

The Conservation of Cis Regulatory Elements Across Development in *Caenorhabditis elegans*

Rachel Woo
BCB330
Dr. Asher Cutter

Abstract

Gene expression is largely mediated by cis-regulatory elements including transcription factor binding sites (Stern and Orgogozo, 2008). Cutter et al. has recently shown that cross-development there are distinct patterns of gene expressions in *C. elegans* (Cutter et al., unpublished). However, the degree of conservation in transcription factor binding sites within profiles is unknown. Here I use bioinformatic approaches to correlate transcription factor binding site (TFBS) occurrence and motif occurrence to gene co-expression modules. Through motif detection and enrichment analysis from MEME suite we found that motif enrichment is loosely correlated with the amount of gene expression change over time. There was an additional relationship between the number of binding sites in *C. elegans* modules and the orthologous modules in *C. briggsae* suggesting some conservation of cis-regulatory elements.

Introduction:

Ontogenetic development in all species requires complex regulation of gene expression over time. Cis regulatory elements (CREs) are a common mode of gene regulation and include enhancers and promoters, which both regulate the transcription of neighbouring coding regions. CREs often correspond to specific transcription factors (TFs) which can bind the CRE's transcription factor binding site (TFBS) to facilitate or prevent

transcription (Riddle et. al., 1997). It has been hypothesized that CRE mutations play a role in evolution of phenotypic change by being more likely to affect phenotype and more prone to selective pressures (Stern and Orgogozo, 2008; Wray, 2007). Previous work has demonstrated the role of CREs in phenotypic divergence, however their distribution and enrichment for genes in co-expression groups across development is unclear (Wittkopp, P. J., and G. Kalay, 2012).

Because of its short developmental times and well-studied genome, *C. elegans* is an ideal system to study the evolution of developmental gene regulation. The development of *C.elegans* begins with embryogenesis, where the body plan is established, the larval stages, L1, L2, L3 and L4, where growth and reproductive functions are developed, and finally adulthood when reproduction takes place (Levin and Yani, 2012).

The differing conditions during ontogenesis may influence the number and type of cis regulatory elements associated with genes expressed at different life stages. Across the *C. elegans* ontogeny there appears to be a mid-embryogenesis phylotypic stages or point of development in which gene expression is tightly conserved (Levin and Yani, 2012). During this period, researchers found that gene duplication is constrained (Castillo-Davis and D. L. Hartl, 2002). These findings offer support for the hourglass model of evolution in which conservation is greatest during mid-embryogenesis. Perhaps this effect is due to body plan establishment being under high selective pressure or the increase in genes involved in a high number of interactions (Artieri et. al., 2009; Cruickshank and Wade, 2008; Domazet-Loso and Tautz, 2010; Kalinka and

Tomancak, 2012; Kalinka et. al., 2010). Additionally, genes expressed after post-reproductive maturity, genes involved in spermatogenetic functions and genes expressed briefly during embryogenesis evolve more quickly (Cutter and Ward, 2005). The low rate of evolution during mid embryogenesis has been additionally documented using proteomic data in *D. melanogaster*, indicating that these results may be widely generalizable evolutionary facts (Davis et. al., 2005). While there is elevated conservation and varying evolutionary constraints across *C. elegans* ontogeny, how periods of differing expression are transcriptionally regulated is unclear.

Questions regarding the regulation and evolution across embryogenesis have prompted research into quantifying the links between gene co-expression networks, or genes with a similar expression pattern over development (Cutter et. al., unpublished). If we identified gene co-expression networks, this data could be used to predict functions, identify cis regulatory elements and their associated transcription factors (Santos-Mendoza, 2008). Genes within the same co-expression profiles often have related function, suggesting that these profiles may contain repeating TFBS motifs (Santos-Mendoza, 2008; Vandepoele et. al., 2009). Correlating TFBS motifs with co-expression modules may offer more confidence when predicting gene functions from bioinformatic approaches. Additionally, if particularly enriched or conserved motifs largely correspond to genes expressed at certain times, this could offer insight into the means and evolution of gene regulation across development.

Given the diverse modes of gene regulation it is unclear how much of the transcriptome can be explained by TFBSs in CREs. Other forms of gene regulation such as the conformation of the chromosomes *in situ* or trans-regulatory elements may also contribute to the varying gene expression seen across development. The strength of selection to conserve CREs can be seen by examining homologous genes across closely related species. Castillo-Davis et. al. (2004) experimentally verified that protein and regulatory evolution is weakly linked in orthologs - genes that evolved from a common ancestor through speciation - but unrelated in paralogs - genes related by duplication. These findings suggest that protein regulatory evolution is divergent between species and that different types of genes experience measurable differences in CREs (Castillo-Davis et. al, 2004). In *D. melanogaster*, 50%-70% of established binding site were conserved in CRE and these sites were not more enriched than other CREs in the genome (Emberly et. al., 2003). However, the degree of conservation for a TFBS associated with co-expressed genes is unknown.

In this analysis, co-expression profiles are used to test if there are more common motifs, or reoccurring patterns of DNA for given profiles compared to the average occurrence for all regions in the analysis. Then, the motifs are compared to known TFBS to determine if the motif found is due to a known TF or if it is satellite DNA.

Materials & Methods:

This study was conducted exclusively through the usage of public *C. elegans* and *C. briggsae* bioinformatic databases and unpublished work by Cutter et al.

1. Gene Co-expression Profiles:

I accessed gene co-expression profiles across developmental time from unpublished data by Cutter et. al. This analysis is based upon RNAseq transcriptome sequences from the modENCODE data repository (<http://data.modencode.org>) that looked at 30 different developmental time points for *C. elegans* with one replicate per developmental time. The RNAseq transcriptome sequences were exported as sam-format files and mapped to the reference genes WS248. Cutter et. al. used featureCounts, to quantify expression for each gene for timer periods ranging from early embryos, L1, L2, L3, L4, and young adult in hermaphrodites. The result and the data used from Cutter et al. are the 14 different co-expression profiles clustering 19,711 genes into patterns of expression across 30 developmental time points (Cutter et. al., unpublished).

2. Verifying TFBS presence:

I focused on 5' upstream region of every gene in every co-expression module found by Cutter et al. to verify the presence of transcription factor binding sites and potential enhancer or promoter regions. Although this method captures most of the CREs present, there is potential to miss CREs not located near the gene of interest. This happens because CREs can be located both in introns that they regulate and far away from the genes they regulate (Wray, 2007).

The most up to date version of the *C. elegans* genome was used (WS269) in order to segment the upstream and downstream regions of genes (www.wormbase.org).

Additionally, other tools such as WormBase Parasite, and WormBase WormMine were used to facilitate the analysis (Howe et al., 2017 and Howe et al., 2016). The region analyzed is the 200 nucleotides upstream from the gene of interest depending upon its distance from neighboring genes. This choice is due to known accuracy of binding site algorithms. According to Hu et. al, for sequences 200 nt long there is 25-35% accuracy for binding site prediction and the algorithm used (MEME) will correctly predict one site more than 90% of the time (Hu et. al., 2005). Another reason for this choice is to optimize the algorithm that will be used for motif detection, MEME suite's Multiple Em for Motif Elicitation (MEME) (<http://meme-suite.org/>). This algorithm works best when the motifs are a larger percentage of the input and when motifs have very few differences between them. MEME FASTA files containing 200 upstream of the genes of interest for each of the co-expression modules were passed to the algorithm. From this algorithm the most commonly seen motifs will be returned (Mathelier and Wasserman, 2013; Valouev et. al., 2008).

MEME works by agnostically searching for sets of similar short ungapped motifs among sequences (Bailey et. al., 2006). In this case, MEME searched for motifs in the 14 co-expression modules in the 200 nt regions adjacent to the gene of interest (Bailey et. al., 2006).

The differential enrichment mode was utilized in order to determine if the motifs found were significantly higher than the rest of the co-expression modules (Bailey et. al., 2006). The controls used were all the co-expression module genes combined, except

for the module being analyzed. The differential enrichment mode automatically runs another MEME Suite tool called Motif Alignment and Search Tool (MAST). MAST detects if motifs found are not significantly different from each other, that is, the same motif (Bailey et. al., 2006). Repeat motifs were excluded from the analysis.

The first order model of sequences was used, in order to decrease the number of single nucleotide repeats that were found as significant motifs (Mathelier and Wasserman, 2013). Additionally, motifs that were significantly similar to dinucleotide repeats were also excluded. MEME search for motifs 6-20 nucleotides in length as most TFBS are short and degenerate (Valouev, 2008).

Next, all significant motifs were compared to known *C. elegans* transcription factor binding sites (Narasimhan et al., 2015 and Ho et al., 2017). This will determine if MEME found sequences that were TFBS rather than DNA satellites. For this, another MEME suite tool was used, TomTom which compares motifs to a provided database using known *elegans* TFBS (Bailey et. al., 2006). This algorithm will help determine if discovered motifs are similar to known TFBS.

4. Motif enrichment

The next analysis focuses on TFBS motif enrichment which searches for the increased presence of known binding motifs (McLeay and Bailey, 2010). For this analysis, the database of comparison is known and predicted TFBS from previous studies (Narasimhan et al., 2015 and Ho et al., 2017).

For this analysis the 0-order model was used since the algorithm compared to previously validated motifs. The motif enrichment is determined by a Fischer's exact test and only E-values above 10 were considered.

This method offers different strengths and weaknesses compared to MEME. The main difference being that AME only searches for known TFBS and MEME will look for any motif regardless of biological significance (Bailey et. al., 2015). Using both algorithms will help increase the signal to noise ratio and yield a more thorough analysis.

AME results that had a large number discovered motifs were clustered by TFBS occurrence similarity in a heatmap and dendrograms based upon a distance matrix. The criterium for sufficient enrichment is at least eight discovered motifs in every module.

5. Ortholog Analysis

The same MEME and AME analyses were then repeated on the corresponding 200 nucleotide upstream orthologues of *Caenorhaditis briggsae*. The homologous regions of *C. elegans* and *C. briggsae* contain interspersed regions of high similarity and non-alignable sequences. However, these high similarity regions had a high number of transversions suggesting unique regulatory functions (Webb et al., 2002).

In order to obtain the orthologs the OMA database was used (Altenhoff et al., 2015). The OMA database infers homology by analyzing the significance of Smith-Waterman

alignments, and then infers orthology by analyzing the evolutionary distance between homologues (Altenhoff et al., 2018). OMA's computational results predict orthology between *C. elegans* and *C. briggsae* and predict the type of orthologs.

For this analysis, only 1:1 orthologs that were deemed significant by OMA were used. This is in order to limit the effects of paralogy which may produce differential regulation via effects such as sub-functionalization (Woollard, 2018). The 200 bp upstream regions of the orthologs of *C. briggsae* were then found using the WS269 WormBase version of the genome and WormBase parasite (Howe et al., 2017 and Howe et al., 2016). Then, the MEME and AME analysis was performed in the same manner described for *C. elegans*.

To control for the smaller number of the orthologous genes, I performed another analysis. For each of the orthologues that was found in *C. briggsae*, the 1:1 orthologue for *C. elegans* was found. The same MEME and AME analysis were then performed on this smaller subset of *C. elegans* genes.

Results:

1. *C. elegans* MEME Results

For each co-expression module 8-15 motifs were discovered. These motifs can be found in Table 1. For each module there were at most 2 motifs that were both significantly different from each other and statistically significant, with an E-value cut off of 10 or less. Very few motifs were found to be significantly similar to known motifs, with

only both module six containing any that were significantly similar to the constructed database. As can be seen in Table 1, motifs in module 4 were the most significant relative to the controls of the other modules with E-values of $8.6\text{e-}33$ and $3.6\text{e-}14$ respectively. Other modules with highly significant motifs include module 10 which has E-values of $1.1\text{e-}16$ and $4.0\text{e-}13$, module 11 which has an E-value of $4.2\text{e-}14$.

2. *C. elegans* AME Results

The *C. elegans* enriched motifs dataset was large enough that it could be clustered for a dendrogram and heatmap.

The TFBS dendrogram had a very weak positive correlation with the co-expression generated dendrogram of 0.01 and an entanglement value of 0.23 (See Figure 1). Next, the modules were grouped according to family to see if there was a stronger correlation between related subfamilies (Figure 2,3 and Table 3). The subfamilies were made by grouping the families whose modules had the smallest changes in expression (3, 6, 9, 12, 13, 14), modules with a parabola-like increase and modules with the largest increases in expression. All subfamily groups had stronger correlations, but they were still quite weak and averaged 0.4. The modules with parabola-like increases and the modules with the smallest changes in expression had no entanglement, meaning they had the same clade formations. The modules with largest increases in expression had the lowest sub-family correlation of 0.37 and the highest entanglement of 0.4. Module 4 was excluded from the subfamily categorization as it is its own subfamily and the only

module that has expression trend downwards. Additionally, module 4 has very different TFBS enrichment than the other modules.

The TFBS occurrence similarity was also modelled using a heatmap in Figure 4.

Modules 11 and 4 had the lowest enrichment for commonly seen motifs. In contrast, modules 12, 3, and 13 had the most enrichment for commonly seen motifs. These modules are all from the same sub-family of constant expression over time.

3. *C. briggsae* MEME Results

The *C. briggsae* MEME results are not as robust as the *C. elegans*, as the genes per module are much fewer. Only 4 significant motifs were found (Table 2 and Supp. Table 2) with 2 in module 1 and 2 in module 3. While all had a significant E-value of less than $1e-2$, none had a highly significant E-value of $1e-10$.

4. *C. briggsae* AME Results:

The effect of the smaller number of genes per module is very evident in the *C. briggsae* AME results. Many modules had very few known motifs found. Module size was a good predictor of number of enriched motifs as can be seen in Figure 5. The 1:1 orthologs AME results for *elegans* had a very similar pattern. The lines of best fit were very similar, between the *C. elegans* and *C. briggsae* plots, $y = 0.019x - 3.36$ and $y = 0.018x - 2.13$ respectively. The motifs that were largely under-enriched, over-enriched and average appear to have held across *C. elegans* and *C. briggsae*. In both species, modules 1, 8, and 14 were under-enriched, modules 3, 4, 10 and 11 were

over-enriched and modules 2, 9 and 12 were averagely enriched. The remaining module 5, 6, 7 and 13 fell into different enrichment categories between *C. elegans* and *C. briggsae*.

Discussion:

1. *C. elegans* MEME

Only module 6 had motifs that were algorithmically determined to be like known TFBS. This suggests that the signal is not high enough to detect motifs non-unique or rare motifs or that the cis-regulation between modules is similar. The modules with highly significant (E-value ≤ 10) were 4, 10, and 11. For module 4 this makes sense as it had the most divergent TFBS enrichment and co-expression (Cutter et al., unpublished). Perhaps in modules 10 and 11 a genuine signal was found, and they contain a novel TFBS. However, further research is needed in order to confirm if these motifs are genuine TFBS or satellite DNA (Henikoff et al., 2001). Additionally, given that modules likely have similar cis-regulation between sub-families the control groups used in analysis should be altered in future to provide a larger signal to noise ratio.

2. *C. elegans* AME

The number of shared motifs between modules suggests that many TFBS are common cis-regulatory elements in *C. elegans* (Shen et al., 2012). The weak positive correlation amongst all modules to the co-expression modules seems to suggest that there is no relation between co-expression and cis-regulatory elements. However, upon closer investigation certain sub-classes of module families have a more significant correlation.

This fact suggests that while not all patterns of cis-regulation relate directly to co-expression they do relate very broadly to amount of expression change. The correlation between the co-expression dendrogram and the TFBS dendrogram may also be artificially smaller due to differences in the distance measurement as the distance for co-expression profiles is based upon three beta best fit values and the TFBS distance is based upon co-enrichment of motifs. Module 4 did not fit into a module sub-category, which suggests that it is very differently regulated than the other modules. This fits with a biological interpretation of Cutter et al.'s previous work and implies that cis-regulation is significantly different in module 4.

In general, motif similarity correlates with the amount of expression change over time. Given that these motifs correspond to hypothesized TFBS it can be said that there is a weak relationship between cis-regulatory elements and co-expression across development.

In the heatmap (Figure 4) the modules which contained over-enriched motifs 3, 12 and 13 are all from the same constant expression subfamily. However, the under-enriched modules 11 and 4 have no obvious connection.

3. *C. briggsae* MEME

The *C. briggsae* MEME results do not appear to have a clear biological interpretation given the weak results. This suggests that the sample size of the number of genes per module is too low and that the large disparity in module size ranging from 50-1000

masked some results. Perhaps a different control group or different algorithmic analysis may reveal the true effects. Additionally, additional transcriptome analysis across development upon *C. briggsae* and other more closely related species would shed light upon the divergence of cis-regulatory elements.

4. *C. briggsae* AME

The results of the *C. briggsae* and smaller *C. elegans* AME have two likely interpretations. The first interpretation is the biological one. Given the two species have very similar enrichment and lines of best fit, the motifs that are enriched in *C. elegans* and *C. briggsae* have held across evolutionary time. The second interpretation is the null results, that the effect of module size is so strong that it masks significant findings. While the full scope of the effect of module size is unknown, the similar partitioning of modules into over-enriched, under-enriched and averagely-enriched that loosely held across species suggests that there is some shared similarity in upstream sequence. The effect of module size is clearly large, however there appears to be a weak macro-pattern suggesting that certain modules *C. elegans* and *C. briggsae* share similar numbers of TFBS. This result implies that there is some amount of conservation of cis-regulatory elements between *C. elegans* and *C. briggsae*.

While this research lacks strong conclusions, most results weakly confirm the findings of Cutter et al. and suggest that similar co-expression modules share similar TFBS. There is no clear TFBS that is shared during a common developmental time. However, there are many shared binding sites across modules in line with existing literature (Castillo-

Davis et. al, 2004). Although the amount of CRE binding sites conserved in *C. briggsae* could not be accurately quantified it appears that there are many similarities across species which fits with the current understanding of orthologous CREs and may imply conservation (Emberly et. al., 2003).

Future research should focus on establishing more transcriptome data for different data point in development in species closely related to *C. elegans* and expanding the transcriptomic data available. Specifically, chip-seq de novo identification using immunoprecipitation of the TBS regions could be valuable.

Figures:








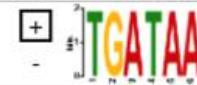






MODULE	LOGO 1	LOGO 2
1		
4		
5		
6		
7		
10		
11		
12		
14		

Table 1. *C. elegans* MEME significant discovered motifs. Motifs in module six were both found to be significantly similar to known motifs (E-value ≤ 10).


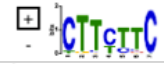


MODULE	LOGO 1	LOGO 2
1		
3		

Table 2. *C. briggsae* MEME significant discovered motifs (E-value ≤ 10).

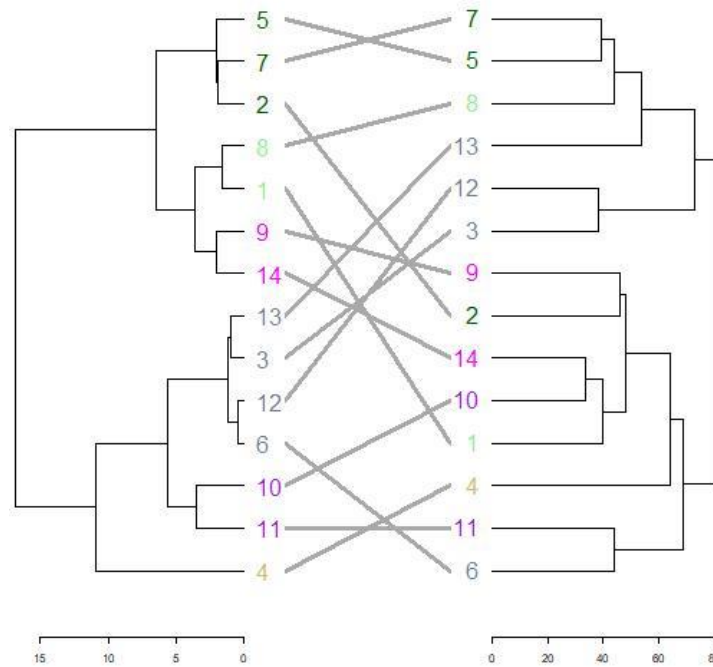


Figure 1. *C. elegans* TFBS derived dendrogram and co-expression derived dendrograms and their comparison for all modules. Co-expression dendrogram is on the left and TFBS dendrogram is on the right.

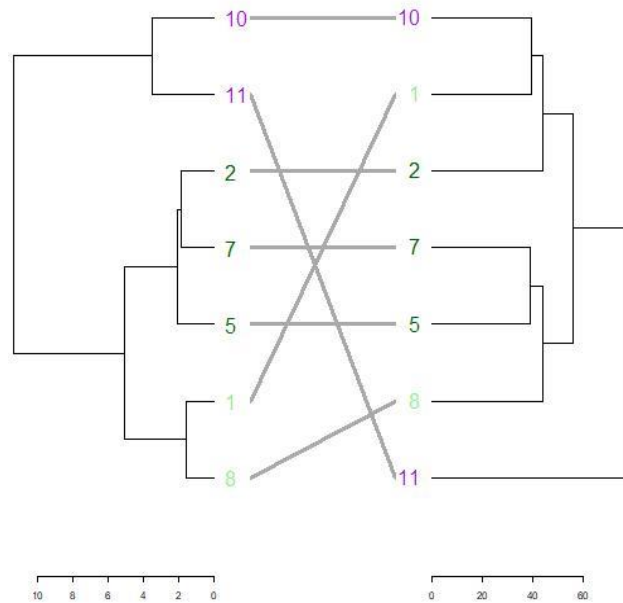


Figure 2. *C. elegans* TFBS derived dendrogram and co-expression derived dendrograms and their comparison for modules 1, 2, 5, 7, 8, 10, and 11. Co-expression dendrogram is on the left and TFBS dendrogram is on the right.

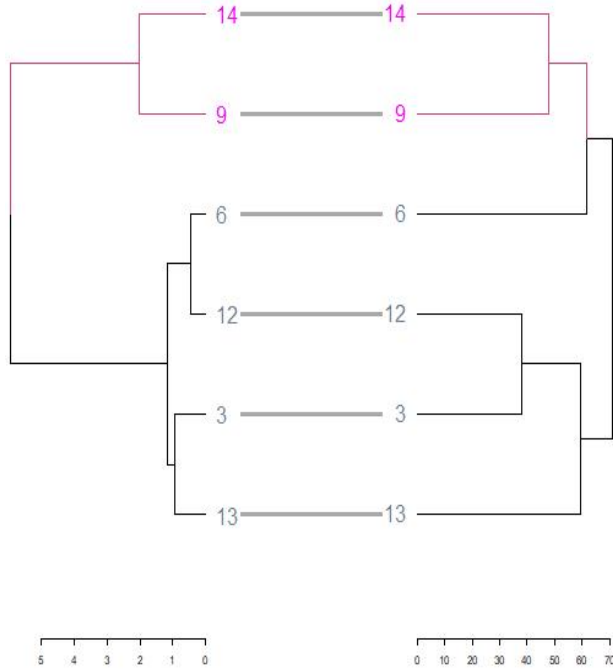


Figure 3. *C. elegans* TFBS derived dendrogram and co-expression derived dendrograms and their comparison for modules 3, 6, 9, 12, 13, and 14. Co-expression dendrogram is on the left and TFBS dendrogram is on the right.

Modules	Description	Coorelation	Entanglement
1-14	All modules	0.01	0.23
3, 6, 9, 12, 13, 14	Modules with constant expression and slight increase (grey and light purple)	0.40	0.00
1, 2, 5, 7, 8	Modules with parabola-like increases (light and dark green)	0.44	0.00
1, 2, 5, 7, 8, 10, 11	Modules with large increases in expression (light green, dark green and dark purple)	0.37	0.40

Table 3. *C. elegans* correlation and entanglement values for TFBS and co-expression dendrograms by module.

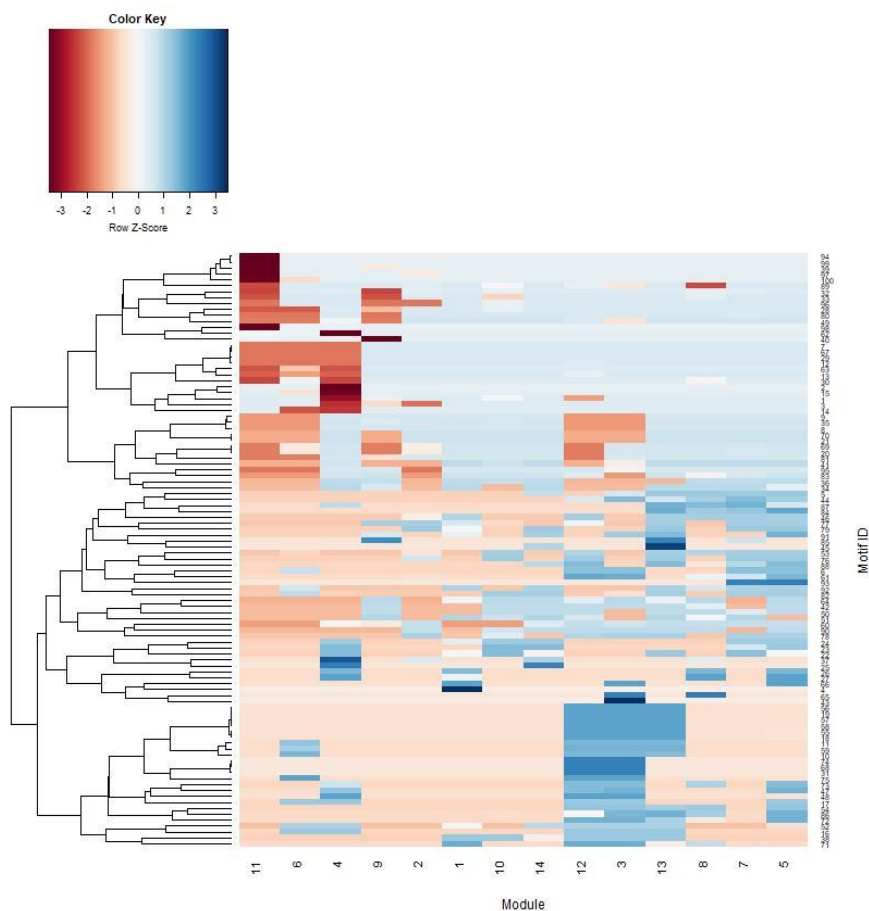


Figure 4. Heatmap of TFBS occurrence for *C. elegans*.

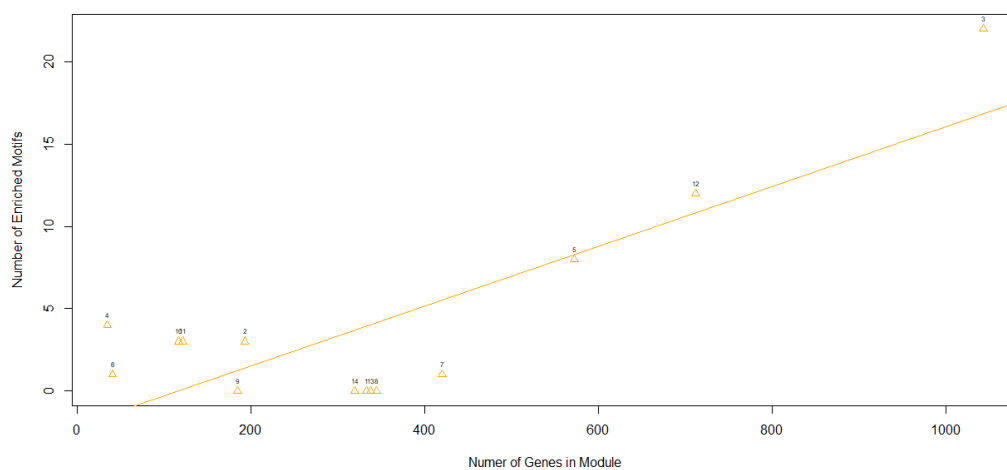


Figure 5. Orthologous *C. briggsae* modules plotted as the number of genes in the module by the number of motifs discovered.

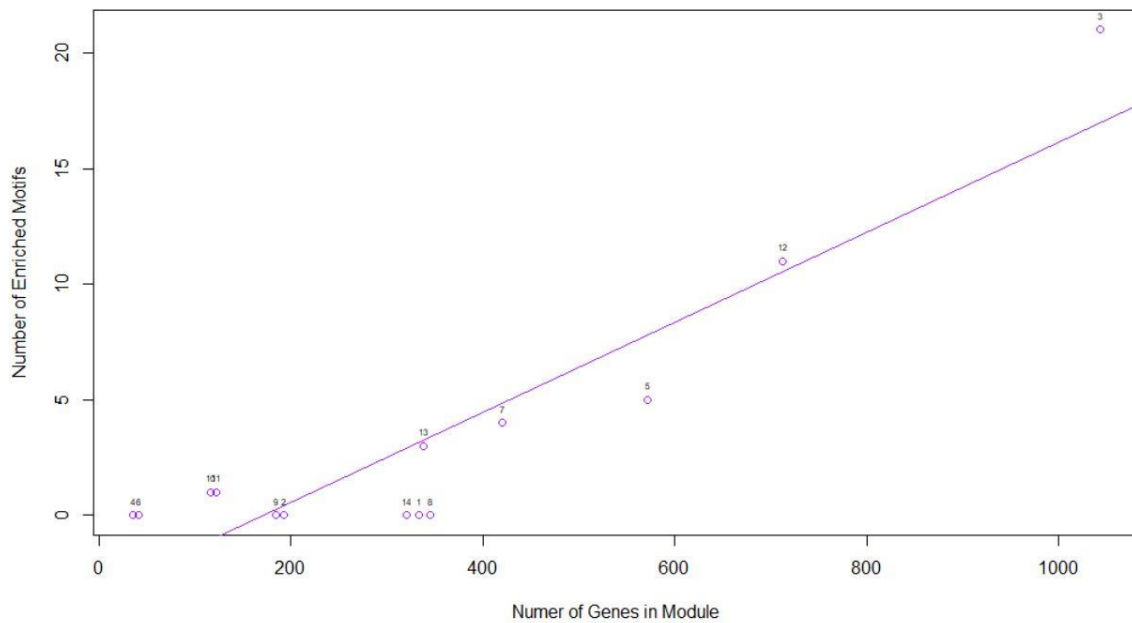


Figure 6. *C. elegans* genes that correspond to the 1:1 orthologs found of plotted as the number of genes in the module by the number of motifs discovered.

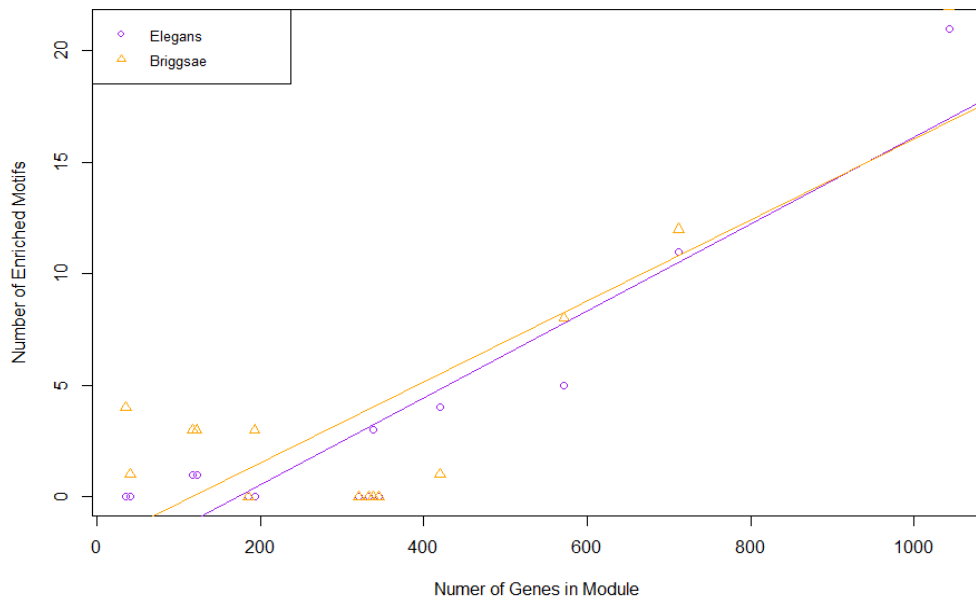


Figure 7. Orthologous *C. briggsae* modules with their *C. elegans* 1:1 orthologs plotted as the number of genes in the module by the number of motifs discovered.

MODULE	LOGO 1	LOGO 2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		

Supp. Table 1. *C. elegans* MEME results. For each module, the two most significant motifs were reported. Lighter coloured text indicates that the motif is not statistically significant. Motifs in module six were both found to be significantly similar to known motifs (E-value ≤ 10).

MODULE	LOGO 1	LOGO 2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		

Supp. Table 2. *C. briggsae* MEME results. For each module, the two most significant motifs were reported. Lighter coloured text indicates that the motif is not statistically significant. (E-value ≤ 10).

Literature Cited:

- Artieri, C. G., W. Haerty and R. S. Singh. 2009. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila*. *BioMed Central Biology* 7: 23-27.
- Altenhoff A. M., Glover N. M, Train C., Kaleb K., Vesztröcy A. W., Dylus D., de Farias T. M., et al. 2018. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*: 46 (D1): D477-D485.
- Altenhoff A. M., Škunca M., Glover N., et al. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view, and other improvements. *Nucleic Acids Research*: 43 (D1): D240-D249.
- Bailey T. L., J. Johnson, C. E. Grant and W. S. Noble. 2015. The MEME Suite. *Nucleic Acids Research*. 43: W39–W49.
- Bailey T. L., N. Williams, C. Mislé and W. W. Li. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. 34: W369–W373.
- Castillo-Davis, C. I., and D. L. Hartl. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Molecular Biology and Evolution* 19: 728-735.

- Castillo-Davis, C. I., D. L. Hartl and G. Achaz. 2004. Cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Research* 14: 1530-1536.
- Cutter A. D., R. Garrett, S. Mark, W. Wang and L. Sun. Unpublished. Molecular evolution across developmental time reveals rapid divergence in early embryogenesis.
- Cutter, A. D., and S. Ward. 2005. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. *Molecular Biology and Evolution* 22: 178-188.
- Cruickshank, T., and M. J. Wade. 2008. Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evolution & Development* 10: 583-590.
- Davis, J. C., O. Brandman and D. A. Petrov. 2005. Protein evolution in the context of *Drosophila* development. *Journal of Molecular Evolution* 60: 774-785.
- Domazet-Loso, T., and D. Tautz. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468: 815-U107.
- Emberly E., N. Rajewshy and E. D. Siggia. 2003. Conservation of regulatory elements between two species of *Drosophila*. *BioMedical Central Bioinformatics*. 4:57

- Henifoff S., Ahmad K., and Malik H. S. 2001 The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*: 293(5532): 1098-1102.
- Ho M. C., Quintero-Cadena P., Sternberg P. W. 2017. Genome-wide discovery of active regulatory and transcription factor footprints in *Caenorhabditis elegans* using DNase-seq. *Genome Res*: 27(12): 2108-2119.
- Howe K. L., Bolt B. J., Shafie M., et al. 2017. WormBase ParaSite – a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology*: 215 2-10.
- Howe K. L., Bolt B. J., Cain S., et al. 2016. WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Research*: 44(D1) D774-D780.
- Hu J., B. Li, and D. Kihara. 2005. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*. 33(15): 4899–4913.
- Kalinka, A. T., and P. Tomancak. 2012. The evolution of early animal embryos: conservation or divergence? *Trends in Ecology & Evolution* 27: 385-393.
- Kalinka, A. T., K. M. Varga, D. T. Gerrard, S. Preibisch, D. L. Corcoran et al. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468: 811-U102.

Langfelder P. and S. Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. *Biomedical Central Bioinformatics*. 9:559.

Levin, M., T. Hashimshony, F. Wagner and I. Yanai. 2012. Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Developmental Cell* 22: 1101-1108.

Mathelier A., and W. W. Wasserman. 2013. The Next Generation of Transcription Factor Binding Site Prediction. *PLOS Computational Biology*. 9(9): 1-18

McLeay R. C. and T. L. Bailey. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BioMedical Central Bioinformatics*: 11: 165-170.

Narasimhan K., Lambert S. A., Yang A. W. et al. 2015. Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities. *Elife*: 23: 4-9.

Riddle DL, T., Blumenthal, B.J. Meyer, et al., editors. *C. elegans* II. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. Section III, Transcription Factors.

Santos-Mendoza M., B. Dubreucq, S. Baud, F. Parcy, M. Caboche and L. Lepiniec 2008. Deciphering gene regulatory networks that control seed development and maturation in *Arabidopsis*. *The Plant Journal*. 54: 608–620.

- Shen Y., Yue F., McCleary D. F. et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature*: 488 116-120.
- Sleumer M. C., Bilenky M., He A., et al. 2009. *Caenorhabditis elegans* cisRED: a catalogue of conserved genomic elements. *Nucleic Acids Res*: 37(4): 1323-1334.
- Stern, D. L., and V. Orgogozo. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62: 2155-2177.
- Vandepoele K., M. Quimbaya, T. Casneuf, L. DeVeylder and Y. Van de Peer. 2009. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology*. 150(2): 535-546.
- Valouev A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers and A. Sidow. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods*. 5(9): 829-34.
- Webb C. T., Shabalina S. A., Yu A. and Kondrashov A. S. 2002 Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Research* 30: 5: 1233-1239.
- Wittkopp, P. J., and G. Kalay. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13:

59-69.

Woollard A. 2018. Gene duplications and genetic redundancy in *C. elegans*. Wormbook: The Online Review of *C. elegans* Biology. Pasadena (CA).

Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8: 206-216.