Title: A GATA factor radiation in *Caenorhabditis* rewired the endoderm specification network

Keywords: GATA factors, gene duplications, endoderm specification network, evolution, *Caenorhabditis*

Abbreviations:
GATA factors: GATA-type transcription factors
EMS: endomesoderm
E: endoderm
smFISH: single molecule fluorescence *in situ* hybridization
DBD: DNA-binding domains
bZIP: basic leucine zipper
ZnF: zinc finger
bps: base pairs
CF: C-terminal GATA zinc finger
NF: N-terminal GATA-like zinc finger
DAPI: 4′,6-diamidino-2-phenylindole
SUMO: small ubiquitin-like modifier
BR: basic region
RT: room temperature
DI: distilled
dN/dS: ratio of the number of non-synonymous substitutions per site to the number of synonymous substitutions per site

Abstract

Although similar developmental regulatory networks can produce diverse phenotypes, different networks can also produce the same phenotype. In theory, as long as development can produce an acceptable end phenotype, the details of the process could be shielded from selection, leading to the possibility of developmental system drift, where the developmental mechanisms underlying a stable phenotype continue to evolve. Many examples exist of divergent developmental genetics underlying conserved traits. However, studies that elucidate how these differences arose and how other features of development accommodated them are rarer. In *Caenorhabditis elegans*, six GATA-type transcription factors (GATA factors) comprise the zygotic part of the endoderm specification network. Here we show that the core of this network - five of the genes - originated within the genus during a brief but explosive radiation of this gene family and that at least three of them evolved from a single ancestral gene with at least two different spatio-temporal expression patterns. Based on analyses of their evolutionary history, gene structure, expression, and sequence, we explain how these GATA factors were integrated into this network. Our results show how gene duplication fueled the developmental system drift of the endoderm network in a phylogenetically brief period in developmentally canalized worms.

Introduction

Six of the 11 GATA factors in *C. elegans* function in the endoderm specification network (Fig. 1; McGhee 2013; Maduro 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 and Rothman 2002). SKN-1 is a basic leucine zipper (bZIP)/homeodomain-like transcription factor (Bowerman et al. 1992) that first activates transcription of the functionally redundant GATA factors *med-1* and *med-2* in the endomesoderm (EMS) cell (Fig. 1; Maduro et al. 2001). MED-1, MED-2, and possibly SPTF-3 (Sullivan-Brown et al. 2016), a specificity protein transcription factor, 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) (Fig. 1; Maduro and Rothman 2002; Baugh et al. 2003; Maduro, Hill, et al. 2005; Maduro et al. 2015). POP-1 and PAL-1 are other maternally provided transcription factors (HMG box and homeoprotein, respectfully) (Lin et al. 1995; Hunter and Kenyon 1996) that make a minor contribution to endoderm specification (Fig. 1). In fact, only when other parts of the network are disrupted does PAL-1 show an effect (Maduro et al. 2001; Maduro, Hill, et al. 2005; Maduro et al. 2007; Maduro et al. 2015). 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). *C. elegans* SKN-1 binding sites (Blackwell et al. 1994) are also found in most *Caenorhabditis end-3* and *end-1* promoters suggesting that SKN-1 also directly activates them (Zhu et al. 1997; Maduro, Kasmir, et al. 2005a; 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 and Rothman 2002; Sommermann et al. 2010;

Boeck et al. 2011) (Fig. 1). ELT-7 and ELT-2 are partially redundant in regulating and directing the differentiation and maintenance of the endoderm from the 2E stage to the final twenty intestinal cells that comprise the entire endoderm of these worms (Sulston et al. 1983; Fukushige et al. 1998; Fukushige et al. 1999; McGhee et al. 2007; McGhee et al. 2009; Sommermann et al. 2010; Dineen et al. 2018). *elt-4,* a likely degenerate duplicate of *elt-2,* is expressed later in the development of the endoderm but does not have any known function (Fukushige et al. 2003). The four other canonical GATA factors in *C. elegans* (*elt-3, elt-1, elt-6,* and *egl*-18) all function during hypodermal (ectoderm) cell development (Gilleard and McGhee 2001; Koh and Rothman 2001).

GATA factors are potent endodermal regulators throughout bilaterians (Patient and McGhee 2002; 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 the genes *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-specifically expressed 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 some GATA factors has been difficult because these proteins have diverged extensively outside of their DNA-binding domains (Lowry and Atchley 2000; Gillis et al. 2008; Eurmsirilerd and Maduro 2020; Maduro 2020; see below) and (other than the MED orthologs (Broitman-Maduro et al. 2005; Lowry et al. 2009; Eurmsirilerd and Maduro 2020; see below) they are all thought to bind to canonical HGATAR DNA sites (Gerstein et al. 2010; Araya et al. 2014; Narasimhan et al. 2015; Du et al. 2016; Wiesenfahrt et al. 2016; Maduro 2020; see below).

Over the last 50 years, many studies have demonstrated that gene duplication is a major mechanism through which new genes with novel functions evolve (e.g., Ohno 1970; Gottlieb 1977; Escriva et al. 2006; Assis and Bachtrog 2013; McKeown et al. 2014). Four possible models of paralog divergence currently dominate the literature: pseudogenization (Ohno 1970; Nei and Roychoudhury 1973; Charlesworth et al. 1994; Lynch and Walsh 1998; Eyre-Walker and Keightley 1999; Denver et al. 2004; Haag-Liautard et al. 2007), neofunctionalization (Ohno 1970), subfunctionalization (Hughes 1994; Force et al. 1999; Lynch and Force 2000), and redundancy (Ohno 1970; Nei et al. 2000; Piontkivska et al. 2002; Kondrashov and Kondrashov 2006). Evidence for each of these evolutionary outcomes of gene duplication can be found in nature (e.g., Jozefowicz et al. 2003; He and Zhang 2005; Gout and 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 and these categories are not necessarily mutually exclusive for specific paralog pairs (Gera et al. 2022).

A recent comparison of nematode GATA factors found that the *elt-3* family of orthologs had undergone the most gene duplications and sequence divergence, suggesting that this gene may have evolved faster than the other GATA factors in the phylum (Eurmsirilerd and Maduro 2020). Maduro (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 species from outside of that supergroup. We took advantage of draft sequences of the genomes of an additional 34 *Caenorhabditis* species (Stevens 2020) to carry out a more comprehensive analysis of GATA factors throughout the *Caenorhabditis* genus and determine the origin of the *C. elegans*-type endoderm specification network.

To identify GATA factors in *Caenorhabditis*, we searched for their characteristic DNA-binding domain in all fifty-eight *Caenorhabditis* species for which genomic sequence assemblies or transcriptomes were available and in the genomes of two outgroup *Diploscapter* species. We estimated the evolutionary history of this gene family using maximum likelihood approaches. We focus here on the effects of an *elt-3* radiation on the developmental genetics of endoderm specification. This study illustrates how gene duplications fueled the evolution and elaboration of an essential regulatory network, all without causing any obvious change in development or morphology.

Results

The preponderance of blue and green in our GATA-domain containing phylogeny (Fig. 2) indicates that the GATA factor family dramatically radiated within the *Elegans* supergroup, expanding from a typical six genes per species to at least 10 (median of 16, maximum of 34). This radiation occurred along two consecutive internal branches in the species tree (Fig. 2), suggesting a concerted burst of gene duplication that affected different developmental genetic networks, including the endoderm specification network[26], during a phylogenetically brief period. In this paper we focus on the expansion highlighted by the black arrow in Figure 2B; analyses of the other expansions will be published elsewhere.

Of the endoderm network GATA factors, ELT-2 is found in every species in the genus, but MED, END-3, END-1, and ELT-7 orthologs are absent from non-*Elegans* supergroup species (Fig. 2; see also Eurmsirilerd and Maduro 2020). This suggests a perplexing developmental question: how do these species specify endoderm when they are missing the genes that comprise the central part of the endoderm specification network that we know about from *C. elegans*? To answer this question, we started by investigating whether the role of ELT-2 in endoderm development is conserved.

The role of ELT-2 is likely conserved throughout the genus

Using single molecule fluorescence *in situ* hybridization (smFISH) (Raj et al. 2008) we examined *elt-2* expression in *C. angaria*, of the *Angaria* group. *C. angaria elt-2* expression resembles that of *C. elegans* (Fig. 3A,C,D); i.e., its expression is restricted to the endoderm, starts in the 4E cell stage, and continues throughout embryonic development. Data from RNA-sequencing in *C. angaria* (Macchietto et al. 2017) corroborates this expression pattern (Supp. Fig. 1A).

C. elegans ELT-2 prefers to bind TGATAA sites. For example, McGhee and colleagues (McGhee et al. 2007; McGhee et al. 2009) identified 197 genes in *C. elegans* expressed specifically or predominantly in the intestine and found that the putative promoters of these genes are enriched with TGATAA sites (McGhee et al. 2007; McGhee et al. 2009). Analysis of one of these genes showed that *C. elegans* ELT-2 interacts with these sites *in vivo* to regulate its targets(Lancaster and McGhee 2020). Using reciprocal BLASTp (Altschul et al. 1990; Camacho et al. 2009), we identified orthologs of these *C. elegans* intestine-expressed genes in 57 other sequenced *Caenorhabditis* species and found that many of their putative intestinal promoters are also enriched for TGATAA sites compared to promoters for orthologs of muscle, hypoderm, and neural genes (Fig. 4; Supp. Fig. 2).

ELT-3 is the closest relative to END-1, END-3, and ELT-7 and has a broader expression pattern outside of the Elegans *supergroup*

Even if ELT-2 organizes endoderm development in non-*Elegans*-supergroup species, its expression starts too late for it to be activated by maternal factors. There must be an intervening gene (or genes) that transmits and refines the positional signal from maternal factors (Wagner 2014) over several cell divisions. This is the role played by the MED-1, MED-2, END-3, END-1, and ELT-7 GATA factors in *C. elegans* (Fig. 1). Our GATA factor phylogeny presents a clue to the possible identity of one such intervening gene: since END-3, END-1, and ELT-7 orthologs group together with ELT-3 orthologs in a well-supported clade (Fig. 2A), *elt-3* might fill that role.

In *C. elegans*, ELT-3 is expressed only in differentiating and differentiated hypoderm cells (Gilleard et al. 1999; Gilleard and McGhee 2001), which makes it an unlikely candidate for involvement in endoderm development. Indeed, even when expressed under the control of the *C. elegans end-1* promoter, in the right place and at the right time (in a *C. elegans end-1* and *end-3* double mutant), *C. elegans elt-3* is unable to initiate endoderm specification (Wiesenfahrt et al. 2016), despite being able to bind to TGATAA sites *in vitro* (Narasimhan et al. 2015). However, unlike its paralogs, ELT-3 is found in all species in the genus raising the possibility that its role in *C. elegans* is not indicative of its ancestral function.

To investigate ELT-3's role in non-*Elegans*-supergroup species we measured *elt-3* expression in *C. angaria* embryos using smFISH (Raj et al. 2008). We found that *C. angaria elt-3* is expressed in two phases. The first phase starts at the 2E stage with expression restricted to the endoderm which erodes after the 4E cell stage (Fig. 3B-D), timing that resembles that of *C. elegans end-1* (Raj et al. 2010). Moreover, single-embryo RNA-sequencing revealed expression of *elt-3* in *C. angaria* that was slightly earlier and at higher levels than in *C. elegans* (Macchietto et al. 2017) and a blip of higher expression in *C. angaria* (Supp. Fig. 1B) that coincides with the timing of *end-1* expression in *C. elegans*. In the second phase, *elt-3* in *C. angaria* is not expressed in endodermal but in hypodermal cells (Fig. 3B,C), resembling *elt-3* expression in *C. elegans* (Gilleard et al. 1999; Gilleard and McGhee 2001).

The organization of the elt-2 *promoter differs markedly between* Caenorhabditis *subclades although regulation by a HGATAR-binding transcription factor is likely conserved throughout the genus*

To investigate how *elt-2* was regulated before the *elt-3* subclade expansion (see above), we searched for conserved transcription factor binding sites in the *elt-2* promoters of non-*Elegans* supergroup species (see Materials and Methods). We found significantly (p-value less than 0.05) more canonical GATA factor binding sites, i.e. HGATAR (Ravagnani et al. 1997); in all but two of these *elt-2* promoters than expected by chance (Fig. 5A). We also examined *Elegans* supergroup *elt-2* promoters and found a striking conservation of HGATAR sites in them (Fig. 5A). There are six highly conserved HGATAR sites in most *Elegans* supergroup *elt-2* promoters (highlighted by the colored dots at the top of Fig. 5A). Three of these sites comprise the sequence TGATAA in all *Elegans* supergroup species, the only exception being *C. elegans* which does not have a HGATAR site that aligns with the most 3' of these three sites (sites under the pink dots in Fig. 5A). TGATAA sites are important for *C. elegans elt-2* expression (McGhee et al. 2007; McGhee et al. 2009; Du et al. 2016; Wiesenfahrt et al. 2016). They comprise the most overrepresented transcription factor binding site in *C. elegans elt-2* target genes(McGhee et al. 2007; McGhee et al. 2009), and TGATAA sites have been found to be the sites that *C. elegans* ELT-7, ELT-6, and ELT-3 GATA factors prefer to bind to (Narasimhan et al. 2015). Moreover, *C. elegans* ELT-2, ELT-7, END-3, and END-1 bind to TGATAA sites *in vitro* (McGhee et al. 2007; McGhee et al. 2009; Du et al. 2016; Wiesenfahrt et al. 2016). Two other conserved HGATAR sites, AGATAG and CGATAA, are found in all *Elegans* supergroup *elt-2* promoters (sites under the yellow and blue dots, respectively, in Fig. 5A). The sixth HGATAR site is the least well conserved; in 23 of 35 *Elegans* supergroup species CGATAG comprises this site, but in four species its sequence is AGATAA, in three species TGATAG, in two species AGATAG, and three species do not have a conserved HGATAR site that aligns at this position (see under the green dots in Fig. 5A). A few of these HGATAR sites similarly align in some non-*Elegans* supergroup species. However, no non-*Elegans* supergroup species contains more than one of these six sites (Fig. 5A). 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. 5A). The organization of the *elt-2* promoter in *Elegans* supergroup species evolved in parallel with the expansion of GATA factors involved in the endoderm specification network (see above) and has remained highly conserved since. We found no evidence of overrepresentation of non-GATA-factor-binding sites among the non-*Elegans* supergroup *elt-2* promoters we analyzed (Fig. 5A).

Elegans *supergroup* elt-2 *orthologs may be regulated by an Sp1 family transcription factor, SPTF-3*

We found significant numbers of Sp1-like (CYCCRCCY (Saito et al. 2013)) and/or SPTF-3 (MCGCCCMY (Narasimhan et al. 2015)) binding sites in the promoters of many (18 of 35) *Elegans* supergroup *elt-2* orthologs (blue squares next to gene names in Fig. 5A). SPTF-3 is a homolog of the Sp1 family in *C. elegans* (Ulm et al. 2011). Moreover, we found that a Sp1/SPTF-3 motif aligns near the middle of the putative promoters of 30 of 35 *Elegans* supergroup *elt-2* orthologs (Fig. 5A). Additionally, using MEME (Bailey and Elkan 1994), an Sp1-like motif was identified in 34 of 35 *Elegans* supergroup *elt-2* promoters, a significant hit rate (data not shown). This suggests that SPTF-3 or another Sp1 transcription factor may directly regulate *Elegans* supergroup *elt-2* orthologs.

*non-*Elegans *supergroup and non-*Guadeloupensis *group* 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 its expansion (see above), we searched for conserved transcription factor binding sites in the promoters of *elt-3* orthologs (see Materials and Methods). We found significant numbers of Sp1-like (CYCCRCCY (Saito et al. 2013)) and/or SPTF-3 (MCGCCCMY (Narasimhan et al. 2015)) binding sites in the promoters of most (13 of 19) *elt-3* orthologs of non-*Elegans* supergroup and non-*Guadeloupensis* group species (blue squares next to gene names in Fig. 5B). Moreover, an Sp1-like motif was the top hit identified using MEME (Bailey and Elkan 1994), which identified a similar motif in 17 of 19 *elt-3* promoters in non-*Elegans* supergroup and non-*Guadeloupensis* group species. On the other hand, similar motifs were identified in the *elt-3* promoters of only five of 37 *Elegans* supergroup and *Guadeloupensis* group species (Fig.

5B). RNAi knockdown of *sptf-3* in *C. elegans* leads to reduced *end-3* and *end-1* reporter expression and incorrectly specified endoderm (Sullivan-Brown et al. 2016). Sp1-like binding sites are also found in the promoters of most *med, end-1*, and *end-3* orthologs (Maduro 2020). Moreover, whole-embryo single-cell RNA sequencing indicates that *C. angaria sptf-3* is expressed during early embryogenesis (Supp. Fig. 1C; Macchietto et al. 2017).

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

In addition to Sp1-like binding sites, SKN-1 binding sites are overrepresented in most *med*, *end-1*, and *end-3* promoters (Zhu et al. 1997; Maduro 2020). The SKN-1 orthologs in *C. elegans* and *C. briggsae* contribute extensively to initiating the endoderm specification network, primarily by activating *med-1* and *med-2* and possibly by directly activating *end-3* (Bowerman et al. 1992; Maduro et al. 2001; Maduro, Kasmir, et al. 2005a; Maduro et al. 2007; Lin et al. 2009). We found at least one SKN-1 core binding site (RTCAT; Blackwell et al. 1994) in every *elt-3* promoter we examined (Fig. 5B); however, none of them have more SKN-1 sites than expected by chance. We did not identify any additional strongly conserved transcription factor binding sites in the promoters of any of the *elt-3* orthologs, such as POP-1 or PAL-1 sites which have both been found to contribute to endoderm specification initiation in *C. elegans* (Maduro, Kasmir, et al. 2005b; Maduro et al. 2015). However, we did identify an invariant motif, TACTATATATAGTGCATGCGCAA, in the promoters of all seven *elt-3* orthologs in the *Angaria* group (Fig. 5B). We then searched the JAPSPAR 2018 core non-redundant database (jaspar.genereg.net) for similar motifs. *Arabidopsis thaliana* FUS3, a B3 DNA-binding domain (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. However, Nrf1 contains a 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 (1 nM; Blackwell et al. 1994), and this exact sequence is not found in the invariant *Angaria* group motif, the specificity of SKN-1 in *C. elegans* may have diverged from that of SKN-1 in other *Caenorhabditis* species; alternatively, this invariant motif could be a secondary binding site for SKN-1 orthologs.

Structures, motifs, and locations of the elt-3 *paralogs hint at their evolutionary history*

Although our data indicate that *end-1, end-3,* and *elt-7* all evolved from an ancestral *elt-3,* it is not clear whether this ancestral *elt-3* duplicated once to produce an *elt-7/end* ancestor or whether two separate duplications produced the *elt-7* and the *end-3/1* subclades (Fig. 6A vs. B). To investigate this, we examined the locations and structures of, and amino acids encoded by, all of these genes. The median protein length of ELT-3 orthologs used in this study is 322 residues, 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. 3A). Ancestral structures of the *Elegans* supergroup *end-1*, *end-3*, *elt-7*, and *elt-3* genes (Supp. Fig. 3A) predicted from the structures of extant orthologs (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022) indicate that most of these genes in this clade have an intron located at the same position in their zinc finger (ZnF) coding sequence, an intron location also found in most *elt-2* orthologs and in *Japonica* group *med* orthologs (Eurmsirilerd and Maduro 2020; Maduro 2020) (Supp. Fig. 3). Moreover, the last two exons of *Elegans* supergroup *end-1*, *end-3*, *elt-7*, and *elt-3* orthologs code for their GATA domains (Supp. Fig. 3A). The combination of this conserved intron position and the GATA domain location is only found in this clade and among the *Japonica* group *meds* (Supp. Fig. 3; Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). Most *end-1* and *elt-7* orthologs have four exons, while most *end-3* and *elt-3* orthologs have three and eight, respectively (Supp. Fig. 3A). Nevertheless, we predict that the *Elegans* supergroup ancestral *end-3* gene was comprised of four exons because of the conservation of a poly-serine motif at the N-terminus of the END proteins, the fact that the ZnFs are coded for in the last two exons of *end-1* and *end-3*, and the many *end-3* orthologs that have four exons (Maduro 2020); all this evidence is consistent with a full-length duplication of an ancestral *end* gene producing the ancestral *end-1* and *end-3* genes.

We found no instances in which an *elt-3* ortholog occurred on the same piece of genomic DNA as an *elt-7*, *end-1*, or *end-3* ortholog (Supp. Fig. 4A). In fact, in the six species for which chromosome-level assemblies were available, *elt-3* orthologs are found on the X chromosome, while *elt-7*, *end-1*, or *end-3* orthologs are found on chromosome 5. Based on an analysis of synteny (see Materials and Methods), these locations are likely consistent throughout the genus (Supp. Fig. 5). Moreover, *elt-7*, *end-1*, and *end-3* orthologs are syntenic in one additional species (Supp. Fig. 4B-D). All of these pieces of evidence suggest that the *elt-7* and the *end* genes share a more recent history with each other than with *elt-3* (i.e., Fig. 6B). An additional piece of evidence is more equivocal. Maduro and colleagues(Maduro, Hill, et al. 2005; Maduro 2020) identified a poly-serine motif near the N-terminus of the END-1 and END-3 orthologs they examined. We found such a motif in 30 of 35 ELT-3 orthologs in the *Elegans* supergroup and four of 23 ELT-3 orthologs in the non-*Elegans* supergroup; on the other hand, we only found this motif in three of the 35 *elt-7* orthologs we examined. Other than their GATA domains and this poly-serine motif, we found no additional sequence homology among the ELT-3, ELT-7, and the END orthologs (Supp. Fig. 6).

Evidence of relaxed selection on one paralog relative to the other

Because gene duplication changes the functional context of genes we tested whether the intensity of selection changed after the *elt-3* expansion using RELAX (Wertheim et al. 2015). Our results indicated that in the *Elegans* supergroup both the *elt-7* orthologs (p<0.0001; k=0.34) and the *end* orthologs (p<0.0001; k=0.44) experienced less intense selection than did the *elt-3*s (Supp. Fig. 7). In turn, selection intensity relaxed on the *end-3* ortholog group after duplication as compared to the *end-1* orthologs (p<0.0001; k=0.47) (Supp. Fig. 7). Additionally, selection on *Elegans* supergroup *elt-3*s has intensified since expansion as compared to selection on non-*Elegans* supergroup *elt-3*s (p<0.0001; k=2.64) (Supp. Fig. 7). All of these patterns of selection intensity are concordant with the differences in branch lengths among these groups that are readily apparent in our phylogenetic tree (Fig. 2A).

The evolutionary history of med *orthologs is opaque due to quick turnover*

While our phylogenetic reconstruction supports a clear hypothesis about the origin of *end-3, end-1,* and *elt-7*, the *med* orthologs sit in a subclade of their own, with no clear phylogenetic connection to other groups (Fig. 2A). *med* genes code for the shortest GATA-domain-containing proteins that we identified in the 60 species included in our study (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). The *Elegans* group *meds* have no introns, while those in the *Japonica* group have one to three introns, including one splice site in the same location in their ZnF as is found in *elt-3* and its paralogs and in the *elt-2* orthologs (Maduro 2020) (Supp. Fig. 3), but which is not found in any of the other GATA-domain-containing proteins we identified (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). Although the *C. elegans* MEDs bind an atypical motif (Broitman-Maduro et al. 2005; Lowry et al. 2009), their DNA-binding domains more closely resemble canonical GATA domains than they do the atypical GATA domains of EGL-27, SPR-1, or RCOR-1 (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022), supporting the hypothesis that the MEDs arose from one of the canonical *Caenorhabditis* GATA factors instead of from a different, more atypical GATA-domain-containing protein. Syntenic *med* paralogs are usually relatively close to each other (range of 319 bp to 13 kb, median of 3.1 kb; Supp. Fig. 8A,B) and similar in sequence (93% median nucleotide identity, compared to 77% for non-syntenic *med* paralog pairs; Supp. Fig. 8C,D). However, most sister species have *med* paralogs on at least two different chromosomes (Supp. Fig. 5; Supp. Fig. 9). Our phylogeny (Fig. 2A) shows that most MEDs are most closely related to paralogs within their own species and have highly variable copy numbers across species, indicating that these genes evolve through rapid duplication and loss (Maduro 2020).

Discussion

Our analysis of the evolution of GATA factors in the 58 *Caenorhabditis* species for which proteome sequences are currently available shows that the genes of most of the GATA factors involved in endoderm development – *end-1*, *end-3*, *elt-7*, and the *med* genes – arose during the course of a larger GATA factor expansion in the genus (Fig. 2). Although this radiation re-wired the endoderm specification network, it was not associated with any known change in the environment, morphogenesis, or anatomy of these animals. Additionally, we found that the role of the most downstream GATA factor in this network, encoded by *elt-2*, is likely conserved across the genus (Fig. 3A,C; Fig. 4). Interestingly, five GATA factor binding sites were likely fixed in *elt-2* promoters at the base of the *Elegans* supergroup concurrent with the elaboration of its *trans*-activating network (highlighted by the colored dots at the top of Figure 5A). The concentration of a single type of transcription factor in a gene regulatory network – especially one as temporally and spatially restricted as the endoderm network – is extraordinarily rare and creates the potential for extensive developmental system drift.

Maduro (Maduro 2020) hypothesized that two *elt-2* duplications in the *Elegans* supergroup ancestor produced an ancestral *end/med* gene and an ancestral *elt-7* gene (Fig. 6E). This hypothesis was supported by the fact that *elt-2*, *elt-7*, *end-1*, *end-3*, and the *med* orthologs all function in the *C. elegans* endoderm specification network (Zhu et al. 1997; Broitman-Maduro et al. 2005; McGhee et al. 2007; McGhee et al. 2009; Sommermann et al. 2010; Dineen et al. 2018), *elt-2* orthologs are conserved throughout the *Caenorhabditis* genus, and *elt-2* orthologs share a conserved intron location within their ZnFs with *end-1*, *end-3*, *elt-7*, and *Japonica* group *meds*. However, *elt-3* orthologs have the same conserved intron location in their ZnFs which has been conserved in all 58 *Caenorhabditis* species we examined (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022; Supp. Fig. 3A). Moreover, our phylogeny indicates that ELT-3 orthologs share a more recent common ancestor with ELT-7, END-1, and END-3 orthologs than with any other *Caenorhabditis* GATA-domain-containing orthologs (Fig. 2A), and our smFISH analysis indicates that *C. angaria elt-3* mRNA is expressed in the early endoderm (Fig. 4B,C). Based on this new evidence, we argue that one or two *elt-3* duplications in the *Elegans* supergroup ancestor produced the ancestors of the *end* and *elt-7* genes.

Consistent with previous results (Maduro, Hill, et al. 2005; Boeck et al. 2011; Maduro 2020), our phylogeny (Fig. 2A) indicates that END-3 and END-1 orthologs share a recent common ancestor; the evolutionary relationship(s) between ELT-3, ELT-7, and the END ancestor is less clear, however. The ELT-7 ortholog group branches off of a shared node between the ELT-3 and the END orthologs (0.2 substitutions per site in Figure 2A), suggesting that END orthologs are more closely related to ELT-3 orthologs than to ELT-7 orthologs. This finding supports a scenario in which separate *elt-3* duplications produced the *elt-7* and *end* ancestors, as opposed to 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 (Fig. 6A vs. B). However, both the structures and chromosome locations of these GATA-factor-encoding genes (Supp. Figs. 3A,4) support a scenario in which a single *elt-3* duplication (along with a single shortening and interchromosomal rearrangement event) gave rise to the *elt-7/end* ancestor (e.g., Fig. 6C or D). If the ELT-3 and END N-terminal poly-serine motifs and/or the SPTF-3 regulatory sites in the promoters of the non-*Elegans* supergroup and non-*Guadeloupensis* group *elt-3* orthologs and *end* genes are homologous, it supports a scenario in which a full-length duplication of the *elt-3* coding sequence produced the *end* ancestor (followed by sequence loss to produce the final four-exon version). If the *elt-7* and *end* ancestors arose from the same *elt-3* duplication, then their ancestor likely quickly duplicated again since these ortholog groups experienced different subsequent 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 poly-serine motif and SPTF-3 binding sites in *elt-7* orthologs we observed (data not shown) could be the result of a partial *elt-3* duplication producing the ancestral *elt-7* gene or due to greater relaxation of selection pressure experienced by *elt-7* orthologs. This ambiguity in the precise birth order of the *elt-7* and *end* gene ancestors may reflect the fact that this 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, albeit in different ways.

The expression of *C. elegans end-3* starts before that of *end-1*, whereas the peaks of *end-3* and *end-1* mRNA expression occur in first (1E) and second (2E) cell stages of endoderm development, respectively (Raj et al. 2010). *C. elegans elt-7* expression starts after that of *end-1*, during 2E (Sommermann et al. 2010). *C. elegans elt-7* continues to be expressed for the lifetime of the worm, whereas the *ends* genes are only expressed transiently during endoderm specification (Raj et al. 2010; Sommermann et al. 2010). Based on our finding that *C. angaria elt-3* is expressed similarly to *C. elegans end-1* in the endoderm (Fig. 3B-D), we predict that this was the endoderm expression pattern of the *Elegans* supergroup ancestral *elt-3* paralog. This suggests that the expression patterns of *end-3* and *elt-7* diverged from that of their ancestor. Despite this apparent divergence in gene expression patterns, the functions of *C. elegans end-3*, *end-1*, and *elt-7* have been found to be interchangeable. For example, the expression of *end-3*, or *end-1*, or *elt-7* in the early endoderm is sufficient for gut specification (Zhu et al. 1998; Maduro, Hill, et al. 2005; Wiesenfahrt et al. 2016) and ectopic expression of any of these GATA factors is sufficient to activate expression of endoderm markers (Maduro and Rothman 2002; Sommermann et al. 2010). This suggests that these paralogs have primarily functionally diverged through *cis*-regulatory changes, while their protein sequence differences have not been found to have functional consequences. However, the DNA binding preferences of *C. elegans* END-3 and END-1 are less specific than those of *C. elegans* ELT-7 (Narasimhan et al. 2015). The ENDs bind to GATA sequences, without much preference regarding the flanking base pairs, while ELT-7 prefers to bind to TGATAA sequences (Narasimhan et al. 2015). Moreover, we found that non-TGATAA HGATAR sites are highly conserved in *Elegans* supergroup *elt-2* promoters (Fig. 5A), suggesting that these non-TGATAA sites are more likely bound by the ENDs while the TGATAA sites are preferably bound by ELT-7 and ELT-2. Therefore, the binding preference of ELT-3 paralogs may have diverged in parallel with the expression pattern of *elt-3*.

Given the resemblance of the invariant Nrf1/bZIP motif in *Angaria* group *elt-3* promoters to a possible SKN-1 binding site (see Results) and the involvement of SKN-1 in the endoderm specification network – at least in *Elegans* supergroup species (Maduro 2020), we hypothesize that *Angaria* group SKN-1 orthologs bind to this invariant sequence to activate *elt-3* expression in early endoderm cells (Fig. 7 left side). If true, and if ELT-3 is indeed part of the endoderm specification network in non-*Elegans* supergroup species, then any regulation of the network involving SKN-1 should be conserved in the initial stages of endoderm specification, despite change in the SKN-1 binding site.

While our phylogeny strongly suggests that an *elt-3* expansion occurred in the *Elegans* supergroup ancestor (Fig. 2A), the tree also suggests that additional *elt-3* duplications occurred elsewhere in the genus. We identified divergent *elt-3* paralogs in at least two of the three *Guadeloupensis* group species and in *C. astrocarya*, a species likely basal to the *Elegans*/*Guadeloupensis* species (Fig. 2); interestingly, most of these divergent *elt-3* paralogs have shorter gene structures (two to six exons and median of 220 amino acids) more like those of *elt-7*, *end-1*, and *end-3* (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022) and, like the *Elegans* supergroup *elt-3*s, their promoters do not have SPTF-3 binding motifs (Fig. 5B). Therefore, our evidence also fits the alternative scenario that *elt-3* duplication, shortening, and subfunctionalization occurred one stage earlier, i.e. in the *Elegans*/*Guadeloupensis*/*astrocarya* ancestor, followed by extensive divergence of *elt-3* paralogs in the different lineages.

In *C. elegans, elt-2* overexpressed under the control of the *end-1* promoter can compensate for loss of *end-3*, *end-1*, *elt-7*, and *elt-4* (Wiesenfahrt et al. 2016). Analogous expression of *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 to ELT-2 orthologs (Fig. 2A). This suggests that *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, our finding that mRNA of the *C. angaria elt-3* ortholog is expressed in early endoderm cells (Fig. 3B,C) in a pattern reminiscent of *end-1* expression, suggests that it probably also functions there. Not only did the *Elegans* supergroup ancestral *elt-3* likely subfunctionalize its expression pattern, something else must have changed about the coding region in its descendants such that one branch preserved its capacity to function in the endoderm while the other branch lost this ability. We did not find any obvious differences between the DNA binding domains in the *elt-3* orthologs in *Elegans* supergroup versus non-*Elegans* supergroup species (Supp. Fig. 6). However, we did find a highly conserved, possible SUMOylation site (Chang et al. 2018) towards the N-terminus of most (19 of 23) ELT-3 orthologs in non-*Elegans* supergroup species, and that this [VIA]KE[ED] motif has been lost from all the ELT-3 orthologs in *Elegans* supergroup species (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). We therefore speculate that ELT-3 orthologs in non-*Elegans* supergroup species could undergo post-translational modification (associated with this putative SUMOylation site) and be involved in an endoderm-specific protein-protein interaction(s). Interestingly, our finding that selection has been more intense on the *elt-3* orthologs in *Elegans* supergroup species compared to those in the non-*Elegans* supergroup species (Supp. Fig. 7), may be reflective of a functional change in the *Elegans* supergroup ELT-3s.

We found no additional evidence for a *med* gene ancestor originating from an *elt-2* duplication (as proposed by Maduro (Maduro 2020)) but, rather, more evidence that the *med* ancestor originated from a duplication within the *end-3/end-1/elt-7/elt-3* clade in the *Elegans* supergroup ancestor. The many species-specific paralogs and long phylogenetic branches found in the MED ortholog group (Fig. 2) suggest that the *med* genes turn over quickly, as previously noted (Maduro 2020). This quick turnover has likely erased additional evidence relating to the relationship between the MED orthologs and other GATA factors. The strongest evidence for the origin of the *med* orthologs (on which Maduro (Maduro 2020) based his 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*, *end-3*, *elt-2*, and *elt-7* orthologs as well as of *elt-3* orthologs, as we have shown (Supp. Fig. 3; Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). The structures of *Japonica* group *med* genes are also most similar to those of *end-3* homologs (Supp. Fig. 3A,B; Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). Since the *Elegans* supergroup ancestral *end-3* gene may have lost an intron after diverging from *end-1* (see Results), a full-length *end-3* duplication could have produced the *med* ancestor. However, a partial duplication of an *end-1*, *elt-2*, *elt-3*, or *elt-7* ortholog, as well as a full duplication of any of these ancestral genes followed by coding sequence loss, are additional possible explanations. Since we have shown that three gene duplications occurred within the *end-3/end-1/elt-7/elt-3* clade and were likely fixed in the *Elegans* supergroup ancestor (Fig. 2), it is plausible that at least one more could have occurred to produce the *med* clade**.**

In conclusion, our data suggest that *elt-2* plays a consistent role throughout the *Caenorhabditis* genus in regulating, through TGATAA binding sites, hundreds of genes expressed specifically or predominantly in intestine (Fig. 3A,C,D; Fig. 4; Fig. 7). As we have shown in *C. angaria* (Fig. 3B-D; Fig. 5; Fig. 7), we predict that *elt-3* orthologs function before *elt-2* orthologs in the endoderm specification network of non-*Elegans* supergroup species and did so in the *Elegans* supergroup ancestor. Evidence also suggests that the *Elegans* supergroup ancestral network may have been initiated by SPTF-3 activating *elt-3* (Fig. 5B). SKN-1 may also play a conserved role in the initiation of this network across the genus, through an as yet unknown gene(s) upstream of *elt-3* that is analogous to the *med* and *end-3* genes which were subsequently displaced by the radiating GATA factors (Fig. 7). Or SKN-1 may directly regulate *elt-3*, even though we did not find significant numbers of SKN-1 binding sites in *elt-3* promoters (Fig. 5B; Fig. 7). Figure 7 summarizes our proposed model of how the *Elegans* supergroup ancestral endoderm specification network evolved in a relatively short period of time, all without any apparent phenotypic change.

Materials and Methods

Phylogenetic analysis

GATA factor homolog identification and an initial phylogenetic analysis was done for a companion paper (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022). We used the same alignment and tree building procedure, with an additional 3000 ultrafast bootstraps (Minh et al. 2013; Hoang et al. 2018), for creating a phylogenetic inference of the 714 protein sequences we deemed “confident” (Supp. Table 1; see companion paper (Darragh AC, Rifkin SA, unpublished data, https://doi.org/10.1101/2022.05.20.492891, last accessed May 23, 2022) for the protocol used to determine “confident” proteins); the resulting maximum likelihood tree is shown in Figure 2A.

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

We classified the 714 proteins we were most confident in as a *singleton* or *paralog*, depending on how many proteins from each species clustered into the 12 ortholog groups in our phylogeny (Fig. 2A). 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 END-3 ortholog group contained the most confident paralogs (Fig. 2A) and, since we focus here on how an *elt-3* radiation produced *end-3*, *end-1*, and *elt-7***,** we further categorized these paralogs as either *representative* or *divergent*. This classification is based on how conserved the paralog's sequence is in relation to the sequences of singletons in the same ortholog group (Supp. Table 1); if an individual paralog within a species had a noticeably higher level 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 exhibited approximately equal levels of conservation to singleton orthologs, they were all considered representative paralogs. For most analyses, we used both representative paralogs and singletons; for most analyses of the MED ortholog group, we used single ZnF paralogs and singletons; and for the other groups we used only singletons for most analyses.

Worm maintenance

C. angaria strain PS1010 was grown at room temperature (RT, approximately 21-22 ºC) 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 (250 mM NaOH, 1% NaOCl, in DI water) and then, following standard *C. elegans* protocols for synchronizing them, 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 one day until they hatched. Synchronized, larval stage one (L1) worms were then pipetted onto NGM Lite agar plates with *E. coli* lawns and grown, using standard procedures, at RT for four days until the by then adult worms started laying eggs. To enrich for early embryos (i.e., those still inside the worms), 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 treated with worm-bleaching solution (as described 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 PBS, placed in 70% ethanol in nuclease-free water (Ambion), and stored at 4 ºC for at least overnight and up to one week. Embryos were then washed in a solution (wash buffer) comprised of 10% formamide, 2x SSC (300 mM NaCl, 30 mM $Na_3C_6H_5O_7$) prepared using nuclease-free water. Hybridizations were carried out as previously described (Raj et al. 2008; Raj et al. 2010) in 100 uL 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 hybridization solution for 16 hours at 30 ºC. Embryos were then washed with wash buffer and their nuclei stained with 5 ug/mL DAPI (4′,6-diamidino-2-phenylindole, Roche) prepared in wash buffer for 10 minutes at 30 ºC.

For imaging, embryos were suspended in RT glox buffer comprised of 20 mM Tris Cl pH 7.5, 2x SSC, 0.4% glucose in nuclease-free water, 37 ug/mL glucose oxidase, and 1 ul catalase (Raj et al. 2010; Sigma-Aldrich). 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 and 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 TGATAA sites in the promoters of orthologous genes expressed gut-, muscle-, neural-, or hypoderm-specifically/enriched

Using a custom Python script, we identified the longest isoform among each of the following groups of *C. elegans* genes that are specifically expressed or enriched for expression (as per McGhee et al. 2007, 2009) in the organ/tissue indicated: 197 intestine, 71 muscle, 47 neural, and 79 hypoderm. We then used reciprocal BLASTp (e-value 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 that had at least two orthologs. Next, we identified the putative proximal promoters (i.e., 2 kb upstream of each coding sequence start, if available) of each ortholog. Some gene start codons are very close to the end of their scaffold/contig, if there was less than 5 bps upstream of the ATG we did not include this sequence as a promoter. We then determined the number of TGATAA sites in each promoter. Next, we used hierarchical clustering with Euclidean distance metrics to organize the genes by number of TGATAA sites in their promoters (and whether the species had a putative ortholog). These data are what is plotted in Figure 4 and Supplemental Figure 2.

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

Using a custom Python script, we extracted the sequences comprising 1200 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 (Bailey and Elkan 1994) 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 DNA strands (revcomp), motif widths between five to 12, expected site distribution, and any number of repetitions (anr, which often identified more repetitive A/T-rich motifs in *elt-3* promoters) or zero or one occurrence (zoops); the program usually stopped finding additional 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 following specific sites identified using MEME: canonical and non-canonical GATA-factor-DNA-binding sites, and binding sites for the *C. elegans* endoderm transcription factors SKN-1, POP-1, SPTF-3, and PAL-1 (Maduro, Kasmir, et al. 2005b; Maduro et al. 2015) (Fig. 1A). 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).

Gene structure comparisons and predictions of ancestral gene structures

Using a custom Python script, the exon sequences listed in the Supplemental Table 1 column “exonSeq”, and the respective scaffold sequence we determined intron lengths in our selected genes (data not shown). Using a custom Python script we also found the locations of the GATA ZnF domains that we had previously identified in each confident protein (Supp. Table 1; see above). Eurmsirilerd and Maduro (Eurmsirilerd and Maduro 2020) defined a poly-S region as at least six serines within ten adjacent residues. Using a custom Python script, we found the locations of any such motifs (Supp. Table 1). Using the exon lengths and the domain/motif location information, we created representations of the gene structures of all the genes we deemed “confident” in this study and for which genomic data was available using a custom R script (data to be reported elsewhere). We also marked the locations of the possible SUMOylation sites ([VIA]KE[ED] ;Chang et al. 2018) that we found in *elt-3* orthologs (Supp. Table 1). (Note: the *C*. sp. *45* and *C*. sp. *47* genes were excluded from this analysis because only transcriptome data were available for these species.)

We visually compared the gene structures of the confident genes (which will be reported elsewhere) 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 and Penny 2006)), then predicted ancestral gene structures (exon number and domain location(s)) for the END-3, END-1, ELT-7, ELT-3, ELT-2, and MED ortholog groups (Supp. Fig. 3). 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.

Predicting chromosome location

Since the genome assemblies of most of the species used in this study lack chromosome-level resolution we also used a custom Python script to identify all annotated genes within 70 kb upstream and downstream of each confident gene (“neighbor genes”). We then used BLASTp (e-value cutoff 0.1) to search for each neighbor gene’s longest isoform “tophit” in the *C. elegans* proteome and then determined which chromosome that *C. elegans* tophit was coded on. This information on neighbor genes and their associated *C. elegans* tophit chromosomes is what is plotted for each confident gene in Supplemental Figure 4.

Testing for extent of selection pressure on paralogs and orthologs

RELAX (Wertheim et al. 2015) compares two sets of branches in a phylogenetic tree and evaluates whether the data better fits a single distribution of a few ratios of the number of non-synonymous substitutions per site to the number of synonymous substitutions per site (dN/dS) as rate categories among all branches, or 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 used RELAX to test four hypotheses about the strength of selection between sets of branches in the clades of *elt-3*, *elt-7*, *end-1*, and *end-3* orthologs. In several cases there were paralogs of the same gene within a species (e.g., two *end-3s*) for which one gene was more conserved and the other(s) more divergent (black bars in Supplemental Figure 7). In these cases, we only included the more conserved gene in our analysis, which made our tests more conservative. We used three possible rate categories and the RELAX default settings for each test.

Data availability

The custom Python, R, and MATLAB scripts used for this article will be shared on reasonable request to the corresponding author.

Acknowledgements

We thank Lewis Stevens for pre-publication access to *Caenorhabditis* genomic data. We gratefully acknowledge The Caenorhabditis Genome Project for *Caenorhabditis* and outgroup datasets. Special thanks to Rifkin lab members for their helpful discussions. This work was supported by the National Institutes of Health (R01 GM103782); and the National Science Foundation (IOS 1936674).

Supplemental Table 1. Additional details on the proteins, and the genes encoding them, included in this analysis. (See separate large csv file).

Supplemental Table 2. Number of TGATAA sites in putative promoters of orthologs specifically or predominantly expressed in gut, muscle, neural, or hypoderm tissues. (See separate large csv file).

Figure legends

Figure 1. *C. elegans* endoderm specification network. The *C. elegans* endoderm specification network is shown on the right and the approximate embryonic stages during which most of the gene expression associated this network takes place is shown on the left. This network is initiated primarily by SKN-1 in the EMS cell (orange cell on bottom of four-cell embryo); however, SPTF-3, POP-1, and possibly PAL-1 also contribute to the activation of this GATA factor cascade, as shown. Six of the 11 *C. elegans* GATA factors (*med-1*, *med-2*, *end-3*, *end-1*, *elt-7*, and *elt-2*) function in this network (as shown). *med-1* and *med-2* expression initiates in the EMS cell and MED-1 and MED-2 regulate genes in both the first endoderm (1E) cell (green cell on bottom right of eight-cell embryo) and the first mesoderm (MS) cell (purple cell to the left of the 1E cell). *end-3* expression starts in the late EMS or early 1E cell while *end-1* expression starts in the late 1E or early two endoderm (2E) cell stage (two green cells in 14-cell embryo). *elt-7* expression starts in the 2E cells and *elt*-2 expression starts near the beginning of the 4E cell stage (not shown). Black arrows indicate well supported regulatory connections, while gray and dashed gray arrows represent weaker and not as well supported interactions, respectively.

Figure 2. Inferred evolutionary history of *Caenorhabditis* GATA-domain-containing proteins. (A) Maximum likelihood phylogeny of 714 “confident” GATA-domain-containing proteins 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 tree includes both canonical GATA factors and EGL-27, SPR-1, and RCOR-1 orthologs which are proteins that contain atypical GATA-binding domains but which scored above our threshold on the PROSITE GATA-type ZnF domain profile. 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 C below). The names of the 12 ortholog groups the 714 proteins were categorized into are indicated in the lighter of the two outer gray rings (with white gaps between groups). Clades comprising 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. 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; color-species designations are the same as used in (B**)** above. The black arrow points to the *Elegans* supergroup ancestral branch where the ancestral *med*, *end-1*, *end-3*, and *elt-7* genes, as we know them from *C. elegans*, likely arose.

Figure 3. Expression of *elt-3* and *elt-2* mRNA in *C. angaria*, a non-*Elegans* supergroup species. (A-C) 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 smFISH. 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). **(D)** Model of *C. angaria* endoderm specification network based on these smFISH results.

Figure 4. Conservation of TGATAA sites in putative promoters of orthologs specifically expressed or enriched for expression in gut and muscle. Heatmaps of the number of TGATAA sites in the promoter regions of orthologs expressed specifically or primarily in (A) gut versus (B) muscle 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 2B. Each row on the y-axis represents the promoter region of a *C. elegans* gene (McGhee et al. 2007; McGhee et al. 2009), 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 indicates species for which we did not find an ortholog for that *C. elegans* gene. **(A)** Promoters of *C. elegans* orthologs specifically expressed or enriched for expression in gut. **(B)** Promoters of *C. elegans* orthologs specifically expressed or enriched for expression in muscle.

Figure 5. 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 the predicted proximal promoter sequence of an *elt-3* (A) or *elt-2* (B) ortholog and they are listed in the same order (top to bottom) as the *Caenorhabditis* species in the phylogeny shown in Figure 2B (left to right). 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* ortholog promoter sequences. Note the highly conserved HGATAR site in the *Elegans* group species (indicated above the panel). **(B)** *elt-2* ortholog promoter sequences. Note the highly conserved HGATAR sites (colored circles) in the *Elegans* supergroup species (as highlighted above each panel).

Figure 6. Scenarios for how initiation of the expansion of endoderm specification GATA factors could have occurred. Comparison of possible gene duplication scenarios for initiating GATA factor expansion, those supported by our results (A-D) and another proposed by Maduro (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. **(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. **(C)** Details of the proposed 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)** Details of a proposed 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 (Maduro 2020) for how two *elt-2* duplications could have produced the *elt-7* and *end* ancestor genes. The key to color-coding of gene domains and expression patterns is located in the upper right corner of the figure.

Figure 7. Evolutionary model of how GATA factors expanded in the endoderm specification network. Data from this study are consistent with this evolutionary model in which, prior to our proposed expansion of the *elt-3* gene in the endoderm specification network (left side of figure), the functioning of this network was initiated by expression of *sptf-3* and/or *skn-1*, which activated *elt-3* (and possibly another transcription factor expressed earlier, depicted as “A non-GATA factor?”). Expression of ELT-3 (and possibly other transcription factors) then activated *elt-2*. ELT-2 then likely regulated hundreds of genes expressed specifically (or primarily) in the intestine and perhaps auto-regulated its own gene expression. This “pre-expansion” network (shown on the left) is expected to be similar to the endoderm specification networks found in non-*Elegans* supergroup and non-*Guadeloupensis* group species, like *C. angaria*. Our data suggest that a duplication(s) of *elt-3* led to the addition of three or four GATA factor paralogs to the endoderm specification network that function between *sptf-3* and/or *skn-1* and *elt-2* resulting in the network shown on the right. This model predicts that during the GATA factor expansion *elt-3* paralogs subfunctionalized into: an *elt-3*-like gene expressed only in the hypoderm (not shown), an

endoderm-specifically expressed *elt-7*, and an ancestor of the *end* genes. (See Figure 6A-D for molecular details of how this subfunctionalization could have occurred). Data from this study also support the previously proposed hypotheses that an additional *end* gene duplication produced the ancestors of *end-1* and *end-3* (Maduro, Hill, et al. 2005; Coroian et al. 2006) and that another *end* gene duplication likely produced the ancestor *med* gene (Maduro 2020). Neither we nor Maduro (Maduro 2020) found POP-1 nor PAL-1 transcription factor binding sites overrepresented in *end-1* (or *end-3*) promoters and therefore they are not included in the network on the right. Black arrows indicate well supported regulatory connections, while gray and dashed gray arrows represent weaker and not as well supported interactions, respectively.

Supplemental Figure 1. RNA-seq analysis of *C. angaria* and *C. elegans* genes of interest. RNA-seq data from embryos at 10 different stages of development and from the first larval stage (L1), in *C. angaria* and *C. elegans* (Macchietto et al. 2017), were used as RNA-seq inputs. The developmental stages that were sampled are listed (in chronological order) on the x-axis. The numbers of transcripts corresponding to each gene, normalized as transcripts per million (TPM), are shown on the y-axis. In all panels, *C. angaria* data are shown in cyan and *C. elegans* data are 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 2. Conservation of TGATAA sites in putative promoters of orthologs expressed or enriched in gut, muscle, neural, or hypoderm tissue. 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 2B (top to bottom). 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 indicates species for which we did not find an ortholog for that *C. elegans* gene.

Supplemental Figure 3. Comparisons of predicted ancestral gene structures. (A) Predicted *Elegans* supergroup ancestral gene structures for *elt-3*, *elt-7*, *end-1*, and *end-3*, respectively. **(B)** Predicted *Elegans* and *Japonica* group ancestral *med* gene structures, respectively. **(C)** Predicted *Elegans* supergroup ancestral *elt-2* gene structure. 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); N-terminal GATA-like zinc fingers (NFs) are in pink; C-terminal GATA zinc fingers (CFs) are in blue; and the basic regions of GATA domains (BR) are in red.

Supplemental Figure 4. Contig/scaffold/chromosome locations of *end-3/end-1/elt-7/elt-3* clade genes. 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 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 Figure 2B. (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)** *elt-3*; **(B)** *elt-7*; **(C)** *end-1*; and **(D)** *end-3*.

Supplemental Figure 5. Chromosome assignments for genes on scaffolds or contigs. To expand our analysis of the chromosome locations of these GATA factors throughout the genus, we assigned scaffolds or contigs to chromosomes based on the *C. elegans* assembly. Each dot corresponds to a neighbor gene (of the gene represented on the x-axis); the dot location along the y-axis shows the chromosome the *C. elegans* homolog of that neighbor is located on. The color of the dot indicates the ortholog group the gene is assigned to; the key to these color codes is shown on the right. The genes are ordered as in the Figure 2A phylogeny. The numbers and X on the y-axis refer the designations of the six *C. elegans* chromosomes.

Supplemental Figure 6. Alignment of ELT-7, ELT-3, END-1, and END-3 orthologs. The consensus sequence is shown at the top. The percent identity to the consensus is plotted underneath the consensus sequence. Intensity of residue shading indicates similarity, more intense representing more similar. Protein domains designated within the sequences are highlighted as colored rectangles above the alignment. The domain color-coding is the same as used in Figure 6 and Supplemental Figure 3 (i.e., ZnF in blue, and BR in red). The gene names are shown on the left of the alignment. A vertical bar to the left of the gene names is colored by ortholog group. The ortholog group color-coding is the same as in Supplemental Figure 5 (i.e., ELT-7s with pink, ELT-3s with purple, END-1s with blue, and END-3s with teal). The species colors are shown to the left of the ortholog group bar (which are the same as in Figure 2B).

Supplemental Figure 7. Comparison of selection intensity on ELT-3, ELT-7, END-1, and END-3 clades after *elt-3* expansion. The RELAX test was used to compare the intensity of selection imposed on these clades of 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 are 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 are 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 8. 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 DNA 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 is 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 9. Contig/scaffold/chromosome locations of *med* ortholog group genes. Contigs/scaffolds/chromosomes (depicted as gray horizontal rectangles) are anchored on a *med* gene (depicted as orange 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 are depicted as filled in or outlined colored squares, respectively. The ortholog group color key is at the top of the 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 Figure 2B. (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.)

References

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, et al. 2014. Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* 512:400–405.

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

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28–36.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Gera T, Jonas F, More R, Barkai N. 2022. Evolution of binding preferences among whole-genome duplicated transcription factors.Landry CR, Struhl K, Kuzmin E, editors. *eLife* 11:e73225.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 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.

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.

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.

He X, Zhang J. 2005. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics* 169:1157–1164.

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.

Hunter CP, Kenyon C. 1996. Spatial and Temporal Controls Target pal-1 Blastomere-Specification Activity to a Single Blastomere Lineage in C. elegans Embryos. *Cell* 87:217–226.

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

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.

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.

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.

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.

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

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.

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, et al. 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, et al. 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.

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

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

Nei M, Rogozin IB, Piontkivska H. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A* 97:10866–10871.

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

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

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.

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.

Piontkivska H, Rooney AP, Nei M. 2002. Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. *Molecular Biology and Evolution* 19:689–697.

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.

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.

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

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.

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.

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.

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.

Stevens L. 2020. Genome evolution in the genus Caenorhabditis. Available from: https://era.ed.ac.uk/handle/1842/36871

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, et al. 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.

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

Wagner GP. 2014. Homology, genes, and evolutionary innovation. Princeton ; Oxford: Princeton University Press

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.

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.

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.

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.

Figure 1

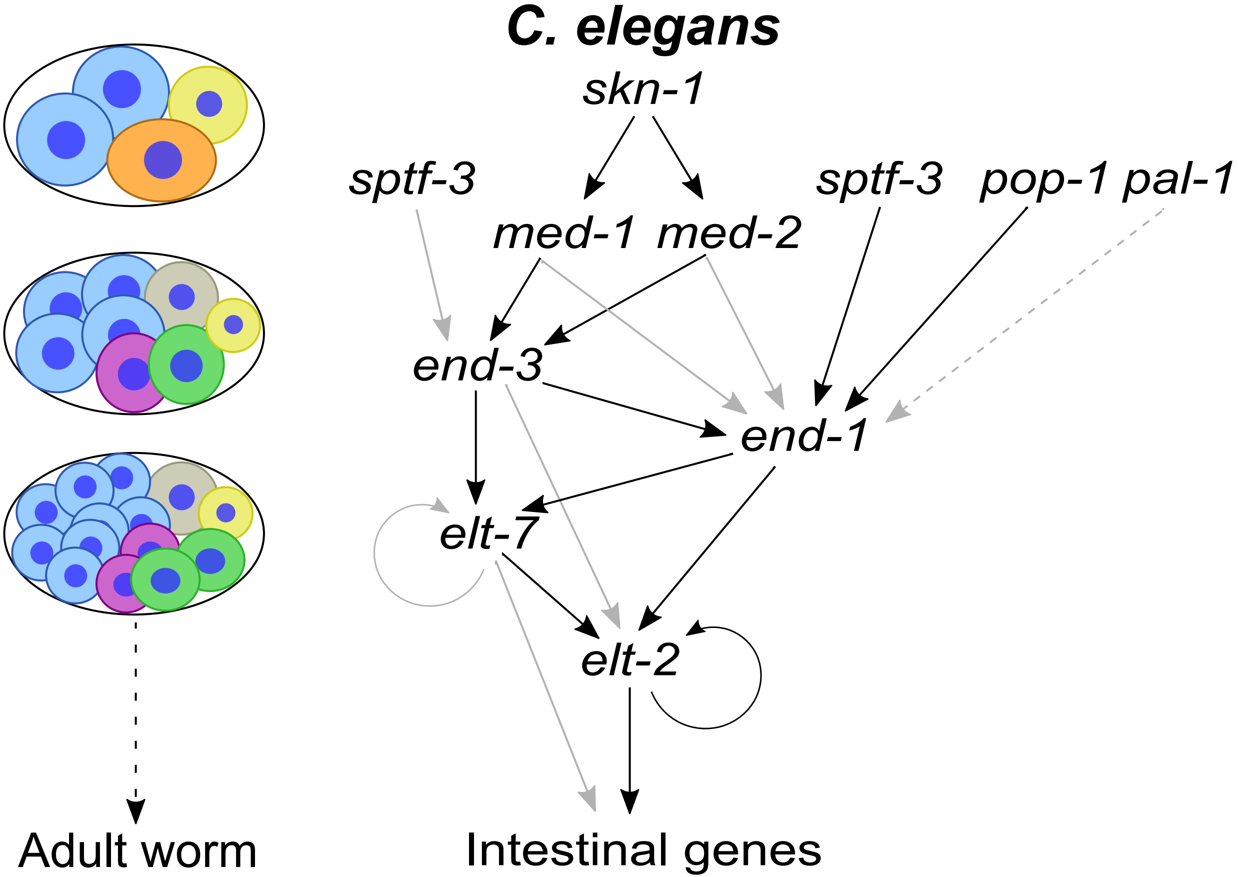

C. elegans
skn-1
sptf-3
med-1
med-2
sptf-3
pop-1
pal-1
end-3
end-1
elt-7
elt-2
Adult worm
Intestinal genes

Figure 2

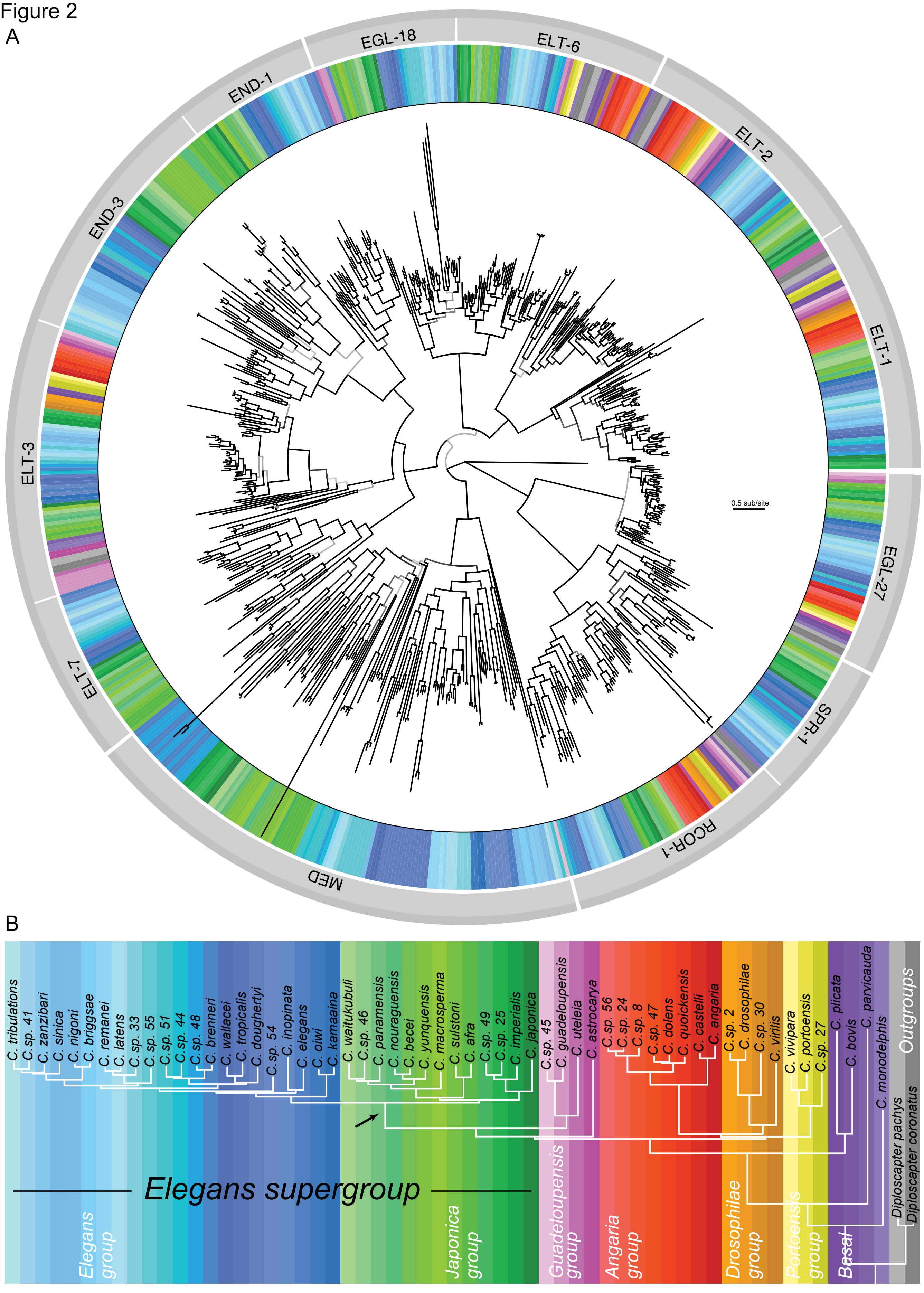

A
EGL-18
ELT-6
ELT-2
ELT-1
EGL-27
SPR-1
RCOR-1
MED
ELT-7
ELT-3
END-3
END-1
0.5 sub/site
B
C. tribulations
C. sp. 41
C. zanzibari
C. sinica
C. nigoni
C. briggsae
C. remanei
C. latens
C. sp. 33
C. sp. 55
C. sp. 51
C. sp. 44
C. sp. 48
C. brenneri
C. wallacei
C. tropicalis
C. doughertyi
C. sp. 54
C. inopinata
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 3

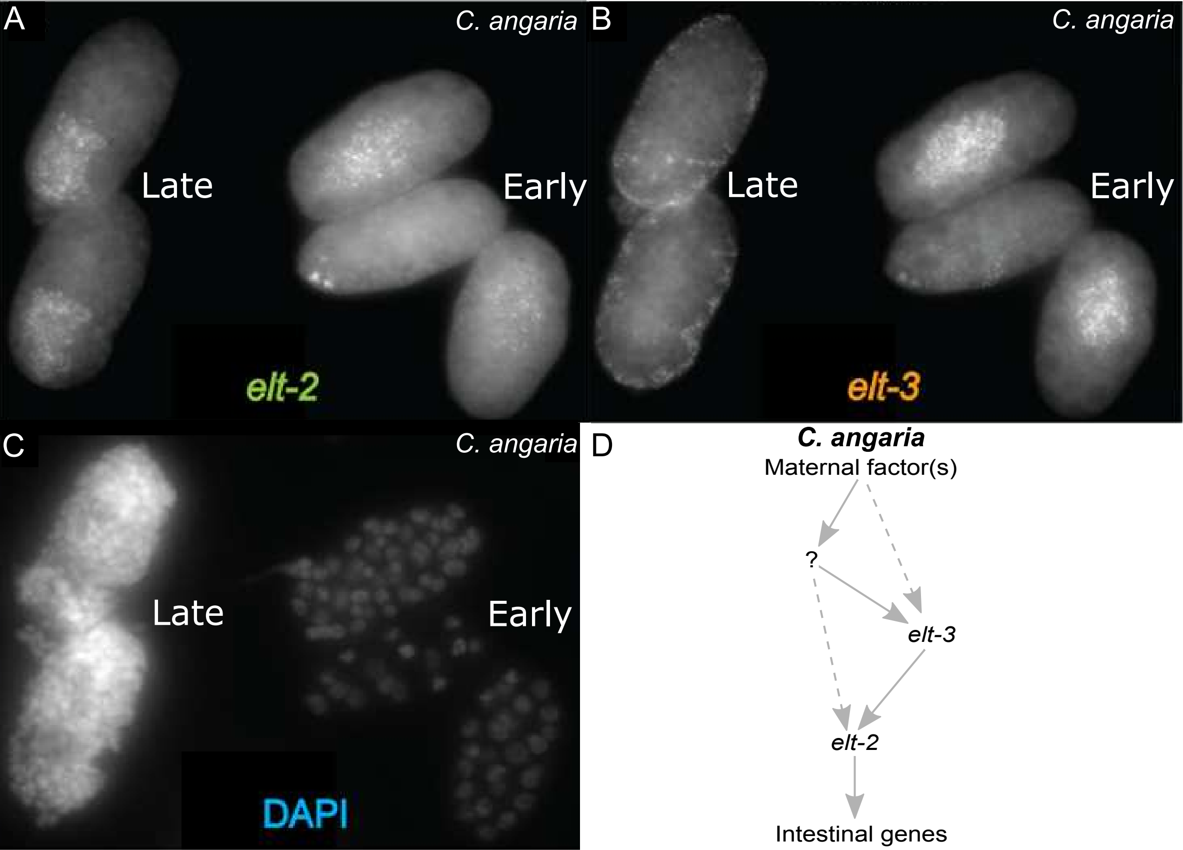

A
C. angaria
Late
Early
elt-2
B
C. angaria
Late
Early
elt-3
C
C. angaria
Late
Early
DAPI
D
C. angaria
Maternal factor(s)
?
elt-3
elt-2
Intestinal genes

Figure 4

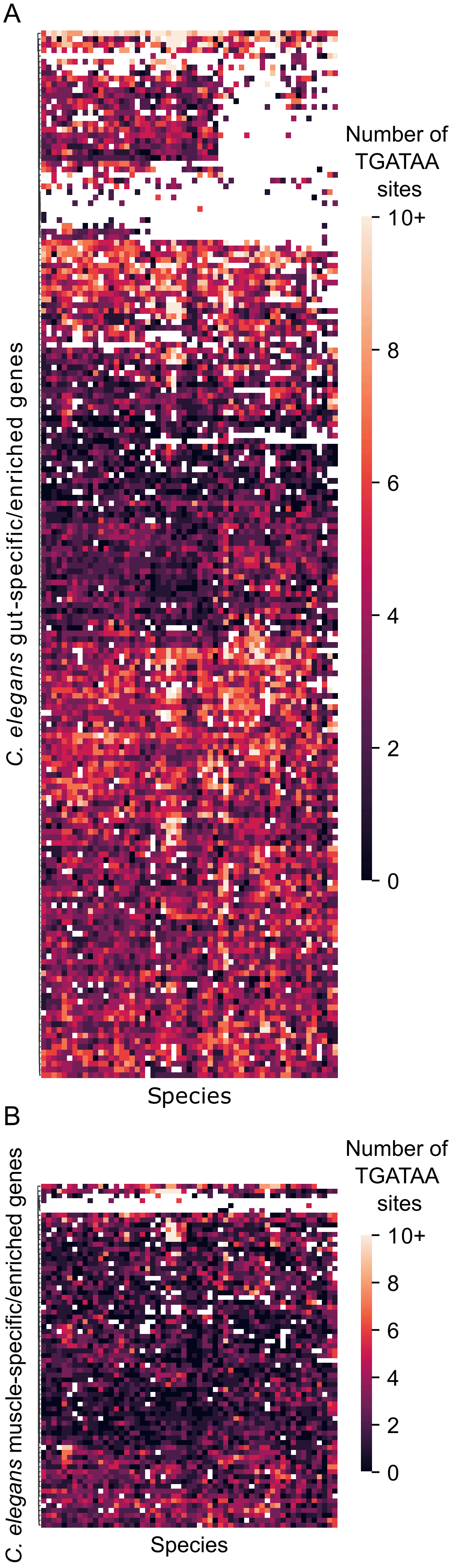

A
C. elegans gut-specific/enriched genes
Species
Number of TGATAA sites
10+
8
6
4
2
0
B
C. elegans muscle-specific/enriched genes
Species
Number of TGATAA sites
10+
8
6
4
2
0

Figure 5

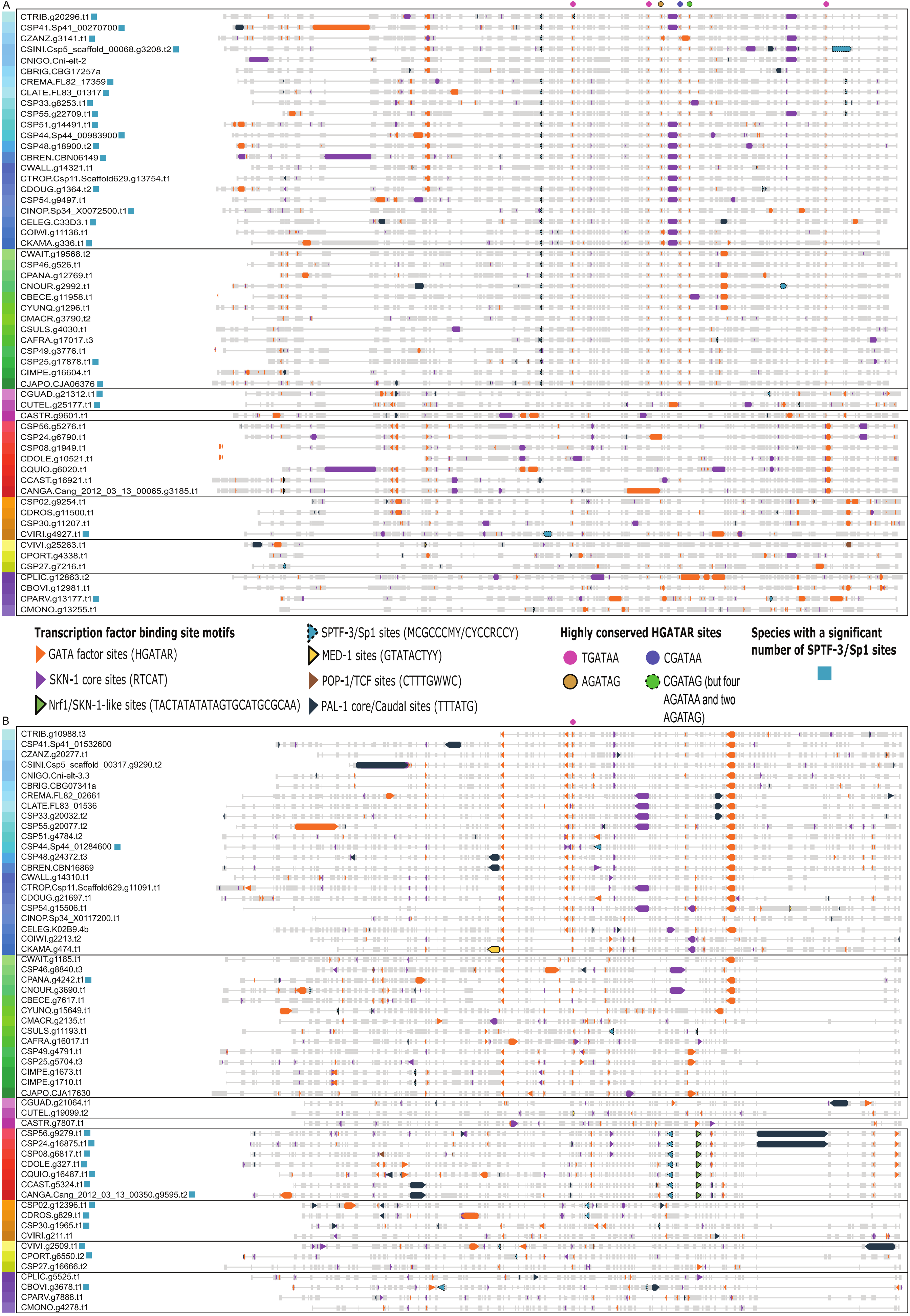

A
CTRIB.g20296.t1
CSP41.Sp41_00270700
CZANZ.g3141.t1
CSINI.Csp5_scaffold_00068.g3208.t2
CNIGO.Cni-elt-2
CBRIG.CBG17257a
CREMA.FL82_17359
CLATE.FL83_01317
CSP33.g8253.t1
CSP55.g22709.t1
CSP51.g14491.t1
CSP44.Sp44_00983900
CSP48.g18900.t2
CBREN.CBN06149
CWALL.g14321.t1
CTROP.Csp11.Scaffold629.g13754.t1
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
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
CGATAG (but four AGATAA and two AGATAG)
Species with a significant number of SPTF-3/Sp1 sites
B
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

Figure 6

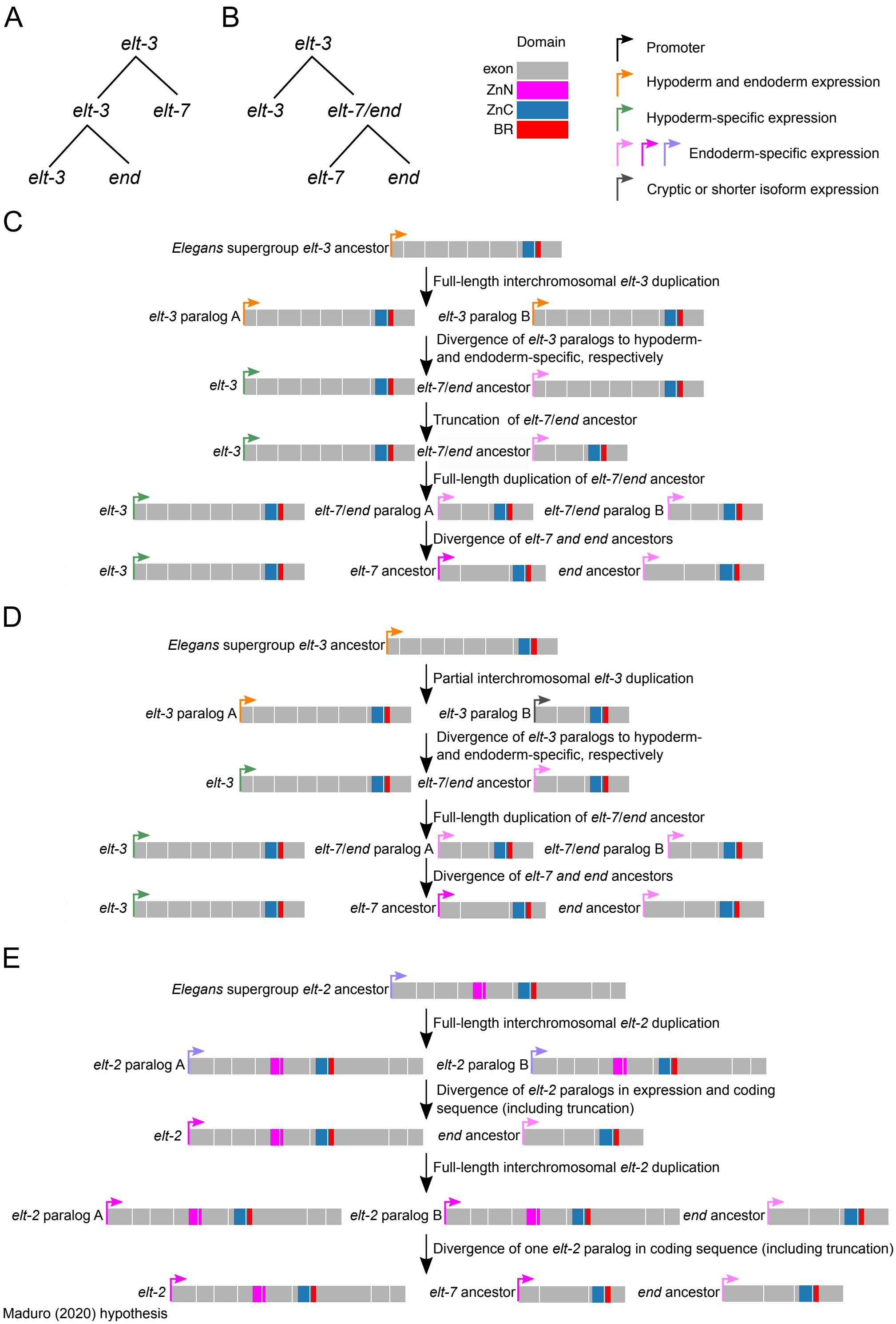
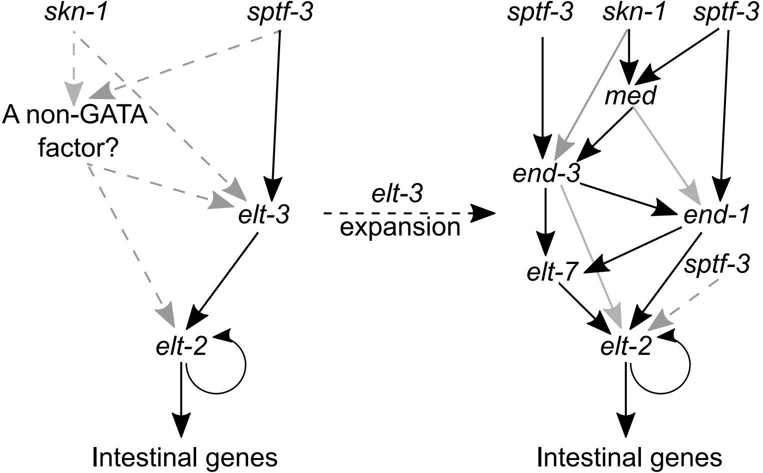
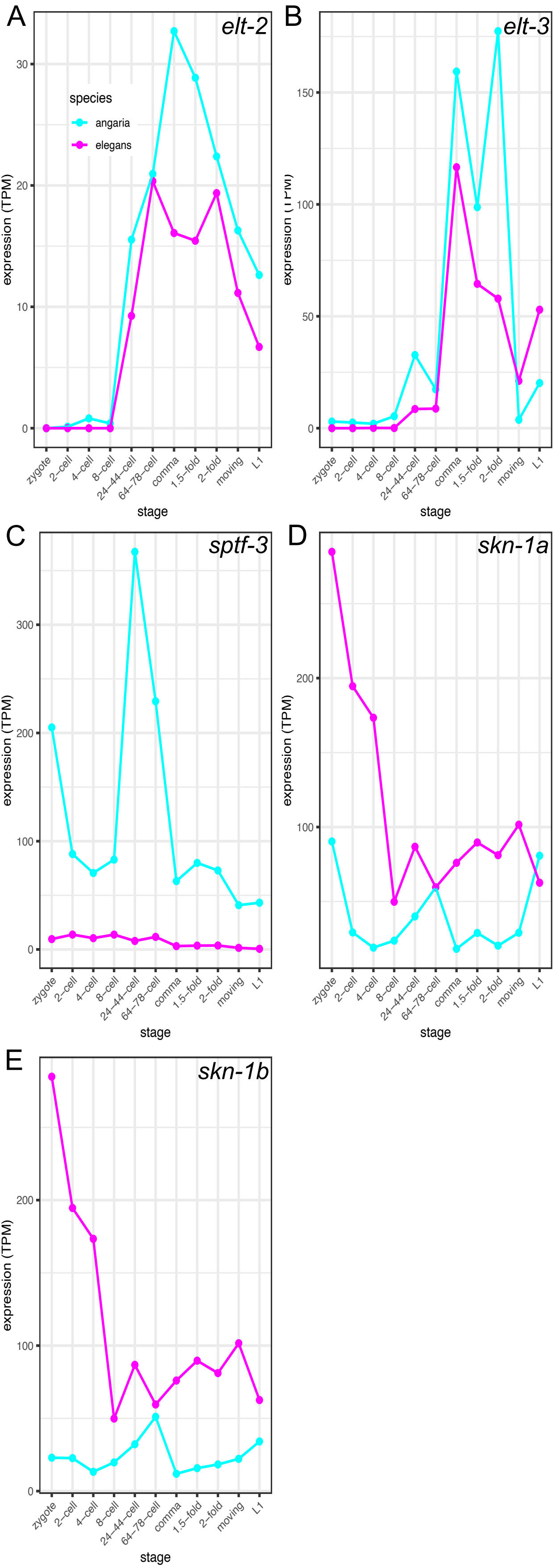

A
elt-3
elt-3
elt-7
elt-3
end
B
elt-3
elt-3
elt-7/end
elt-7
end
Domain
exon
ZnN
ZnC
BR
Promoter
Hypoderm and endoderm expression
Hypoderm-specific expression
Endoderm-specific expression
Cryptic or shorter isoform expression
C
Elegans supergroup elt-3 ancestor
Full-length interchromosomal elt-3 duplication
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

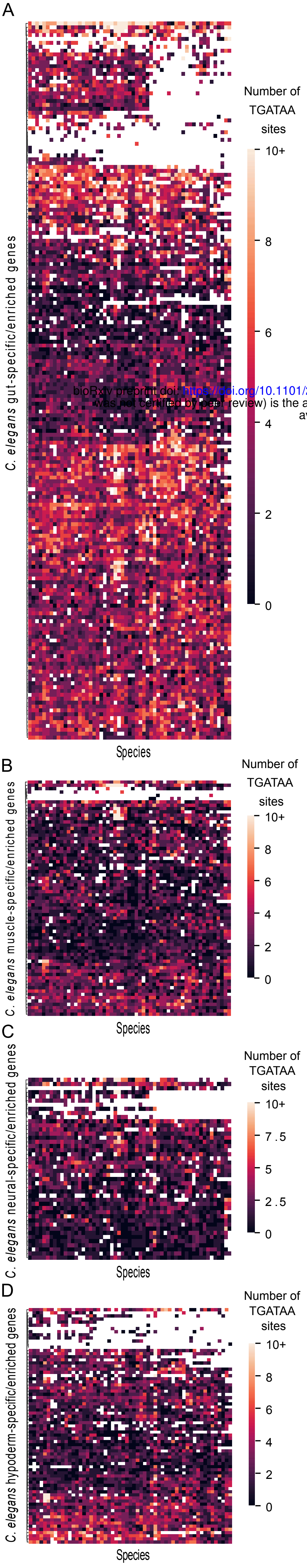
Figure 7


skn-1
sptf-3
A non-GATA factor?
elt-3
elt-2
Intestinal genes
elt-3
expansion
sptf-3
skn-1
sptf-3
med
end-3
end-1
elt-7
sptf-3
elt-2
Intestinal genes

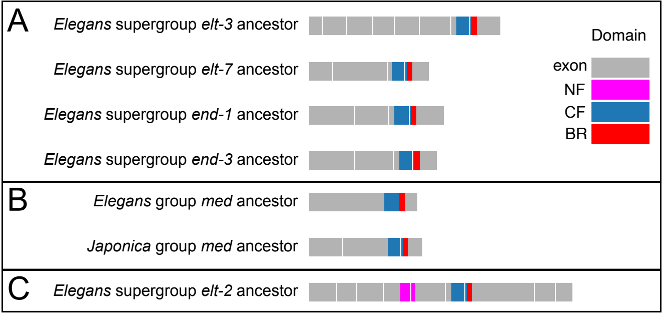
Supplemental Figure 1



Supplemental Figure 2

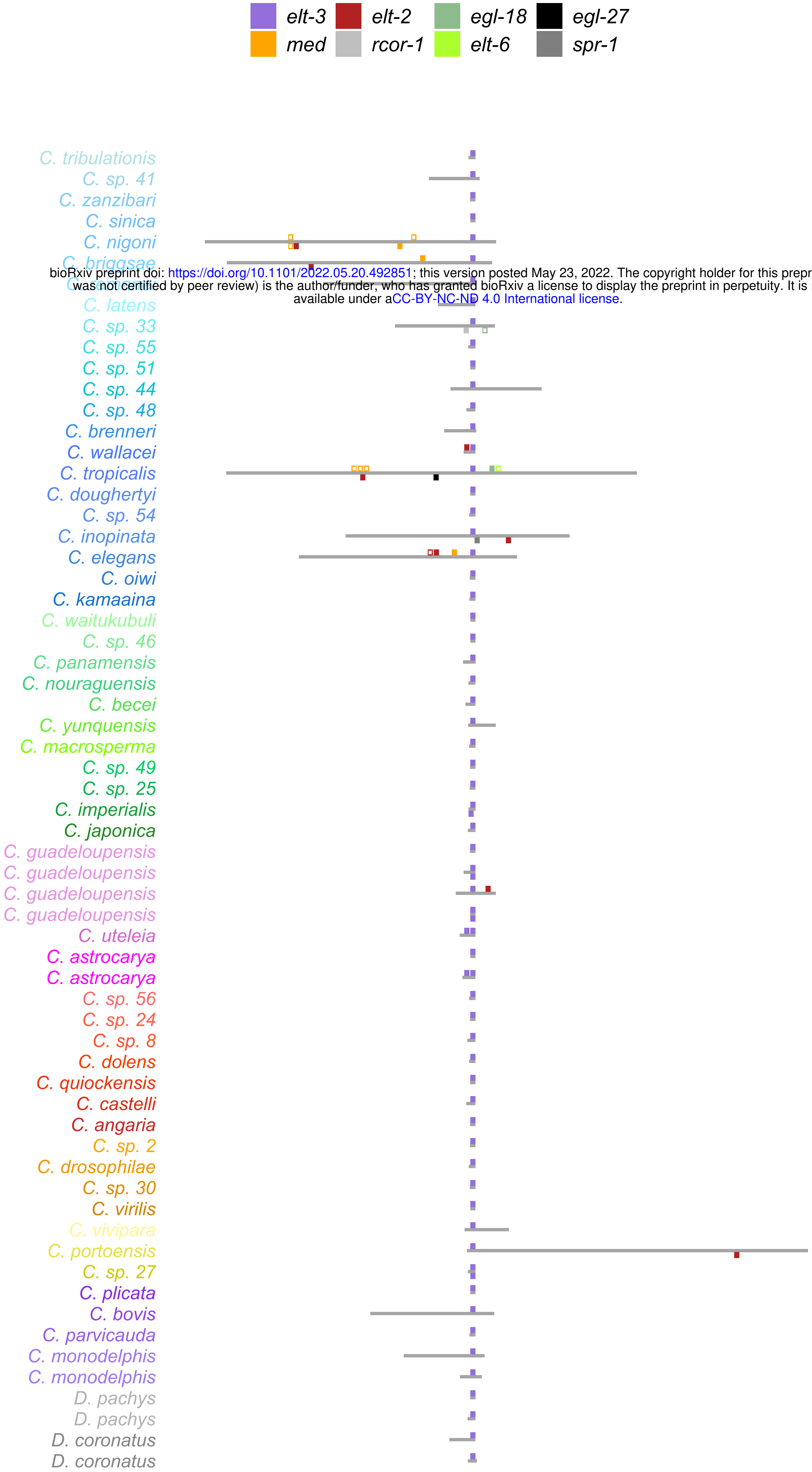


Supplemental Figure 3


A
Elegans supergroup *elt-3* ancestor
Elegans supergroup *elt-7* ancestor
Elegans supergroup *end-1* ancestor
Elegans supergroup *end-3* ancestor
Domain
exon
NF
CF
BR
B
Elegans group *med* ancestor
Japonica group *med* ancestor
C
Elegans supergroup *elt-2* ancestor

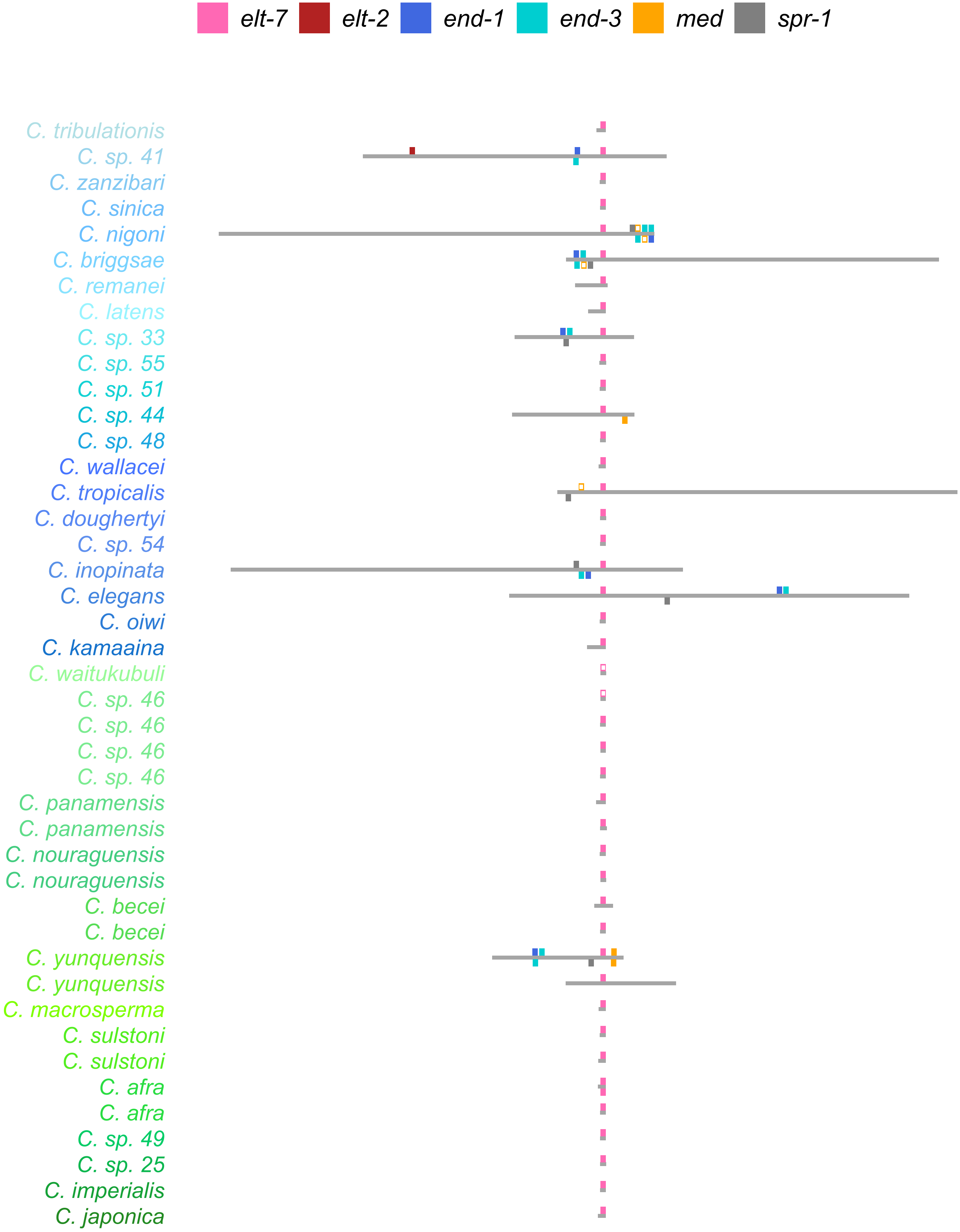
Supplemental Figure 4

A



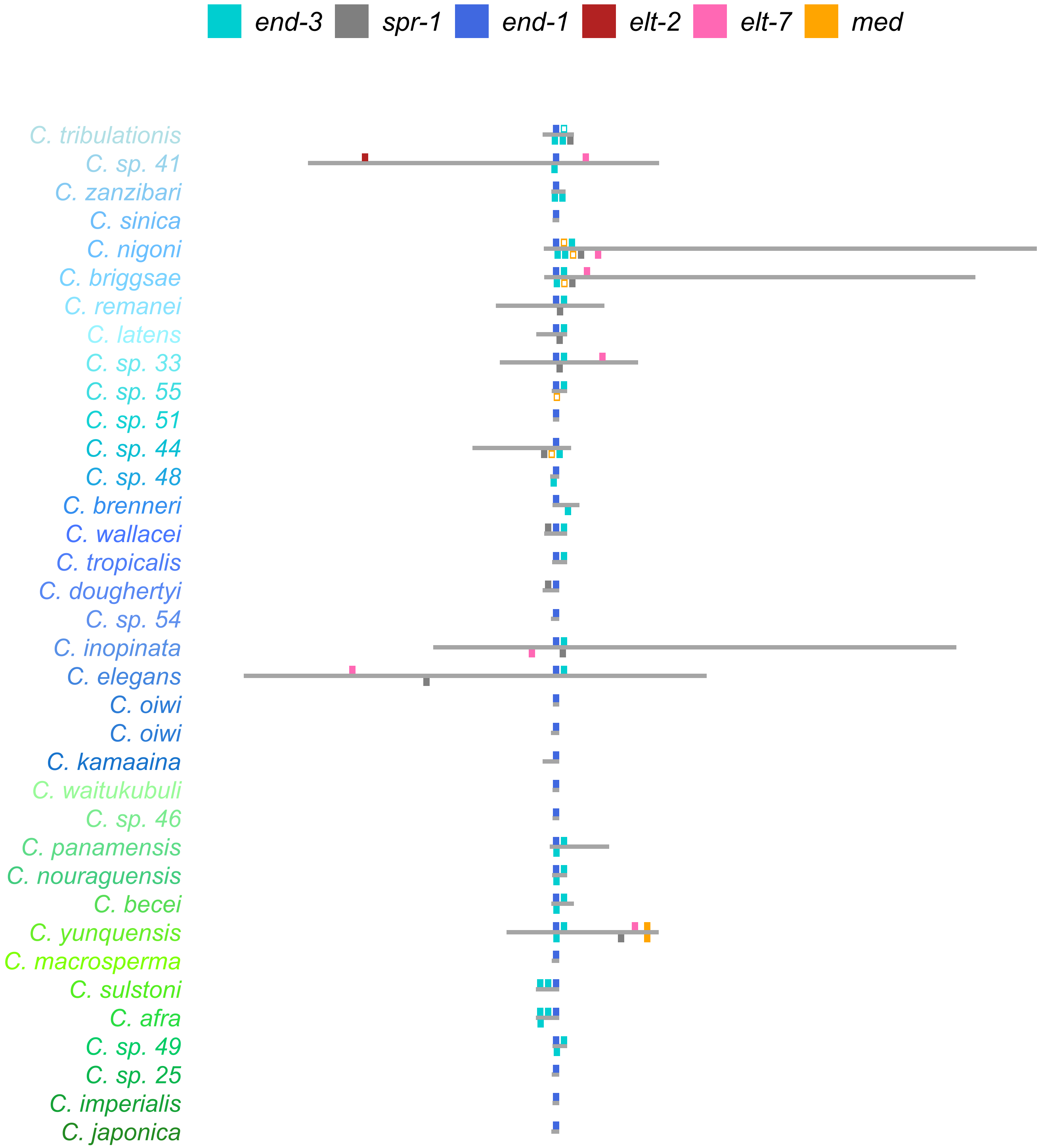
Supplemental Figure 4

B



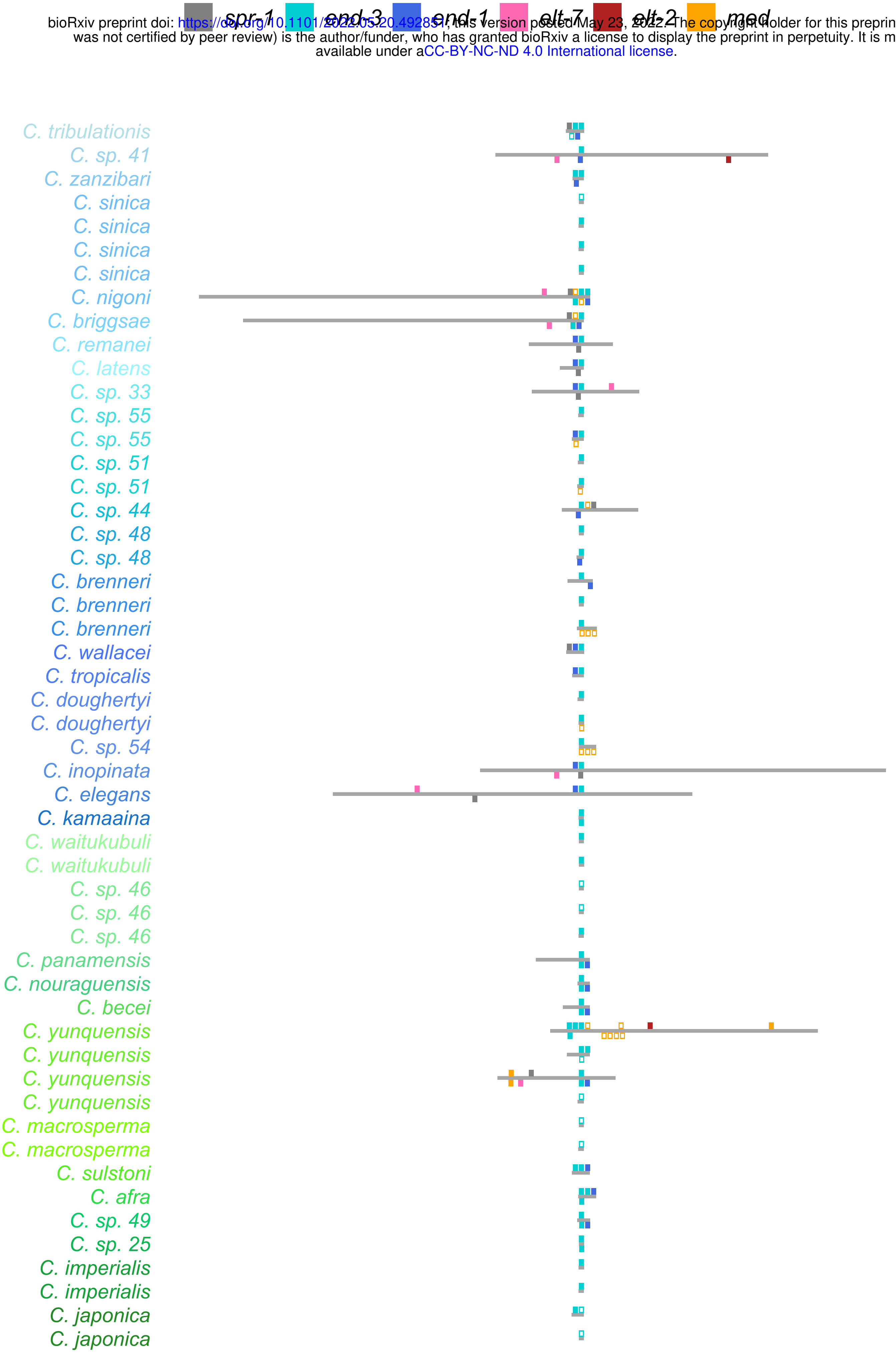
Supplemental Figure 4

C

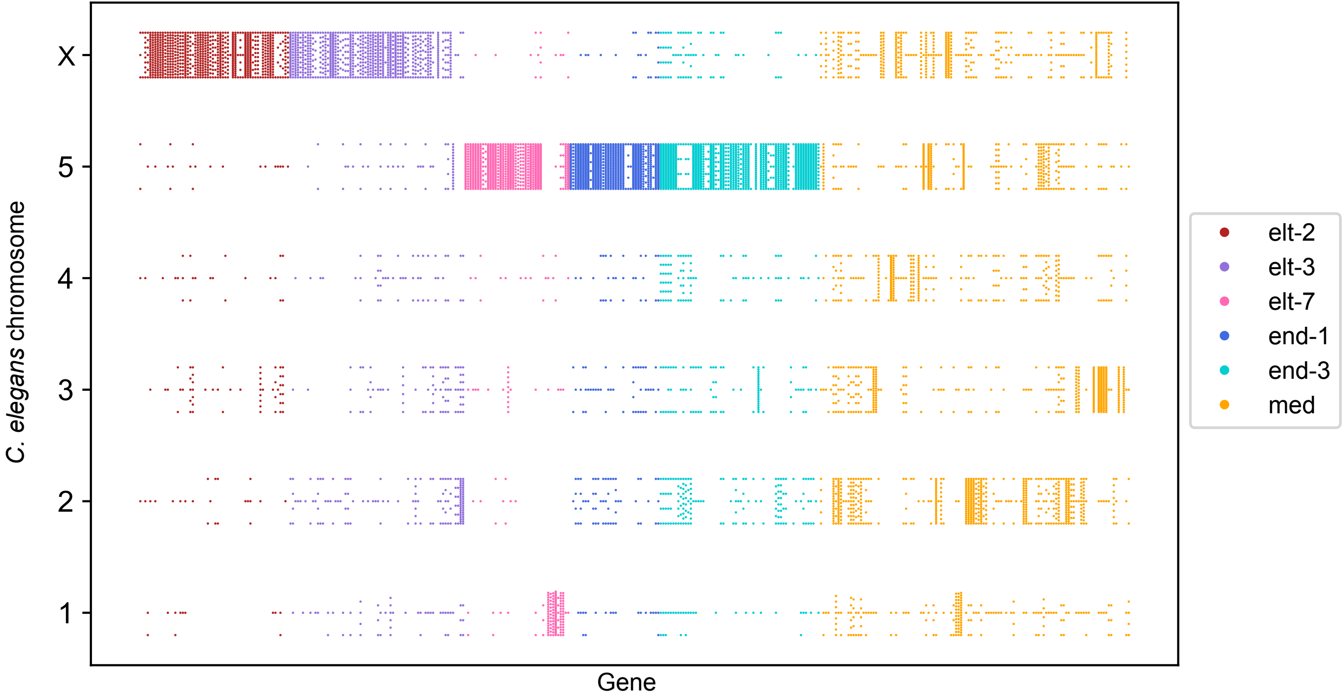


Supplemental Figure 4

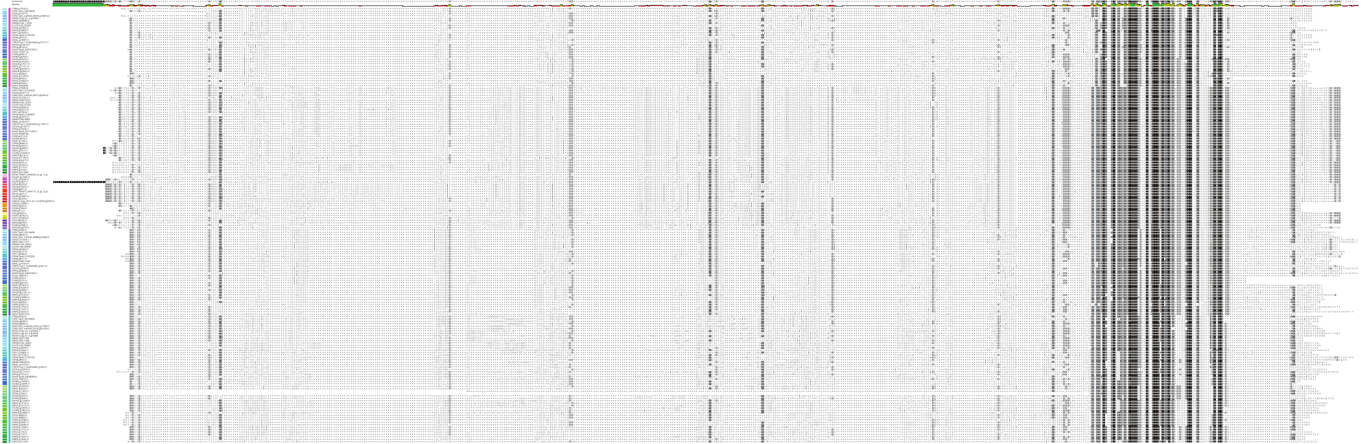
D



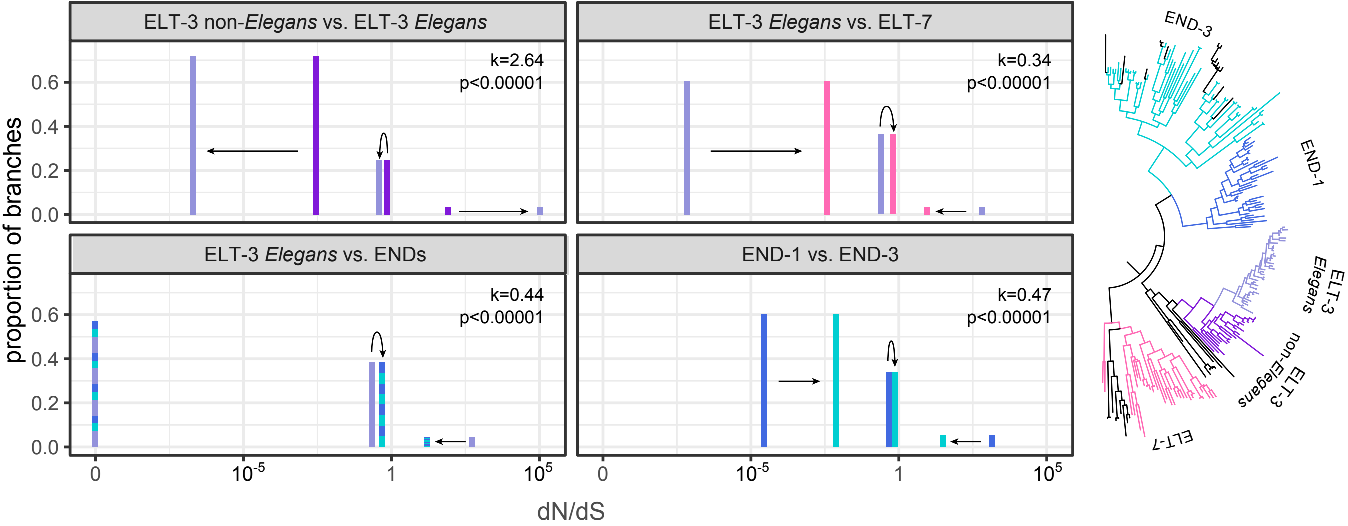
Supplemental Figure 5



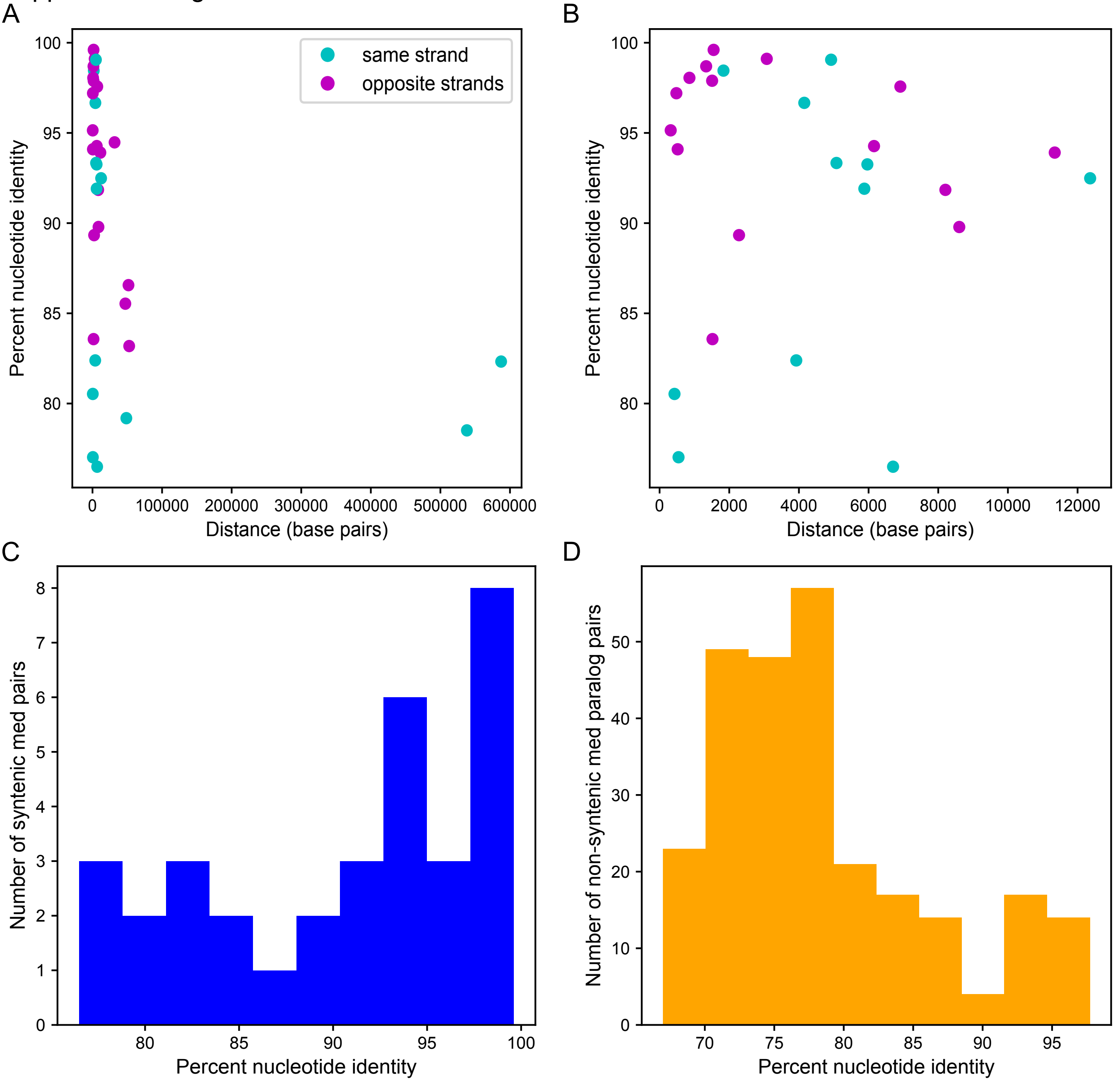
Supplemental Figure 6



Supplemental Figure 7



Supplemental Figure 8



Supplemental Figure 9

med
egl-18
elt-6
elt-1
end-1
end-3
spr-1
elt-7
elt-2
elt-3
egl-27
C. tribulationis
C. sp. 41
C. sp. 41
C. zanzibari
C. zanzibari
C. zanzibari
C. zanzibari
C. zanzibari
C. zanzibari
C. zanzibari
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. sinica
C. nigoni
C. nigoni
C. nigoni
C. nigoni
C. nigoni
C. briggsae
C. briggsae
C. briggsae
C. briggsae
C. remanei
C. remanei
C. remanei
C. remanei
C. latens
C. latens
C. latens
C. latens
C. latens
C. sp. 33
C. sp. 33
C. sp. 33
C. sp. 33
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 55
C. sp. 51
C. sp. 51
C. sp. 51
C. sp. 51
C. sp. 51
C. sp. 44
C. sp. 44
C. sp. 44
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. sp. 48
C. brenneri
C. brenneri
C. brenneri
C. brenneri
C. brenneri
C. brenneri
C. brenneri
C. wallacei
C. wallacei
C. wallacei
C. tropicalis
C. tropicalis
C. tropicalis
C. tropicalis
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. doughertyi
C. sp. 54
C. sp. 54
C. sp. 54
C. sp. 54
C. sp. 54
C. sp. 54
C. inopinata
C. elegans
C. elegans
C. oiwi
C. kamaaina
C. waitukubuli
C. waitukubuli
C. waitukubuli
C. sp. 46
C. panamensis
C. nouraguensis
C. nouraguensis
C. nouraguensis
C. nouraguensis
C. nouraguensis
C. becei
C. becei
C. becei
C. becei
C. becei
C. becei
C. becei
C. becei
C. yunquensis
C. yunquensis
C. yunquensis
C. yunquensis
C. yunquensis
C. yunquensis
C. yunquensis
C. macrosperma
C. macrosperma
C. macrosperma
C. macrosperma
C. macrosperma
C. sulstoni
C. afra
C. afra
C. afra
C. afra
C. afra
C. sp. 49
C. sp. 49
C. sp. 25
C. sp. 25
C. sp. 25
C. imperialis
C. japonica
C. uteleia
C. uteleia
C. uteleia
C. uteleia
C. uteleia
C. portoensis