 DBD containing both the NF and CF did not bind to the CTGT-containing sequence but still bound to an otherwise identical GATA-containing sequence (Fig. 2.5). The ELT-7 DBD also did not bind to the CTGT-containing sequence under these conditions of higher ionic strength (Fig. 2.6).

Due to complications caused by the COVID-19 pandemic I was not able to further pursue the determination of whether these GATA factors were specifically binding to TGAC sequences or not.

The NF of ELT-2 may contribute to a longer residence time on TGAC-containing sequences

ELT-2 is one of the two *C. elegans* GATA factors that contain two zinc-finger motifs (Hawkins & McGhee 1995). The N-terminal zinc finger motif (NF) in some GATA factors in animals can bind to DNA independently (Pedone et al. 1997; Newton et al. 2001), and/or increase the stability of the protein's binding to DNA (Martin & Orkin 1990; Yang & Evans 1992), and/or modulate the specificity of DNA binding (Whyatt et al. 1993; Shim et al. 1995; Trainor et al. 2000; Waltzer et al. 2002), and/or interact with other proteins (e.g., Tsang et al. 1997; Waltzer et al. 2002). Relative to the NFs of other animal GATA factors, the ELT-2 NF in animals has lost many conserved features and was deemed non-functional (Hawkins & McGhee 1995). However, the "degenerate" NF in ELT-2 DBDs is relatively well conserved across the *Caenorhabditis* genus (Darragh AD & Rifkin SA, unpublished data) and the nematode phylum (Eurmsirilerd & Maduro 2020), suggesting it may, in fact, serve some function. I therefore compared binding by the *C. elegans* ELT-2 DBD with only its CF to binding by the same ELT-2 DBD with both zinc fingers using BLI (compare Tables 2.1,2.3,2.6; see Methods). I found that the ELT-2 NF did not have a strong influence on the binding of the ELT-2 DBD to a canonical TGATA site (Table 2.1), nor did it strongly influence binding to a non-canonical CGATT-containing site (Table 2.3). Interestingly, however, I found that the presence of the NF increased the residence time of the ELT-2 DBD bound to a non-canonical TTGAC-containing site (compare the longer residence time (lower k_{off}) observed for the ELT-2 DBD containing both of its zinc fingers to that of the ELT-2 DBD with only its CF in Table 2.6).

Table 2.1. *in vitro* binding kinetics of the *C. elegans* DBDs of ELT-2, END-3, END-1, and ELT-7 to the important TGATA site in the *elt-2* promoter. The respective DNA-binding domain (DBD) proteins utilized for the binding measurements reported in each row are indicated in the first column. Dissociation rates (k_{off}) are reported in units of seconds^{-1}. Association rates (k_{on}) are given in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are reported in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are listed in the fifth column.

DBD	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 NF & CF	0.00142	53,700	26.4	0.9995
ELT-2 CF	0.0021	82,500	25.5	0.9967
ELT-2 CF	0.00115	62,100	18.6	0.9981
ELT-2 CF	0.00183	22,700	80.4	0.998
END-3	0.00312	95,100	32.8	0.9988
END-1	0.0117	109,000	107	0.9914
ELT-7	0.0236	117,000	202	0.9945

Table 2.2. *in vitro* binding kinetics of the *C. elegans* DBDs of ELT-2, END-3, END-1, and ELT-7 to an AGATA site. The respective protein DNA-binding domain (DBD) proteins utilized for the binding measurements reported in each row are indicated in the first column. Dissociation rates (k_{off}) are given in units of seconds^{-1}. Association rates (k_{on}) are reported in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are given in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are listed in the fifth column.

DBD	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 NF & CF	0.00453	225,000	20.1	0.9893
END-3	0.0127	185,000	68.7	0.9936
END-1	0.00907	34,000	267	0.985
ELT-7	0.0286	79,042	362	0.9791

Table 2.3. *in vitro* binding kinetics of *C. elegans* ELT-2, END-3, END-1, and ELT-7 DBDs to a CGATT site in the *elt-2* promoter. The respective DNA-binding domain (DBD) proteins utilized for the binding measurements reported in each row are listed in the first column. Dissociation rates (k_{off}) are given in units of seconds^{-1}. Association rates (k_{on}) are listed in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are given in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are reported in the fifth column.

DBD	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 NF & CF	0.014	117,000	119	0.9914
ELT-2 NF & CF	0.012	132,000	94.1	0.9914
ELT-2 CF	0.0153	110,000	139	0.9982
ELT-2 CF	0.0108	132,000	137	0.9972
END-3	0.0193	107,000	180	0.9993
END-1	0.0531	126,000	423	0.9901
ELT-7	Not one to one			

Table 2.4. Pico-molar *in vitro* binding affinity of the *C. elegans* ELT-2 CF to different DNA sites under conditions of "physiological" ionic strength and without competitor DNA. The respective DNA sequence utilized for the binding measurements reported in each row is listed in the second column. Dissociation rates (k_{off}) are given in units of seconds^{-1}. Association rates (k_{on}) are reported in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are listed in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are given in the fifth column.

DBD	DNA	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 CF	AGATA	0.00015	345,000	0.445	0.9995
ELT-2 CF	AGATA	0.00019	436,000	0.443	0.999
ELT-2 CF	AGATA	0.00011	180,000	0.598	0.9998
ELT-2 CF	TGATA	0.000082	144,000	0.572	0.9998
ELT-2 CF	CGATT	0.00013	129,000	0.976	0.9997
ELT-2 CF	TTGAC	0.00018	228,000	0.802	0.9991

Table 2.5. Nano- and micro-molar *in vitro* binding kinetics of *C. elegans* ELT-2 CF, END-3, and ELT-7 DBDs to different DNA sequences under conditions of "physiological" ionic strength and with competitor DNA. The respective protein DNA-binding domains (DBD) utilized for the binding measurements reported in each row are listed in the first column. The respective DNA sequences utilized for the binding measurements reported in each row are listed in the second column. Dissociation rates (k_{off}) are given in units of seconds^{-1}. Association rates (k_{on}) are given in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are listed in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are reported in the fifth column.

DBD	DNA	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 CF	AGATA	0.00062	119,000	5.22	0.9991
ELT-2 CF	TTGAC	0.00066	94,700	6.95	0.9979
END-3	TGATA	0.00072	9,420	76.1	0.9983
ELT-7	TGATA	0.00788	79.1	99,700	0.8819
ELT-7	AGATA	0.0055	76.8	71,800	0.9576
ELT-7	CGATT	0.0059	72.7	81,100	0.9123
ELT-7	TTGAC	0.0044	301	14,600	0.966

Table 2.6. *in vitro* binding kinetics of the *C. elegans* ELT-2 and END-3 DBDs to a non-canonical TTGAC sequence. The respective DNA-binding domain (DBD) proteins utilized for the binding measurements reported in each row are given in the first column. Dissociation rates (k_{off}) are given in units of seconds^{-1}. Association rates (k_{on}) are listed in units of Molarity^{-1}seconds^{-1}. Equilibrium dissociation constants (K_D) are reported in nanomoles per liter (nM). The R^2 values (curve-fitting coefficient of determination) of the curves fit to the BLI association and dissociation sensorgrams to determine these parameters are given in the fifth column.

DBD	k_{off} (s^{-1})	k_{on} ($M^{-1}s^{-1}$)	K_D (nM)	R^2 value
ELT-2 NF & CF	0.00856	336,000	25.5	0.9937
ELT-2 NF & CF	0.00854	415,000	20.6	0.9928
ELT-2 NF & CF	0.00772	578,000	13.4	0.9917
ELT-2 CF	0.0414	477,000	86.7	0.9821
ELT-2 CF	0.0291	590,000	49.4	0.986
END-3	0.0111	374,000	29.6	0.9747

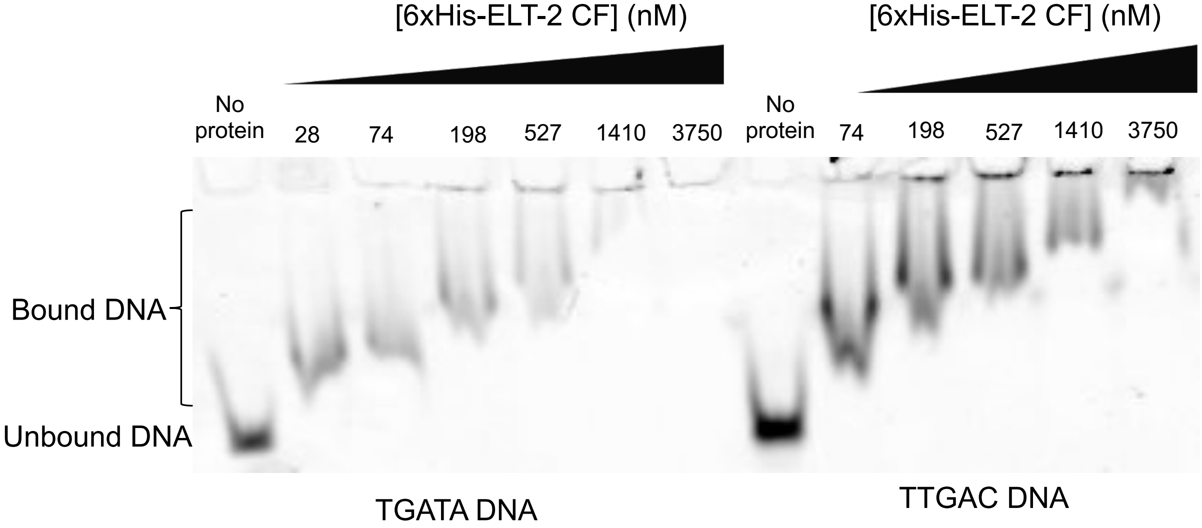


Figure 2.2. EMSA of the ELT-2 CF DBD binding to DNA containing a non-canonical TTGAC site or a canonical TGATA site. The concentration of the 6xHis-ELT-2 CF DBD in each assay, in nanomoles per liter (nM), is given at the top of the depiction of the gel. The DNA sequence being assayed is indicated beneath the gel. Lanes labelled “no protein” contain only unbound DNA. Bands that did not migrate as far into the gel as the unbound DNA band were slowed down by being bound by the 6xHis-ELT-2 CF DBD.

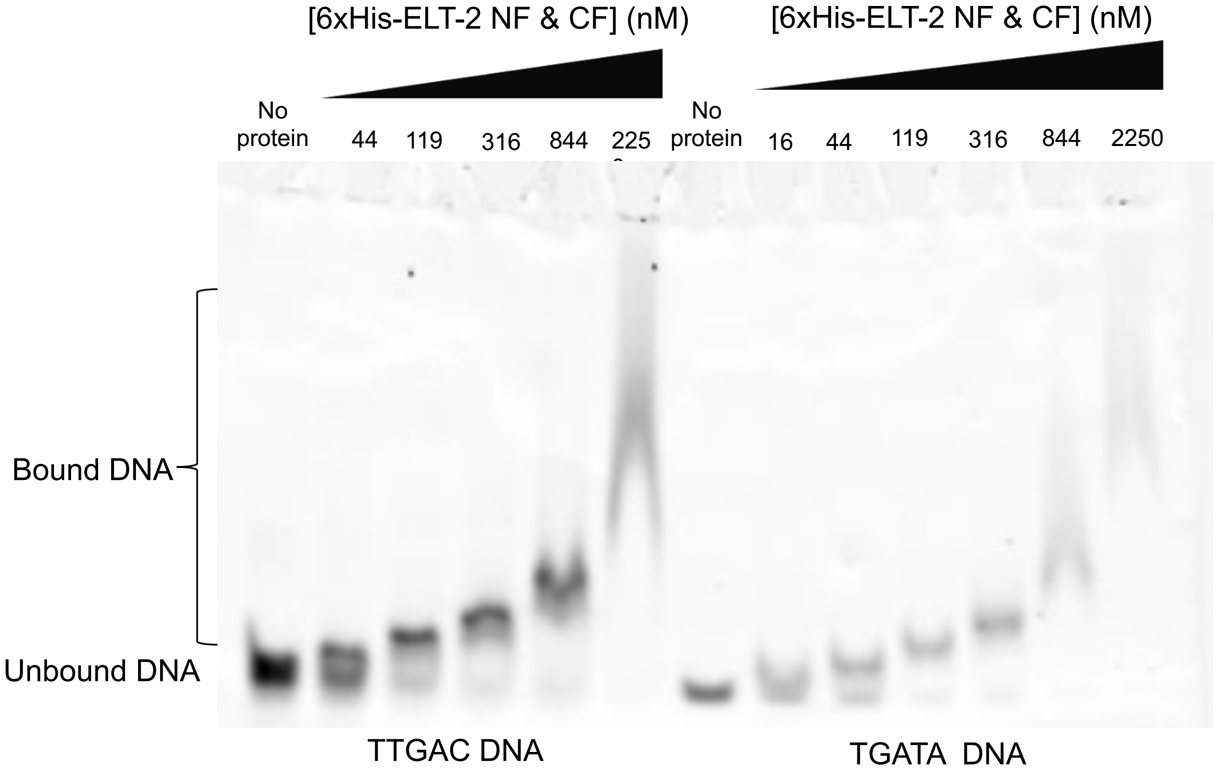


Figure 2.3. EMSA of the ELT-2 DBD (both NF and CF) binding to DNA containing a non-canonical TTGAC or a canonical TGATA site. The concentration of the 6xHis-ELT-2 DBD in each assay in nanomoles per liter (nM) is given at the top of the depiction of the gel. The DNA sequence being assayed is indicated beneath the gel. Lanes labelled “no protein” contain only unbound DNA. Bands that did not migrate as far into the gel as the unbound DNA band were slowed down by being bound by the 6xHis-ELT-2 DBD.

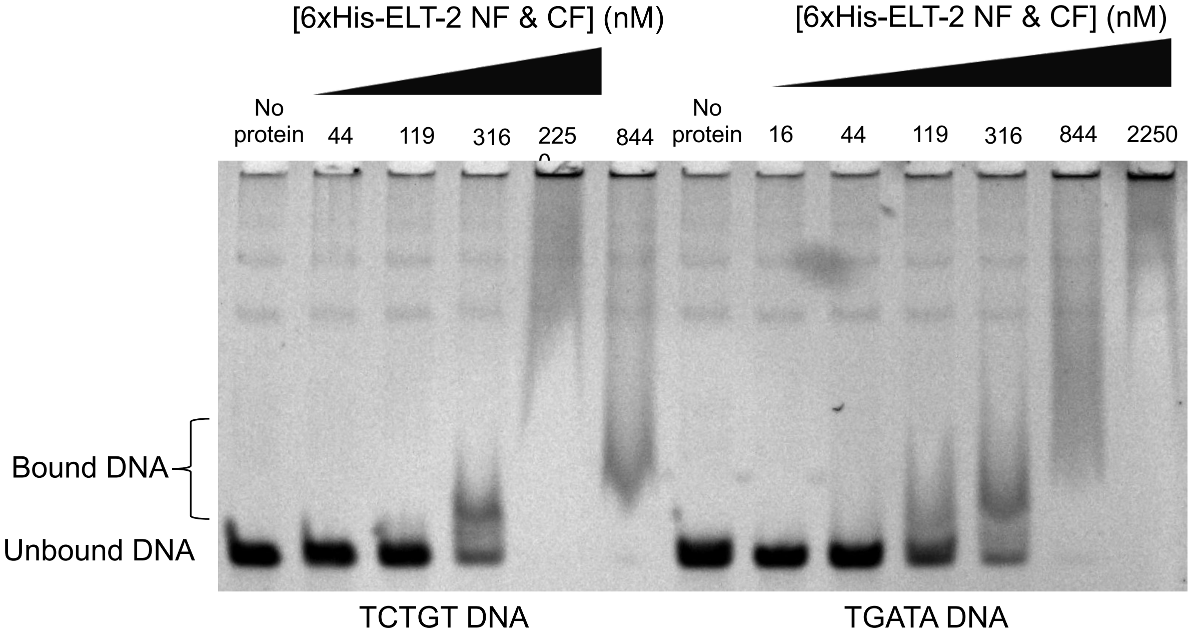


Figure 2.4. EMSA of the ELT-2 DBD (both NF and CF) binding to DNA containing a non-canonical TCTGT or a canonical TGATA site. The concentration of the 6xHis-ELT-2 DBD in each assay in nanomoles per liter (nM) is given at the top of the depiction of the gel. The DNA sequence being assayed is indicated beneath the gel. Lanes labelled “no protein” contain only unbound DNA. Bands that did not migrate as far into the gel as the unbound DNA band were slowed down by being bound by the 6xHis-ELT-2 DBD.

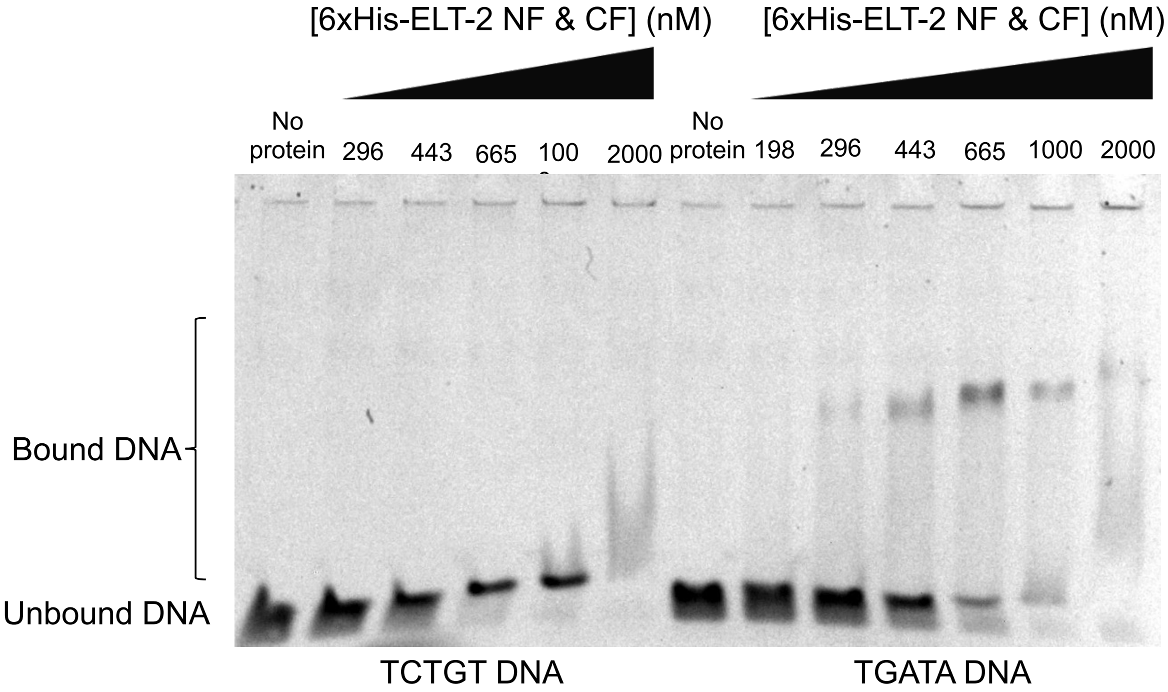


Figure 2.5. EMSA of the ELT-2 DBD (both NF and CF) showing binding to DNA containing a canonical TGATA site but not to a non-canonical TCTGT site. The concentration of the 6xHis-ELT-2 DBD in each assay in nanomoles per liter (nM) is given at the top of the depiction of the gel. The DNA sequence being assayed is indicated beneath the gel. Lanes labelled “no protein” contain only unbound DNA. Bands that did not migrate as far into the gel as the unbound DNA band were slowed down by being bound by the 6xHis-ELT-2 DBD.

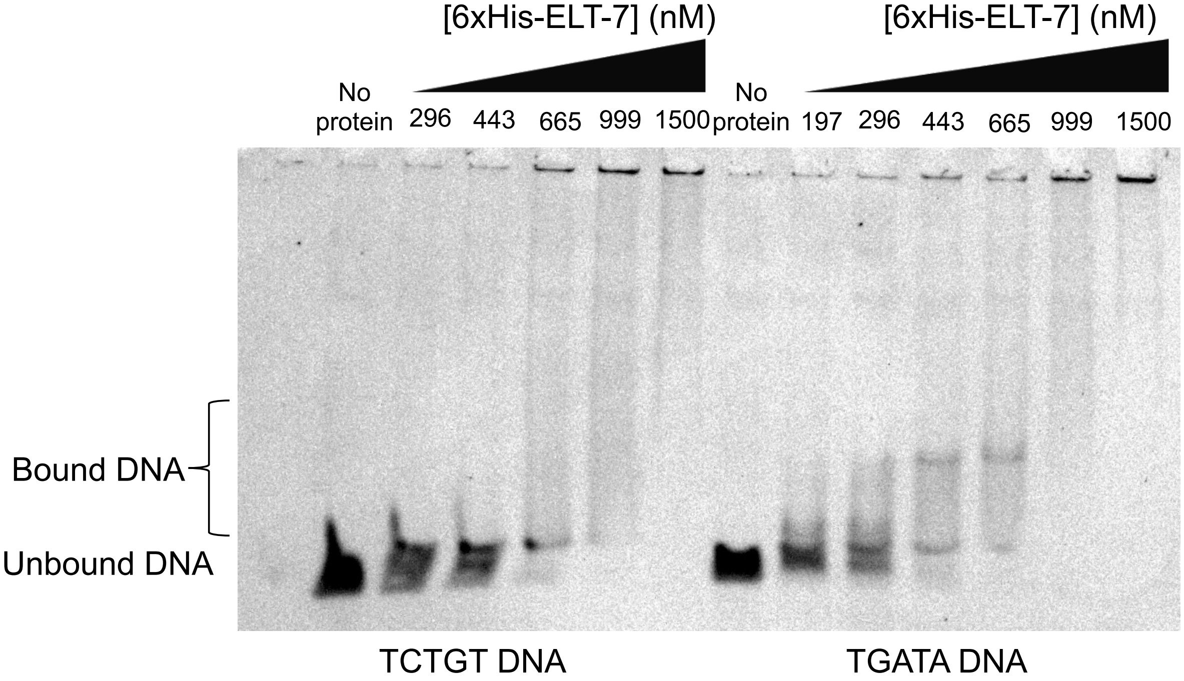


Figure 2.6. EMSA of the ELT-7 DBD binding to DNA containing a canonical TGATA site but not to a non-canonical TCTGT site. The concentration of 6xHis-ELT-7 DBD in each assay, in nanomoles per liter (nM), is given at the top of the depiction of the gel. The DNA sequence being assayed is indicated beneath the gel. Lanes labelled "No protein" contain only unbound DNA. Bands that did not migrate as far into the gel as the unbound DNA band were slowed down by being bound by the 6xHis-ELT-7 DBD.

2.3 Discussion

The results presented in this chapter are preliminary. I initially wanted to use the same binding conditions for each DBD and DNA interaction to ensure consistency from experiment to experiment. However, differences in binding and the stability (or lack thereof) of the DBDs used for these binding assays meant that optimized conditions for most interactions were needed. For example, measurements of weaker binding interactions could benefit from loading more DNA onto the biosensor tips. Future studies in which optimized conditions for weaker interactions are established and sufficient repetitions of all measurements provide the statistical power to handle experiment-to-experiment variability should provide accurate comparisons of the binding kinetics of these GATA factor DBDs to different DNA binding sequences.

One goal of this study was to contribute to the creation a mathematical model of the molecular dynamics of the initial activation of *elt-2* during development. Toward that end, other members of the Rifkin lab have quantified the mRNA expression levels of *end-3*, *end-1*, *elt-7*, and *elt-2* during early endoderm development when the crucial activation of *elt-2* occurs (Wu AC and Rifkin SA personal communication). These mRNA expression levels were measured not only in the wildtype *C. elegans* N2 strain but also in mutants in which the functions of various of the five upstream GATA factors (*med-1*, *med-2*, *end-3*, *end-1*, and *elt-7*) had been eliminated and/or reduced. These measurements give insights into the timing of the transactivation of downstream regulators contributed by each of these transcription factors. Moreover, this information is useful not only for establishing the timing of expression of these genes but also as a proxy for the amounts of the corresponding GATA factor proteins present during this critical period of endoderm development. The combination of these *in vivo* mRNA expression data and accurate comparisons of the binding kinetics of these GATA factor DBDs to different DNA binding sequences *in vitro* (described above) can be used to predict the GATA factor binding dynamics to the important site in the *elt-2* promoter during *C. elegans* endoderm development.

I was also interested in putting the molecular dynamics of the *C. elegans* endoderm developmental gene regulatory network in an evolutionary context. Results comparing *Caenorhabditis* GATA factors indicate that an *elt-3* duplication in the ancestor of the *Elegans* supergroup likely produced the ancestor of *end-3*, *end-1*, and *elt-7* (see Chapter 1). Evidence for the subfunctionalization of *elt-3* and *end-3*/*end-1*/*elt-7* paralogs was also found in that study. Future studies comparing the binding kinetics of *C. elegans* ELT-3 with that of an ELT-3 from a *Caenorhabditis* species from outside of the *Elegans* supergroup and with those of *C. elegans* END-3, END-1, and ELT-7 could help reveal how the binding kinetics of paralogous transcription factors diverged in a developmental gene regulatory network.

2.4 Methods

Protein expression and purification

The protocol for expression and purification of the proteins used for binding experiments is similar to the protocol used by Du et al. (2016). DNA binding domain (DBD) sequences were expressed from the pET His6 TEV LIC (1B) plasmid (Du et al. 2016; Addgene plasmid #29653) in Rosetta 2(DE3) pLysS competent *Escherichia coli* cells (Novagen #71403) grown at 37°C in Luria broth (LB) supplemented with 50 ug/ml kanamycin, 34 ug/ml chloramphenicol, and 20 uM $ZnCl_2$ to an OD_{600} of 0.47-0.8. Expression was induced with 0.1 mM IPTG at 14-24°C for 7-15.5 hours. Cells were pelleted, snap frozen in liquid nitrogen, and stored at -80°C. The day before an experiment, cells were thawed on ice, washed in cold (4°C) 1X PBS (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, 2 mM KH_2PO_4, pH 7.4), resuspended in lysis buffer, and sonicated with an ultrasonic microtip probe to homogenize cell lysate at 4°C (used two to four one minute cycles at 10-17% amplitude, pulsing for one second on and two seconds off, with at least four minutes of rest between cycles). For initial experiments lysis buffer consisted of 10 mM N-2-hydroxyethyl-piperazine-N'-2-ethanesulfonicacid

(HEPES), 300 mM NaCl, 100 uM $ZnSO_4$, 10% glycerol, 10 mM imidazole, 5 mM β-mercaptoethanol, adjusted to pH 7 with about 1.7 mM NaOH, plus 1 tablet of EDTA-free protease inhibitor cocktail (Roche 11-836-170-001). For experiments conducted later, to create more "physiologically relevant" binding conditions, lysis buffer consisted of 10 mM HEPES, 200 mM $NH_4CH_3CO_2$, 300 mM NaCl, 10 mM imidazole, 0.5 uM $Zn(CH_3CO_2)_2$, 10% glycerol, 5 mM β-mercaptoethanol, adjusted to pH 7.4 with about 1.6 mM HCl, plus 1 tablet EDTA-free protease inhibitor cocktail. Lysates were clarified at 18,000 g for 30 minutes at 4°C.

The His6-DBDs were then bound to Talon cobalt metal affinity beads (Clontech #635501), washed once in lysis buffer, washed a second time in modified lysis buffer: 100 mM NaCl, 20 mM imidazole, 3/6 mM HCl (initial/later), and eluted with modified lysis buffer: 100 mM NaCl, 400/200 mM imidazole (initial/later),1 mM tris(2-carboxyethyl)phosphine (TCEP), instead of 5 mM β-mercaptoethanol, 145/70 mM HCl (initial/later). Initial procedure exchanged the purified His6-DBDs buffer using a gravity desalting column (GE Healthcare) was used to exchange the buffer of the purified His6-DBDs for binding buffer (i.e., 10 mM HEPES, 150 mM NaCl, 100 uM $ZnSO_4$, 10% glycerol, 1 mM TCEP, adjusted to pH 7 with about 4 mM NaOH) at 4°C. Purified His6-DBDs were kept at 4°C or on ice until use.

The DNA binding domain (DBD) protein sequences from each of the *C. elegans* GATA factors used for the experiments reported in this chapter (with 1 being the methionine encoded by the start codon in each corresponding gene) were as follows:

END-1: amino acids 114-221

END-3: amino acids 156-242

ELT-7: amino acids 118-198

ELT-2 NF and CF: amino acids 129-336

ELT-2 CF: amino acids 203-336

Determinations of protein concentration

Purified His6-DBDs were spun down at 14000 rotations per minute (RPM) for four minutes at 4°C to pellet any precipitated/unstable DBDs, then their absorbance at 280 nm (A_{280}) was measured using a Nanodrop machine. The concentration (c, in moles/liter) of His6-DBDs was then determined using Beer's Law ($A_{280} = \varepsilon * c * l$, where l is the path length (usually 1 cm), ε is the protein's extinction coefficient, and the asterisks (*) represent multiplication). The extinction coefficients (ε) of the tagged DBDs (assuming all cystines are reduced) are (expasy.org):

END-1: 8480 $M^{-1}cm^{-1}$

END-3: 19480 $M^{-1}cm^{-1}$

ELT-7: 11460 $M^{-1}cm^{-1}$

ELT-2 NF and CF: 15930 $M^{-1}cm^{-1}$

ELT-2 CF: 8480 $M^{-1}cm^{-1}$

Concentrations of purified DBD concentrations used for experiments ranged from 10 to 55 uM. Adjusted the concentration of protein by the purity of the purification (based on Coomassie staining of the protein band of interest compared to total protein staining in that lane in a sodium dodecyl sulphate (SDS)–polyacrylamide gel run using electrophoresis (PAGE)). Only proteins that were at least 89% pure were used for results here.

Bio-layer interferometry binding kinetics

Initial experiments were carried out on an Octet RED96e instrument (ForteBio/Molecular Devices) at 15°C and shaken at 1000 RPM (Tables 2.4,2.5). The binding buffer for initial experiments was 10 mM HEPES, 150 mM NaCl, 100 uM $ZnSO_4$, 10% glycerol, 1 mM TCEP, 0.2 mg/ml bovine serum albumin (BSA), adjusted to pH 7 with 8 mM NaOH. Later experiments were carried out on an Octet RED96 instrument (ForteBio/Molecular Devices) at 30°C, 1000 RPM, and in a binding buffer comprised of 10mM HEPES, 200mM $NH_4CH_3CO_2$, 30mM NaCl, 1.5mM $MgCl_2$, 0.5uM $Zn(CH_3CO_2)_2$, 0.2 mg/ml BSA, 1mM TCEP, 8% glycerol, adjusted to pH 7 with 0.4 mM HCl and 5.6 mM NaOH (Tables 2.1,2.2,2.3,2.6). Later experiments also included 0.5 ug/ml salmon sperm DNA in the binding buffer as a competitor DNA to try to decrease non-specific binding interactions (Tables 2.1-2.3,2.5,2.6). The instrument used for the later experiments could not maintain a temperature lower than the ambient temperature and so, for the sake of consistency among the later experiments, the instrument was set at 30°C, above the ambient temperature. A preliminary trial of binding, conducted at 15°C, 20°C, and 25°C, revealed only small differences in the binding kinetics of the ELT-2 (CF only) DBD to a DNA sequence containing AGATA (data not shown).

All BLI experiments were conducted in black 96-well plates (Greiner) and sample volumes were 200 ul. Biotinylated DNA sequences (25 bps long, 100 nM) were attached to biosensor tips derivatized with streptavidin (SA) (Molecular Devices). The forward strands of the double-stranded DNA sequences used are:

TGATA DNA: 5’-Biotin(C6spacer)-CGCTGAAACTGATAAGAATAGTCGA-3’

AGATA DNA: 5’-Biotin(C6spacer)-CGCTGAAACAGATAAGAATAGTCGA-3’

CGATT DNA: 5’-Biotin(C6spacer)-GCTGGCTAGCGATTATGGCGTTAGT-3’

TTGAC DNA: 5’-Biotin(C6spacer)-CGCTGAAACTTGACAGAATAACCGA -3’.

The BLI assay was performed as follows: biosensor tips were hydrated for at least 10 minutes in binding buffer, a baseline of biosensor tips in binding buffer signal was reached (60 seconds), biotinylated DNA (100 nM in binding buffer) was loaded onto the biosensor tips (120 seconds), unbound streptavidin on biosensor tips was quenched with 10ug/ml biocytin (Sigma) in binding buffer for 120 seconds, a baseline of loaded and quenched biosensor tips in binding buffer was reached (60 seconds), association of the DBD of interest (at different concentrations) with the DNA on biosensor tips occurs for 120 seconds, dissociation of the DBD from the DNA occurs (60-600 seconds). DBD solutions were kept on ice until titrated into the 96-well plate.

Electromobility shift assays to confirm binding

DBDs, over a series of concentrations, were added to microcentrifuge tubes containing DNA (see sequences below) in binding buffer. For initial experiments the binding buffer was 10 mM HEPES, 150 mM NaCl, 100 uM $ZnSO_4$, 10% glycerol, 1 mM TCEP, 0.2 mg/ml BSA, adjusted to pH 7 with 8 mM NaOH (Figs. 2.2-2.4). For later experiments the binding buffer used was 10 mM HEPES, 200 mM $NH_4CH_3CO_2$, 30mM NaCl, 1.5mM $MgCl_2$, 0.5uM $Zn(CH_3CO_2)_2$, 0.2 mg/ml BSA, 1mM TCEP, 8% glycerol, adjusted to pH 7 with 0.4 mM HCl and 5.6 mM NaOH (Figs. 2.5,2.6).

Initial buffer conditions, 5 nM of double-stranded DNA labeled with Alexa 488 fluorophore (A488, ThermoFisher; see below), and DBDs at various concentrations were used for the electromobility shift assays (EMSAs) depicted in Figures 2.2 and 2.3. Reactions were allowed to equilibrate on ice for about 30 minutes before being run on a gel (8% 29:1 acrylamide/bis-acrylamide and 8% glycerol in 1X TB) at 100 volts (V) for 60 minutes at 4°C. Binding was then visualized with a Typhoon imager using a Fluorescein (FITC) filter.

The forward strands of the double-stranded DNA sequences used for experiments depicted in Figures 2.2 and 2.3 were:

TGATA DNA: 5’- A488-CGCTGAAAC**TGATA**AGAATAGTCGA-3’

TTGAC DNA: 5’- A488-CGCTGAAAC**TTGAC**AGAATAACCGA -3’

Initial buffer conditions, 70 nM of double-stranded DNA (Integrated DNA Technologies; see below), and various concentrations of the 6xHis-ELT-2 NF and CF DBD were used for the EMSA depicted in Figure 2.4. Reactions were allowed to equilibrate on ice for about 30 minutes before being run on a gel (8% 29:1 acrylamide/bis-acrylamide and 8% glycerol in 1X TB), at 100 V for 60 minutes at 4°C. The gel was then stained with SYBR Green I nucleic acid gel stain (Molecular Probes S7563) for about 30 minutes at room temperature. Binding was visualized on gels with a blue-light transilluminator imager.

The buffer conditions used for later experiments, 70 nM of double-stranded DNA (Integrated DNA Technologies; see below), and various concentrations of the DBDs were used for the experiments depicted in Figures 2.5 and 2.6. Reactions were allowed to equilibrate at room temperature for about 30 minutes before being run on a gel (8% 29:1 Acrylamide/Bis-acrylamide and 8% glycerol in 0.25X TBE buffer [TB buffer which also contains EDTA]), at 100 V for 60 minutes at room temperature. The gels were then stained with SYBR Green I nucleic acid gel stain (Molecular Probes S7563) for about 30 minutes at room temperature. Binding was visualized on gels with a blue-light transilluminator imager.

The forward strands of the DNA sequences used for the experiments depicted in Figures 2.4-2.6 were:

TGATA DNA:

AGCGACAGAGGTCGGGGCTGAAAC**TGATA**AGAATAGTCGACACTAACGCC

TCTGT DNA:

AGCGACAGAGGTCGGGGCTGAAAC**TCTGT**AGAATAGTCGACACTAACGCC

Acknowledgements

The dissertation author would like to thank Dr. Sharon Tracy for her guidance in protein expression, purification, and assays. The dissertation author would like to thank the laboratory of Dr. Martin Yanofsky, especially Dr. Juan-José Ripoll, for use of their sonicator. The dissertation author would like to thank the laboratory of Dr. Yishi Jin for use of their centrifuge. The dissertation author would also like to thank the laboratory of Dr. Joel Mackay, especially Dr. Joel Mackay, Dr. Jason Low, Karishma Patel, Dr. Mario Torrado Del Rey, and Ngaio C. Smith, for their advice on the stability of the protein DNA-binding domains, assay conditions, and training on binding assays. The dissertation author would also like to thank Dr. M.L. Richard Yip, for his Octet RED96e instrument demonstration. The dissertation author would also like to thank Dr. Peter Teriete for use of his labs Octet RED96 instrument.

Works Cited

Bartlett ME. 2017. Changing MADS-Box Transcription Factor Protein-Protein Interactions as a Mechanism for Generating Floral Morphological Diversity. *Integr Comp Biol* 57:1312–1321.

Bates DL, Chen Y, Kim G, Guo L, Chen L. 2008. Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. *J Mol Biol* 381:1292–1306.

Carey M, Carey MF, Smale ST. 2000. Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques. CSHL Press.

Chen Y, Bates DL, Dey R, Chen P-H, Machado ACD, Laird-Offringa IA, Rohs R, Chen L. 2012. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep* 2:1197–1206.

Concepcion J, Witte K, Wartchow C, Choo S, Yao D, Persson H, Wei J, Li P, Heidecker B, Ma W, Varma R, Zhao L-S, Perillat D, Carricato G, Recknor M, Du K, Ho H, Ellis T, Gamez J, Howes M, Phi-Wilson J, Lockard S, Zuk R, Tan H. 2009. Label-free detection of biomolecular interactions using BioLayer interferometry for kinetic characterization. *Comb Chem High Throughput Screen* 12:791–800.

Crocker J, Noon EP-B, Stern DL. 2016. The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. *Curr Top Dev Biol* 117:455–469.

Du L, Tracy S, Rifkin SA. 2016. Mutagenesis of GATA motifs controlling the endoderm regulator elt-2 reveals distinct dominant and secondary cis-regulatory elements. *Dev Biol* 412:160–170.

Esadze A, Kemme CA, Kolomeisky AB, Iwahara J. 2014. Positive and negative impacts of nonspecific sites during target location by a sequence-specific DNA-binding protein: origin of the optimal search at physiological ionic strength. *Nucleic Acids Res* 42:7039–7046.

Eurmsirilerd E, Maduro MF. 2020. Evolution of Developmental GATA Factors in Nematodes. *J Dev Biol* 8:E27.

Expasy. SIB Swiss Institute of Bioinformatics. https://www.expasy.org/.

Ferris SD, Whitt GS. 1979. Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol* 12:267–317.

Figueira ACM, Polikarpov I, Veprintsev D, Santos GM. 2010. Dissecting the Relation between a nuclear receptor and GATA: binding affinity studies of thyroid hormone receptor and GATA2 on TSHβ promoter. *PLoS One* 5:e12628.

Fukushige T, Hendzel MJ, Bazett-Jones DP, McGhee JD. 1999. Direct visualization of the elt-2 gut-specific GATA factor binding to a target promoter inside the living Caenorhabditis elegans embryo. *Proc Natl Acad Sci U S A* 96:11883–11888.

Gao R, Stock AM. 2015. Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. *mBio* 6:e00686-00615.

Gu Z, Nicolae D, Lu HH-S, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18:609–613.

Halford SE. 2009. An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem Soc Trans* 37:343–348.

Hawkins MG, McGhee JD. 1995. elt-2, a second GATA factor from the nematode Caenorhabditis elegans. *J Biol Chem* 270:14666–14671.

IUPAC. 1997. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by McNaught AD and Wilkinson A. Blackwell Scientific Publications, Oxford. Online version (2019-) created by Chalk SJ. ISBN 0-9678550-9-8.

Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20:207–220.

Ko LJ, Engel JD. 1993. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol* 13:4011–4022.

Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. 2019. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu Rev Cell Dev Biol* 35:357–379.

Lancaster BR, McGhee JD. 2020. How affinity of the ELT-2 GATA factor binding to cis-acting regulatory sites controls Caenorhabditis elegans intestinal gene transcription. *Development* 147:dev190330.

Lewis GN, Randall M. 1921. The activity coefficient of strong electrolytes. *J. Am. Chem. Soc.* 5:1112-1154.

Maduro MF. 2015. Developmental robustness in the Caenorhabditis elegans embryo. *Mol Reprod Dev* 82:918–931.

Maduro MF. 2017. Gut development in C. elegans. *Semin Cell Dev Biol* 66:3–11.

Martin DI, Orkin SH. 1990. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev* 4:1886–1898.

McGhee JD. 2013. The Caenorhabditis elegans intestine. *Wiley Interdiscip Rev Dev Biol* 2:347–367.

Merika M, Orkin SH. 1993. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol* 13:3999–4010.

Narasimhan K, Lambert SA, Yang AWH, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi HS, Reece-Hoyes JS, Fuxman Bass JI, Walhout AJM, Weirauch MT, Hughes TR. 2015. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. *Elife* 4.

Newton A, Mackay J, Crossley M. 2001. The N-terminal zinc finger of the erythroid transcription factor GATA-1 binds GATC motifs in DNA. *J Biol Chem* 276:35794–35801.

Ohno S. 1970. Evolution by Gene Duplication. Springer-Verlag.

Omichinski JG, Trainor C, Evans T, Gronenborn AM, Clore GM, Felsenfeld G. 1993. A small single-"finger" peptide from the erythroid transcription factor GATA-1 binds specifically to DNA as a zinc or iron complex. *Proc Natl Acad Sci U S A* 90:1676–1680.

Pedone PV, Omichinski JG, Nony P, Trainor C, Gronenborn AM, Clore GM, Felsenfeld G. 1997. The N-terminal fingers of chicken GATA-2 and GATA-3 are independent sequence-specific DNA binding domains. *EMBO J* 16:2874–2882.

Ptashne M. 1967. Specific Binding of the λ Phage Repressor to λ DNA. *Nature* 214:232–234.

Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. 2010. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* 79:233–269.

Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. 2009. The role of DNA shape in protein-DNA recognition. *Nature* 461:1248–1253.

Sevall JS, Cockburn A, Savage M, Bonner J. 1975. DNA-protein interactions of the rat liver non-histone chromosomal protein. *Biochemistry* 14:782–789.

Shim YH, Bonner JJ, Blumenthal T. 1995. Activity of a C. elegans GATA transcription factor, ELT-1, expressed in yeast. *J Mol Biol* 253:665–676.

Trainor CD, Ghirlando R, Simpson MA. 2000. GATA zinc finger interactions modulate DNA binding and transactivation. *J Biol Chem* 275:28157–28166.

Tsang AP, Visvader JE, Turner CA, Fujiwara Y, Yu C, Weiss MJ, Crossley M, Orkin SH. 1997. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* 90:109–119.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A* 97:6579–6584.

Waltzer L, Bataillé L, Peyrefitte S, Haenlin M. 2002. Two isoforms of Serpent containing either one or two GATA zinc fingers have different roles in Drosophila haematopoiesis. *EMBO J* 21:5477–5486.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng, H, Goity A, van Bakel H, Lozano J-C, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJM, Bouget F-Y, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158:1431–1443.

Whyatt DJ, deBoer E, Grosveld F. 1993. The two zinc finger-like domains of GATA-1 have different DNA binding specificities. *EMBO J* 12:4993–5005.

Wiesenfahrt T, Berg JY, Osborne Nishimura E, Robinson AG, Goszczynski B, Lieb JD, McGhee JD. 2016. The function and regulation of the GATA factor ELT-2 in the C. elegans endoderm. *Development* 143:483–491.

Yang HY, Evans T. 1992. Distinct roles for the two cGATA-1 finger domains. *Mol Cell Biol* 12:4562–4570.