



## Genomes &amp; Developmental Control

## The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine

James D. McGhee<sup>a,\*</sup>, Monica C. Sleumer<sup>b</sup>, Mikhail Bilenky<sup>b</sup>, Kim Wong<sup>b</sup>, Sheldon J. McKay<sup>b,1</sup>, Barbara Goszczynski<sup>a</sup>, Helen Tian<sup>a</sup>, Natisha D. Krich<sup>a</sup>, Jaswinder Khattri<sup>b</sup>, Robert A. Holt<sup>b</sup>, David L. Baillie<sup>c</sup>, Yuji Kohara<sup>d</sup>, Marco A. Marra<sup>b</sup>, Steven J.M. Jones<sup>b</sup>, Donald G. Moerman<sup>e</sup>, A. Gordon Robertson<sup>b</sup>

<sup>a</sup> Department of Biochemistry and Molecular Biology, University of Calgary, 3330 Hospital Drive N.W., Calgary, Alberta, Canada T2N 4N1

<sup>b</sup> Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia, Canada V5Z 1L3

<sup>c</sup> Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada V5A 1S6

<sup>d</sup> National Institute of Genetics, 1111 Yata, Mishima 411-8540, Japan

<sup>e</sup> Department of Zoology, University of British Columbia, 2329 West Mall Vancouver, British Columbia, Canada V6T 1Z4

Received for publication 10 August 2006; revised 8 October 2006; accepted 14 October 2006

### Abstract

A SAGE library was prepared from hand-dissected intestines from adult *Caenorhabditis elegans*, allowing the identification of >4000 intestinally-expressed genes; this gene inventory provides fundamental information for understanding intestine function, structure and development. Intestinally-expressed genes fall into two broad classes: widely-expressed “housekeeping” genes and genes that are either intestine-specific or significantly intestine-enriched. Within this latter class of genes, we identified a subset of highly-expressed highly-validated genes that are expressed either exclusively or primarily in the intestine. Over half of the encoded proteins are candidates for secretion into the intestinal lumen to hydrolyze the bacterial food (e.g. lysozymes, amoebapores, lipases and especially proteases). The promoters of this subset of intestine-specific/intestine-enriched genes were analyzed computationally, using both a word-counting method (RSAT oligo-analysis) and a method based on Gibbs sampling (MotifSampler). Both methods returned the same over-represented site, namely an extended GATA-related sequence of the general form AHTGATAARR, which agrees with experimentally determined *cis*-acting control sequences found in intestine genes over the past 20 years. All promoters in the subset contain such a site, compared to <5% for control promoters; moreover, our analysis suggests that the majority (perhaps all) of genes expressed exclusively or primarily in the worm intestine are likely to contain such a site in their promoters. There are three zinc-finger GATA-type factors that are candidates to bind this extended GATA site in the differentiating *C. elegans* intestine: ELT-2, ELT-4 and ELT-7. All evidence points to ELT-2 being the most important of the three. We show that worms in which both the *elt-4* and the *elt-7* genes have been deleted from the genome are essentially wildtype, demonstrating that ELT-2 provides all essential GATA-factor functions in the intestine. The SAGE analysis also identifies more than a hundred other transcription factors in the adult intestine but few show an RNAi-induced loss-of-function phenotype and none (other than ELT-2) show a phenotype primarily in the intestine. We thus propose a simple model in which the ELT-2 GATA factor directly participates in the transcription of all intestine-specific/intestine-enriched genes, from the early embryo through to the dying adult. Other intestinal transcription factors would thus modulate the action of ELT-2, depending on the worm’s nutritional and physiological needs.

© 2006 Elsevier Inc. All rights reserved.

**Keywords:** *C. elegans*; Intestine; Transcriptional regulation; ELT-2; GATA factor

### Introduction

Our aim is to understand how a transcriptional program unfolds during the process of organogenesis. What are the transcription factors that drive primordium specification,

\* Corresponding author.

E-mail address: jmcghee@ucalgary.ca (J.D. McGhee).

<sup>1</sup> Current address: Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 117240, USA.

cellular differentiation and organ morphogenesis? How do these factors coordinate transcription of the genes encoding the structural proteins and enzymes that produce the form and function of the final organ? Is there a single dominant transcription factor per organ? Or, in the other extreme, are there multiple quasi-independent regulatory networks, each making its own particular contribution to organogenesis?

The intestine lineage of the nematode *Caenorhabditis elegans* provides an experimental system in which the above questions can be addressed (for a recent review, see McGhee, in press). The intestine (the entire worm endoderm) constitutes roughly one-third of the adult soma but is clonally derived from a single cell (the E cell) in the eight cell embryo (Deppe et al., 1978; Sulston et al., 1983). The assimilative and synthetic capacities of the intestine are remarkable: an adult hermaphrodite can convert her body mass into oocytes roughly once per day (Hirsh et al., 1976) and all of this material must pass through the intestine. In addition, the intestine is the seat of complex behaviours such as rhythmic defecation (Dal Santo et al., 1999; Espelt et al., 2005) and is intimately involved in the control of lifespan and aging (Libina et al., 2003; Berman and Kenyon, 2006). Thus, one purpose of the current paper is to produce a complete transcript inventory of the adult intestine, an important and fundamental step towards understanding how the intestine functions in food digestion, macromolecular synthesis and storage, and the overall coordination of the worm's physiology.

The second purpose of the present paper is to investigate the global regulation of gene transcription in the mature intestine. Over the past 20 years, the promoters of a number of intestine-specific genes in *C. elegans* have been analyzed experimentally; in all cases, intestine genes have turned out to be controlled by critical *cis*-acting GATA-related sequences (MacMorris et al., 1992, 1994; Egan et al., 1995; Britton et al., 1998; Moilanen et al., 1999; An and Blackwell, 2003; Luersen et al., 2004; Fukushige et al., 2005; Oskouian et al., 2005; Pauli et al., 2006). Computational analysis of *C. elegans* promoters is turning out to be quite successful in identifying candidate regulatory sequences in coordinately-controlled genes (Ao et al., 2004; Bigelow et al., 2004; Gaudet et al., 2004; Portman and Emmons, 2004; Wenick and Hobert, 2004; McCarroll et al., 2005; Pauli et al., 2006) and a variety of algorithms are available; it is not yet clear which algorithm is optimal but combinations of independent methods appear to produce more reliable results (Tompa et al., 2005). Thus, to bridge the gap between the small numbers of intestinal promoters that have been (and can be) investigated experimentally and the much larger number of intestine-specific/intestine-enriched promoters provided by the current SAGE analysis, we select a set of 74 highly-expressed highly-validated intestine-specific/intestine-enriched genes, analyze their promoters by two independent computational methods (Van Helden et al., 1998; Thijs et al., 2001) and identify an extended GATA sequence (consensus=AHTGATAARR) that agrees well with the experimentally determined motif. Further analysis suggests that the majority (perhaps all) of intestine-specific/intestine-enriched genes transcribed in the *C. elegans* intestine are controlled by a *cis*-acting extended GATA site.

ELT-2, ELT-4 and ELT-7 are the only three GATA-type transcription factors expressed in the post-embryonic intestine and thus candidates for interacting with the extended GATA site in intestinal promoters (Hawkins and McGhee, 1995; Fukushige et al., 1998; Maduro and Rothman, 2002; Fukushige et al., 2003; McGhee, in press). We show that worms in which both the *elt-4* and *elt-7* genes have been deleted from the genome are essentially wildtype. We thus propose that ELT-2 is involved in all acts of transcription of intestine-specific/intestine-enriched genes, beginning from the 4–8E cell stage of embryogenesis and continuing until the worm dies several weeks later. ELT-2 may also be involved in regulating the intestinal component of genes expressed ubiquitously. To be sure, the action of ELT-2 is likely to be modulated by other transcription factors and the SAGE inventory identifies more than 100 such factors expressed in the adult intestine. However, none of these other transcription factor genes appear to have a loss-of-function phenotype that primarily involves the intestine.

In summary, the evidence of the current paper suggests that the ELT-2 GATA factor is the major transcription factor directly controlling the many “effector” genes that produce the *C. elegans* intestine. We believe that this work represents a significant step towards understanding the complete oocyte-to-adult transcriptional pathway controlling formation and function of this simple organ.

## Results

### *Production and characterization of the SAGE libraries*

As described in detail in the Methods section, a total of 1863 intestine fragments were hand-dissected from adult *C. elegans glp-4(bn2)* hermaphrodites (*glp-4(bn2)* worms raised at 25°C have no gonad to interfere with the dissection; Beanan and Strome, 1992). Two SAGE libraries were prepared from unamplified RNA, one library from the isolated intestines and a second “total soma” library from intact *glp-4 (bn2)* adults harvested in exact parallel. These are among 35 such libraries, each sequenced to a depth of ~100,000 tags, prepared from different stages, sexes and tissues of *C. elegans* (McKay et al., 2003; Wong et al., submitted for publication). All of this information is publicly available at <http://www.elegans.bcgsc.ca/home/sage.html>. We estimate that the purity of the intestine library is >95% (see Methods).

Analysis of the adult intestine library identified 4247 individual genes (tag counts >1); 5867 individual genes were identified in the control library from the adult soma using the same criteria (see Methods). If singleton tags are ignored, 2830 and 3797 individual genes are identified in the intestine and total soma library, respectively. As expected, the majority of the genes identified in the intestine library are also found in the soma library (77%, 87% and 96% for intestinal tag counts >1, >2 and >10 respectively); less than perfect inclusion is reasonably ascribed to sampling error at low tag counts.

Of the genes identified in the two libraries, 60–70% can currently be assigned a KOG classification (Tatusov et al., 2003; Koonin et al., 2004). Fig. 1A shows how the genes in the

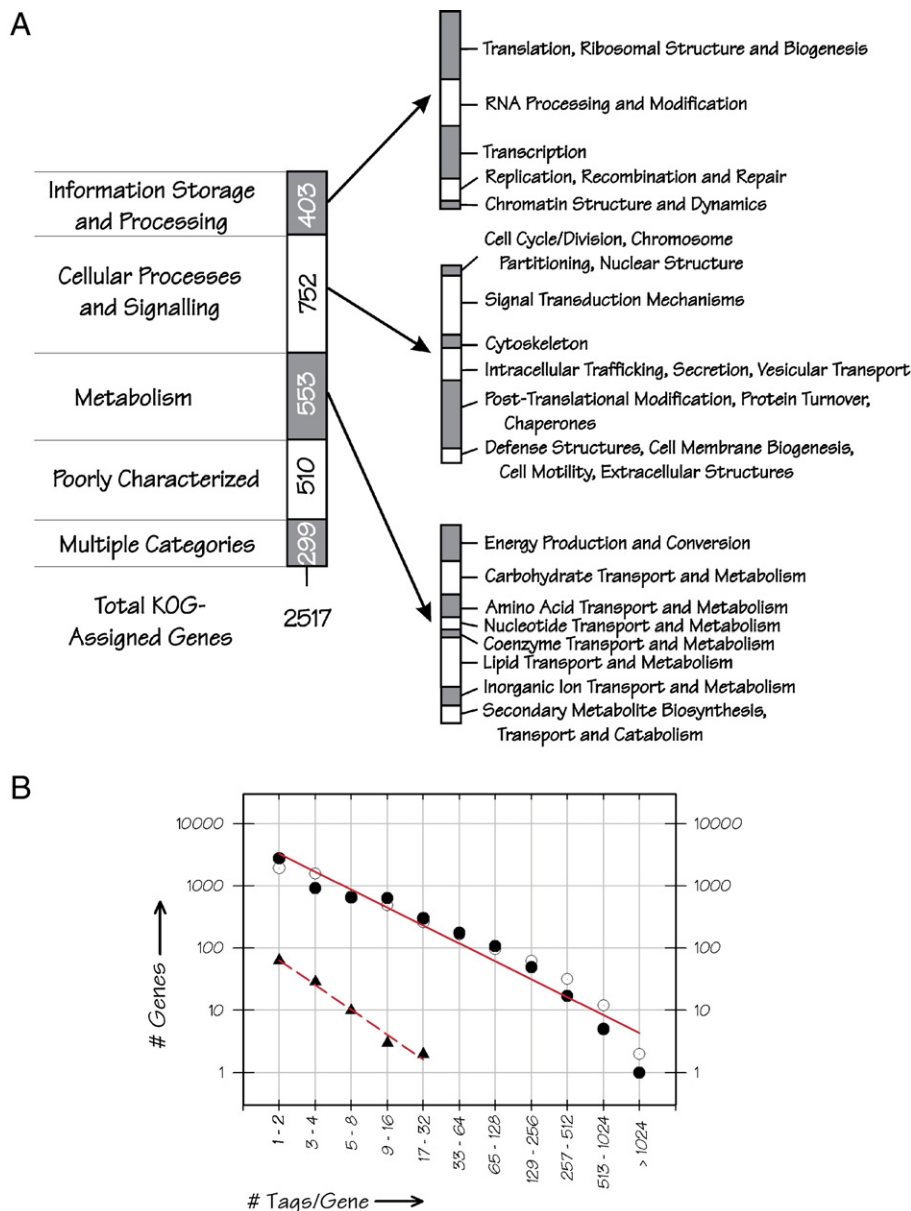


Fig. 1. Distribution and classification of transcripts in the SAGE libraries from adult *C. elegans* intestine and total soma. (A) KOG classification and sub-classification for genes identified in the SAGE library from the adult *C. elegans* intestine. The distribution of KOG classifications for genes identified in the total soma SAGE library is highly similar (data not shown). (B) The distribution of tag counts in both the intestine (open circles) and total soma (closed circles) SAGE library obeys a power law. The closed triangles represent the distribution of tag counts for transcription factors identified in the intestine SAGE library.

intestine library are distributed into KOG categories, as well as into the major KOG subcategories of “Information Storage and Processing”, “Cellular Processes and Signalling” and “Metabolism”. Fig. 1B shows that transcript frequencies in both libraries obey a power law, i.e. the logarithm of the tag number/gene is linearly related to the logarithm of the number of genes with this particular number of tags. Within the overall transcript distribution, individual classes of gene transcripts are also linear on the same plot (Fig. 1B); for example, few of the transcription factor genes identified in the intestine library are transcribed even at modest levels (tag counts in the range of 20–30), whereas a considerably greater number (approaching 100) are transcribed at low levels (tag counts of 1–2).

#### *Identification of a subset of genes expressed exclusively or primarily in the intestine*

To study transcriptional regulation in the intestine, we wish to identify a relatively small number of genes (<100) that are expressed highly and specifically in the adult intestine. Just as the intestine is wholly contained within the worm’s soma, the intestine library should be wholly contained within the somatic library (in the absence of sampling error; see above). However, because both libraries are normalized to 100,000 total tags, transcripts for a particular gene expressed only in the intestine should be present at a lower tag level in the soma library than in the intestine library, i.e. intestine transcripts have been diluted by non-intestinal transcripts. An initial estimate of this dilution

factor would be  $\sim 3$ , since the volume of the intestine is roughly one-third of the total somatic volume of the adult worm (Hirose et al., 2003). A more accurate estimate of this dilution factor is 2.6, based on tag counts of the several genes known to be highly expressed only in the intestine; (a total of 3909 tags identified for *cpr-1* (Britton et al., 1998), *mtl-1* (Moilanen et al., 1999) and all the vitellogenins (Kimble and Sharrock, 1983; Blumenthal et al., 1984) in the normalized intestine library, compared to 1526 total tags for the same genes in the normalized somatic library). This “I/S tag ratio” is the key parameter that will be used to identify intestine-specific genes.

Fig. 2A shows a histogram plotting the distribution of the I/S tag ratio (log scale) for all genes for which the tag counts in the intestine library are  $>9$  (to minimize sampling error). The distribution is clearly bimodal and, as seen in the inset to Fig. 2A, can be approximately decomposed as the sum of two individual (normal) distributions, each containing similar numbers of genes: one distribution has a peak I/S tag ratio  $\sim 1$  and the second distribution has a peak I/S tag ratio  $\sim 2-3$ . We interpret the first distribution as containing genes expressed with no bias toward the intestine and interpret the second distribution as containing genes expressed exclusively or primarily in the intestine. In other words, the bimodal distribution of Fig. 2A implies that roughly half of all genes expressed in the adult intestine produce the large majority of their transcripts in the intestine.

To investigate how tag counts are distributed within the two libraries, Fig. 2B plots the tag count for a particular gene in the intestine library (X-axis) against the tag count for the same gene in the somatic library (Y-axis). The data points on the resulting (log–log) scatter plot represent all 2054 genes that have  $>2$  tags in both (normalized) libraries; the distribution of these data points within the plane reveals how a particular gene is expressed in the intestine and outside of the intestine. For genes that are expressed ubiquitously and uniformly, the data points should distribute around the mid-diagonal (I/S tag ratio = 1); for genes that are expressed only in a somatic tissue other than the intestine, the data points should lie towards the Y-axis; following the argument from the previous paragraph, the data points associated with genes expressed only in the intestine should lie below the diagonal, clustering around a line corresponding to an I/S tag ratio  $\sim 2.6$ . To select a set of genes that are candidates to be expressed strongly and exclusively in the intestine, we consider only genes for which the I/S tag ratio  $\geq 3$  and for which the tag number in the intestine library is  $\geq 50$  (influence of sampling error is considered in the legend to Fig. 2B). One hundred genes meet these two criteria. Twenty of these 100 genes encode ribosomal proteins or are involved in ribosome assembly, suggesting that the intestine persists as the major site of ribosome synthesis in the adult worm. However, because most ribosomal protein genes are unique in the genome and therefore must be widely expressed at other developmental stages, these genes were removed from the list to leave the set of 80 highly-expressed intestine-specific/intestine-enriched (non-ribosomal) genes collected in Table 1. The majority of these genes are also expressed in other stages; for example,  $72/80=90\%$  of the genes can be identified in the L1 SAGE library.

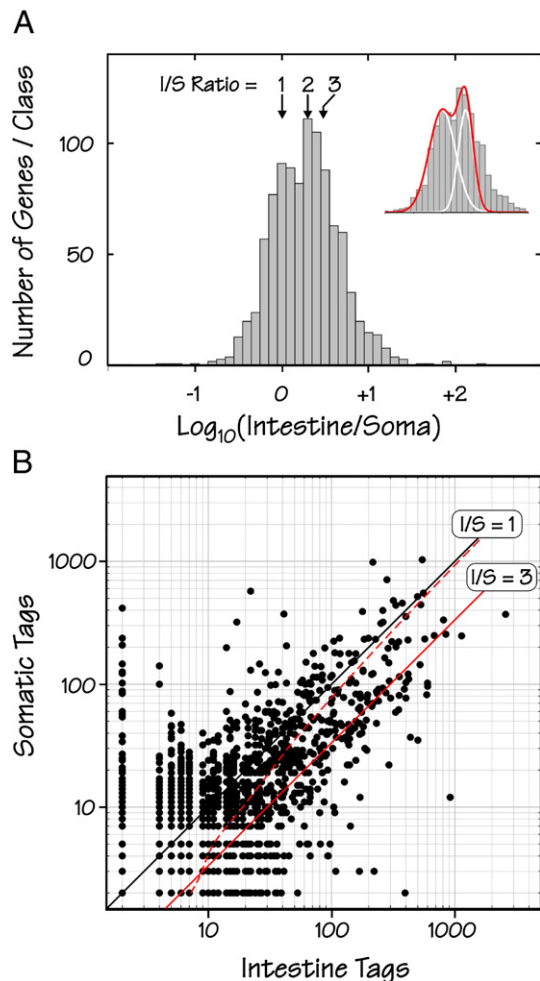


Fig. 2. Distribution of transcript levels between the adult intestine and soma. (A) Histogram showing the distribution of the I/S tag ratio i.e. the ratio of the number of tag counts for a gene in the intestine library to the number of tag counts for the same gene in the somatic library. The data correspond to 1043 genes with tag counts  $\geq 9$  in the intestine library and  $>0$  in the soma library. In the inset, the bimodal distribution was reconstructed (by trial and error) as the sum of two normal distributions (white curves); the qualitative best fit (red curve) was produced by one distribution with a peak I/S ratio  $\sim 1$  and a second distribution with a peak I/S ratio  $\sim 2.6$ . (B) Scatter plot showing the distribution of tag counts for individual genes between the intestine SAGE library (X-axis; logarithmic scale) and somatic SAGE library (Y-axis; logarithmic scale). The data points represent 2054 genes having tag counts  $\geq 2$  in both the intestine and the somatic library; (many of the data points are superimposed and cannot be individually distinguished). As explained in the text, data points for ubiquitously expressed genes should fall along the mid-diagonal (black line; I/S tag ratio  $\sim 1$ ); data points for genes expressed only in the intestine should cluster around the red diagonal (I/S tag ratio  $\sim 3$ ). The dashed red line provides a measure of sampling error, based on the statistical analysis of (Audic and Claverie, 1997); under the null hypothesis that the intestine and the somatic library are identical, for a particular tag count of a gene sampled from the somatic library, there is a 95% probability that the tag count for the same gene sampled from the intestine library will lie to the left of the dashed red line. The set of intestine-specific (or at least highly intestine-enriched) genes were selected to have a I/S tag ratio  $\geq 3$  and a tag count in the intestine library  $\geq 50$ .

Of the 80 genes collected in Table 1, 63 have had their expression pattern characterized in adult worms, either by *in situ* hybridization (Nematode Expression Pattern DataBase: <http://www.nematode.lab.nig.ac.jp/db2/index.php>) or by transgenic

reporters ([http://www.elegans.bcgsc.ca/home/ge\\_consortium.html](http://www.elegans.bcgsc.ca/home/ge_consortium.html), <http://bgypc059.leeds.ac.uk/~web/databaseintro.htm> or individual literature references cited in Table 1). Available expression patterns were classified as follows: Class I=intestine is the only (obvious) site of expression in adult worms; Class II=intestine is the major site of expression (with, say, <10% of total expression intensity detected in other cell types); Class III=intestine is one expression site among others in the adult worm, and; Class IV= gene is not expressed in the adult intestine (i.e. we have selected a false positive). In addition, we assign a reliability estimate to the expression data, ranging from “±” (perhaps, usually due to low signal intensity) to “+++” (certain). Fig. 3 shows eight examples of “Class I (+++)” expression patterns associated with genes in Table 1. There are 28 expression patterns in Table 1 assigned a reliability of “+++”: 23 Class I; 2 Class II; 2 Class III, and 1 Class IV, i.e. 82% (23/28) to 89% (25/28) of the genes are expressed, respectively, only in the intestine or in a pattern that is highly enriched in the intestine. If the gene expression patterns of all reliabilities are considered, the estimated degree of intestinal enrichment is similar (86% are in Class I and II). We interpret this degree of intestinal enrichment as validation of the procedure used to identify intestine-specific/intestine-enriched genes, as well as a validation of the present data set. The data set is not perfect, however, and several false positives (Class IV) can be detected, e.g. T20G5.7 is expressed in pharyngeal gland cells (J. Gaudet, personal communication). Our overall assessment of the reliability of the present SAGE data is described in the Methods section.

#### *Function of highly-expressed intestine-specific/intestine-enriched genes*

The 80 highly-expressed intestine-specific/intestine-enriched (non-ribosomal) genes listed in Table 1 have been ordered and annotated with the aim of understanding their role in intestine function. More than half of these genes appear to participate in early stages of digestion, i.e. the genes often encode simple signal-peptide-containing hydrolytic enzymes, likely to be secreted into the intestinal lumen. Other genes appear to function in detoxification and stress response, while still other genes would appear to be associated with a high level of metabolic activity (both catabolism and anabolism) in the mature intestine. Most (~70%) of the genes collected in Table 1 show a wildtype RNAi phenotype but many of these genes are members of multi-gene families where redundancies are to be expected.

We briefly draw attention to the following points of interest associated with the genes in each functional class.

#### *Bacterial lysis*

Based on the rate of oocyte production (Hirsh et al., 1976), an adult hermaphrodite must ingest and process on the order of a thousand bacteria every minute. Average residence time of a fluorescent latex bead in the *C. elegans* intestine is <2 min (Ghafouri and McGhee (in press) see also Avery and Shtonda, 2003) and hence bacterial lysis and degradation must be both rapid and efficient. Although the first step in bacterial lysis is

undoubtedly the physical damage inflicted by the pharyngeal grinder (Avery and Thomas, 1997), little attention has been paid to the plausible second step, biochemical degradation of the bacterial cell wall. There are no hen egg-white type lysozymes encoded in the *C. elegans* genome but several amoeba-type lysozymes have been suggested to form an innate immunity pathway to protect against pathogens (Mallo et al., 2002). Three of these lysozyme genes (*lys-1*, *lys-2* and *lys-4*) are highly expressed in the intestine and several other lysozymes of the same family are expressed at lower levels. Thus, we propose that the normal *in vivo* function of these genes is more likely to be the constitutive lysis of their customary bacterial food. Table 1 also contains two genes (C45G7.3 and F22A3.6) annotated as “destabilase-like”. Destabilase is a leech peptidase with lysozyme activity (Zavalova et al., 2000), related to the wider class of invertebrate lysozymes (Bachali et al., 2002).

Following degradation of the bacterial cell wall, the plasma membrane must be breached and Table 1 lists two saposin genes, *spp-1* and *spp-8*, that are excellent candidates to perform this function. Saposins are usually associated with breakdown of sphingolipid-containing membranes, acting either as a domain attached to a lipid-modifying enzyme (e.g. a lipase) or as a separate sphingolipid-activating protein (Bruhn, 2005). However, an additional class of saposins are small proteins similar to pore-forming amoebapores that perforate cell membranes (Bruhn, 2005). Both *spp-1* and *spp-8* have been proposed to encode amoebapores (Banyai and Patthy, 1998; Zhai and Saier, 2000), not customary saposins, and indeed, SPP-1 protein has been demonstrated to be bacteriolytic (Banyai and Patthy, 1998).

#### *Luminal degradation of macromolecules*

A striking feature of the intestine-enriched genes collected in Table 1 is the high proportion and high expression levels of peptidases (18 different genes producing ~7% of all tags in the intestine library). All four peptidase classes (Rawlings and Barrett, 1993) are represented but aspartic and cysteine proteases have especially high tag levels. Several of these genes have been previously shown to be intestinal specific (Britton et al., 1998; Tcherepanova et al., 2000). Greater than half of the peptidases listed in Table 1 appear to be secreted and their primary function is thus likely to be luminal hydrolysis of bacterial proteins. Several peptidases also appear to have functions outside of the intestine, e.g. *cpz-1* in moulting (Hashmi et al., 2004) and *asp-3* in necrosis (Syntichaki et al., 2002); however, these are likely to be minor (and probably earlier) expression components compared to the massive expression in the adult intestine.

Table 1 contains five lipase genes that together produce ~1% of all tags in the intestine library. Four of these enzymes are predicted to be secreted, suggesting that their function is to hydrolyse bacterial lipids within the intestinal lumen. Table 1 also identifies *nuc-1* as a highly-expressed intestine-enriched gene whose product appears to be secreted, consistent with the fact that *nuc-1* was originally identified because of failure to degrade bacterial DNA in the intestinal lumen (Sulston, 1976); unpublished results of J.E. Sulston and of P. Babu, cited in

Table 1  
Set of 80 intestine-specific/intestine-enriched genes identified in the adult *C. elegans* intestine

Gene ID	Locus	Description <sup>a</sup>	Signal peptide <sup>b</sup>	RNAi <sup>c</sup>	Expression pattern <sup>d</sup>		
					<i>In situ</i>	GFP	Reliability
<i>(1) Bacterial lysis</i>							
Lysozymes							
F58B3.1	<i>lys-4</i>	Lysozyme (amoeba)	Y	WT	–	–	–
Y22F5A.4	<i>lys-1</i>	Lysozyme (amoeba)	Y	WT	I	I <sup>e</sup>	+++
Y22F5A.5	<i>lys-2</i>	Lysozyme (amoeba)	Y	WT	I	–	+++
C45G7.3	.	Lysozyme (invertebrate type, destabilase)	Y	WT	I	–	+/-
F22A3.6	.	Lysozyme (invertebrate type, destabilase)	Y	Emb	I	–	+++
Saposins, amoebapores							
T07C4.4	<i>spp-1</i>	Amoebapore	Y	WT	I	–	+/-
C28C12.5	<i>spp-8</i>	Amoebapore	Y	WT	I	–	++
<i>(2) Luminal degradation of macromolecules</i>							
Proteinases							
H22K11.1	<i>asp-3</i>	Aspartic peptidase	Y	WT, Ced	I	–	+++
C15C8.3	.	Aspartic peptidase	Y	WT	I?	–	+/-
F21F8.3	<i>asp-5</i>	Aspartic peptidase	Y	WT, Ced	I	I <sup>f</sup>	+++
F28A12.4	.	Aspartic peptidase (~gastricisin)	Y	WT	I	–	++
K10C2.3	.	Aspartic peptidase (~gastricisin)	N	WT	I	–	+++
C52E4.1	<i>cpr-1</i>	Cysteine peptidase	Y	WT	I	I <sup>g</sup>	+++
F44C4.3	<i>cpr-4</i>	Cysteine peptidase	Y	WT	I	–	++
F57F5.1	.	Cysteine peptidase	N	WT, Emb, Lva, Unc	–	–	–
F32B5.8	<i>cpz-1</i>	Cysteine peptidase	Y	Emb	I	III <sup>h</sup>	+++
M04G12.2	<i>cpz-2</i>	Cysteine peptidase	Y	WT	I	–	++
R07E3.1	.	Cysteine peptidase	N	WT	I	–	++
R09F10.1	.	Cysteine peptidase	N	WT	I	–	+/-
C41C4.6	<i>ulp-4</i>	Cysteine peptidase (ubiquitin-like protease)	N	Dpy, Egl, Gro, Pch	I	–	++
K09F5.3	<i>spp-14</i>	Cysteine peptidase (saposin-like domain)	N	WT	I	–	++
F54F11.2	.	Metallopeptidase (neprilysin similarity)	Y	WT	I	–	+++
C10C5.4	.	Metallopeptidase (aminoacylase)	N	WT	I?	–	+/-
R57.1	.	Metallopeptidase (glutamate carboxypeptidase)	Y	WT	I	–	+
K12H4.7	.	Serine peptidase	Y	WT	I	–	+++
Nucleic acids							
C07B5.5	<i>nuc-1</i>	Deoxyribonuclease (DNaseII-like)	Y	WT	I?	–	+/-
Carbohydrate							
ZK1320.2	.	Pectin lyase (?)	Y	WT	–	–	–
C50B6.7	.	Amylase	Y	WT	–	I <sup>f</sup>	+
Lipases							
T21H3.1	.	Lipase	Y	?	I	I <sup>f</sup>	+++
Y49E10.16	.	Lipase	Y	WT	?	–	+/-
T10B5.7	.	Lipase	Y	WT	I	–	+/-
F28H7.3	.	Lipase	Y	WT	I	–	++
Y54F10AM.8	.	Phospholipase	Y	?	I	–	+
Lectins, extracellular binding proteins, etc.							
B0218.8	.	Lectin (C-type)	Y	WT	–	? <sup>b</sup>	+/-
ZK896.7	.	Lectin (C-type)	Y	WT	–	–	–
C14A6.1	.	Lectin (C-type)	Y	WT	–	–	–
F57F4.4	.	Multiple (19) ET domains	Y	Gro, Sck	I	–	+++
T01D3.6	.	EGF-like domains; trypsin inhibitor like domain	Y	WT	–	–	–
C49C3.4	.	Similarity to human intestinal mucin	Y	WT	?	–	+/-
<i>(3) Detoxification and stress response</i>							
Cytochrome P450							
C03G6.15	<i>cyp-35A2</i>	Cytochrome P450	Y	WT	–	I <sup>i</sup>	++
K09D9.2	<i>cyp-35A3</i>	Cytochrome P450	Y	WT	–	–	–
K07C6.4	<i>cyp-35B1</i>	Cytochrome P450	Y	WT	I	–	++
C06B3.3	<i>cyp-35C1</i>	Cytochrome P450	Y	WT	I	–	++
Other							
F28D1.5	<i>thn-2</i>	Thaumatococcus-like (antifungal ?)	Y	WT	–	–	–
F14F4.3	<i>mrp-5</i>	ABC transporter	N	Clr, Gro, Bli, Slu	I	–	++
C30G12.2	.	Alcohol dehydrogenase (short chain type)	N	WT	–	–	–
Y69F12A.2	<i>alh-12</i>	Aldehyde dehydrogenase	N	WT	I	–	++
C07D8.6	.	Aldo/keto reductase	N	WT	?	–	+/-
M88.1	<i>ugt-62</i>	UDP-glucuronosyltransferase	Y	WT	–	–	–

Table 1 (continued)

Gene ID	Locus	Description <sup>a</sup>	Signal peptide <sup>b</sup>	RNAi <sup>c</sup>	Expression pattern <sup>d</sup>		
					<i>In situ</i>	GFP	Reliability
<b>(4) Metabolism</b>							
Energy metabolism, glycolysis, etc.							
Y82E9BR.3	.	F0F1 type ATP Synthase (proteolipid)	N	Emb, Lva	–	–	–
C34E10.6	<i>atp-2</i>	F0F1 type ATP synthase (Beta subunit)	N	Emb, Ste, Lva	–	III <sup>f</sup>	+++
ZK593.1	.	Pyruvate kinase	N	WT	–	–	–
C36A4.9	.	Acetyl-coenzyme A synthetase	N	WT	–	I <sup>f</sup>	+++
T20G5.2	<i>cts-1</i>	Citrate synthase	N	Emb	–	II <sup>f</sup>	+++
Miscellaneous metabolism							
K07H8.6	<i>vit-6</i>	Vitellogenin	Y	?	I	I <sup>f,j</sup>	+++
K11D2.2	<i>asah-1</i>	Acid ceramidase	Y	WT	I	I <sup>f</sup>	++
F41H10.8	<i>elo-6</i>	Polyunsaturated fatty acid elongase	N	WT, Gro	I	I/II <sup>k,1</sup>	+++
M02D8.4	.	Asparagine synthase	N	WT	II	I <sup>f</sup>	++
C34F11.3	.	AMP deaminase	N	WT, Gro, Unc, Lva	I	II <sup>f</sup>	++
R09B5.6	.	3-Hydroxyacyl-CoA dehydrogenase	N	WT	I	–	+++
M03A8.1	<i>dhs-28</i>	17-Beta-hydroxysteroid dehydrogenase	N	Gro, Sck	I	–	+++
T15B7.2	.	Protein tyrosine phosphatase-like	Y	Adl, Gro, Lvl, Lva, Unc	–	–	–
F46E10.1	.	Acyl coA synthetase, long chain	N	Emb, Sck, Ste	I	I <sup>f</sup>	+++
F41H10.7	<i>elo-5</i>	Polyunsaturated fatty acid elongase	N	Gro, Sck	?	I/II <sup>k</sup>	+++
C08H9.2	.	Vigilin (lipoprotein and/or RNA binding protein)	N	Bmd, Dpy, Gro, Pvl, Slu, Unc	?	–	+/-
<b>(5) Unknown functions</b>							
F53A9.8	.	~ <i>Plasmodium lophurae</i> His-rich glycoprotein	N	WT	I	I <sup>f</sup>	+++
T20G5.7	<i>dod-6</i>	~Sea anemone toxin; ~metallopeptidase (?)	Y	WT	IV	–	+++
Y119D3B.21	.	~ <i>Plasmodium falciparum</i> hypothetical protein	N	Sck	I	–	++
R06C1.4	.	RNA recognition motif	N	WT	–	–	–
H06I04.4	<i>ubl-1</i>	Ubiquitin/40S ribosomal protein S27a fusion	N	Ste, Lon, Sck, Lva	I	–	++
C39B10.3	.	~ <i>Staphylococcus aureus</i> hypothetical protein	N	WT	–	–	–
D2096.8	.	~Nucleosome assembly protein	N	Emb, Gro, Pvl, Rup, Stp, Unc	?	–	+/-
R02E12.6	.	~ <i>Brachydanio rerio</i> hypothetical protein	N	?	?	–	+/-
F15E11.12	.	~ <i>Wolinella succinogenes</i> hypothetical protein	N	WT	–	–	–
K03A1.2	.	Multiple (9) leucine-rich repeats	Y	WT	I	–	+++
F58G1.4	.	Endoplasmic reticulum targeting sequence	Y	WT, Clr, Gro, Sck	I	I <sup>f</sup>	+++
C03B1.12	<i>lmp-1</i>	Lysosome-associated membrane protein	Y	Clr, Gro, Bli, Slu	I	I <sup>m</sup>	+++
F42A10.6	.	~ <i>Macaca radiata</i> NADH dehydrogenase	Y	WT	–	I <sup>f</sup>	+++
Y39B6A.1	.	~ <i>Plasmodium lophurae</i> His-rich glycoprotein	N	WT	I	–	+++
C05D2.8	.	~ <i>Drosophila melanogaster</i> protein	N	WT	I?	–	+

<sup>a</sup> Brief description of possible gene functions, emphasizing plausible roles in digestion; tag counts in intestine and somatic libraries (SWAG1 and SWAG2 respectively) are available at <http://elegans.bcgs.cca/home/sage.html>. The designation “~” is used to signify “shows sequence similarity to”.

<sup>b</sup> The presence of a signal peptide (from Wormbase annotations) is consistent with protein secretion into the intestinal lumen; in each case, however, other locations (e.g. endoplasmic reticulum, membrane surface, etc.) must be assessed based on other data (not always available).

<sup>c</sup> RNAi phenotypes collected from Wormbase: Adl=Adult lethal; Bli=Blistered; Bmd=Body morphology defect; Ced=Cell death abnormality; Clr=Clear; Dpy=Dumpy; Egl=Egg laying defective; Emb=Abnormal embryogenesis; Gro=Abnormal growth rate; Lva=Larval arrest; Lvl=Larval lethal; Pch=Pachytene checkpoint; Pvl=Protruding vulva; Sck=Sick; Slu=Sluggish; Ste=Sterile; Unc=Uncoordinated; WT=Wild type.

<sup>d</sup> Expression patterns based on NextDB compilation of *in situ* hybridization patterns (left column) or on GFP-based transgenic reporters (middle column); references to GFP reporters are provided as footnotes. The assignment of four different classes of expression patterns (and the reliabilities of these assignments) is explained in the text. In cases of discrepancies, *in situ* hybridizations are given priority. Indeterminate expression patterns, usually because of weak signals, are designated as “?”.

<sup>e</sup> (Mallo et al., 2002).

<sup>f</sup> [http://elegans.bcgs.cca/home/ge\\_consortium.html](http://elegans.bcgs.cca/home/ge_consortium.html).

<sup>g</sup> (Britton et al., 1998).

<sup>h</sup> (Hashmi et al., 2002).

<sup>i</sup> (Menzel et al., 2001).

<sup>j</sup> (Kimble and Sharrock, 1983).

<sup>k</sup> (Kniazeva et al., 2004).

<sup>l</sup> (Pauli et al., 2006).

<sup>m</sup> (Kostich et al., 2000).

Hevelone and Hartman (1988). *nuc-1* also plays a role in clearing apoptotic corpses (Wu et al., 2000).

#### Detoxification and stress response

Considering the worm lifestyle, the four cytochrome P450 genes listed in Table 1 (cyp-35 A2, A3, B1 and C1) probably

function to detoxify ingested xenobiotics, although synthetic roles cannot be excluded. Cytochrome 35A2 has previously been shown to be strongly induced in the intestine in response to naphthoflavone (Menzel et al., 2001). Table 1 also lists a number of genes coding for phase II detoxification enzymes, such as UDP-glucuronosyltransferase (Bock, 2003; Gregory



Fig. 3. Expression patterns in adult *C. elegans* of eight genes identified in Table 1. Left column represents *in situ* hybridization to endogenous transcripts; right column represents fluorescence produced by transgenic GFP-reporters. Gene identifiers are shown for each panel (as are gene loci if available). Briefly, the genes encode the following types of enzymes: F21F8.3=*asp-5* aspartic protease; K12H4.7=serine peptidase; F54F11.2=neprilysin-like metalloproteinase; T21H3.1=lipase; C36A4.9=acetyl coA synthetase; F41H10.7=*elo-5* polyunsaturated fatty acid elongase; F46E10.1= long chain acyl coA synthetase, and; F42A10.6= similarity to NADH dehydrogenase. The *in situ* images are taken from the Nematode Expression Pattern DataBase: (<http://nematode.lab.nig.ac.jp/db2/index.php>). The *elo-5*::GFP strain is described in Kniazeva et al. (2004); production of the other GFP-expressing transgenic strains is described in [http://elegans.bcgsc.ca/home/ge\\_consortium.html](http://elegans.bcgsc.ca/home/ge_consortium.html). As explained in the text, expression patterns for these genes are classed as “I+++”, i.e. reliable and intestine-specific.

et al., 2004), alcohol dehydrogenase, aldehyde reductase and aldo/keto reductase (Sladek, 2003; Vasiliou et al., 2004; Srivastava et al., 2005) as well as an ABC type transporter (*mnp-5*) that could potentially export conjugated xenobiotics back to the intestinal lumen (Homolya et al., 2003).

### Metabolism

Table 1 lists seven enzymes involved in later stages of glycolysis (pyruvate kinase), the citric acid cycle (citrate synthase), early stages of fatty acid oxidation (3-OH-CoA dehydrogenase and long chain acyl-CoA synthetase) and the mitochondrial respiratory chain (two subunits of ATP synthase), suggesting that the adult worm intestine is likely to be a major

site of energy production. As with the ribosomal protein genes discussed earlier, genes encoding general metabolic enzymes are likely to be expressed in other tissues at other developmental stages, even though they appear enriched in the adult intestine. Several additional genes listed in Table 1 are involved in (intracellular) lipid metabolism. For example, the *elo-5* and *elo-6* genes are involved in elongation of polyunsaturated fatty acids (Kniazeva et al., 2004) and acid ceramidase is presumably involved in degradation of sphingolipids.

### “Other”

Table 1 lists 10 or so genes for which the sole annotation is similarity to an unknown gene in another species. We have little idea how these genes might function in intestinal development or physiology but they are highly conserved in the related nematode *C. briggsae* (median BLAST Probability  $\sim e^{-84}$ ).

As a final comment in this section, we could find no evidence that the genes listed in Table 1 are expressed at significantly lower levels in males compared to hermaphrodites (Jiang et al., 2001; Reinke et al., 2004). This is curious: males are more physically active than hermaphrodites but do not need the hermaphrodite’s prodigious capacity to convert bacteria into oocytes. One could have imagined that males use the sex determination pathway to repress expression of (high levels) of digestive enzymes in their intestines, just as they repress intestinal expression of vitellogenins (Shen and Hodgkin, 1988; Yi and Zarkower, 1999; Yi et al., 2000). In one particular case, we were able to verify similar male-hermaphrodite expression levels, using Western blots probed with an anti-ASP-1 antibody (Tcherepanova et al., 2000); data not shown); the results of this experiment also argue against significant post-transcriptional regulation of the *asp-1* gene.

### Computational analysis of intestinal promoters

As noted in the Introduction, experimental analysis of *C. elegans* intestine-specific promoters has in all cases identified a *cis*-acting GATA-type sequence that is critical for correct gene expression. Table 2 collects all such sequences of which we are aware, together with the summarizing “Sequence Logo” (Schneider and Stephens, 1990) in Fig. 4A. In the current section, we wish to analyze the promoters of the highly-expressed intestine-specific/intestine-enriched genes collected in Table 1 to determine: (1) whether this presumptive GATA-site-dependence extends to other (possibly all) intestine promoters, and; (2) whether non-GATA-related sequences can also be identified as over-represented in intestinal promoters.

“Promoters” (i.e. 5′-flanking regions) were compiled for 74 of the intestine-specific/intestine-enriched genes listed in Table 1 (omitting five genes associated with universal aspects of metabolism, as well as the single gene that is likely part of an operon). Each promoter corresponded to 1500 bps upstream of the ATG translation initiation codon or up to the closest 5′-gene, whichever distance was shorter; repeats were masked. Motif discovery was performed on the entire set of sequences with two independent methods: the word-counting oligo-analysis algo-

Table 2  
Experimentally-determined *cis*-acting sequences important for the expression of intestine-specific genes in *C. elegans*

Sequence <sup>a</sup>	Gene	Position <sup>b</sup>	Ref.
TTCTGATAAGGG	<i>vit-2</i> vitellogenin	–159	<sup>c</sup>
CATTGATAAGCT	<i>vit-2</i> vitellogenin	–105	<sup>c</sup>
AACTGATAGCAA	<i>ges-1</i> carboxylesterase	–1135	<sup>d</sup>
AACTGATAAGGG	<i>ges-1</i> carboxylesterase	–1123	<sup>d</sup>
TACTGATAAGAA	<i>cpr-1</i> cysteine protease	–175	<sup>e</sup>
GATTGATAAGAC	<i>cpr-1</i> cysteine protease	–79	<sup>e</sup>
AACTGATAAAAT	<i>mtl-1</i> metallothionein	–319	<sup>f</sup>
AACTGATAAAGG	<i>mtl-2</i> metallothionein	–305	<sup>f</sup>
AGCTGATAACAG	<i>mtl-2</i> metallothionein	–90	<sup>f</sup>
TGATGATAAAGT	<i>gcs-1</i> glutamyl-cysteine synthetase	–116	<sup>g</sup>
TGTTGATAAGAT	<i>S</i> -adenosylmethionine decarboxylase	–874	<sup>h</sup>
CACTGATAACGA	<i>S</i> -adenosylmethionine decarboxylase	–860	<sup>h</sup>
GGTAGATAGAAC	<i>S</i> -adenosylmethionine decarboxylase	–795	<sup>h</sup>
AGGTGATAAGAT	Spermidine synthase	–133	<sup>h</sup>
TAGTGATAATGG	Spermidine synthase	–119	<sup>h</sup>
CAGTGATAATAG	Spermidine synthase	–110	<sup>h</sup>
AGTTGATAGTGA	Spermidine synthase	–97	<sup>h</sup>
TTGTGATAATGA	<i>spl-1</i> sphingosine-1-phosphate lyase	–320	<sup>i</sup>
AACTGATAAAAG	<i>pho-1</i> acid phosphatase	–122	<sup>j</sup>

<sup>a</sup> A 12 bp sequence is shown centred on each GATA site. Each of these sites has been experimentally mutated in the various promoters and the effects range from significant diminution to complete abolition of gene expression; the original papers should be consulted for details. The “Sequence Logo” representing the information at each sequence position is shown in Fig. 4A.

<sup>b</sup> The coordinates of the mutated GATA sites were recalculated as the number of base pairs between the “G” of the GATA site (or the “C” of the TATC site) and the A residue of the translation initiation codon; in several cases, this changes the description significantly from that reported in the original paper. We note that Pauli et al. (2006) mutated TGATAA sites in three genes (*elo-6*, *gst-42* and D2030.5) and reported significant lowering of reporter gene expression; however, the coordinates of the mutated sites were not reported.

<sup>c</sup> (MacMorris et al., 1992; MacMorris et al., 1994).

<sup>d</sup> (Egan et al., 1995).

<sup>e</sup> (Britton et al., 1998).

<sup>f</sup> (Moilanen et al., 1999).

<sup>g</sup> (An and Blackwell, 2003); note that these authors mutated this site in order to prevent binding of the SKN-1 factor; however, the SKN-1 binding site overlaps with the GATA sequence reported in the table.

<sup>h</sup> (Luersen et al., 2004).

<sup>i</sup> (Oskouian et al., 2005); note that an adjacent ACTGATAAGA at –293 was not investigated.

<sup>j</sup> (Fukushige et al., 2005).

rithm of the RSAT toolset (Van Helden et al., 1998; hereafter referred to simply as RSAT), and the Gibbs-sampling-based program MotifSampler (Thijs et al., 2001).

When presented with the promoters of the 74 intestine-specific/intestine-enriched genes, both RSAT and MotifSampler consistently return the highly similar extended GATA motif shown as Sequence Logos in Figs. 4B and C, respectively. Three important observations can immediately be made. First, at least one such motif was detected in 100% (74/74) of the intestine-specific/intestine-enriched promoters. Secondly, the computationally-identified motifs look highly similar to the experimental motif summarized in Fig. 4A. Thirdly, the extended GATA motif is essentially the only significantly over-represented site: whereas MotifSampler returned 136 instances of an extended GATA site in the 74 promoters, the next strongest signal was a weak AGAGA-like motif with only 6 instances in 5 genes. As

controls, the same two independent analyses were performed on three sets of 74 randomly selected genes. RSAT returned no over-represented motif with the control promoters and, compared to the results from the intestinal promoters, the control MotifSampler results were poorer quality in every way: each returned motif had fewer instances, lower scores and was found on fewer genes; for example, GATA-like motifs were identified in only ~3–5% of the control promoters.

As an independent verification of the above analyses, the same procedures were applied to the promoters of the 57 *C. briggsae* orthologs and 39 *C. remanei* orthologs of the genes in Table 1. As for *C. elegans*, the two algorithms identified a dominant TGATAA motif, as well as much weaker secondary motifs (data not shown). Notably, the secondary motifs differed between the three species.

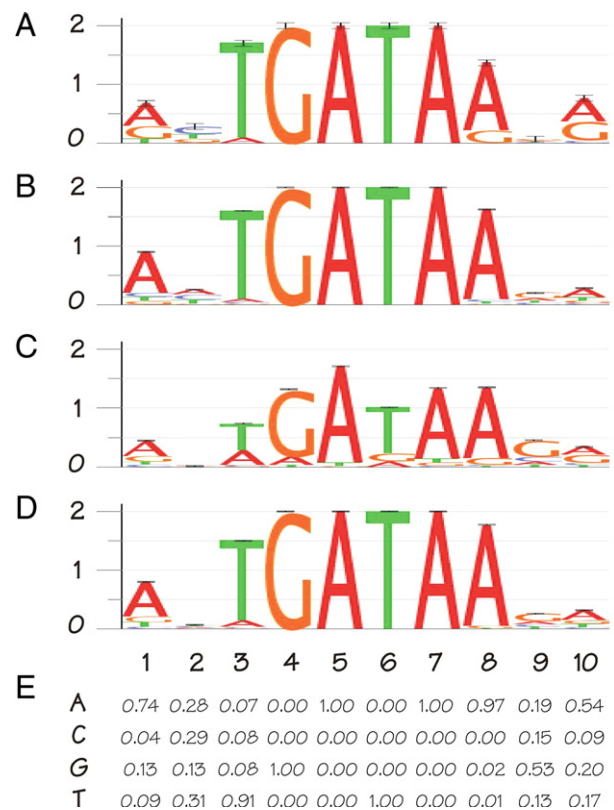


Fig. 4. *cis*-regulatory sites in the promoters of intestinal genes. Sequence logos summarizing: (A) the experimentally important *cis*-acting sequences collected in Table 2; (B) all motifs discovered by RSAT oligo-analysis (width 8 bp) aligned with ClustalW and adjusted to the width 10 bp; (C) all motifs discovered by MotifSampler algorithm (widths 6, 8, 10 and 12 bp), aligned with ClustalW and adjusted to the width 10 bp; only motifs identified in at least 7 of 100 Gibbs sampling iterations were included in the analysis, and; (D) our best estimate of the most important *cis*-acting sequence regulating *C. elegans* intestine-specific/intestine-enriched genes, produced by a combination of experimental sites (Table 2) with the most significant OPTICS-based cluster of 111 motifs from 59 promoters determined by RSAT and MotifSampler. Units for the Y-axis are information bits. (E) Position Frequency Matrix (PFM) describing our current best estimate of the GATA-like sequence regulating *C. elegans* intestine-specific/intestine-enriched genes, produced from the combined experimental and computational sequence motifs (shown as a sequence logo in D). For convenience in the text, we refer to the motif “consensus” as AHTGATAARR, where H=A or C or T and R=A or T. Each consensus designation corresponds to >70% of all PFM entries for that particular position.

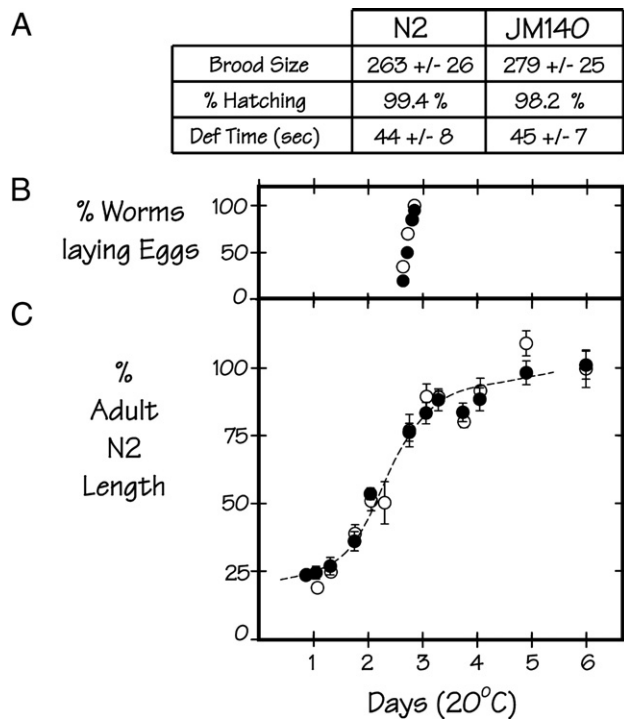


Fig. 5. Phenotype of the *C. elegans* strain in which both the *elt-4* and *elt-7* genes have been deleted: JM140 (*elt-7(tm840); elt-4(ca16)*). (A) Comparison of brood size, % embryo hatching and defecation interval (20°C) between N2 control worms and strain JM140 (*elt-7(tm840); elt-4(ca16)*). Errors represent standard deviations. Complete brood sizes measured on 10 adults; % hatching measured on >400 embryos; defecation interval measured as either pBoc-to-pBoc or Exp-to-Exp and the >150 individual measurements pooled. (B) Comparison of egg-to-egg interval (20°C) for N2 worms (open symbols) and JM140 (*elt-7(tm840); elt-4(ca16)*) worms (closed symbols). For both strains, laid eggs were collected over a 2-h period, individual F1 progeny distributed to separate plates and inspected at intervals for the appearance of the first egg on the plate. Data points represent the cumulative fraction of the egg-laying progeny. (C) Comparison of overall growth rate (20°C) for N2 worms (open symbols) and JM140 (*elt-7(tm840); elt-4(ca16)*) worms (closed symbols). Error bars represent standard deviations from length measurements on 10±2 worms for each time point. Length measurements are normalized to length of mature N2 adults.

#### *ELT-2 is the only (post-specification) GATA factor necessary for normal intestinal development and function*

There are seven GATA-type transcription factors expressed in the *C. elegans* endoderm (reviewed in McGhee in press). Four of these factors (MED-1, MED-2, END-1 and END-3) are expressed early and transiently during the specification phase of the endoderm, and transcripts can no longer be detected past the ~4E–8E cell stage (Zhu et al., 1997; Maduro et al., 2001; Robertson et al., 2004; Baugh et al., 2005). This leaves the three GATA factors ELT-2, ELT-4 and ELT-7 (C18G1.2) expressed in the intestine during the remaining 95% of the worm lifespan. Tag counts in the intestine SAGE library for *elt-2*, *elt-4* and *elt-7* are 25, 4 and 1 respectively. (Contrary to the assertion of Pauli et al., 2006, the GATA factor ELT-3 is not expressed in the intestine but in the hypodermis, with a minor component in the pharyngeal–intestinal and rectal–intestinal valve cells (Gilleard et al., 1999; Gilleard and McGhee, 2001); *elt-3* has 4 tags in the somatic library but none in the intestine library.)

Of the three post-specification intestinally-expressed GATA factors, ELT-2 is by far the best candidate to be the major regulator of intestinal transcription (discussed in more detail below). To test the importance of ELT-2 in a definitive manner, the recently available *elt-7(tm840)* deletion (which removes the zinc-finger DNA-binding domain and is thus likely to be a null) was combined with the *elt-4(ca16)* mutation (also a null; Fukushige et al., 2003) to produce strain JM140 (*elt-7(tm840); elt-4(ca16)*), which appears essentially wildtype. The results are shown in Fig. 5A (for brood sizes, hatching rates and defecation intervals), in Fig. 5B for egg-to-egg interval and in Fig. 5C for the overall growth curve. We conclude that ELT-4 and ELT-7 together are largely dispensable and that (following endoderm specification) ELT-2 is sufficient for all obvious GATA-factor-related functions of both the developing and the mature intestine.

#### *Non-GATA-type transcription factors in the adult intestine*

A strong advantage of SAGE is the ability to identify genes expressed at low levels, such as transcription factors (see Fig. 1B above). Thus 108 different transcription factors (tag counts >1) were identified in the adult intestine SAGE library. The most highly expressed of these factors (tag counts >5) are collected in Table 3. ELT-2 has the highest tag number (25) of any recognizable transcription factor in the intestine library. Of the 21 transcription factors listed in Table 3, 15 are either nuclear hormone receptors or have significant sequence similarity to nuclear hormone receptors (for example, to a ligand binding domain). Perhaps the least expected entry in Table 3 is UNC-62, represented by 24 tags in the intestine library, one fewer tag than ELT-2. The *unc-62* gene encodes a homeobox protein most similar to *Drosophila* homothorax; *unc-62* is expressed widely in the embryo and *unc-62* loss-of-function mutants die during embryogenesis (Van Auken et al., 2002). (Interestingly, *ceh-20*, the *C. elegans* homolog of extradenticles (Liu and Fire, 2000) is also present in the adult intestine library (4 tags).) The forkhead factor PHA-4 is expressed at an intermediate level in the adult intestine (6 tags). PHA-4 has been well studied as a factor critical for embryonic formation of the pharynx and the rectum (Mango et al., 1994; Azzaria et al., 1996; Horner et al., 1998; Kalb et al., 1998; Gaudet and Mango, 2002; Ao et al., 2004; Gaudet et al., 2004). However, the *pha-4* gene is known also to be expressed in the intestines of both embryos and adults, and *pha-4* loss of function mutants show a mild phenotype in the embryonic intestine (Azzaria et al., 1996; Kalb et al., 1998). It is not known whether either *unc-62* or *pha-4* mutants would show a phenotype in the adult intestine.

Do any of the transcription factors listed in Table 3 have an observable function? All RNAi-induced loss-of-function phenotypes available from the literature (Gonczy et al., 2000; Ashrafi et al., 2003; Kamath et al., 2003; Sonnichsen et al., 2005) are listed as wildtype (with the exception of *elt-2*, *unc-62*, *pha-4* and the two nuclear hormone receptor genes *nhr-68* and *nhr-8*, for which RNAi induces a modest alteration in lipid content of the worm; Ashrafi et al., 2003). Large scale

Table 3  
Recognized transcription factors identified in the SAGE library of the adult *C. elegans* intestine

Gene ID	Gene	Protein <sup>a</sup>	Tag counts <sup>b</sup>	RNAi phenotype <sup>c</sup>	% Hatch <sup>d</sup>	% Length (100 h) <sup>e</sup>	Brood size <sup>f</sup>
C33D3.1	<i>elt-2</i>	GATA factor	25	Lva	99	21	0
T28F12.2	<i>unc-62</i>	Homeobox	24	Emb			
H12C20.3	<i>nhr-68</i>	Nuclear hormone receptor	17	WT <sup>g</sup>		116	
F26D12.1	<i>fkf-7</i>	Forkhead/winged helix	12	WT	99	109	75–100%
K08H2.8	<i>nhr-32</i>	Nuclear hormone receptor	11	WT			
C28D4.9	<i>nhr-138</i>	Nuclear hormone receptor	10	WT			
T01B10.4	<i>nhr-14</i>	Nuclear hormone receptor	10	WT	97	95	75–100%
C30G4.7	.		9	WT			
F33D4.1	<i>nhr-8</i>	Nuclear hormone receptor	9	WT <sup>g</sup>	98	101	75–100%
F58G6.5	<i>nhr-34</i>	Nuclear hormone receptor	7	WT			
T24H10.7	.	bZIP	7	WT			
C56E10.1	.	Steroid zinc finger	6	WT	97	97	75–100%
F38A6.1	<i>pha-4</i>	Forkhead/winged helix	6	Emb			
F44C8.4	<i>nhr-103</i>	Nuclear hormone receptor	6	WT	97	103	75–100%
Y41D4B.20	.	NHR ligand binding domain	6	WT	97	93	75–100%
B0280.8	<i>nhr-10</i>	Nuclear hormone receptor	5	WT	99	95	
C05G6.1	<i>nhr-76</i>	Nuclear hormone receptor	5	WT	98	93	
C33G8.12	.	Nuclear hormone receptor	5	WT	98	98	
F09C6.9	<i>nhr-116</i>	Nuclear hormone receptor	5	WT			75–100%
F16B4.12	<i>nhr-117</i>	Nuclear hormone receptor	5	WT			
R02C2.4	.	Nuclear hormone receptor	5	WT	99	103	75–100%

<sup>a</sup> Class of transcription factor.

<sup>b</sup> Number of SAGE tags identified in the intestine library (normalized to a total tag count of 100,000).

<sup>c</sup> RNAi phenotypes are collected from Wormbase and represent a number of independent genome wide screens. Only the most severe RNAi phenotype is listed (e.g. Embryonic lethal for *pha-4*).

<sup>d</sup> Adult hermaphrodites were injected with double stranded RNA (~1 mg/ml) and % hatching of F1 embryos was measured.

<sup>e</sup> Adult hermaphrodites were injected with double stranded RNA (~1 mg/ml), F1 embryos were collected over a 2-h time period and larval length was measured 100 h later (20°C), normalized to the length of N2 control worms.

<sup>f</sup> L4 worms were transferred to RNAi feeding plates and re-transferred periodically over the next few days (20°C). The brood size is listed only as 75–100% of the control N2 worms; thick growth of the various bacterial strains made accurate egg counts difficult but overall, we could detect no significant differences from the N2 control.

<sup>g</sup> RNAi is reported to cause a moderate increase in fat content of the worm (Ashrafi et al., 2003).

RNAi screens by necessity concentrate on major qualitative phenotypes and may not have detected a partial loss of intestinal function. Thus, for the subset of genes in Table 3 for which clones are present in the Kamath et al. (2003) library, we measured three parameters quantitatively: (1) % hatching; (2) body length at 100 h (20°C) after laying and; (3) brood size. Overall, no significant quantitative phenotypes were observed (Table 3).

Of the remaining 87 transcription factors expressed at lower levels (tag counts ≤ 4), 11 show RNAi phenotypes. However, the reported phenotypes are generally weak, variable and, with the possible exception of *nhr-80* (Miyabayashi et al., 1999), not intestine specific.

## Discussion

Understanding how the adult *C. elegans* intestine functions must, at some level, reduce to understanding the properties of the >4000 intestinally-expressed genes identified in the present paper. As shown above in Fig. 1B, transcript levels within the intestine (as well as within the total worm soma) follow a power-law or scale-free distribution, encapsulating the observations that both tissues and whole organisms contain few genes with high transcript levels (only one or two genes have tag counts >1000), many genes with low transcript levels (thousands of

genes with 1–2 tags), but no typical gene. However, it is doubtful if any theoretical meaning can be attached to this behaviour; Keller (2005) has clearly pointed out that many different rules and network architectures can produce such distributions.

Among the several thousand intestinally-expressed genes, we identified a select subset of 80 highly-expressed intestine-specific (or at least highly intestine-enriched) genes (Table 1), whose properties begin to elucidate intestinal functions. For example, the majority of the 80 genes encode proteins that are good candidates to be secreted into the intestinal lumen and to function in digestion. One striking feature of the gene list is the high proportion (and high expression levels) of a wide variety of proteases, a feature held in common with blood-eating parasitic nematodes (Jasmer et al., 2001, 2004).

### Comparison to previous analyses of intestinal transcription

Pauli et al. (2006) have used a clever affinity-tagging protocol (Roy et al., 2002) to isolate a fraction of mRNA enriched for transcripts from the L4 stage intestine; the contained sequences were then identified using spotted microarrays. A list of 1938 intestinally-expressed genes were selected based on an arbitrary level of significance for enrichment; 53% (1020/1938) of these genes are also identified in the intestine SAGE library, including 45% (36/80) of the highly expressed

intestine-specific/intestine-enriched genes listed in Table 1. Pauli et al. (2006) removed all genes that are also expressed in L1 muscle and/or L4/adult germline and this slightly increases agreement with our results: now 58% (361/624) of their intestine enriched genes can be identified in our SAGE intestine library. A number of reasons could be suggested for the incomplete overlap of the two data sets: for example, the different stages (L4 vs. mature adults in the present study), the fact that our worms were mildly starved prior to dissection, and the inevitable under-representation in our data set of genes expressed in the intestine anterior and posterior. Overall, however, we expect that most of the differences simply reflect the vastly different technologies used in the two studies. Among the intestine-enriched genes selected by Pauli et al. (2006), only 109 have tag counts >9 in the intestine and also appear in the somatic library, allowing them to be plotted (with low sampling error) on a scatter plot such as Fig. 2B; a simple sign test shows that the distribution of data points corresponding to the Pauli et al. (2006) intestine-enriched genes are not significantly different from the distribution of our starting un-enriched SAGE data.

Kim et al. (2001) have analyzed the combined data from hundreds of microarray experiments, using clustering algorithms to sort genes into “mountains”, which have then been interpreted to reflect tissue-specificity or some particular biochemical feature of the associated genes. Genes in mountain #8 have been proposed to be intestine specific. Of the 80 non-ribosomal intestine-specific/intestine-enriched genes collected in Table 1, 25 (31%) have been assigned to mountain #8. Thus, the clustering algorithms clearly extracts some significant “intestine” signal from the microarray data; (~5% would be expected by chance). However, overall, this approach would seem to be an imperfect predictor of intestinal expression.

#### *Are all intestine-specific/intestine-enriched genes controlled by cis-acting GATA-like sequences?*

Within the precision allowed by the current high-throughput data, we distinguish two broad categories of genes expressed in the *C. elegans* intestine, corresponding to the two peaks observed in Fig. 2A above. The first category (corresponding to the peak for which the I/S tag ratio ~ 1) contains genes expressed in many, perhaps all, tissues in the worm; for lack of more detailed knowledge, such genes are often called “housekeeping” but we will also refer to them as widely-expressed. The second category (corresponding to the peak for which the I/S ratio = 2 to 3), contains genes we refer to as intestine-specific/intestine-enriched. It is possible that many of the genes in this second category could also be expressed outside of the intestine. However, the intestine is so massive compared to non-intestinal cells where such a gene might be expressed (e.g. a small number of neurons) that the majority of the gene’s transcripts should still derive from the intestine (and this is presumably the reason that the secondary peak in Fig. 2A is distinct).

The intestinal genes for which important (frequently critical) cis-acting GATA sites have been found experimentally (Table 2) are intestine-specific/intestine-enriched, not housekeeping/widely-expressed. Although these 10 genes represent a small

sample from the many hundreds of intestine-specific/intestine-enriched genes predicted from the SAGE data, the tentative conclusion must be that, if any other gene expressed only or mainly in the intestine is investigated experimentally, it too is likely to be controlled by a cis-acting GATA-like sequence. The computational analyses of the 74 intestine-specific/intestine-enriched promoters performed in the present paper fully support this conclusion: two different algorithms detected an extended GATA-like sequence (approximate consensus = AHTGATAARR) in 100% of the promoters of this gene set, compared to <5% for control sets of genes chosen randomly from the genome. (Pauli et al., 2006 found a TGATAA site in ~53% of the promoters from their gene set, consistent with the lower degree of intestinal enrichment discussed above.)

The important question remains: what fraction of all intestine-expressed genes (either intestine-specific/intestine-enriched or housekeeping/widely-expressed) contain a critical GATA-related sequence in their promoter? To approach this question quantitatively, the experimentally-identified (Table 2; Fig. 4A) and computationally-identified (Figs. 4B,C) GATA-like motifs were combined into an overall best estimate, shown in Fig. 4D as the Sequence Logo (Schneider and Stephens, 1990) and in Fig. 4E as the normalized position frequency matrix (PFM). Before searching various sets of promoters for high-scoring matches to the PFM, it is instructive to consider the range of scores that is likely to be biologically relevant. The maximum score possible for a match to the PFM of Fig. 4E is 0.80, corresponding to the sequence ATTGATAAGA. For the tandem pair of GATA sites controlling expression of the *ges-1* gene (Egan et al., 1995), the downstream site (ACTGATAAGG) scores 0.77 and the upstream site (ACTGATAGCA) scores 0.67. Experimentally, the upstream site has a significantly weaker influence on *ges-1* expression than does the downstream site (Egan et al., 1995). Likewise, the critical site in the promoter of the *pho-1* gene (ACTGATAAAA) scores 0.76 but a single residue alteration that lowers the score to 0.66 completely destroys transcriptional activity (Fukushige et al., 2005). Furthermore, the *pho-1* promoter has two additional GATA sites, both with scores of 0.66, but mutation of either of these sites has no obvious effect on *pho-1* expression (Fukushige et al., 2005). Thus, we suggest that the lower limit of biologically relevant scores is likely to be ~0.65 to 0.70.

We then searched various sets of promoters for matches to the PFM, recording the highest scoring site in each promoter. The results are shown in Fig. 6 as a cumulative distribution function, i.e. the highest scoring match to the PFM in a particular promoter (*X*-axis) is plotted against the fraction of all promoters (in this particular set) that have site scores up to and including this particular score (*Y*-axis). We consider four sets of promoters corresponding to genes that are increasingly intestine-specific and shown by the coloured lines in Fig. 6: (1) all promoters in the genome (black line); (2) all promoters from genes identified in the intestine SAGE library (tag counts >2; magenta line); (3) all promoters from genes in the secondary peak of Fig. 2A (i.e. tag counts in the intestine library >9 and I/S ratio >2; blue line), and; (4) promoters from the set of 74 intestine-specific/intestine-enriched genes of Table 1 (red line). Clearly, as the intestinal

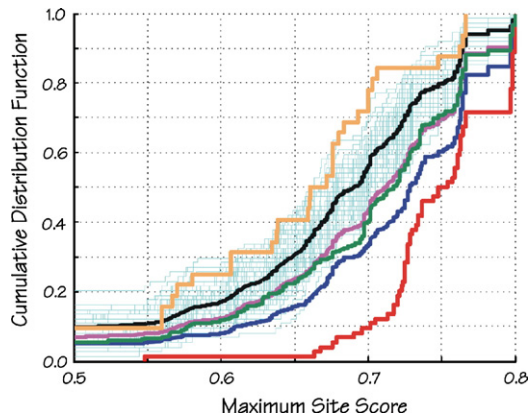


Fig. 6. Cumulative distribution of the maximum PFM score found in each promoter for sets of genes with varying degrees of expression in the adult *C. elegans* intestine. Lines are colour coded as follows: Black=all promoters in the *C. elegans* genome (total of 19,574 promoters included in the analysis); Magenta=promoters from genes expressed in the intestine (intestine tag count >1; 2816 promoters); Blue=promoters from genes in the intestine-specific/intestine-enriched peak of Fig. 2A (intestine tag count  $\geq 9$ ; somatic tag count >0; I/S tag ratio  $\geq 2$ ; 534 promoters); Red=74 highly expressed intestine-specific/intestine-enriched promoters (Table 1) used in the computational analysis; Orange=promoters from ribosomal protein genes expressed in the adult intestine (somatic tag count >0; I/S tag ratio  $\geq 2$ ; 33 promoters); Green=promoters from genes in the housekeeping/widely-expressed peak of Fig. 2A (Intestine tag count  $\geq 9$ ; somatic tag count >0;  $0.67 \leq$  I/S tag ratio  $\leq 1.5$ ; 291 genes). Thin cyan lines in the background represent 100 independent random samplings of 74 promoters from the entire genome. As discussed in the text, the biologically relevant scores are likely to lie in the range from 0.65 to 0.80.

specificity of the promoter set increases, the entire distribution shifts to higher scores until, for the most intestine-specific promoter set, all but one of the scores lie within the range likely to be biologically relevant. Furthermore, we found that the greater the degree of intestinal specificity, the greater the average number of high scoring sites per promoter (data not shown). Overall, we interpret the curves of Fig. 6 to be consistent with a model in which all intestine-specific/intestine-enriched genes have at least one critical AHTGATAARR-like sequence in their promoters. On the other hand, it is difficult for such computational analyses to be more definitive: as can be seen, there is a significant probability that any gene in the *C. elegans* genome has a reasonably high-scoring sequence in its promoter, either because the genome includes intestine genes and GATA-regulated hypodermal genes (Page et al., 1997; Gilleard and McGhee, 2001; Smith et al., 2005) or simply because of chance.

Although we are proposing that the high-scoring AHTGATAARR-like sequences identified in any particular intestine-specific/intestine-enriched promoter are indeed functional, this must ultimately be verified experimentally. Nonetheless, it is worth recalling the results of Table 2, namely that all intestine promoters experimentally examined to date have indeed depended on *cis*-acting GATA sites.

*Is the expression of housekeeping/widely-expressed genes also regulated in the intestine by cis-acting GATA-like sequences?*

To approach this question, we collected two different sets of promoters: (1) to represent housekeeping genes expressed in the

intestine, 33 promoters from genes encoding ribosomal proteins (for which the I/S tag ratio  $\geq 2$  and for many of which the *in situ* hybridization data clearly shows intestine-enriched transcripts in adult worms), and; (2) to represent widely-expressed genes, a set of 291 promoters from genes corresponding to the main peak of Fig. 2A (i.e. intestinal tag count  $\geq 9$ ;  $0.67 \leq$  I/S  $\leq 1.5$ ). The data are plotted as the orange and green lines, respectively, in Fig. 6 and suggest the following conclusions. The promoters of ribosomal protein genes are depleted in extended AHTGATAARR sites, consistent with a model in which expression of “true housekeeping” genes in the *C. elegans* intestine may not fall under GATA-factor control. In contrast, the promoters of widely-expressed genes are clearly enriched in the extended AHTGATAARR sites relative to genes selected at random from the genome. Such behaviour is consistent with a model in which ubiquitous expression results from the piecemeal assembly of tissue-specific or cell-specific controls (see, for example, Hwang and Lee, 2003; Wenick and Hobert, 2004). In other words, widely-expressed (non-housekeeping) genes may fall under GATA-factor control, just like the intestine-specific/intestine-enriched genes discussed in the previous section. As one particular example of a widely-expressed gene, the *sur-5* gene is expressed intensely in the *C. elegans* intestine but also in almost every other cell in the worm (Yochem et al., 1998). Inspection of the *sur-5* promoter shows a tandem pair of relatively high scoring (0.64 and 0.75) extended GATA-like sequences lying between 113 and 133 bps upstream of the *sur-5* ATG.

#### *Genetic and biochemical properties of ELT-2*

All evidence, both from the current paper and from the previous literature, indicates that the ELT-2 GATA-factor is likely to be the dominant transcription factor in the *C. elegans* intestine (following the events of endoderm specification that occur in the early embryo). The *elt-2* gene is necessary for correct intestinal development and deletion mutants die as newly hatched larvae with malformed intestines (Fukushige et al., 1998). In contrast, deletion of *elt-4* (Fukushige et al., 2003) or RNAi-induced loss-of-function in *elt-7* (K. Strohmaier and J. Rothman, personal communication; Fukushige et al., 2005; Oskouian et al., 2005) produces no obvious phenotype. In the current paper, we demonstrated that a strain of worms deleted for both *elt-4* and *elt-7* genes is essentially wild type. Since ELT-2 is thus the only GATA-type transcription factor remaining in the (post-specification) intestine of these worms (and in the absence of some other non-GATA factor that binds to the same sequence), ELT-2 must be the factor that binds *in vivo* to the AHTGATAARR site identified as important for the control of intestinal genes. (A more accurate statement would be that the *in vivo* binding of ELT-2 to the AHTGATAARR site must be sufficient for normal intestinal development and function, recognizing the possibility that some of these sites could normally be occupied by ELT-4 or ELT-7 or, in the early embryo, by END-1 or END-3).

ELT-2 has biochemical properties consistent with the proposed major role regulating intestinal transcription. ELT-2

was cloned by virtue of its binding to the ACTGATAAGG sequence from the *ges-1* gene (Hawkins and McGhee, 1995), ELT-2 binds tightly to this (and similar) sequences *in vitro* (Hawkins and McGhee, 1995; Moilanen et al., 1999; Fukushige et al., 2003, 2005; Oskouian et al., 2005) and ELT-2 is a strong activator of GATA-site-dependent reporter constructs in yeast (Kalb et al., 2002; Fukushige et al., 2003; Oskouian et al., 2005). In contrast, neither ELT-4 nor ELT-7 have been found to bind to DNA or to be able to activate GATA-site-dependent transcription in yeast (Fukushige et al., 2003; Oskouian et al., 2005). It will be interesting to see how closely the intrinsic sequence preferences of the ELT-2 protein, measured *in vitro*, match the PFM shown in Fig. 4E. Are the conserved bases flanking the central GATA sequence indeed preferred by ELT-2 or, alternatively, could they be evidence for an auxiliary binding factor?

At the present moment, ELT-2 has been shown experimentally to be necessary for the expression of only the *mtl-2* (Moilanen et al., 1999), *pho-1* (Fukushige et al., 2005) and *spl-1* genes (Oskouian et al., 2005). These three ELT-2 targets are first expressed in the second half of embryogenesis, after the early phase of ELT-2 redundancy (see below) and prior to the point of arrest of *elt-2* mutants. A far more comprehensive analysis of predicted ELT-2 targets will be necessary before the proposed dominant role for ELT-2 can be accepted. We note that Pauli et al. (2006) reported decreased expression levels of a variety of transgenic intestinal markers when adult worms were fed dsRNA corresponding to seven different *C. elegans* GATA factors, even factors expressed only in the early embryo (*end-1* and *end-3*) (Zhu et al., 1997; Maduro et al., 2005a) or only in the hypodermis plus a few non-intestinal cells (*elt-1*, *elt-3*, *elt-5* (*egl-18*) and *elt-6*) (Page et al., 1997; Gilleard et al., 1999; Gilleard and McGhee, 2001; Koh and Rothman, 2001; Koh et al., 2002; Smith et al., 2005). We have found it difficult to induce a reliable and penetrant *elt-2* null phenotype by RNAi-feeding, possibly because of ELT-2 protein stability (data not shown) and we agree with one of the interpretations offered by Pauli et al. (2006), namely that their reported effects could well be indirect. In particular, we do not regard their results as a serious challenge to the present evidence for the predominance of ELT-2 in controlling intestinal gene expression.

#### *How do other transcription factors in the C. elegans intestine cooperate with ELT-2?*

Although the SAGE inventory identified a total of 108 transcription factors expressed in the adult intestine, RNAi to the large majority of these transcription factor genes has little effect and in no case (except for *elt-2*) did RNAi induce a severe loss-of-function phenotype centred on the intestine. On the one hand, these results support the proposed dominant role of ELT-2 in the *C. elegans* intestine, even though the evidence is largely negative. On the other hand, it seems likely that at least some of these factors will act in conjunction with ELT-2 to regulate subsets of intestinal genes for particular digestive or physiological purposes. We note the following possibilities: (1) it now seems certain that ELT-2 is the principal activator of

vitellogenin gene transcription in the hermaphrodite intestine and that ELT-2 activation is repressed by the MAB-3 protein in the male intestine (Yi and Zarkower, 1999; Yi et al., 2000); (2) a metal-responsive transcription factor has been proposed to repress ELT-2 activation of the *mtl-2* gene in the absence of toxic metals in the environment (Moilanen et al., 1999); (3) ELT-2 may act jointly with SKN-1 in controlling stress-response genes (for example, potential antioxidant response elements have been identified in the promoters of *C. elegans* stress response genes (An and Blackwell, 2003) but many of these sequences are also TGATAA sites); (4) three high scoring GATA sequences (PFM scores 0.70, 0.72 and 0.73) have been identified immediately adjacent to a Notch-pathway target sequence in the promoter of the *ref-1* gene, implicated in regulating intestinal twist (Neves and Priess, 2005), and; (5) although the *pho-1* intestinal acid phosphatase gene is activated by ELT-2, *pho-1* is not expressed in the intestine anterior, suggesting that ELT-2 activation must be suppressed, directly or indirectly, by the zygotic Wnt pathway patterning the intestine (Fukushige et al., 2005). We suggest that the transcription of these other intestinal transcription factors may also be regulated by ELT-2; in the one example that has been investigated experimentally, *pha-4* does appear to be controlled by ELT-2, at least in the embryonic intestine (Kalb et al., 1998).

We offer two reasons why our computational analysis of intestine-specific/intestine-enriched promoters failed to identify significant secondary *cis*-acting sites that could be candidates for the binding of non-GATA-type transcription factors: (1) such sites are likely to be associated with only particular subsets of intestinal promoters, and; (2) the motif discovery algorithms are designed to detect only the most prevalent over-represented sequences. We fully expect that other classes of *cis*-acting sequences, the potential targets of factors acting combinatorially with ELT-2, will be identified by alternative experimental and bioinformatic strategies.

#### *The role of ELT-2 in the overall pathway producing the C. elegans endoderm*

Fig. 7 summarizes our view of where ELT-2 fits into the core regulatory pathway controlling the *C. elegans* endoderm, extending from maternal genes through to the production of vitellogenins for the next generation. One can distinguish three phases in the *C. elegans* endoderm pathway. The first, which we will not discuss in detail, concerns the early events in specification of the endoderm, beginning with maternal effect genes such as *skn-1* (Bowerman et al., 1992, 1993) and *pop-1* (Lin et al., 1995, 1998) and ending with activation of the two genes encoding the END-1 and END-3 GATA factors (Zhu et al., 1997) in the 1E blastomere, the clonal progenitor of the entire intestine (for recent discussion, see Goszczynski and McGhee, 2005; Maduro et al., 2005a,b, and references therein).

The next step in forming the *C. elegans* endoderm is activation of the *elt-2* gene. This begins in the mid-2E cell stage, one cell cycle after the endoderm has been specified, (Fukushige et al., 1998) and is almost certainly due to the END-1 and END-3 proteins acting directly on the *elt-2* pro-

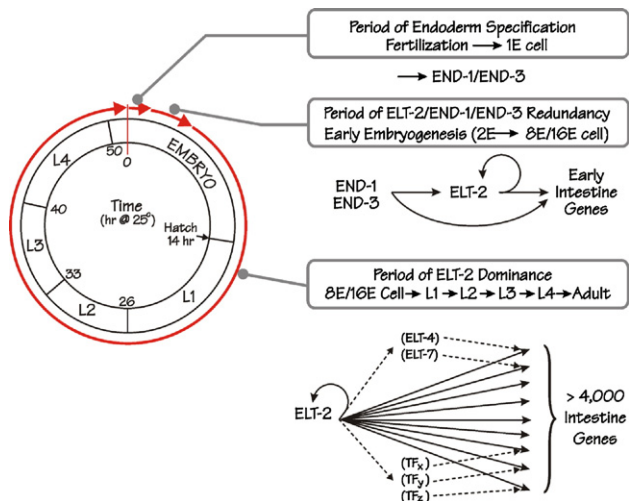


Fig. 7. The proposed role(s) of the ELT-2 GATA factor in the overall pathway forming the *C. elegans* endoderm. The successive life stages and the approximate times (hours at 25°C following fertilization) of the *C. elegans* life cycle are depicted by the circle on the left (adapted from Wood et al., 1980). As described in more detail in the text, the first stage in endoderm formation occupies ~1.5 h following fertilization, ending with production of END-1 and END-3 in the E blastomere, the defining event in endoderm specification. The second phase (“Period of ELT-2/END-1/END-3 Redundancy”) begins at the 2E cell stage when ELT-2 is first produced and ends at the 8E–16E cell stage (~4 h after fertilization) when END-1 and END-3 levels have decayed. We suggest that ELT-2, END-1 and END-3 all participate in the transcriptional activation of intestine genes during this early phase of intestine development. The third phase (“Period of ELT-2 Dominance”) begins at the 8E–16E cell stage and continues throughout all subsequent larval stages including the adult. In this phase, we propose that ELT-2 is directly and necessarily involved in all acts of transcription in the intestine, including transcription of genes encoding other transcription factors (e.g. ELT-4, ELT-7, TF<sub>x</sub>, TF<sub>y</sub> etc), which in turn may cooperate with ELT-2 in mounting particular transcriptional responses. ELT-7 (and possibly ELT-4) may provide redundant backup for a minor fraction of genes regulated by ELT-2, i.e. an *elt-7*; *elt-2* double knockout has a slightly more severe phenotype than does an *elt-2* knockout by itself (unpublished results of K. Strohmaier and J. Rothman; our unpublished results).

moter (Zhu et al., 1998; Maduro et al., 2005a; Berg, 2006). Thereafter, ELT-2 autoregulates its own production (Fukushige et al., 1998, 1999). Although ectopic ELT-2 induces ectopic expression of early markers of intestinal differentiation, these markers are still expressed in the *elt-2*(null) mutant (Fukushige et al., 1998). As evidence that this early redundancy likely involves GATA factors i.e. END-1 and/or END-3, we can cite two observations: (i) the *ges-1* gene is still expressed in the absence of ELT-2 (Fukushige et al., 1998, 1999) but deletion of the critical GATA sites from the *ges-1* promoter nonetheless abolishes all *ges-1* intestinal expression (Egan et al., 1995), and; (ii) multiple copies of the ACTGATAA site from the *pho-1* gene (Fukushige et al., 2005) drive reporter expression in the early endoderm; the strength of this expression is decreased several fold by *elt-2* RNAi but is not abolished (J. Yan and J.D.M., unpublished observations). Overall, however, it remains an important question whether ELT-2 is redundant with other early transcription factors besides END-1 and END-3.

The third phase of endoderm formation depicted in Fig. 7 begins at the ~8E to 16E cell stage when END-1/END-3 levels have decayed (Zhu et al., 1997; Baugh et al., 2003; Baugh et al.,

2005) to the point where they no longer provide ELT-2 backup. We refer to this phase as the period of ELT-2 dominance and propose that ELT-2 is necessary for and directly participates in all acts of intestinal transcription during the remaining weeks of the worm’s lifespan. We propose that all other intestinal transcription factors are subsidiary. In particular, ELT-4 and ELT-7 GATA factors together appear to be completely dispensable, although they might provide backup for a minor fraction of the genes controlled by ELT-2 (see Legend to Fig. 7) or control genes whose loss produces no phenotype. We also suggest that the hundred or so other intestinal transcription factors may also fall under ELT-2 control and then cooperate with ELT-2 (Fig. 7), for example, in mounting the worm’s transcriptional response to some particular nutritional or environmental situation.

Finally, it is instructive to compare the role proposed for ELT-2 in formation of the *C. elegans* intestine to the roles proposed for the PHA-4/FoxA factor in formation of the pharynx, the adjacent module of the *C. elegans* digestive tract. Both ELT-2 and PHA-4 can make claims to the status of “organ identity factor” or “organ selector gene” (Mango et al., 1994; Horner et al., 1998; Kalb et al., 1998; Gaudet and Mango, 2002; Ao et al., 2004; Gaudet et al., 2004) but their modes of action are distinct. The most obvious difference is in the early phase of organ specification. ELT-2 appears only after the endoderm has been specified but PHA-4 is critically involved in the act of pharynx specification itself. Like the role of ELT-2 proposed above, PHA-4 too has been proposed to participate in all acts of transcription within the pharynx, often in collaboration with PHA-4 regulated transcription factors. Whereas ELT-2 regulates the temporal unfolding of a transcriptional program within an (almost) spatially homogenous and one-dimensional clone of cells, PHA-4 must oversee a vastly more complex program that specifies five major cell types and >80 individual cells, all within an intricate three-dimensional structure assembled from two distinct cell lineages. The biochemical properties of ELT-2 fit with this simpler role; from our limited experimental comparisons (Kalb et al., 1998, 2002), ELT-2 appears far more robust and active than PHA-4 in site-specific binding *in vitro*, in the ability to drive transcription in yeast, and in the ability to induce ectopic expression of differentiation markers inside the *C. elegans* embryo. As we understand more about the development and function of the *C. elegans* digestive tract, it will be important to re-examine this comparison between the complex nuanced roles for PHA-4 in the pharynx and the simpler more straightforward requirements for ELT-2 in the intestine.

## Materials and methods

### Isolation of adult intestines

The worm strain SS104 *glp-4*(*bn2*) was propagated at 15°C on NGM agar seeded with *E. coli* OP50. To initiate the tissue isolation process, 10 healthy young adults were transferred to seeded NGM-agarose plates and placed at 25°C to lay eggs overnight. The next morning, adults were removed and the plate incubated a further 5 days (±6 h) at 25°C. Immediately prior to dissection, 10–20 (gonadless) adults were transferred to an unseeded NGM agarose plate and

maintained at room temperature for a maximum of 2 h (to digest intestinal bacteria and to remove bacteria adhering to the cuticle). Five starved worms at a time were transferred to a well slide containing the following solution: 100  $\mu$ l of PBS–EDTA–ATA (125 mM NaCl, 16.6 mM Na<sub>2</sub>HPO<sub>4</sub>, 8.4 mM NaH<sub>2</sub>PO<sub>4</sub>, 0.1 mM EDTA, 1 mM aurin tricarboxylic acid (diluted from a 100 mM pH-adjusted stock; Hallick et al., 1977) treated with diethylpyrocarbonate), 25  $\mu$ l of 10 mM levamisole in PBS–EDTA–ATA and 0.5  $\mu$ l RNAGuard (Amersham; 10–20 units). Worms were cut just behind the pharynx or just in front of the rectum, using a 27 gauge needle; most bisected worms immediately extruded their intestines, which could then be detached from the carcass with the needle. “Intestines” (i.e. fragments corresponding to the central half to three-quarters of the full intestine) were collected with baked drawn-out capillaries, rinsed several times in each of four separate wash volumes (200–300  $\mu$ l of PBS–EDTA–ATA in baked multiwell slides), transferred to RNA Later (Ambion) on ice and ultimately lysed in Trizol for RNA isolation. Control worms for the “total soma” library were treated in exact parallel but without the dissection. Worm strains containing the *glp-4(bn2)* mutation have been used extensively to distinguish somatic and germline transcripts (Reinke et al., 2004) and have been shown to have normal lifespan (McElwee et al., 2003) and wildtype levels of (intestine-specific) vitellogenin transcripts (Reinke et al., 2000); we thus feel that it is unlikely that the presence of the *glp-4(bn2)* mutation influences the present results.

#### Production and analysis of the SAGE libraries

Two SAGE libraries (intestines and total soma) were prepared by standard methods and analyzed as described in detail elsewhere (McKay et al., 2003; Wong et al., submitted for publication). Mapping of SAGE tags to *C. elegans* genes used Wormbase freeze WS140 (March 2005). Gene identification criteria were: removal of ditags, sequence quality >99%; only coding RNA; only position 1). With these criteria, there were 80,489 and 91,888 tags identified in the intestine and soma libraries, respectively; for all comparisons in the present paper, library sizes were normalized to 100,000 total tags. We estimate that the purity of the isolated intestine library is >95%: 12 genes annotated as “cuticle (or cuticular) collagen” can be identified in the somatic library and are associated with a total of 93 SAGE tags; the same 12 genes are associated with only three tags in the intestine library. In general, tissue specific transcripts encoding major structural proteins (e.g. muscle myosin), which one would ordinarily be used to estimate purity, are present at low levels in the somatic library, probably reflecting the stability of such proteins in the adult worm, as well as the fact that the worms used for the library have attained maximal body size.

We have adopted the following quasi-objective assessment of data reliability. For the majority of the data (say >80%), we feel we can trust the tag numbers within a factor of 1–2 at high tag numbers and perhaps within a factor of 2–4 at low tag numbers. For a minority of the data (say 10% or perhaps 20%), the tag count in either the intestine or the soma library appears to be “spurious”, undoubtedly due to the complexities and particularities of library preparation. While these tag counts provide evidence for presence in the library, they cannot be used to interpret expression quantitatively (e.g. the outliers seen in the I/S ratio plotted in Fig. 2A). In other libraries of this series, the majority of single SAGE tags are indeed validated when investigated by targeted RT-PCR (D.G.M. unpublished). Overall, we remain convinced of the high quality of the present SAGE data relative to the data produced by other high-throughput platforms (see Wong et al., submitted for publication).

#### Bioinformatic methods

##### Motif discovery

Upstream regions of the 74 genes in *C. elegans* (from WS140) and their orthologues in *C. briggsae* (from Cb25) and *C. remanei* (determined using WABA; Kent and Zahler, 2000) were collated into three species-specific files. Promoters were taken as the lesser of 1500 bp (not including repeats) or the distance to the end of the nearest upstream gene. MotifSampler (widths 6, 8, 10 and 12 bp; Thijs et al., 2002) and RSAT Oligo-analysis (width 8; Van Helden, 2003) were used to detect motifs. Species-specific backgrounds were generated for both methods. Detected motifs were aligned with ClustalW (Thompson et al., 1994) and clustered with OPTICS (Ankerst et al., 1999), using a base mismatch counter as a distance function between pairs of aligned motifs.

##### GATA-site analysis

A Position Frequency Matrix (PFM) was generated using the combined results from the largest OPTICS cluster of motifs from *C. elegans* upstream regions and experimentally-determined sites (127 sequences in total). All but one of these sequences were variations on the pattern NNNGATARNN (the exception was AATGATATAT). The upstream regions of all genes in the genome were scanned for instances of this pattern; instances were scored by summing respective frequencies in each position and normalizing to the number of base pairs in the site (10).

##### Miscellaneous

The images showing GFP fluorescence (Fig. 3) were collected as follows. Transgenic adults were raised at 25°C and transferred to a small volume of 0.2% Tricaine, 0.02% tetramisole in M9 buffer on an agarose pad. Fluorescent images were taken at 3–5 different focal planes (20 $\times$  lens; Zeiss Axioplan 2i microscope equipped with a Hamamatsu OrcaER digital camera), projected onto a single plane, processed at high gain in order to emphasize sites of weak expression and superimposed on images taken with differential interference contrast optics. Finally, overlapping images were assembled using Adobe Photoshop.

The production, outcrossing and analysis of the *cal6* deletion of the entire *elt-4* coding sequence was previously described (Fukushige et al., 2003). The *elt-7(tm840)* mutation was obtained from Dr. Shohei Mitani (Tokyo Women’s Medical University School of Medicine) and outcrossed five times to wild type worms. The *elt-7(tm840)* mutation deletes essentially all of the ELT-7 zinc-finger DNA-binding domain and we assume it is a null. Strain JM140 (*elt-7(tm840); elt-4(cal6)*) was produced by standard genetic crosses and all deletions were verified by PCR. It was also verified, both by Southern blotting and by PCR with primers closely flanking the DNA-binding domain, that JM140 does not contain an unexpected wildtype copy of *elt-7*.

#### Acknowledgments

The authors would like to thank Drs. Min Han (University of Colorado), Ian Hope (University of Leeds) and Robert Johnsen (Simon Fraser University) for providing several transgenic reporter strains, Rebecca Newbury (University of British Columbia) for analysis of expression patterns, Dr. Shohei Mitani (Tokyo Women’s Medical University School of Medicine) for providing the *elt-7(tm840)* deletion allele and Dr. Jonathan Freedman (Duke University) for providing the ASP-1 antibody. The worm strain SS104 (*glp-4(bn2)*) was obtained from the Caenorhabditis Genetics Stock Center (funded by the National Center for Research Resources). This work was supported by an operating grant from the Canadian Institutes of Health Research (to J.D.M) and from Genome Canada and Genome British Columbia (to D.G.M., D.L.B., M.A.M. and S.J.J.). J.D.M is a Medical Scientist of the Alberta Heritage Foundation for Medical Research and a Canada Research Chair in Developmental Biology; M.A.M. and S.J.J. are scholars of the Michael Smith Research Foundation for Health Research. M.A.M. is a Terry Fox Young Investigator.

#### References

- An, J.H., Blackwell, T.K., 2003. SKN-1 links *C. elegans* mesendodermal specification to a conserved oxidative stress response. *Genes Dev.* 17, 1882–1893.
- Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J., 1999. Ordering Points to Identify the Clustering Structure. ACM SIGMOD Int. Conf. on Management of DATA (SIGMOD ’99). Philadelphia, PA.
- Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., Mango, S.E., 2004. Environmentally

- induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 305, 1743–1746.
- Ashrafi, K., Chang, F.Y., Watts, J.L., Fraser, A.G., Kamath, R.S., Ahringer, J., Ruvkun, G., 2003. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature* 421, 268–272.
- Audic, S., Claverie, J.M., 1997. The significance of digital gene expression profiles. *Genome Res.* 7, 986–995.
- Avery, L., Shtonda, B.B., 2003. Food transport in the *C. elegans* pharynx. *J. Exp. Biol.* 206, 2441–2457.
- Avery, L., Thomas, J.H., 1997. Feeding and defecation. In: Riddle, D.L., Blumenthal, T., Meyer, B.J., Priess, J.R. (Eds.), *C. elegans* II. Cold Spring Harbor Laboratory Press, pp. 679–716.
- Azzaria, M., Goszczynski, B., Chung, M.A., Kalb, J.M., McGhee, J.D., 1996. A fork head/HNF-3 homolog expressed in the pharynx and intestine of the *Caenorhabditis elegans* embryo. *Dev. Biol.* 178, 289–303.
- Bachali, S., Jager, M., Hassanin, A., Schoentgen, F., Jolles, P., Fiala-Medioni, A., Deutsch, J.S., 2002. Phylogenetic analysis of invertebrate lysozymes and the evolution of lysozyme function. *J. Mol. Evol.* 54, 652–664.
- Banyai, L., Pathy, L., 1998. Amoebapore homologs of *Caenorhabditis elegans*. *Biochim. Biophys. Acta* 1429, 259–264.
- Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., Hunter, C.P., 2003. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* 130, 889–900.
- Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., Hunter, C.P., 2005. The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. *Development* 132, 1843–1854.
- Beanan, M.J., Strome, S., 1992. Characterization of a germ-line proliferation mutation in *C. elegans*. *Development* 116, 755–766.
- Berg, J.Y. (2006). Transcriptional regulation of the *elt-2* gene in the nematode *Caenorhabditis elegans*. M.Sc. thesis, Department of Biochemistry and Molecular Biology, University of Calgary.
- Berman, J.R., Kenyon, C., 2006. Germ-cell loss extends *C. elegans* life span through regulation of DAF-16 by kri-1 and lipophilic-hormone signaling. *Cell* 124, 1055–1068.
- Bigelow, H.R., Wenick, A.S., Wong, A., Hobert, O., 2004. CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinformatics* 5, 27.
- Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, M., Spieth, J., Sharrock, W., 1984. Cloning of a yolk protein gene family from *Caenorhabditis elegans*. *J. Mol. Biol.* 174, 1–18.
- Bock, K.W., 2003. Vertebrate UDP-glucuronosyltransferases: functional and evolutionary aspects. *Biochem. Pharmacol.* 66, 691–696.
- Bowerman, B., Eaton, B.A., Priess, J.R., 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.
- Bowerman, B., Draper, B.W., Mello, C.C., Priess, J.R., 1993. The maternal gene *skn-1* encodes a protein that is distributed unequally in early *C. elegans* embryos. *Cell* 74, 443–452.
- Britton, C., McKerrow, J.H., Johnstone, I.L., 1998. Regulation of the *Caenorhabditis elegans* gut cysteine protease gene *cpr-1*: requirement for GATA motifs. *J. Mol. Biol.* 283, 15–27.
- Bruhn, H., 2005. A short guided tour through functional and structural features of saposin-like proteins. *Biochem. J.* 389, 249–257.
- Dal Santo, P., Logan, M.A., Chisholm, A.D., Jorgensen, E.M., 1999. The inositol trisphosphate receptor regulates a 50-second behavioral rhythm in *C. elegans*. *Cell* 98, 757–767.
- Deppe, U., Schierenberg, E., Cole, T., Krieg, C., Schmitt, D., Yoder, B., von Ehrenstein, G., 1978. Cell lineages of the embryo of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* 75, 376–380.
- Egan, C.R., Chung, M.A., Allen, F.L., Heschl, M.F., Van Buskirk, C.L., McGhee, J.D., 1995. A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans ges-1* gene centers on two GATA sequences. *Dev. Biol.* 170, 397–419.
- Espelt, M.V., Estevez, A.Y., Yin, X., Strange, K., 2005. Oscillatory Ca<sup>2+</sup> signaling in the isolated *Caenorhabditis elegans* intestine: role of the inositol-1,4,5-trisphosphate receptor and phospholipases C beta and gamma. *J. Gen. Physiol.* 126, 379–392.
- Fukushige, T., Hawkins, M.G., McGhee, J.D., 1998. The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine. *Dev. Biol.* 198, 286–302.
- Fukushige, T., Hendzel, M.J., Bazett-Jones, D.P., McGhee, J.D., 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.
- Fukushige, T., Goszczynski, B., Tian, H., McGhee, J.D., 2003. The evolutionary duplication and probable demise of an endodermal GATA factor in *Caenorhabditis elegans*. *Genetics* 165, 575–588.
- Fukushige, T., Goszczynski, B., Yan, J., McGhee, J.D., 2005. Transcriptional control and patterning of the *pho-1* gene, an essential acid phosphatase expressed in the *C. elegans* intestine. *Dev. Biol.* 279, 446–461.
- Gaudet, J., Mango, S.E., 2002. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* 295, 821–825.
- Gaudet, J., Muttumu, S., Horner, M., Mango, S.E., 2004. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol.* 2, e352.
- Ghafouri, S., McGhee, J.D., in press. Bacterial residence time in the intestine of *Caenorhabditis elegans*. *Nematology*.
- Gilleard, J.S., McGhee, J.D., 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, J.S., Shafi, Y., Barry, J.D., McGhee, J.D., 1999. ELT-3: a *Caenorhabditis elegans* GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.* 208, 265–280.
- Gonczy, P., Echeverri, C., Oegema, K., Coulson, A., Jones, S.J., Copley, R.R., Duperon, J., Oegema, J., Brehm, M., Cassin, E., Hannak, E., Kirkham, M., Pichler, S., Flohrs, K., Goessen, A., Leidel, S., Alleaume, A.M., Martin, C., Ozlu, N., Bork, P., Hyman, A.A., 2000. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* 408, 331–336.
- Goszczynski, B., McGhee, J.D., 2005. Reevaluation of the role of the *med-1* and *med-2* genes in specifying the *Caenorhabditis elegans* endoderm. *Genetics* 171, 545–555.
- Gregory, P.A., Lewinsky, R.H., Gardner-Stephen, D.A., Mackenzie, P.I., 2004. Regulation of UDP glucuronosyltransferases in the gastrointestinal tract. *Toxicol. Appl. Pharmacol.* 199, 354–363.
- Hallick, R.B., Chelm, B.K., Gray, P.W., Orozco Jr., E.M., 1977. Use of aurintricarboxylic acid as an inhibitor of nucleases during nucleic acid isolation. *Nucleic Acids Res.* 4, 3055–3064.
- Hashmi, S., Britton, C., Liu, J., Guiliano, D.B., Oksov, Y., Lustigman, S., 2002. Cathepsin L is essential for embryogenesis and development of *Caenorhabditis elegans*. *J. Biol. Chem.* 277, 3477–3486.
- Hashmi, S., Zhang, J., Oksov, Y., Lustigman, S., 2004. The *Caenorhabditis elegans* cathepsin Z-like cysteine protease, Ce-CPZ-1, has a multifunctional role during the worms' development. *J. Biol. Chem.* 279, 6035–6045.
- Hawkins, M.G., McGhee, J.D., 1995. *elt-2*, a second Gata factor from the nematode *Caenorhabditis elegans*. *J. Biol. Chem.* 270, 14666–14671.
- Hevelone, J., Hartman, P.S., 1988. An endonuclease from *Caenorhabditis elegans*: partial purification and characterization. *Biochem. Genet.* 26, 447–461.
- Hirose, T., Nakano, Y., Nagamatsu, Y., Misumi, T., Ohta, H., Ohshima, Y., 2003. Cyclic GMP-dependent protein kinase EGL-4 controls body size and lifespan in *C. elegans*. *Development* 130, 1089–1099.
- Hirsh, D., Oppenheim, D., Klass, M., 1976. Development of the reproductive system of *Caenorhabditis elegans*. *Dev. Biol.* 49, 200–219.
- Homolya, L., Varadi, A., Sarkadi, B., 2003. Multidrug resistance-associated proteins: Export pumps for conjugates with glutathione, glucuronate or sulfate. *Biofactors* 17, 103–114.
- Horner, M.A., Quintin, S., Domeier, M.E., Kimble, J., Labouesse, M., Mango, S.E., 1998. *pha-4*, an HNF-3 homolog, specifies pharyngeal organ identity in *Caenorhabditis elegans*. *Genes Dev.* 12, 1947–1952.
- Hwang, S.B., Lee, J., 2003. Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode *Caenorhabditis elegans*. *J. Mol. Biol.* 333, 237–247.
- Jasmer, D.P., Roth, J., Myler, P.J., 2001. Cathepsin B-like cysteine proteases and *Caenorhabditis elegans* homologues dominate gene products expressed in

- adult *Haemonchus contortus* intestine. *Mol. Biochem. Parasitol.* 116, 159–169.
- Jasmer, D.P., Mitreva, M.D., McCarter, J.P., 2004. mRNA sequences for *Haemonchus contortus* intestinal cathepsin B-like cysteine proteases display an extreme in abundance and diversity compared with other adult mammalian parasitic nematodes. *Mol. Biochem. Parasitol.* 137, 297–305.
- Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., Kim, S.K., 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 218–223.
- Kalb, J.M., Lau, K.K., Goszczynski, B., Fukushige, T., Moons, D., Okkema, P.G., McGhee, J.D., 1998. *pha-4* is Ce-*flh-1*, a fork head/HNF-3 homolog that functions in organogenesis of the *C. elegans* pharynx. *Development* 125, 2171–2180.
- Kalb, J.M., Beaster-Jones, L., Fernandez, A.P., Okkema, P.G., Goszczynski, B., McGhee, J.D., 2002. Interference between the PHA-4 and PEB-1 transcription factors in formation of the *Caenorhabditis elegans* pharynx. *J. Mol. Biol.* 320, 697–704.
- Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P., Ahringer, J., 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237.
- Kent, W.J., Zahler, A.M., 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*–*C. elegans* genomic alignment. *Genome Res.* 10, 1115–1125.
- Kimble, J., Sharrock, W.J., 1983. Tissue-specific synthesis of yolk proteins in *Caenorhabditis elegans*. *Dev. Biol.* 96, 189–196.
- Kniazeva, M., Crawford, Q.T., Seiber, M., Wang, C.Y., Han, M., 2004. Monomethyl branched-chain fatty acids play an essential role in *Caenorhabditis elegans* development. *PLoS Biol.* 2, E257.
- Koh, K., Rothman, J.H., 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.
- Koh, K., Peyrot, S.M., Wood, C.G., Wagmaister, J.A., Maduro, M.F., Eisenmann, D.M., Rothman, J.H., 2002. Cell fates and fusion in the *C. elegans* vulval primordium are regulated by the EGL-18 and ELT-6 GATA factors—Apparent direct targets of the LIN-39 Hox protein. *Development* 129, 5171–5180.
- Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Rogozin, I.B., Smimov, S., Sorokin, A.V., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5, R7.
- Kostich, M., Fire, A., Fambrough, D.M., 2000. Identification and molecular-genetic characterization of a LAMP/CD68-like protein from *Caenorhabditis elegans*. *J. Cell Sci.* 113, 2595–2606.
- Libina, N., Berman, J.R., Kenyon, C., 2003. Tissue-specific activities of *C. elegans* DAF-16 in the regulation of lifespan. *Cell* 115, 489–502.
- Lin, R., Thompson, S., Priess, J.R., 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.
- Lin, R., Hill, R.J., Priess, J.R., 1998. POP-1 and anterior–posterior fate decisions in *C. elegans* embryos. *Cell* 92, 229–239.
- Liu, J., Fire, A., 2000. Overlapping roles of two Hox genes and the *exd* ortholog *ceh-20* in diversification of the *C. elegans* postembryonic mesoderm. *Development* 127, 5179–5190.
- Luersen, K., Eschbach, M.L., Liebau, E., Walter, R.D., 2004. Functional GATA- and initiator-like-elements exhibit a similar arrangement in the promoters of *Caenorhabditis elegans* polyamine synthesis enzymes. *Biol. Chem.* 385, 711–721.
- MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T., Spieth, J., 1992. Regulation of vitellogenin gene expression in transgenic *Caenorhabditis elegans*: short sequences required for activation of the vit-2 promoter. *Mol. Cell Biol.* 12, 1652–1662.
- MacMorris, M., Spieth, J., Madej, C., Lea, K., Blumenthal, T., 1994. Analysis of the VPE sequences in the *Caenorhabditis elegans* vit-2 promoter with extrachromosomal tandem array-containing transgenic strains. *Mol. Cell Biol.* 14, 484–491.
- Maduro, M.F., Rothman, J.H., 2002. Making worm guts: the gene regulatory network of the *Caenorhabditis elegans* endoderm. *Dev. Biol.* 246, 68–85.
- Maduro, M.F., Meneghini, M.D., Bowerman, B., Broitman-Maduro, G., Rothman, J.H., 2001. Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3 $\beta$  homolog is mediated by MED-1 and -2 in *C. elegans*. *Mol. Cell* 7, 475–485.
- Maduro, M.F., Hill, R.J., Heid, P.J., Newman-Smith, E.D., Zhu, J., Priess, J.R., Rothman, J.H., 2005a. Genetic redundancy in endoderm specification within the genus *Caenorhabditis*. *Dev. Biol.* 284, 509–522.
- Maduro, M.F., Kasmir, J.J., Zhu, J., Rothman, J.H., 2005b. 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.
- Mallo, G.V., Kurz, C.L., Couillault, C., Pujol, N., Granjeaud, S., Kohara, Y., Ewbank, J.J., 2002. Inducible antibacterial defense system in *C. elegans*. *Curr. Biol.* 12, 1209–1214.
- Mango, S.E., Lambie, E.J., Kimble, J., 1994. The *pha-4* gene is required to generate the pharyngeal primordium of *Caenorhabditis elegans*. *Development* 120, 3019–3031.
- McCarroll, S.A., Li, H., Bargmann, C.I., 2005. Identification of transcriptional regulatory elements in chemosensory receptor genes by probabilistic segmentation. *Curr. Biol.* 15, 347–352.
- McElwee, J., Bubbs, K., Thomas, J.H., 2003. Transcriptional outputs of the *Caenorhabditis elegans* forkhead protein DAF-16. *Aging Cell* 2, 111–121.
- McGhee, J.D. (in press). The *C. elegans* Intestine. In “WormBook” (The *C. elegans* Research Community, Eds.). <http://www.wormbook.org>.
- McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., Halfnight, E., Hollebakk, R., Huang, P., Hung, K., Jensen, V., Jones, S.J., Kai, H., Li, D., Mah, A., Marra, M., McGhee, J., Newbury, R., Pouzyrev, A., Riddle, D.L., Sonhammer, E., Tian, H., Tu, D., Tyson, J.R., Vatcher, G., Warner, A., Wong, K., Zhao, Z., Moerman, D.G., 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harbor Symp. Quant. Biol.* 68, 159–169.
- Menzel, R., Bogaert, T., Achazi, R., 2001. A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Arch. Biochem. Biophys.* 395, 158–168.
- Miyabayashi, T., Palfreyman, M.T., Sluder, A.E., Slack, F., Sengupta, P., 1999. Expression and function of members of a divergent nuclear receptor family in *Caenorhabditis elegans*. *Dev. Biol.* 215, 314–331.
- Moilanen, L.H., Fukushige, T., Freedman, J.H., 1999. Regulation of metallothionein gene transcription. Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from *Caenorhabditis elegans*. *J. Biol. Chem.* 274, 29655–29665.
- Neves, A., Priess, J.R., 2005. The REF-1 family of bHLH transcription factors pattern *C. elegans* embryos through Notch-dependent and Notch-independent pathways. *Dev. Cell* 8, 867–879.
- Oskouian, B., Mendel, J., Shocron, E., Lee Jr., M.A., Fyrst, H., Saba, J.D., 2005. Regulation of sphingosine-1-phosphate lyase gene expression by members of the GATA family of transcription factors. *J. Biol. Chem.* 280, 18403–18410.
- Page, B.D., Zhang, W., Steward, K., Blumenthal, T., Priess, J.R., 1997. ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in *Caenorhabditis elegans* embryos. *Genes Dev.* 11, 1651–1661.
- Pauli, F., Liu, Y., Kim, Y.A., Chen, P.J., Kim, S.K., 2006. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in *C. elegans*. *Development* 133, 287–295.
- Portman, D.S., Emmons, S.W., 2004. Identification of *C. elegans* sensory ray genes using whole-genome expression profiling. *Dev. Biol.* 270, 499–512.
- Rawlings, N.D., Barrett, A.J., 1993. Evolutionary families of peptidases. *Biochem. J.* 290, 205–218.
- Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J., Davis, E.B., Scherer, S., Ward, S., Kim, S.K., 2000. A global profile of germline gene expression in *C. elegans*. *Mol. Cell* 6, 605–616.
- Reinke, V., Gil, I.S., Ward, S., Kazmer, K., 2004. Genome-wide germline-

- enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* 131, 311–323.
- Robertson, S.M., Shetty, P., Lin, R., 2004. Identification of lineage-specific zygotic transcripts in early *Caenorhabditis elegans* embryos. *Dev. Biol.* 276, 493–507.
- Roy, P.J., Stuart, J.M., Lund, J., Kim, S.K., 2002. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* 418, 975–979.
- Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. *Nucleic. Acids Res.* 18, 6097–6100.
- Shen, M.M., Hodgkin, J., 1988. *mab-3*, a gene required for sex-specific yolk protein expression and a male-specific lineage in *C. elegans*. *Cell* 54, 1019–1031.
- Sladek, N.E., 2003. Human aldehyde dehydrogenases: potential pathological, pharmacological, and toxicological impact. *J. Biochem. Mol. Toxicol.* 17, 7–23.
- Smith, J.A., McGarr, P., Gilleard, J.S., 2005. The *Caenorhabditis elegans* GATA factor *elt-1* is essential for differentiation and maintenance of hypodermal seam cells and for normal locomotion. *J. Cell Sci.* 118, 5709–5719.
- Sonnichsen, B., Koski, L.B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A.M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Stamford, J., Winzi, M., Heinkel, R., Roder, M., Finell, J., Hantsch, H., Jones, S.J., Jones, M., Piano, F., Gunsalus, K.C., Oegema, K., Gonczy, P., Coulson, A., Hyman, A.A., Echeverri, C.J., 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* 434, 462–469.
- Srivastava, S.K., Ramana, K.V., Bhatnagar, A., 2005. Role of aldose reductase and oxidative damage in diabetes and the consequent potential for therapeutic options. *Endocr. Rev.* 26, 380–392.
- Sulston, J.E., 1976. Post-embryonic development in the ventral cord of *Caenorhabditis elegans*. *Philos. Trans. R. Soc. London—Ser B. Biol. Sci.* 275, 287–297.
- Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N., 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 100, 64–119.
- Syntchaki, P., Xu, K., Driscoll, M., Tavernarakis, N., 2002. Specific aspartyl and calpain proteases are required for neurodegeneration in *C. elegans*. *Nature* 419, 939–944.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.
- Tcherepanova, I., Bhattacharyya, L., Rubin, C.S., Freedman, J.H., 2000. Aspartic proteases from the nematode *Caenorhabditis elegans*. Structural organization and developmental and cell-specific expression of *asp-1*. *J. Biol. Chem.* 275, 26359–26369.
- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., Moreau, Y., 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17, 1113–1122.
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., Moreau, Y., 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* 9, 447–464.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids Res.* 22, 4673–4680.
- Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavese, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z., 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23, 137–144.
- Van Auken, K., Weaver, D., Robertson, B., Sundaram, M., Saldi, T., Edgar, L., Elling, U., Lee, M., Boese, Q., Wood, W.B., 2002. Roles of the Homothorax/Meis/Prep homolog UNC-62 and the Exd/Pbx homologs CEH-20 and CEH-40 in *C. elegans* embryogenesis. *Development* 129, 5255–5268.
- van Helden, J., 2003. Regulatory sequence analysis tools. *Nucleic. Acids Res.* 31, 3593–3596.
- van Helden, J., Andre, B., Collado-Vides, J., 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827–842.
- Vasilio, V., Pappa, A., Estey, T., 2004. Role of human aldehyde dehydrogenases in endobiotic and xenobiotic metabolism. *Drug Metab. Rev.* 36, 279–299.
- Wenick, A.S., Hobert, O., 2004. Genomic *cis*-regulatory architecture and *trans*-acting regulators of a single interneuron-specific gene battery in *C. elegans*. *Dev. Cell* 6, 757–770.
- Wong, K., McKay, S.J., Khattraj, J., Chan, S., Asano, J., Go, A., Pandoh, P., MacDonald, H., Huang, P., Ruzanov, P., Mills, C., Warner, A., Bailie, D.L., Holt, R.A., Jones, S.J.M., Marra, M.A., and Moerman, D.G. submitted for publication. Comparative analysis of SAGE and microarray technologies for global transcription profiling of development in *Caenorhabditis elegans*.
- Wood, W.B., Hecht, R., Carr, S., Vanderslice, R., Wolf, N., Hirsh, D., 1980. Parental effects and phenotypic characterization of mutations that affect early development in *Caenorhabditis elegans*. *Dev. Biol.* 74, 446–469.
- Wu, Y.C., Stanfield, G.M., Horvitz, H.R., 2000. NUC-1, a *Caenorhabditis elegans* DNase II homolog, functions in an intermediate step of DNA degradation during apoptosis. *Genes Dev.* 14, 536–548.
- Yi, W., Zarkower, D., 1999. Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAB-3 and *Drosophila melanogaster* DSX suggests conservation of sex determining mechanisms. *Development* 126, 873–881.
- Yi, W., Ross, J.M., Zarkower, D., 2000. *Mab-3* is a direct *tra-1* target gene regulating diverse aspects of *C. elegans* male sexual development and behavior. *Development* 127, 4469–4480.
- Yochem, J., Gu, T., Han, M., 1998. A new marker for mosaic analysis in *Caenorhabditis elegans* indicates a fusion between *hyp6* and *hyp7*, two major components of the hypodermis. *Genetics* 149, 1323–1334.
- Zavalova, L.L., Baskova, I.P., Lukyanov, S.A., Sass, A.V., Snezhkov, E.V., Akopov, S.B., Artamonova, I.I., Archipova, V.S., Nesmeyanov, V.A., Kozlov, D.G., Benevolensky, S.V., Kiseleva, V.I., Poverenny, A.M., Sverdlov, E.D., 2000. Destabilase from the medicinal leech is a representative of a novel family of lysozymes. *Biochim. Biophys. Acta* 1478, 69–77.
- Zhai, Y., Saier Jr, M.H., 2000. The amoebapore superfamily. *Biochim. Biophys. Acta* 1469, 87–99.
- Zhu, J., Hill, R.J., Heid, P.J., Fukuyama, M., Sugimoto, A., Priess, J.R., Rothman, J.H., 1997. *end-1* encodes an apparent GATA factor that specifies the endoderm precursor in *Caenorhabditis elegans* embryos. *Genes Dev.* 11, 2883–2896.
- Zhu, J., Fukushige, T., McGhee, J.D., Rothman, J.H., 1998. Reprogramming of early embryonic blastomeres into endodermal progenitors by a *Caenorhabditis elegans* GATA factor. *Genes Dev.* 12, 3809–3814.