



## A novel candidate *cis*-regulatory motif pair in the promoters of germline and oogenesis genes in *C. elegans*

Chaim Linhart, Yonit Halperin, Amir Darom, et al.

*Genome Res.* 2012 22: 76-83 originally published online September 19, 2011

Access the most recent version at doi:[10.1101/gr.115626.110](https://doi.org/10.1101/gr.115626.110)

---

**References** This article cites 35 articles, 14 of which can be accessed free at:  
<http://genome.cshlp.org/content/22/1/76.full.html#ref-list-1>

### License

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Copyright © 2012 by Cold Spring Harbor Laboratory Press

## Research

# A novel candidate *cis*-regulatory motif pair in the promoters of germline and oogenesis genes in *C. elegans*

Chaim Linhart,<sup>1,3</sup> Yonit Halperin,<sup>1,3</sup> Amir Darom,<sup>2</sup> Shahar Kidron,<sup>2</sup> Limor Broday,<sup>2,4</sup> and Ron Shamir<sup>1,4</sup>

<sup>1</sup>School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel; <sup>2</sup>Department of Cell and Developmental Biology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

In this study we report on a novel pair of *cis*-regulatory motifs in promoter sequences of the nematode *Caenorhabditis elegans*. The motif pair exhibits extraordinary genomic traits: The order and the orientation of the two motifs are highly specific, and the distance between them is almost always one of two frequent distances. In contrast, the sequence between the motifs is variable across occurrences. Thus, the motif pair constitutes a nearly combinatorial sequence configuration. We further show that this module is conserved among, and unique to, the entire *Caenorhabditis* genus. By analyzing several gene expression data sets, our data suggest that this motif pair may function in germline development, oogenesis, and early embryogenesis. Finally, we verify that the motifs are indeed functional *cis*-regulatory elements using reporter constructs in transgenic *C. elegans*.

[Supplemental material is available for this article.]

The transcriptional program is one of the main mechanisms by which the cell controls gene activity. Regulation of transcription is governed by transcription factors (TFs)—proteins that activate or repress gene expression through binding to short sequences called *cis*-regulatory elements or binding sites (BSs). Each TF binds, with varying affinities, to many different sequences that reflect a common pattern, or *motif*, characteristic of the factor. Most BSs are found in the promoter, the region upstream of the gene's transcription start site (TSS). Some TFs cooperate in the regulation of genes, resulting in more complex and specific transcription profiles. Classical examples are the *cis*-regulatory modules (CRMs) that control the development of early *Drosophila* embryos (Berman et al. 2002). We previously demonstrated that the NFY and CHR elements comprise a regulatory module that is conserved among vertebrates and dictates expression that is restricted to the G<sub>2</sub> and G<sub>2</sub>/M cell cycle phases (Linhart et al. 2005). In some cases, the BSs within a module display some biases in terms of the order of the BSs or the distance between them. For example, studies on several promoters have shown that the SP1 binding site must be located close to the TATA-box in order to initiate transcription of the downstream genes (see, e.g., Wu and Berk 1988). Another example is periodic spacing between BSs, which is usually attributed to full turns of the DNA helix (see, e.g., Makeev et al. 2003). Many computational tools have been developed to date for discovering CRMs with various structural constraints (see survey in Sandve and Drablos 2006).

Due to its simple anatomy, optical properties, short life cycle, invariant cell lineage, powerful genetics (Brenner 1974), and its sequenced genome (The *C. elegans* Sequencing Consortium 1998), *Caenorhabditis elegans* has been widely used as a model system to

understand how regulation of gene transcription controls the development of a multicellular organism. Development from an embryo to a self-fertilizing adult hermaphrodite proceeds through four larval stages. Hermaphrodites produce ~300 embryos by fertilization of oocytes with self-sperm stored in the spermatheca. The reproductive system consists of the somatic gonad, the germline, and the egg-laying apparatus. The U-shaped gonad contains the germline, which differentiates into sperm and oocytes. Along the gonad, the germ cells are ordered in various differentiation states (Fig. 3B–D below). The distally located germ cells that are closest to the distal tip cell (DTC) are mitotic stem cells. The proximal germ cells undergo meiosis to produce haploid gametes. The first germ cells that enter meiosis produce sperm (late L4 stage), whereas the remaining germ cells (adult stage) develop into oocytes (Hubbard and Greenstein 2005).

In this study, we report on a pair of motifs with striking features in the promoter sequences of the nematode *C. elegans*. This motif pair is enriched in the promoters of genes that play a role in the regulation of germline stem cell proliferation, meiosis entry and progression, and female gametogenesis.

## Results

We performed a genome-wide analysis of *C. elegans* promoter sequences to search for motifs enriched in specific chromosomes. Among the discovered motifs we identified one striking pair that co-occurs significantly more often than expected by chance, and almost always in the same order and orientation. Focusing on this motif pair, we analyzed its genomic features and evolutionary conservation, and used several expression data sets to reveal its functional roles. Specifically, the motifs are enriched in the promoters of germline-specific genes. Finally, we analyzed transgenic worms expressing promoter GFP reporters to verify our computational findings. Of note, some of the phenomena we uncovered are, to the best of our knowledge, unique and have never been observed in any other genome.

<sup>3</sup>These authors contributed equally to this work.

<sup>4</sup>Corresponding authors.

E-mail [broday@post.tau.ac.il](mailto:broday@post.tau.ac.il).

E-mail [rshamir@post.tau.ac.il](mailto:rshamir@post.tau.ac.il).

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.115626.110>.

## Genomic features

The consensus sequences of the two motifs we discovered are GAGACCY and GYGCCTT (where Y stands for C or T) (Fig. 1A). We denote the motifs M1 and M2, respectively. A search in the TF databases Transfac and Jaspar found no similar motifs. Out of 20,151 promoter sequences in the genome of *C. elegans*, motifs M1 and M2 appear in 1804 and 969 promoters, respectively, and the two occur together in 285 promoters ( $p = 10^{-67}$ , hypergeometric [HG] test). Although we analyzed promoters spanning 1000 bases, in most of these co-occurrences, the distance between the motifs is short—in 208 promoters they appear at a distance of less than 100 bases. Considering both the order and orientation of the two motifs, there are four possible arrangements on the two DNA strands: M1 → M2 (which means that the promoter sequence contains GAGACCY...GYGCCTT, or its reverse complement: AAAGGCRC...RGGTCTC), M2 → M1, M2 → r(M1) [where r(M1) denotes the reverse-complement of M1] and r(M1) → M2. Strikingly, out of the 208 promoters with co-occurrence at distance <100, in 154 cases the configuration is M1 → M2 ( $p = 10^{-49}$ , binomial test) (Fig. 1B). This nearly combinatorial arrangement is highly exceptional: We are not aware of any other *cis*-regulatory elements that display a similar spatial bias. Scanning the entire *C. elegans* genome for occurrences of M1 → M2, we identified only 14 hits inside exons of protein-coding sequences (one hit per 2.4 Mbp, compared to one hit per 98 kbp in promoters), supporting the hypothesis that the motif pair, indeed, constitutes a *cis*-regulatory module (in fact, several of these hits are inside alternative promoters, so they may actually represent functional BSs). Moreover, the location of the M1 → M2 hits along the promoters is not distributed evenly: The motifs occur more frequently 100–300 bases upstream of

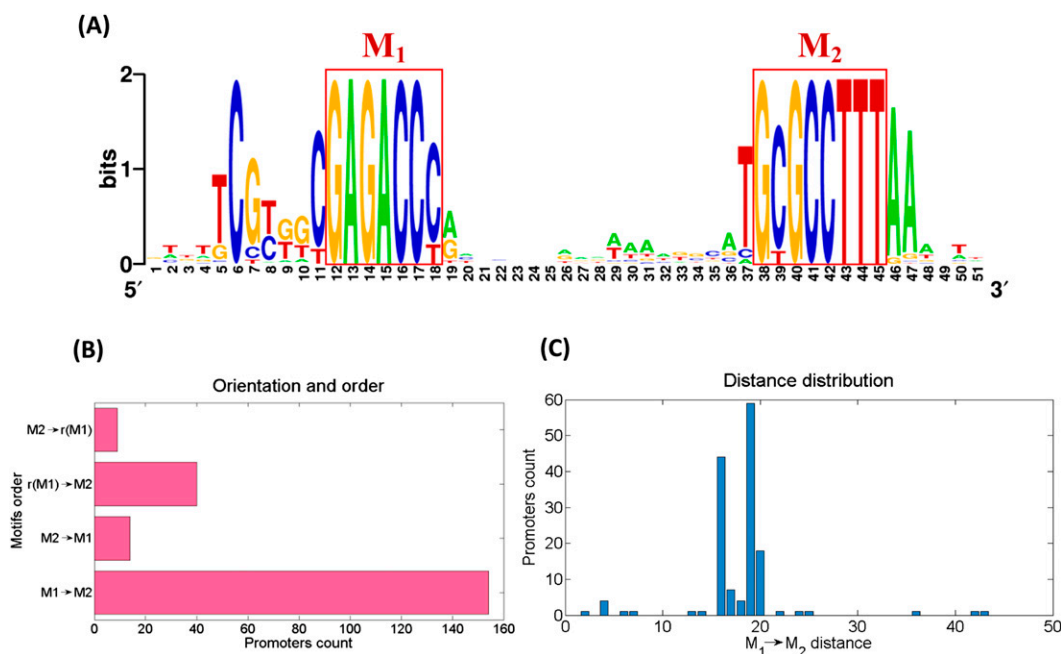
the transcript, a common characteristic of TF BSs (Supplemental Fig. 1).

Henceforth, we refer to the set of 154 promoters that contain the two motifs with orientation M1 → M2 and at distance <100 as the motif pair's candidate target promoters. These promoters regulate 205 genes (51 of the promoters are divergent, i.e., they are shared by two adjacent genes), of which 55 are in operons. These operons contain 85 additional genes, so the total number of genes whose promoters harbor the motif pair is 290. The full list of genes, their expression patterns (WormBase; <http://www.wormbase.org/>), and germline expression (Y. Kohara, NEXTDB in situ hybridization data set; <http://nematode.lab.nig.ac.jp/>) are presented in Supplemental Table 1.

We next examined the distance between the two motifs in their candidate target promoters. Remarkably, in almost all cases, the distance is between 16 and 20 bases (Fig. 1C). The two most common distances are 19 (59 occurrences) and 16 (44 occurrences). This additional spatial constraint is, again, unique and novel, and may stem from structural mechanisms that underlie the relevant regulators.

In light of the above observations, one may suggest that the two motifs are, in fact, part of one long repetitive sequence. However, this is not the case—the sequence of the gap between the two motifs is not conserved. Apart from a few bases on either side of each motif, the gap is highly variable, suggesting that it has no functional role (Fig. 1A).

Finally, we checked the chromosomal distribution of the motif pair's putative target promoters. We found that they are over-represented on chromosomes I ( $p = 2 \times 10^{-5}$ , HG test) and III ( $p = 2 \times 10^{-5}$ ), and under-represented on chromosomes X ( $p = 6 \times 10^{-9}$ ) and V ( $p = 8 \times 10^{-5}$ ) (Supplemental Fig. 2). (Note that these *P*-values are biased, since chromosomal bias was the initial clue to



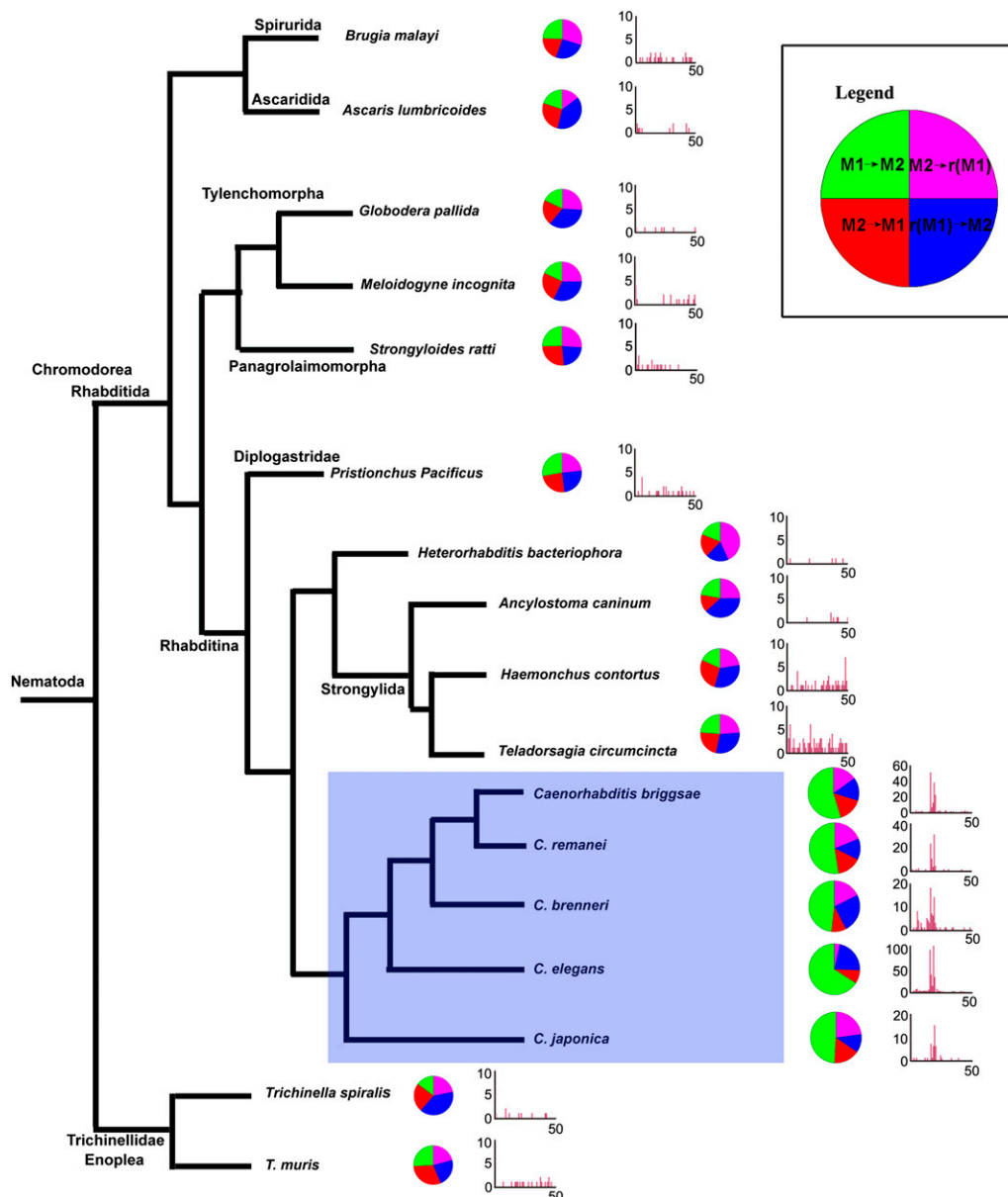
**Figure 1.** Properties of the motif pair (M1, M2). We refer only to occurrences in which the distance between the two motifs is <100. (A) Alignment of the 59 occurrences of M1 → M2 with gap length 19. The motifs are enclosed in red boxes. Several positions flanking the two motifs are conserved. In contrast, the gap between them is highly variable. The alignment for gap length 16 exhibits the same features (data not shown). (B) Number of occurrences of each of the possible arrangements of the two motifs. (C) The distribution of the distance between the two motifs for the M1 → M2 arrangement. The two motifs are almost always 16–20 bases apart. The graph shows all gaps of length 1–50; only four occurrences have a gap of length 51–99.

finding the motif pair.) These chromosomal preferences match those observed by Reinke et al. (2004) for genes that are expressed in the germline and related to oogenesis.

### Evolutionary conservation

To trace the evolutionary span of the motif pair, we scanned genomic sequences of several nematode species for occurrences of the motifs. Note that the total number of occurrences detected in each genome depends on various factors, such as its nucleotide composition and the number of bases scanned. However, the dis-

tribution of hits among the four possible structural arrangements and the distance between the motifs are insensitive to these variables. In all five *Caenorhabditis* species—*Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Caenorhabditis brenneri*, *Caenorhabditis elegans*, and *Caenorhabditis japonica*—the motifs exhibited the aforementioned order and orientation bias ( $p < 10^{-7}$  in all species, binomial test) (Fig. 2; Supplemental Table 2). Moreover, the pair has similar distance distributions; in all five species, the two most common distances are 16 and 19. Remarkably, we detected neither the order/orientation bias nor the distance bias in 12 other nematodes outside the *Caenorhabditis* genus, for which we could obtain



**Figure 2.** Evolutionary conservation of the motif pair in genomic sequences of 17 nematode species. All occurrences of the two motifs with distance  $<100$  between them were considered. The distributions of the four possible arrangements are shown in pie charts. Large pie charts indicate a statistically significant large number of occurrences of  $M1 \rightarrow M2$ . A histogram of the distances between the two motifs in occurrences of  $M1 \rightarrow M2$  is shown next to each pie chart (only distances  $<50$  are shown). All *Caenorhabditis* genomes contain a large number of occurrences of  $M1 \rightarrow M2$ , nearly always with 16–20 bases between the two motifs. All of the other nematodes we analyzed do not exhibit these phenomena. For tree construction we used the published phylogenies in De Ley (2006), Kiontke et al. (2007), and Jex et al. (2010).

genomic sequences ( $p > 0.2$  in all species). It therefore seems that the motif pair is distinctive of the *Caenorhabditis* genus.

### Functional analysis

Our next goal was to determine the regulatory role of the pair of motifs. First, we manually noticed that many of the 290 candidate target genes have well-established roles in germline development (Supplemental Table 1; Discussion). We next identified several Gene Ontology (GO) categories that are enriched in these 290 genes: embryonic development (GO:0009790,  $p = 10^{-26}$ , HG test), reproduction (GO:0000003,  $p = 10^{-22}$ ), gene expression (GO:0010467,  $p = 2 \times 10^{-9}$ ), and organ development (GO:0048513,  $p = 10^{-8}$ ).

The SAGE (Serial Analysis of Gene Expression) data set contains genome-wide expression profiles of 33 cell types and developmental stages of *C. elegans* (<http://elegans.bcgsc.ca>). We checked whether the putative target genes of the motif pair are highly expressed in some of these 33 conditions. To this end, for each condition we computed the overlap between the 290 candidate target genes and the set of genes whose expression in that condition was at least 1.5-fold higher than their average expression among all conditions. We then calculated the *P*-value of the overlap using the HG test. The motif pair's putative targets are highly expressed in the dissected gonad ( $p = 8 \times 10^{-23}$ ), as well as in N2 young adults ( $p = 4 \times 10^{-7}$ ), "fer-15 6 day" ( $p = 8 \times 10^{-7}$ ), and FACS sorted pharyngeal gland cells ( $p = 3 \times 10^{-6}$ ).

In light of these findings and the germline-related chromosomal bias mentioned above, we analyzed the expression profiles of the candidate target genes in two developmental and germline-specific microarray data sets. Baugh et al. (2003) profiled embryonic expression in precisely staged *C. elegans* embryos. Reinke et al. (2004) analyzed germline and sex-regulated genes by measuring 23 expression profiles of males and hermaphrodites during their development and also of mutant strains with defects in germline proliferation or gamete production. The 290 putative target genes of the motif pair exhibit several salient properties in these two data sets (see Supplemental Fig. 3; Supplemental Table 3): (1) Their expression decreases during the first 2 h of embryogenesis ( $p = 4 \times 10^{-9}$ , HG score), as well as during early larval development, and increases in L4 and to a greater extent in adults ( $p = 10^{-16}$ ). (2) They are down-regulated in mutants without germline relative to wild-type ( $p = 5 \times 10^{-21}$ ). (3) They are up-regulated in hermaphrodites relative to males ( $p = 3 \times 10^{-21}$ ). (4) Their expression is higher in hermaphrodite mutants that produce only oocytes than in mutants that produce only sperm ( $p = 4 \times 10^{-19}$ ).

These results indicate that the motif pair is highly concentrated in upstream sequences of genes required for the regulation of germline development and oogenesis in the late larval and adult hermaphrodite stages, although it may have additional somatic roles.

We repeated the functional analysis tests for the genes whose promoter contains only one of the motifs or the motif pair in one of the other three orientations. Some of these gene sets exhibited similar trends to those we reported above for the GO categories and the gene expression data sets (Supplemental Table 4). However, these trends are considerably less significant, both in terms of the fraction of genes that overlap between the examined sets, and in terms of the statistical scores they attained.

### Analysis of the motif pair in transgenic worms

To test the functionality of the motif pair experimentally, we focused on the *pas-5* gene, whose promoter contains the motif pair,

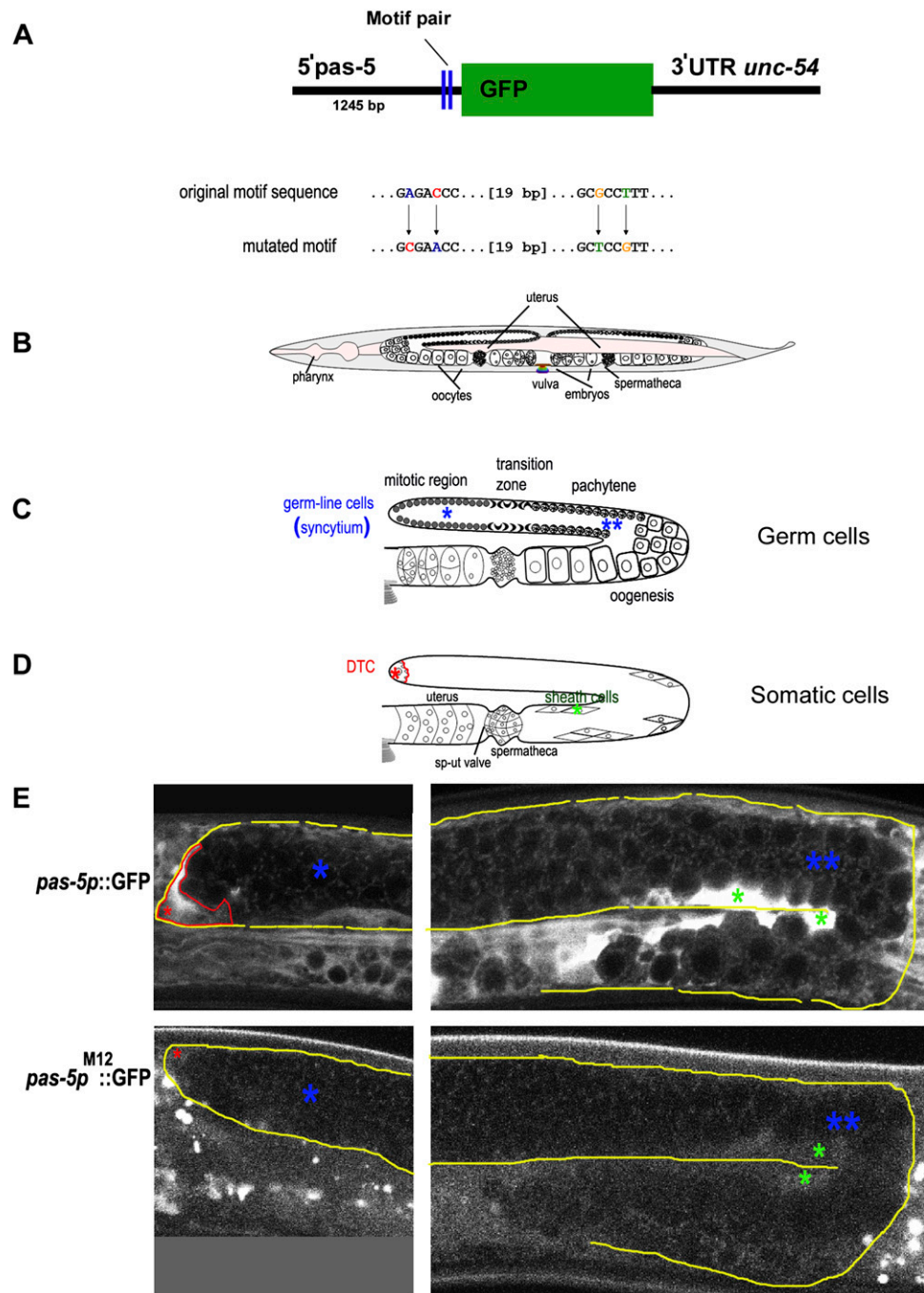
and carried out functional analysis in transgenic worms. *pas-5* encodes a proteasome subunit alpha type-5 and has been recently shown to have a role in maintaining the balance between proliferation and meiotic entry in the *C. elegans* germline as well as somatic functions (Macdonald et al. 2008). Previous *in situ* hybridization analysis showed *pas-5* mRNA accumulation in the germline (Y. Kohara, NEXTDB; <http://nematode.lab.nig.ac.jp/db2/ShowCloneInfo.php?clone=479g6>).

The motifs are located 186 bases upstream of the *pas-5* ATG start site, with a 19-base gap between them (Supplemental Table 1). We generated a transcriptional reporter containing 1245 bp upstream of the *pas-5* ATG translation start codon fused to GFP (Fig. 3A). The wild-type (WT) reporter, which contains an intact motif pair, is expressed in the germline as well as in the somatic gonad, including the DTCs and the sheath cells (Fig. 3C–E). This is consistent with the *in situ* hybridization data from the Y. Kohara laboratory. We next generated a mutated reporter by two purine-to-pyrimidine substitutions in each motif (Fig. 3A). In these animals, fluorescence levels are low and weakly detected in the germline and in the somatic gonad (Fig. 3E; Supplemental Figs. 4, 6). These observations indicate that the motif pair is required for the expression of *pas-5* both in the germline and in the somatic gonad, and therefore suggest that it serves as the binding site of one or more germline-specific transcriptional complexes.

### Discussion

Among the motif pair's putative target genes, 137 (47%) have a role in germline control and gametogenesis or are expressed in the germline according to the WormBase and NEXTDB *in situ* hybridization data (Supplemental Table 1A,B). One hundred twenty-eight (44%) genes showed sterile and/or embryonic lethal phenotypes in RNAi experiments, supporting our analysis of the published expression profiles (Supplemental Table 3; Baugh et al. 2003; Reinke et al. 2004). In addition, 72 (25%) of the genes are expressed or function in the soma, suggesting that the motif pair also functions outside the germline (Supplemental Table 1A,B). Many of the germline-expressed genes were shown to function in meiotic processes, such as chromosome pairing, recombination, and segregation (e.g., *zim-1*, *zim-2*, *zim-3*, *him-3*, *him-8*). Additional known genes identified are *gld-2*, which encodes a translational activator (Suh et al. 2006), and *nos-3*, which encodes a member of the Nanos family of RNA-binding proteins (Hansen et al. 2004)—both are key regulators of entry into meiosis (Kimble and Crittenden 2007).

The germ cells near the distal tip cells of the gonad serve as the stem cell niche for the adult germline stem cells. Maintenance of germ cell proliferation is controlled by their interaction with the DTCs (Kimble and Crittenden 2007). Expression of the *pas-5p::GFP* WT reporter in the DTCs suggests a role for the motifs in the regulation of the stem cell niche and germline proliferation. Analysis of transgenic lines expressing the extrachromosomal array of *pas-5p::GFP*, which is silenced in the germline (see Methods) (Supplemental Fig. 5), showed similar expression in the somatic gonad and additional weak expression in the vulva, muscle, and epidermis, suggesting somatic roles for the motif pair. Expression of the WT reporter in the germline itself along the gonad supports its role in the regulation of meiotic entry decision (Macdonald et al. 2008) and gametogenesis. This experimental evidence is in full agreement with our computational findings, which showed enrichment of the motif pair in promoters of genes that are germline- and oogenesis-specific, and that are expressed mainly in L4 and adult hermaphrodites. The temporal expression profiles of the 290 genes



**Figure 3.** The motif pair in the *pas-5* promoter is required for germline and somatic gonad expression. (A) Structure of the *pas-5p::GFP* reporter. A region of 1245 bp of the *pas-5* promoter was fused to GFP. The sequences of the wild-type and mutated motifs are shown. (B) A scheme of an adult *C. elegans* hermaphrodite. The gonad is made of two symmetric U-shaped arms. Each of the arms terminates at the spermatheca. Oocytes are fertilized in the spermatheca, and embryos mature in the uterus and are laid through the vulva. (C) The germline cells of the gonad. Germ cells proliferate (mitotic zone) in the distal region of the gonad, and as they move proximally (transition zone), they enter meiosis and differentiate into sperm or oocyte. (D) The somatic gonad is composed of the distal tip cell (DTC), gonadal sheath, spermatheca (sp), spermathecal-uterine (sp-ut) valve, and uterus. The sheath and DTCs are physically associated with the germline cells and have critical roles in their development and function. The DTCs control gonad migration and regulate entry of the germline cells into mitosis versus meiosis (Kimble and White 1981). The sheath cells are smooth muscle-like cells necessary for oocyte maturation and ovulation (McCarter et al. 1999). (E) Expression pattern of wild-type *pas-5p::GFP* versus *pas-5p<sup>M12</sup>::GFP* reporters in the adult gonad. Expression is detected in the mitotic region of the germline (blue asterisk), the meiotic region (pair of blue asterisks), the DTC (red asterisk and curve), and the gonadal sheath cells (green asterisks). Additional images are shown in Supplemental Figure 6.

putatively regulated by the motif pair (Supplemental Fig. 3) show a decrease in their levels during embryogenesis, as expected from this group of transcripts.

Since expression of the GFP reporter tested was mainly detected in the adult stage (in parallel to the sperm/oocyte switch), we can further narrow its function in gametogenesis to female

gametogenesis. Two of the *Caenorhabditis* species shown in Figure 2 are hermaphrodites (*C. elegans* and *C. briggsae*), while the others are gonochoristic species (*C. remanei*, *C. brenneri*, and *C. japonica*). Hermaphrodites produce sperm at the L4 stage and then switch to oocyte generation, whereas females of gonochoristic species generate oocytes in a more continuous mode. The conservation of the motif pair in the *Caenorhabditis* species supports one of its roles in female gametogenesis, which is a common process to all these species, although its timing is different between hermaphrodites and gonochoristic species.

Scanning genomic sequences of multiple nematode species for occurrences of the motif pair revealed that this *cis*-regulatory module is unique to the *Caenorhabditis* genus. By testing the expression pattern of enhancers from *Drosophila* in *C. elegans*, it has been shown that changes in *cis*-regulatory sequences accumulate relatively fast, while expression patterns remain conserved along animal evolution (Ruvinsky and Ruvkun 2003). This may explain why the motif pair was observed here only in the *Caenorhabditis* genus. Yet, there are known examples in which *cis*-regulatory sequences are highly conserved between *C. elegans* and mammalian promoters. For example, the target genes of DAF-19, an RFX-type TF, contain conserved X boxes of 13–14 bp of imperfect inverted repeats that are similar to the mammalian X boxes (Swoboda et al. 2000). Shostak et al. (2004) showed that most BSs of the *C. elegans* intracellular receptor DAF-12 contain conserved sites of the mammalian DR5 motifs (retinoic acid–responsive element) sufficient for transcriptional activation by DAF-12. Thus, although the developmental programs and TFs that play key roles in them are highly conserved along evolution, some *cis*-regulatory elements remain conserved along evolution, while others are not. The concrete motif pair that we identified is not observed beyond the *Caenorhabditis* genus. However, we cannot rule out the possibility that the motifs do appear in some of the non-*Caenorhabditis* species in a diverged form and therefore were not detected. Testing this would require development of novel methods for identifying less conserved motif pairs.

Many *cis*-regulatory elements have been identified and characterized in *C. elegans*. Among them are the promoters of terminal differentiation/selector genes of individual neurons (Hobert 2008), the transcriptional network underlying vulval development (Inoue et al. 2005), and motifs in target genes of the TF PHA-4 that specifies pharyngeal identity (Gaudet and Mango 2002). In terms of length and distance from the TSS, the motifs we found resemble many known TFBS patterns in *C. elegans* and in other species. However, as a pair, they exhibit spatial features (order, orientation, and gap length) that, as far as we know, have not been observed with similar nearly combinatorial constraints in any other genome.

It has been shown that germline genes in *C. elegans* rely primarily on 3' UTRs for regulation during germline development (Merritt et al. 2008). Post-transcriptional regulation through the 3' UTR is likely to be responsible for establishing a spatially restricted domain of expression, since expression driven by the promoter motif pair described here does not appear sufficient.

Since *pas-5* reporter expression is detected both in the somatic gonad and in the germline, it is possible that at least two different transcriptional or repressor complexes are involved in germline control in different cells through binding to the motifs. Additional experiments are required in order to infer the exact function of the motif pair and to uncover the TF or TFs (or perhaps other regulatory molecules) that bind to it. These findings may also explain the unique structural arrangement of the motif pair.

In summary, we have identified a pair of *cis*-regulatory motifs in promoter sequences of *C. elegans* with some extraordinary ge-

netic traits: The order, orientation, and distance between the two motifs are highly specific. Genes whose promoter contains only one of the motifs, or both motifs in one of the other configurations, exhibit similar functional trends, albeit at much lower statistical significance levels. One possible explanation of this finding is that we do not identify the real motifs entirely accurately (and, hence, some of the genes that we identify as containing only one of the motifs actually contain the pair). Alternatively, or in addition, the single motifs may be functional as well (in the same processes as the pair), either on their own or with other partners. The pair M1 → M2 not only yields by far the most statistically significant functional enrichment in the various data sets we checked, but it also occurs at a much higher frequency than the other configurations across all *Caenorhabditis* species. The predominance of gaps of lengths 16 and 19 may have structural significance for the binding of TFs to the DNA double helix. On a broader scale, it would be interesting to conduct a comprehensive search for *cis*-regulatory modules with similar characteristics in various genomes.

## Methods

### Genomic sequences

Promoter sequences of *C. elegans* were downloaded from Ensembl (release 56, Sep 2009) (Flicek et al. 2010); 1000 bp upstream of each transcript was used as promoters (transcription start sites in Ensembl are based on WormBase annotations). Repetitive sequences were masked out using the RepeatMasker program (version 3.2.6) (Smit et al. 1996). Whole-genome assemblies for evolutionary conservation analysis were downloaded from the UCSC Bioinformatics website (Rhead et al. 2010). The following genomes were downloaded: *C. brenneri* (caePb2, Feb 2008), *C. briggsae* (cb3, Jan 2007), *C. elegans* (ce6, May 2008), *C. japonica* (caeJap1, Mar 2008), *C. remanei* (caeRem3, May 2007), and *Pristionchus pacificus* (priPac1, Feb 2007). For species without assemblies, trace records (raw single-pass reads of DNA sequence) were downloaded from the Taxonomy Browser in the NCBI website (Sayers et al. 2010). Only nematode species with total sequence length of >10 Mbp were analyzed.

### Computational analyses

We carried out a genome-wide search for motifs that exhibit a chromosomal bias in *C. elegans*, using the Amadeus/Allegro software (Halperin et al. 2009). Among the identified motifs, we manually spotted a pair of motifs that seemed to co-occur in promoters at a preserved distance. To investigate the phenomenon further, we analyzed the motifs individually and as a pair using custom-made Perl scripts.

GO functional analysis was performed using the TANGO algorithm via the Expander software (Ulitsky et al. 2010). All other enrichment tests were computed using the hypergeometric (HG) test. Briefly, the HG test checks whether the overlap between two sets of genes is significantly large (or small), given the size of each set and the total number of genes in the genome. Order and orientation bias of the motif pair was computed using the binomial test. Assuming equal probability for each of the four possible spatial arrangements of two motifs on two strands, the probability for observing  $k$  or more occurrences of a specific arrangement out of a total of  $n$  occurrences of the motif pair can be computed using the binomial distribution with  $p = 0.25$ . The HG and binomial scores are described in detail in Halperin et al. (2009) and Linhart et al. (2008).

SAGE data were obtained from the Genome BC *C. elegans* Gene Expression Consortium (<http://elegans.bcgsc.bc.ca>). Expression

values of each gene were transformed to the fold-change with respect to its average expression in all the conditions. The SAGE data include a total of 17,123 distinct genes and 33 conditions (tissue types and developmental stages).

### Cloning and mutagenesis of *pas-5* upstream regulatory sequences

Genomic sequence (1245 bp) upstream of the *pas-5* start codon was amplified from wild-type N2 genomic DNA using the primers F\_pas5p: 5'-CTCAGCTGTCTCGGCTCTCT, and R\_pas5p: 5'-GGAAGATGACTGGCTGAAATG. The PCR product was fused to GFP and to the 3' UTR of *unc-54* (pPD95.75 A. Fire) and cloned into a pAD010 vector containing the *unc-119* gene (Margalit et al. 2007) by replacing the *baf-1* promoter region.

Site-directed mutagenesis of the motif pair was performed using the primers F\_pas5p\_M12 (mutagenized bases are in bold):

5'-CGCGTCGCGA**ACC**AGTAAGCGATTAAAAGAAAG**CTCC**GTTA  
AAAACC  
and R\_pas5p\_M12:  
5'-GGTTTTTAACGG**AG**CTTCTTTTAATCGCTTACTGG**TTCC**GGC  
ACGCG.

### Generation and analysis of transgenic worms

Transgenic lines were generated using the DNA-coated micro-particle bombardment technique (Praitis 2006). In this technique, the gold-coated DNA construct is integrated directly into the chromosome, thus allowing germline expression (Green et al. 2008). Existence of the transgene was verified by PCR. Transgenic lines showed 100% heritability when tested. Analysis was done at the F<sub>1</sub> generation. Screening for transgenes was done at 25°C and isolated transgenic worms were grown at 20°C. Confocal analysis of *pas-5p::GFP* (two transgenic lines) and *pas-5p<sup>M12</sup>::GFP* (three transgenic lines) expression was performed using a Zeiss LSM 5 EXCITER confocal scanning microscope using a ×63 NA1.4 objective lens. At least 20 animals were analyzed from each line. Quantification of the average fluorescent intensity in the DTCs of 10 worms (one DTC/worm) was done by the Zeiss LSM 5 EXCITER confocal program and is shown in Supplemental Figure 4. Representative images are shown in Figure 3 and Supplemental Figure 6.

### Acknowledgments

We thank Marian Walhout, Inmaculada Barrasa, Natalia Martinez, and Maya Mouler-Rechtman for their help with the experimental testing; and Karin Kiontke for her help in phylogenetic tree construction and comments. This work was supported by the Israel Science Foundation (ISF 980/06 to L.B.), the Converging Technologies Program of the Israel Science Foundation (1767.07 to R.S.), and the European Community's Seventh Framework Programme—the TRIREME project (HEALTH-F4-2009-223575 to R.S.).

### References

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.  
Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc Natl Acad Sci* **99**: 757–762.  
Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77**: 71–94.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**: 2012–2018.  
De Ley P. 2006. A quick tour of nematode diversity and the backbone of nematode phylogeny. *WormBook* **25**: 1–8.  
Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, et al. 2010. Ensembl's 10th year. *Nucleic Acids Res* **38**: D557–D562.  
Gaudet J, Mango SE. 2002. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* **295**: 821–825.  
Green RA, Audhya A, Pozniakovsky A, Dammermann A, Pemble H, Monen J, Portier N, Hyman A, Desai A, Oegema K. 2008. Expression and imaging of fluorescent proteins in the *C. elegans* gonad and early embryo. *Methods Cell Biol* **85**: 179–218.  
Halperin Y, Linhart C, Ulitsky I, Shamir R. 2009. Allegro: Analyzing expression and sequence in concert to discover regulatory programs. *Nucleic Acids Res* **37**: 1566–1579.  
Hansen D, Wilson-Berry L, Dang T, Schedl T. 2004. Control of the proliferation versus meiotic development decision in the *C. elegans* germline through regulation of GLD-1 protein accumulation. *Development* **131**: 93–104.  
Hobert O. 2008. Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc Natl Acad Sci* **105**: 20067–20071.  
Hubbard EJ, Greenstein D. 2005. Introduction to the germ line. *WormBook* **1**: 1–4.  
Inoue T, Wang M, Ririe TO, Fernandes JS, Sternberg PW. 2005. Transcriptional network underlying *Caenorhabditis elegans* vulval development. *Proc Natl Acad Sci* **102**: 4972–4977.  
Jex AR, Hall RS, Littlewood DTJ, Gasser RB. 2010. An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Res* **38**: 522–533.  
Kimble J, Crittenden SL. 2007. Controls of germline stem cells, entry into meiosis, and the sperm/oocyte decision in *Caenorhabditis elegans*. *Annu Rev Cell Dev Biol* **23**: 405–433.  
Kimble JE, White JG. 1981. On the control of germ cell development in *Caenorhabditis elegans*. *Dev Biol* **81**: 208–219.  
Kiontke K, Barriere A, Kolotuev I, Podbilewicz B, Sommer R, Fitch DH, Felix MA. 2007. Trends, stasis, and drift in the evolution of nematode vulva development. *Curr Biol* **17**: 1925–1937.  
Linhart C, Elkon R, Shiloh Y, Shamir R. 2005. Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis. *Cell Cycle* **4**: 1788–1797.  
Linhart C, Halperin Y, Shamir R. 2008. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Res* **18**: 1180–1189.  
Macdonald LD, Knox A, Hansen D. 2008. Proteasomal regulation of the proliferation vs. meiotic entry decision in the *Caenorhabditis elegans* germ line. *Genetics* **180**: 905–920.  
Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA. 2003. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res* **31**: 6016–6026.  
Margalit A, Neufeld E, Feinstein N, Wilson KL, Podbilewicz B, Gruenbaum Y. 2007. Barrier to autointegration factor blocks premature cell fusion and maintains adult muscle integrity in *C. elegans*. *J Cell Biol* **178**: 661–673.  
McCarter J, Bartlett B, Dang T, Schedl T. 1999. On the control of oocyte meiotic maturation and ovulation in *Caenorhabditis elegans*. *Dev Biol* **205**: 111–128.  
Merritt C, Rasoloson D, Ko D, Seydoux G. 2008. 3' UTRs are the primary regulators of gene expression in the *C. elegans* germline. *Curr Biol* **18**: 1476–1482.  
Praitis V. 2006. Creation of transgenic lines using microparticle bombardment methods. *Methods Mol Biol* **351**: 93–107.  
Reinke V, Gil IS, Ward S, Kazmer K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131**: 311–323.  
Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al. 2010. The UCSC Genome Browser database: Update 2010. *Nucleic Acids Res* **38**: D613–D619.  
Ruvinsky I, Ruvkun G. 2003. Functional tests of enhancer conservation between distantly related species. *Development* **130**: 5133–5142.  
Sandve GK, Drablos F. 2006. A survey of motif discovery methods in an integrated framework. *Biol Direct* **1**: 11. doi: 10.1186/1745-6150-1-11.  
Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvermin V, Church DM, Dicuccio M, Federhen S, et al. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **38**: D5–D16.

- Shostak Y, Van Gilst MR, Antebi A, Yamamoto KR. 2004. Identification of *C. elegans* DAF-12-binding sites, response elements, and target genes. *Genes Dev* **18**: 2529–2544.
- Smit AFA, Hubley R, Green P. 1996. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.
- Suh N, Jedamzik B, Eckmann CR, Wickens M, Kimble J. 2006. The GLD-2 poly(A) polymerase activates *gld-1* mRNA in the *Caenorhabditis elegans* germ line. *Proc Natl Acad Sci* **103**: 15108–15112.
- Swoboda P, Adler HT, Thomas JH. 2000. The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in *C. elegans*. *Mol Cell* **5**: 411–421.
- Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, Tanay A, Sharan R, Shiloh Y, Shamir R. 2010. Expander: From expression microarrays to networks and functions. *Nat Protoc* **5**: 303–322.
- Wu L, Berk A. 1988. Constraints on spacing between transcription factor binding sites in a simple adenovirus promoter. *Genes Dev* **2**: 403–411.

Received September 22, 2010; accepted in revised form September 12, 2011.