

***cis*-Regulatory Control of Three Cell Fate-Specific Genes in Vulval
Organogenesis of *C. elegans* and *C. briggsae***

Thesis by

Martha Kirouac

In partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2003

(submitted August 26th, 2002)

©2003

Martha Kirouac

All Rights Reserved

ACKNOWLEDGEMENTS

I have been very fortunate to have Dr. Paul Sternberg as my advisor. His enthusiasm and dedication toward science are an inspiration in my life, and his personal support has meant much to me.

I would like to thank my committee members, Dr. Scott Fraser, Dr. Ellen Rothenberg, Dr. Raymond Deshaies, and Dr. Elliot Meyerowitz, for their time and guidance.

To the Sternberg lab members, past and present, I offer thanks for discussions, guidance, and friendship. A special thanks to Yen, Aidyl, Lisa, Cheryl, Gladys, Rene, and Shawn with whom I have had the joy of sharing a room, and who know my penchant for singing while I work. Thanks to Dave, Takao, Bhagwati, Byung, Rene, Nadeem and Minqin, who have helped me shape my work into a story, and to Mary, for always caring how I was and showing me the ropes.

Thanks to Mr. Bandura, who first inspired my interest in biology. I still remember the inspiration I felt with your description of why leaves change color and how corn is fertilized. To Dr. Jill Salvo, who believed in my abilities long before I did, and who pushed me to achieve, I thank you for a great many things, not the least of which is your friendship.

To all things Prufrock, I owe great thanks. Members past and present, you have been my bedrock, you have helped me maintain sanity in the face of adversity, and we've had fun to boot. You have been my family, my housemates and my friends for the past six years. To my confidants and special friends, Yen, Jen-Jen, Tom, Beth, Cheryl, Chi,

Jasper, Keith, Lisa, Big Paul and Aidyl: with each of you I have shared special moments and each of you has a place in my heart. You have made my time in graduate school more enjoyable and fuller than I could have hoped.

To those people in my life who have been steadfast in their never-ending support and love: Mom, Dad, Ian, Gabrielle, and Aunty-Lou. Thanks for your love, patience, encouragement and support.

To Shanti, thanks for things that have been and things that are to be. You have made me smile when I most needed it, and you have supported me throughout our evolving friendship.

ABSTRACT

The great-grandprogeny of the *Caenorhabditis elegans* vulval precursor cells (VPCs) adopt one of the final vulA, B1, B2, C, D, E and F cell types in a precise spatial pattern. Formation of the pattern of vulval cell types is likely to depend upon the *cis*-regulatory regions of the transcriptional targets of these intercellular signals in vulval development. The outcome of such differential activation will result in individual cell types. *egl-17*, *zmp-1*, *cdh-3* are expressed differentially in the developing vulva cells, providing a potential readout for different signaling pathways. To understand how different signaling pathways interact to specify unique vulval cell types in a precise pattern, I have identified upstream *cis*-regulatory regions that are sufficient for their ability to confer vulval cell type-specific regulation when fused in *cis* to the basal *pes-10* promoter. In the *egl-17* promoter, I have identified a 143 base pair (bp) region that drives vulC and vulD expression, and a 102 bp region that is sufficient to drive the early expression in presumptive vulE and vulF cells. In the *zmp-1* promoter, I have identified a 300 bp region that is sufficient to drive expression in vulE, vulA and the anchor cell. In the *cdh-3* promoter, I have identified a 689 bp region sufficient to drive expression in the anchor cell and vulE, vulF, vulD and vulC, a 155 bp region sufficient to drive only anchor cell expression, and a separate 563 bp region that was also sufficient to drive expression in these vulval cells. I have identified the *C. briggsae* homologs of these three genes, and the corresponding control regions, and tested these regions in both *C. elegans* and *C. briggsae*. I find that these regions of similarity in *C. elegans* and *C. briggsae* upstream of *egl-17*, *zmp-1*, and *cdh-3* promote expression in vulval cells and the anchor cell. Using

the regions defined by the sufficiency analysis and phylogenetic footprinting, I have been able to isolate over-represented sequences that may play important roles in conferring vulval and anchor cell expression.

TABLE OF CONTENTS

Acknowledgments	iii
Abstract	v
Table of Contents	vii
Chapter 1: Transcriptional <i>cis</i>-Regulation	I-1
Introduction	I-2
<i>cis</i> -acting regulatory elements of transcription in eukaryotes	I-2
Transcriptional regulation in <i>C. elegans</i>	I-6
Conservation of trans-acting transcriptional regulators in <i>C. elegans</i>	I-7
Vulva cell specification and intracellular signaling pathways	I-8
Genomic regulatory network analysis	I-14
<i>egl-17</i> and the FGF family	I-15
<i>zmp-1</i> and the Matrix Metalloproteinases	I-17
<i>cdh-3</i> and the Cadherins	I-18
Regulatory analysis in <i>C. elegans</i>	I-19
Dissection of co-regulated genes	I-25
Phylogenetic footprinting	I-29
Thesis overview	I-32
References	I-33
Figure 1: Vulva formation in <i>C. elegans</i>	I-49
Figure 2: Available vulval marker gene expression patterns in <i>C. elegans</i>	I-51
Figure 3: <i>egl-17::GFP</i>	I-53
Figure 4: <i>zmp-1::GFP</i>	I-55
Figure 5: <i>cdh-3::GFP</i>	I-57
Chapter 2: <i>cis</i>-Regulatory Control of Cell Fate-Specific Genes in <i>Caenorhabditis elegans</i> Vulval Organogenesis	II-1
Abstract	II-2

Introduction	II-3
Materials and Methods	II-4
Generation of <i>C. elegans</i> promoter GFP constructs	II-4
Generation of <i>C. elegans</i> promoter deletion GFP constructs	II-6
Sequencing of constructs	II-6
Microinjection of promoter GFP constructs into <i>C. elegans</i>	II-6
Microscopy of transgenic animals	II-7
Prediction of binding sites using Transfac database	II-7
AlignACE predictions of over-represented sequences	II-8
Results	II-8
Vulval specificity in the <i>egl-17</i> cis-regulatory region in <i>C. elegans</i>	II-9
Vulva and anchor cell specificity in the <i>zmp-1</i> cis-regulatory region in <i>C. elegans</i>	II-11
Vulva and anchor cell specificity in the <i>cdh-3</i> cis-regulatory region in <i>C. elegans</i>	II-13
Transfac putative binding site predictions in upstream sequences	II-16
AlignACE predictions of over-represented sequences	II-17
Discussion	II-21
<i>egl-17</i>	II-24
<i>zmp-1</i>	II-25
<i>cdh-3</i>	II-26
Distance of elements from translational start sites in <i>egl-17</i> , <i>zmp-1</i> , and <i>cdh-3</i>	II-28
Analysis of putative <i>trans</i> -acting factors	II-28
Analysis of over-represented sequences in regions of sufficiency	II-29
Conclusions	II-30
Acknowledgments	II-31
References	II-32
Figure 1: Marker gene expression summary	II-36
Figure 2: Initial dissection of <i>egl-17</i> , <i>zmp-1</i> , and <i>cdh-3</i> regulatory regions	II-38
Figure 3: Upstream regions that direct <i>egl-17</i> expression	II-40

Figure 4: Upstream sequences of mk84-148, mk50-51, mk96-134 and mk66-67	II-42
Figure 5: Multiple regions direct <i>zmp-1</i> expression	II-44
Figure 6: Regions that direct <i>cdh-3</i> expression	II-46
Table 1: Transfac binding site predictions for regions that confer cell-specific expression	II-48
Table 2: AlignACE predictions of over-represented sequences	II-51
Supplemental Table 1: PCR primers	II-55
Supplemental Figure 1: <i>egl-17</i> cis-regulatory deletion analysis	II-59
Supplemental Figure 2: <i>zmp-1</i> cis-regulatory deletion analysis	II-61
Supplemental Figure 3: <i>cdh-3</i> cis-regulatory deletion analysis	II-64

**Chapter 3: Three Genes, Two Species: A Comparative Analysis of Upstream
Regulatory Sequences Sufficient to Direct Vulval Expression in
C. elegans and *C. briggsae***

Abstract	III-2
Introduction	III-3
Materials and Methods	III-5
Protein prediction of EGL-17, ZMP-1 and CDH-3 homologs in <i>C. briggsae</i>	III-5
Analysis of homologous upstream sequences in <i>C. elegans</i> and <i>C. briggsae</i>	III-5
Generation of <i>egl-17</i> , <i>zmp-1</i> , and <i>cdh-3</i> <i>C. briggsae</i> promoter GFP constructs	III-6
Microinjection of promoter GFP constructs into <i>C. elegans</i>	III-7
Microinjection of promoter GFP constructs into <i>C. briggsae</i>	III-7
Microscopy of transgenic animals	III-7
Prediction of binding sites using Transfac database	III-8
AlignACE predictions of over-represented sequences	III-8
Results	III-9
<i>C. briggsae</i> homologs of <i>egl-17</i> , <i>zmp-1</i> , and <i>cdh-3</i>	III-9

Comparative sequence analysis	III-11
Analysis of <i>C. briggsae</i> upstream regions	III-14
Transfac binding site prediction in conserved regions	III-18
AlignACE predictions of over-represented sequences	III-19
Discussion	III-22
Phylogenetic footprinting	III-23
Potential for specific isolation of <i>trans</i> -acting factor binding sites by phylogenetic footprinting between <i>C. elegans</i> and <i>C. briggsae</i>	III-24
Analysis of putative <i>trans</i> -acting factors using the Transfac database	III-25
Analysis of over-represented sequences in regions of sufficiency	III-26
Implications of cross-species comparison of <i>egl-17</i> , <i>zmp-1</i> and <i>cdh-3</i>	III-26
Conclusions	III-28
Acknowledgments	III-29
References	III-30
Figure 1: EGL-17 clustalW alignment in <i>C. elegans</i> and <i>C. briggsae</i>	III-34
Figure 2: ZMP-1 clustalW alignment in <i>C. elegans</i> and <i>C. briggsae</i>	III-36
Figure 3: CDH-3 clustalW alignment in <i>C. elegans</i> and <i>C. briggsae</i>	III-38
Figure 4: Seqcomp and Family Relations predictions for <i>egl-17</i> , <i>zmp-1</i> , and <i>cdh-3</i> upstream sequences	III-40
Table 1: Summary of construct expression patterns	III-42
Figure 5: <i>egl-17</i> nucleotide sequences of important regions	III-44
Figure 6: <i>zmp-1</i> nucleotide sequences of important regions	III-46
Figure 7: <i>cdh-3</i> nucleotide sequences of mk96-134 and mk162-163	III-48
Figure 8: <i>cdh-3</i> nucleotide sequences of mk66-67 and mk164-165	III-50
Figure 9: <i>C. briggsae</i> upstream regions injected in <i>C. elegans</i>	III-52
Table 2: Transfac binding site predictions in regions of similarity between <i>C. elegans</i> and <i>C. briggsae</i>	III-54
Table 3: AlignACE predictions of over-represented sequences	III-58

Chapter 4: Summary	IV-1
Thesis Summary	IV-2
Sufficiency analysis	IV-3
Determining the necessity of regions defined by sufficiency analysis	IV-4
Phylogenetic footprinting studies of <i>cis</i> -regulatory sequences	IV-4
Practical considerations when identifying phylogenetic footprints	IV-6
Combining the results of sufficiency testing and phylogenetic footprinting studies	IV-6
Analysis of putative <i>trans</i> -acting factors	IV-7
Genomic analysis	IV-11
References	IV-12
Figure 1: Selection of nematode species for comparative genomic analysis	IV-14
Figure 2: Combined results of <i>egl-17</i> sufficiency and phylogenetic analyses	IV-16
Figure 3: Combined results of <i>zmp-1</i> sufficiency and phylogenetic analyses	IV-18
Figure 4: Combined results of <i>cdh-3</i> sufficiency and phylogenetic analyses	IV-20
Table 1: Effect of genetic background on marker expression	IV-22

Chapter 1

Transcriptional *cis*-Regulation

Introduction

The process of differential gene expression, or the selective activation of different subsets of genes, leads to unique populations of cells that are terminally differentiated. Selective activation is carefully regulated and, ultimately, controls all functions of cells, tissues and organs. Central to the process of differential gene expression and cell fate specification are the *cis*-regulatory elements of genes that are responsible for determining the temporal and spatial domains of gene expression. These *cis*-regulatory elements are part of the larger transcriptional machinery that controls the production of gene products that establish and maintain unique cell populations.

Caenorhabditis elegans is a free-living, soil-dwelling nematode. All 959 somatic cells of its transparent, 1mm-long body are visible with a microscope. It has a rapid life cycle (14-hour embryogenesis and 36-hour postembryonic development through four larval stages, L1-L4, to the adult) (reviewed in Riddle *et al.*, 1997). The development and function of this organism is encoded by an estimated 19,476 genes (www.wormbase.org; release WS84). Within this genome are the genes that encode the developmental program of the vulva. The vulva of *C. elegans* provides an excellent system to study the mechanisms by which *cis*-regulatory controls are utilized in establishing differential gene expression and terminal differentiation.

***cis*-acting regulatory elements of transcription in eukaryotes**

The typical eukaryotic gene consists of up to four distinct *cis*-regulatory transcriptional control elements: the promoter itself, the upstream promoter elements (UPEs), elements

adjacent to the promoter that are interspersed with the UPEs, and distinct enhancer elements (reviewed in Latchman, 1998).

Upstream elements contain two types of sequences. The first type are those sequences, which are found in many genes that exhibit distinct patterns of regulation, and are likely to be involved in the basic process of transcription. These are referred to as the basal transcription machinery. The second type of sequences are those that are only in genes transcribed in a particular tissue, or in response to a specific signal. This type of transcription is referred to as regulated transcription (reviewed in Latchman, 1998).

Several sequences characterize the typical eukaryotic basal transcription machinery. The first is the TATA box element. This TATA sequence is found 25-30 bp upstream of the transcriptional start site in most genes, although it is sometimes absent, as in many housekeeping genes. The region delimited by the TATA box and the sites of transcriptional initiation (the cap site) has been defined as the gene promoter (reviewed in Latchman, 1998). The promoter probably binds several proteins essential for transcription, as well as RNA polymerase II, the enzyme that is responsible for the transcription of the genes (reviewed in Sentenac, 1985). Genes may also contain UPEs, such as the CCAAT and Sp1 boxes, which, if found, are typically upstream of the TATA box (reviewed in McKnight and Tjian, 1986). In every instance that they have been found, they are essential for the transcription of the genes (reviewed in Latchman, 1998).

The binding of particular proteins to specific upstream sequences in order to confer on a gene the ability to respond to particular stimuli is known as regulated transcription. To prove that an element found in one group of common genes is important for that group's transcriptional activity, the sequence must confer the same response or

expression to an unrelated gene. A classic example of regulated transcription was characterized in the *hsp70* gene. In this case, the heat-shock element, when transferred to an unrelated gene, the non-heat-shock inducible thymidine kinase gene, conferred on it the ability to respond to a heat-shock stimulus (Pelham, 1982). Such DNA sequence elements in the promoters of tissue-specific genes play a critical role in producing their tissue-specific pattern of expression.

These tissue-specific elements are not confined to the promoters of genes; they may be found at great distances from the transcriptional start sites (Grosschedl and Birnstiel, 1980). Even at great distances and, in any orientation with respect to the transcriptional start site, these elements may affect the level of gene expression whether located upstream, downstream, or within the coding region. Although they lack promoter activity by themselves, these sequences act by increasing or decreasing the activity of a promoter, and hence are referred to as enhancers (reviewed in Muller *et al.*, 1988). Enhancers may increase the activity of a promoter in all cell types, or they may activate a particular promoter only in a select cell type (reviewed in Latchman, 1998). Enhancers usually contain multiple binding sites for transcription factors that cooperatively act to alter gene transcription (reviewed in Carey, 1998). These combinations of binding sites may be found in similarly regulated enhancers and promoters (co-regulation), and may also be present in multiple copies (e.g. Sen and Baltimore, 1986).

The balance between positive- and negative-acting transcription factors that bind to these regulatory regions determines the rate of the gene's transcription. One piece of the puzzle that effects this balance is the access of a transcription factor to its appropriate binding site. This in turn is affected by the manner in which that site is packaged in the

chromatin. A nucleosome, the fundamental unit of chromatin, consists of eight histone molecules around which the DNA wraps. Genes that are about to be transcribed undergo a reorganization of the chromatin (reviewed in Felsenfeld, 1996; Latchman, 1998). While the regulation of chromatin structure is necessary for proper gene expression, it is not sufficient. Distinct multiprotein complexes are needed to alter chromatin structure, to bind to promoters and enhancers, and to communicate between the activators and repressors (reviewed in Narlikar *et al.*, 2002). There are two classes of complexes that regulate the accessibility of the DNA to these various factors. The first class is ATP-dependent complexes that can move the nucleosome positions to expose or hide specific DNA sequences. The second class is those complexes that covalently modify the nucleosomes by adding or removing chemical moieties: acetylation, phosphorylation, and methylation of histone N-termini (reviewed in Narlikar *et al.*, 2002). One of the most studied chromatin-remodeling complexes that utilizes ATP hydrolysis is the SWI/SNF complex in yeast (reviewed in Pazin and Kadonaga, 1997; Tsukiyama and Wu, 1997). The most studied modification of the histone tail involves its acetylation, which *in vitro* has been shown to enhance accessibility of the DNA to restriction enzymes and transcription factors. There are several hypotheses as to why acetylation may have this effect. The first is that the lowered positive charge on the acetylated N-termini may cause a decrease in the stability of interaction with the DNA (Sewack *et al.*, 2001). The second is that the histone acetylation may decrease the compaction of the nucleosomes by interrupting the internucleosomal interactions made via the histone tails (Tse *et al.*, 1998). Finally, a third hypothesis is that these tail modifications might interact and physically recruit additional transcription factors (Strahl and Allis, 2000). Evidence

indicates that some transcription factors may bind directly to both ATP-dependent chromatin remodeling and histone acetyltransferase complexes, to "target" these activities to specific locations (reviewed in Narlikar *et al.*, 2002).

Gene transcription is initiated through the recruitment of RNA polymerase II (Pol II) to the promoters of target genes, the modification of nucleosomes, and the remodeling of chromatin. This occurs in conjunction with the assembly of multiple components of the basal transcription machinery, including the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIIF and TFIIH, and the transcriptional mediator complex (reviewed in Rachez and Freedman, 2001).

Transcriptional regulation in *C. elegans*

When *C. elegans* transcription is compared to other eukaryotic organisms, there are two major differences; the ability to *trans*-splice and the arrangement of some genes into operons (Krause and Hirsh, 1987; Zorio *et al.*, 1994). Many of the basics of the transcriptional machinery, like RNA polymerase II and the TATA-binding protein function, appear to be well conserved between *C. elegans* and other species (Bird and Riddle, 1989; Roberts *et al.*, 1987, 1989; Sanford *et al.*, 1983, 1985; Sanicola *et al.*, 1990; Dantonel, *et al.*, 2000; Vanfleteren and Van, 1983; Vanfleteren *et al.*, 1989). While the details of chromatin structure re-organization are not known, proteins like *dpy-27* belong to a family of chromosome-condensation proteins (Chuang *et al.*, 1994), and studies on dosage compensation have provided a link between chromatin structure and transcriptional activity (Meyer, 2000). Additionally, the complexes involved in nucleosome remodeling appear to have been conserved in *C. elegans*. For example, the

nucleosome remodeling and histone deacetylase (NURD) complex antagonizes vulval development (Solari and Ahringer, 2000), which is induced by the Ras signal transduction pathway (see discussion below). Inhibition of Ras signaling occurs in part through the action of the synthetic multivulval (*synMuv*) genes, which comprise two functionally redundant pathways (*synMuvA* and *synMuvB*) (Ferguson and Horvitz, 1989). The *synMuvA* and *synMuvB* pathways function redundantly to recruit or activate a core NURD complex, which has been hypothesized to repress vulval developmental target genes by local histone deacetylation (Solari and Ahringer, 2000).

The gene-specific function of the Mediator as an integrator of transcriptional regulatory signals between multiple inputs and the RNA Polymerase is conserved, and is essential for *C. elegans* development. RNA interference assays have shown that the *CeMed6*, *CeMed7*, and *CeMed10/CeNut2* gene products form two mediator complexes, and both interact with Pol II via its largest subunit. These components are required *in vivo* for the transcriptional activation of several genes, including *ceh-13* and *nhr-2*, during specific stages of development in the worm, but are not required for the expression of two ubiquitously expressed genes, *rps-5* and *sur-5* (Kwon and Lee, 2001).

In addition, *SOP-1/TRAP230* may be a Mediator target of pathways regulating transcriptional response to the Wnt pathway. Widely expressed *sop-1* appears to block action of the Wnt signal transduction pathway, suggesting that its effect must be relieved wherever the Wnt pathway acts (Zhang and Emmons, 2000).

Conservation of *trans*-acting transcriptional regulators in *C. elegans*

Most of the traditional transcription factor families have been identified and characterized in *C. elegans*. In the homeodomain superfamily, members of the HOX, POU, LIM, Paired, and NK subclasses have all been identified (Burglin *et al.*, 1991; Chisholm and Horvitz, 1995; Finney *et al.*, 1988; Herr *et al.*, 1988; Hobert *et al.*, 1998; Hunter and Kenyon, 1995; Okkema and Fire, 1994; Wang *et al.*, 1993; Way and Chalfie, 1988). The zinc finger family (including GATA family members), the helix-loop-helix family, the hormone receptor family, the forkhead family, the bzip family, the ETS family, and a variety of other families of transcription factors are all represented in *C. elegans* (Beitel *et al.*, 1995; Bowerman *et al.*, 1992; Kostrouch *et al.*, 1995; Krause *et al.*, 1990; Labouesse *et al.*, 1994; Miller *et al.*, 1993; Spieth *et al.*, 1991b; reviewed in McGhee and Krause, 1997).

Vulva cell specification and intracellular signaling pathways

Like more complicated organisms, *C. elegans* has a vulva connecting its uterus to the outside world to allow egg laying, and copulation with males. The development of this organ provides an excellent opportunity to study how cell-fate specification is controlled during development.

It is clear that pattern formation of the vulva involves the initiation, integration, and termination of many signals that work in concert to produce a final invariant lineage. In the *C. elegans* vulval ectoderm, at least three known intercellular signaling pathways, the inductive (EGF), lateral (NOTCH), and the WNT pathways, induce six multipotential Vulval Precursor Cells (VPCs) to generate an invariant spatial pattern of cell fates. These signaling pathways stimulate both the division of the VPC cells and the emergence of a

precise pattern (reviewed in Greenwald, 1997; Sternberg and Han, 1998). The VPCs are of three types: 1° and 2° VPCs, which can be distinguished by their division pattern and differential expression of marker genes, and 3° VPCs, which generate non-vulval epidermis (Burdine *et al.*, 1997; Greenwald, 1997; Kimble *et al.*, 1979; Sternberg and Horvitz, 1986; Sulston and Horvitz, 1977). The morphogenetic interactions of the 1° and 2° VPCs lead to the development of seven toroidal cells that connect the endothelium of the uterus to the external epithelium. These seven toroidal cells are the terminally differentiated VPCs: vulF, E, D, C, B2, B1, and A (Figure 1; Sharma-Kishore *et al.*, 1999).

The formation of competent multipotential cells is the first step in vulva formation. The twelve P cells that are present at hatching divide once; the anterior cells become neuroblasts and the posterior cells other than P3-P8.p fuse with the hypodermal syncytium in the L1 stage (Horvitz and Sternberg, 1991). Members of the homeotic gene family, the HOM-C gene cluster, are thought to play a critical role in establishing VPC competency (Clandinin *et al.*, 1997). In loss of function *lin-39* mutants, a Hom-C gene, P3-P8.p cells fuse with the hypodermal syncytium (Maloof and Kenyon, 1998). Since P3-P8.p cells have the ability to assume any of the vulval fates in response to an inductive signal LIN-3, all six cells must be competent to assume these cellular fates, and are considered developmentally equivalent (Katz *et al.*, 1995; Sternberg and Horvitz, 1986; Sulston and White, 1980). Therefore there is no strong intrinsic difference that pre-ordains the cells to a particular fate, and it does not appear as if cell fate specification in the vulva is dependent on some initial bias in competency. If it is not some initial bias

built into the cell that specifies the terminal cell fate, then there must be a mechanism that distinguishes the P3-P8.p cells such that an invariant lineage of cell fates is established.

We know that three signaling pathways, EGF, Notch and Wnt, play a critical role in specifying the cell fate of the Pn.p cells. In a canonical RAS signaling pathway, a growth factor stimulates a receptor tyrosine kinase (RTK) to activate Ras GTPase and the downstream kinases Raf, MEK, and MAP kinase/ERK, ultimately regulating the activities of transcription factors in the nucleus (reviewed in Sternberg and Alberola-Ila, 1998). In *C. elegans*, the receptor-tyrosine kinase LET-23 is stimulated by the growth factor ligand LIN-3 (Aroian *et al.*, 1990; Ferguson and Horvitz, 1985; Ferguson *et al.*, 1987; Hill and Sternberg, 1992; Horvitz and Sulston, 1980). The anchor cell (AC) serves as the source of the inductive signal, LIN-3 (Hill and Sternberg, 1992; Katz *et al.*, 1995; Kimble, 1981). Following stimulation of the RTK, LET-60 RAS activates the downstream kinases LIN-45 (RAF), MEK-2 (MAP kinase kinase) and MPK-1/SUR-1 (MAP kinase) (Church *et al.*, 1995; Han *et al.*, 1993; Kornfeld *et al.*, 1995; Lackner *et al.*, 1994; Wu and Han, 1994; Wu *et al.*, 1995), which ultimately alter the activities of transcription factors like LIN-1 (ETS), LIN-31 (a winged-helix transcription factor), and LIN-25 (a novel protein) (Beitel *et al.*, 1995; Miller *et al.*, 1993; Tan *et al.*, 1998; Tuck and Greenwald, 1995). There are many downstream positive regulators of *let-60 ras* signaling, including *ptp-2* (a SH2-containing protein tyrosine phosphatase), *ksr-1* (a novel protein kinase), *sur-6* (a subunit of the protein phosphatase 2A PPP2A-B), and *sur-8/soc-2* (a novel protein containing a leucine-rich repeat) (Gutch *et al.*, 1998; Kornfeld *et al.*, 1995; Sieburth *et al.*, 1998, 1999; Sundaram and Han, 1995). There are also several downstream negative regulators of EGF pathway, including the synthetic multivulva

genes (synMuv genes), *unc-101*, *sli-1*, *gap-1*, *ark-1* and *sur-5* (Beitel *et al.*, 1990; Clark *et al.*, 1994; Ferguson and Horvitz, 1985, 1989; Gu *et al.*, 1998; Hajnal *et al.*, 1997; Horvitz and Sulston, 1980; Hsieh *et al.*, 1999; Huang *et al.*, 1994; Jongeward *et al.*, 1995; Lee *et al.*, 1994; Lu and Horvitz, 1998; Solari and Ahringer, 2000; Thomas and Horvitz, 1999; Yoon *et al.*, 1995).

In the canonical model for Notch signaling, a number of proteolytic cleavages within NOTCH release the Notch intracellular domain (NICD) from the plasma membrane following ligand binding. This regulated intramembrane proteolysis allows NOTCH to function as a receptor in ligand binding, and also as a signal transducer, since the NICD translocates to the nucleus to directly interact with the DNA binding factor CSL (CBF-1, Suppressor of Hairless, LAG-1, also known as RBP-J) to regulate Notch target genes. In the absence of NICD, CSL acts as a transcriptional repressor (reviewed in Baron *et al.*, 2002). The existence of a lateral (NOTCH) signaling pathway in *C. elegans* vulva development between the VPCs was suggested of multivulva animals, in which all the VPCs adopt vulval fates independent of the inductive pathway (Sternberg, 1988). LIN-12/NOTCH appears to perform two functions during vulval induction that are separated by the phase of the VPC cell cycle (Ambros, 1999). Before completion of the S phase, LIN-12 is thought to inhibit the specification of the 1° fate and maintain the VPCs in an uncommitted state. After completion of the S phase, LIN-12 promotes the specification of the 2° fate. *Anotch*-like mediated *lin-12* signal induces secondary fate (vulA, B1, B2, C, and D), and prevents any two adjacent VPCs from becoming primary (vulE and F; Sternberg, 1988; Sternberg and Horvitz, 1989). It was recently discovered that the MAP kinase phosphatase LIP-1 appears to mediate this lateral inhibition of the

primary fate (Berset *et al.*, 2001). MAP kinase phosphatases inactivate different types of MAP kinases by dephosphorylating the critical phosphotyrosine and phosphothreonine residues of the kinases (Camps *et al.*, 2000). LIP-1 is initially expressed at a low level in all VPCs. The inductive signal is thought to overcome this constitutive inhibition in P6.p to induce the 1° fate, whereas in P5.p and P7.p, LIN-12/NOTCH appears to up-regulate *lip-1* transcription, and this might inactivate MAP kinase and inhibit primary fate specification (Berset *et al.*, 2001). There are both positive regulators (*sup-17*, which encodes a metalloprotease of the ADAM family, and *sel-12*, which encodes presenilin), and negative regulators (*sel-1*, which encodes a novel extracellular protein, and *sel-10*, which encodes an F-box/WD40 repeat-containing protein) of this pathway (Grant and Greenwald, 1996; Hubbard *et al.*, 1997; Levitan and Greenwald, 1995; Sundaram and Greenwald, 1993; Tax *et al.*, 1997; Wen *et al.*, 1997).

The canonical Wnt pathway involves a WNT ligand that stimulates Frizzled (Fz) receptors to antagonize axin and GSK3 and stabilize β -catenin, ultimately regulating the activities of transcription factors of the TCF/LEF family (Cadigan and Nusse, 1997). In *C. elegans*, analysis of the WNT signaling mutants *bar-1* (a β -catenin-related protein) (Eisenmann *et al.*, 1998), *apr-1* (an APC-related protein) (Hoier *et al.*, 2000; Rocheleau *et al.*, 1997), and *mig-1* (which appears to function in many Wnt-mediated processes) (Eisenmann and Kim, 2000; Harris *et al.*, 1996; Thorpe *et al.*, 1997), shows that P4.p–P8.p can fuse instead of adopting the normal 1°, 2°, or 3° fates. Additionally, P5.p–P7.p can adopt the 3° fate instead of the 1° and 2° fates, resulting in too few VPCs adopting induced fates. Maintenance of the Hox gene *lin-39* in VPCs requires *bar-1* and *apr-1*, and cells that lose *lin-39* expression fuse (Eisenmann *et al.*, 1998; Hoier *et al.*,

2000). *lin-39* acts twice in vulval development, first in the L1 stage during generation of the VPCs (Clark *et al.*, 1993; Wang *et al.*, 1993), and later in the L3 stage during adoption of induced cell fates by the VPCs, when LIN-39 protein levels increase in response to activation of the RTK/Ras pathway (Clandinin *et al.*, 1997; Maloof and Kenyon, 1998). These results suggest that a Wnt pathway utilizing MIG-14, BAR-1, and APR-1 is active in the VPCs, and that one target of this pathway is *lin-39*.

Hyperactivation of the Wnt pathway via a *pry-1* (axin homolog) (Korswagen *et al.*, 2002) loss-of-function mutation, or expression of an activated BAR-1 protein, leads to a Muv phenotype in which extra VPCs adopt induced cell fates (Gleason *et al.*, 2002). This indicates that *pry-1* may negatively regulate Wnt signaling in the VPCs, and that hyperactivation of the Wnt pathway may cause cells to adopt vulval fates that would not normally do so. However, the hyper-induced phenotype caused by Wnt pathway hyperactivation is not dependent on signaling through the Ras pathway (Gleason *et al.*, 2002).

In the final step of vulval development, the morphogenetic interactions of the primary and secondary VPCs, which migrate relative to their neighbors generate seven rings of toroidal cells (vulF, E, D, C, B2, B1, and A; Figure 1) that join the endothelium of the uterus to the external epithelium. The vulval muscles are attached to these rings, and specific cell attachments are made to lateral epithelial cells. Finally, the vulva partially everts to block the transit of eggs until it is opened by activation of the vulval muscles (Sharma-Kishore *et al.*, 1999). The genetics behind what drives these morphogenetic interactions is not well understood, and is currently being studied.

Historically, the only way to distinguish that a cell is terminally differentiated in the worm is by use of lineage analysis and observation of morphological changes. The advent of reporter constructs that reflect a particular cell type or fate is invaluable in figuring out cell fate specification, as well as cell termination mechanisms. We have several vulval cell fate-specific markers, that allow us to determine the identity of the vulva cells (Figure 2) (Burdine *et al.*, 1998; Struhl *et al.*, 1993; Williams-Masson *et al.*, 1998). Little is known about the individual roles of these vulva cells following their terminal differentiation, and what cell-specific functions they possess. Formation of the pattern of vulval cell types is likely to depend upon the *cis*-regulatory regions of the transcriptional targets of these intercellular signals in vulval development. The outcome of such differential activation will result in individual cell types. As in vulval development, we know few of the transcriptional regulators that control anchor cell gene expression. The isolation of response elements used by the anchor cell will facilitate biochemical and bioinformatic identification of major transcriptional factors that control cell-specific gene expression.

Genomic regulatory network analysis

It is not known how the inductive signal, lateral signal, and inhibitory signal are integrated on downstream targets resulting in an invariant pattern of cell-fate specification. However, because these signaling pathways are used elsewhere in the animal's development, there must be a vulva-specific response mechanism. Additionally, since the same pathway appears to be used to specify multiple vulval cell fates, there may

be some branch in the pathway, or there may be key regulators that play a role in distinguishing these distinct fate specifications.

While a number of transcription factors are known to be involved in vulval development (e.g. *lin-1*, *lin-29*, *egl-38*, *lin-31*), little is known of their targets or interactions (Beitel *et al.*, 1995; Bettinger *et al.*, 1997; Chang *et al.*, 1999; Euling *et al.*, 1999; Tan *et al.*, 1998). The identification of *cis*-regulatory regions that confer cell specificity and respond to the inductive EGF pathway would be very helpful in determining such relationships. Three such target genes are: a fibroblast growth factor family member, *egl-17* (Figure 3; Burdine *et al.*, 1998); a FAT-like cadherin gene, *cdh-3* (Figure 4; Burdine *et al.*, 1998); and a zinc metalloproteinase gene, *zmp-1* (Figure 5; J. Butler and J. Kramer personal communication). These genes offer the opportunity to find response regions for multiple vulval cell types: vulE, F, C, D, and A, as well as the anchor cell. In addition, *egl-17* is an early cell-fate marker for the response to the inductive signal; the isolation of a *cis*-regulatory element that drives this early expression, and the identification of genes that regulate this expression, would be informative in determining the hierarchy of gene activation in this pathway.

***egl-17* and the FGF family**

The fibroblast growth factor receptor (FGFR) family plays a major role in how cells communicate with their environment. FGFR signaling is crucial for normal development, and its misregulation in human beings is linked to developmental abnormalities, and has been implicated in tumor progression. The cell-cell communication events mediated by

the FGFRs are used for the proper organization of cells into functional units during development (reviewed in Borland *et al.*, 2001).

In *C. elegans*, there are two putative FGFs, *egl-17* and *let-756*, and there is only one putative FGFR, *egl-15*. EGL-17 has been shown to be the instructive guidance cue in the attraction of a pair of bilaterally symmetric sex myoblasts (SMs: that express the EGL-15 FGFR) from the posterior of the animal to their final positions flanking the precise center of the developing gonad (Branda and Stern, 2000a). The SMs then divide and differentiate into the muscles required for egg laying (Sulston and Horvitz, 1977). The loss of function mutation of *egl-17*, *e1313*, has a severe posterior displacement of hermaphrodite sex muscles due to the improper migration of the SMs (Burdine *et al.*, 1998). This displacement of the muscles disrupts the egg laying machinery, and causes the phenotypic bloating that is seen in some animals. In the vulva, *egl-17* is expressed in vulC and vulD as well as the presumptive vulE, and vulF cells (Figure 3). Besides vulva expression, *egl-17::GFP* is expressed in a variety of other tissue types (Burdine *et al.*, 1998). More recently, a reporter construct with an expanded upstream region of 10.5 kb showed additional expression that includes the dorsal uterine (DU) cells of the somatic gonad and, on rare occasions, weak expression was seen in the anchor cell and the ventral uterine cells (Branda and Stern, 2000b). This expanded region of expression has been shown to produce the gonadal attractive cue that could not be explained fully by the expression of EGL-17 in just the vulva cells; animals that do not have vulva cells due to genetic manipulation can position the SMs correctly. The expression in the descendants of P6.p is thought to play a redundant role in the positioning of the SMs. It has been hypothesized that the later expression of EGL-17 in vulC and vulD cells may play a role

in the precise positioning of the attachment of the vulva muscles between these two cells (Branda and Stern, 2000b).

***zmp-1* and the Matrix Metalloproteinases**

The Matrix Metalloproteinase Family, also called the Matrixins, is a family of zinc-dependent metalloendopeptidases, which collectively are capable of degrading essentially all extracellular matrix components. This family has been shown to play critical roles in embryonic development, morphogenesis, reproduction, and tissue resorption and remodeling through the degradation of specific extracellular matrix components (reviewed in Matrisian, 2000). The expression of most matrixins is tightly regulated at the transcriptional level by growth factors, hormones, cytokines and cellular transformation (reviewed in Matrisian, 2000). Three genes encoding novel matrix metalloproteinases (MMPs) were recently identified and cloned by sequence similarity searching of the *Caenorhabditis elegans* genome database (Wada *et al.*, 1998). One of these three MMPs is *zmp-1*.

In *C. elegans*, a complete dissection of the expression pattern of the zinc metalloproteinase, *zmp-1*, has not been done. However, in hermaphrodites, in addition to vulA, vulE and anchor cell expression (Figure 4), it is expressed in a variety of other cell types from multiple lineages, including uterine and tail cells. The deletion of *zmp-1*, *cg115*, has no apparent phenotype and overexpression of ZMP-1 leads to a slight general degradation of the extracellular matrix components (J. Butler and J. Kramer, personal communication). While the role of this gene is unclear, it is interesting to note that at the time of ZMP-1 expression in the anchor cell, vulE and vulA there seem to be functional

rearrangements of the ECM, which must take place such that: the anchor cell can fuse with the vulF cells; vulE cells can attach to lateral epithelial seam cells; and the vulA cells can make junctions with the syncytial hypodermal cell, *hyp7*.

***cdh-3* and the Cadherins**

A third family of genes, the Cadherin superfamily of cell adhesion molecules, is involved in multiple morphogenetic events in animal development. Specifically, the Cadherin family plays a role in epithelial morphogenesis that is dependent upon coordinated control of changes in cell shape, proliferation, recognition and adhesion (reviewed in Tepass, 1999). It is a large family with many sub-groups that are divided by characteristic protein domains. Cadherin superfamily genes encode variable numbers of an extracellular domain termed the cadherin domain. These domains mediate intermolecular interactions and are dependent on calcium ions, which bind at sites between adjacent cadherin domains to produce a rigid structure. The extracellular domains are linked via a transmembrane helix to a cytoplasmic domain, which is known in some cases to interact with certain classes of intracellular proteins (reviewed in Tepass, 1999).

There are twelve predicted cadherin superfamily members in *C. elegans*. Of these, only *hmr-1* and *cdh-3* have been defined by experimental work on their structure and function (Hill *et al.*, 2001). CDH-3 is a member of the FAT-like cadherin sub-group. FAT-like cadherins are very large proteins with multiple cadherin domains, EGF-like, and laminin-AG domain repeats. It remains unclear whether the FAT-like cadherins operate in adhesion, signaling or both. The FAT-like cadherin family is predominantly expressed in epithelial cells (Hill *et al.*, 2001). In hermaphrodites, *cdh-3::GFP* is

expressed in the seam cells, the buccal and rectal epithelia, the excretory cell, two hypodermal cells in the tail, the uterine epithelium closest to the invaginating vulval cells followed by the multinucleated uterine seam cell (utse), the developing vulva, and associated neurons. Specifically, in the vulva, the reporter construct is expressed in vulA, E, F, C and D, as well as the anchor cell (Figure 5; Pettitt *et al.*, 1996). In *C. elegans* it is clear that CDH-3 is required for the morphogenesis of a single cell that forms the tip of the tail in the hermaphrodite. The other cells that express the *cdh-3* reporter appear to be unaffected by a probably null allele, raising the possibility that other genes can compensate for the loss of CDH-3 (Pettitt *et al.*, 1996). The genesis of the egg-laying system requires several sets of cell-cell recognition events, all of which occur during the expression of *cdh-3::GFP*. First, the anchor cell must invaginate between the two vulF cells, an event that takes place soon after GFP expression is observed in the cells involved. Second, the vulval epidermal cells must invaginate and form a connection with the uterus, and third the utse cell must make contacts with the seam cells. In addition, during the formation of the seven toroidal rings of the vulva, the vulva cells interact with one another (Pettitt *et al.*, 1996).

Regulatory analysis in *C. elegans*

A detailed analysis of *cis*-regulatory elements has been performed for only a few *C. elegans* genes. Like other multicellular organisms it appears that there are a variety of regulatory mechanisms. Genes, such as the vitellogenin gene *vit-2* (MacMorris *et al.*, 1992), the myosin gene *myo-2* (Okkema and Fire, 1994), the cuticle gene *dpy-7* (Gilleard *et al.*, 1997), the NK-2 homeobox gene *ceh-24* (Harfe and Fire, 1998), and the

acetylcholinesterase gene *ace-1* (Culetto *et al.*, 1999), are regulated in a relatively simple fashion by a tissue-specific basal promoter whose activity is enhanced by separate activator elements that can lie in the promoter, or within an intronic sequence (see discussion below). Other genes, such as the carboxylesterase gene *ges-1* (Egan *et al.*, 1995) and *mec-3* (Wang and Way, 1996b), require both activator and repressor elements to establish proper expression (see discussion below).

Upstream sequences of *dpy-7* were characterized in *C. elegans* by comparing the entire intergenic region to *C. briggsae* using a dot-matrix comparison. A single region of homology, 147 bp, was isolated. This corresponds with the minimal functional promoter region defined by deletion analysis in *C. elegans*. When 1kb of upstream sequences, and the *C. briggsae dpy-7* homolog were injected into a *dpy-7 C. elegans* strain, rescue was observed. Additionally, when two translational fusions of the *C. elegans dpy-7* gene (one with and one without the region of homology) were injected, only the translational fusion containing this region showed expression in *C. briggsae*. Contained in this conserved region is a predicted GATA site transcription factor, but no further experiments were performed to decipher a potential role for GATA factor transcription in the regulation of the *dpy-7* gene. These results provide evidence that regulated tissue- and stage-specific expression of *dpy-7* is achieved by a compact tissue-specific promoter element close to the 5' end of the gene, and appears to involve no repressor elements (Gilleard *et al.*, 1997).

The myosin heavy chain *myo-2* gene contains at least two independent tissue-specific regulatory elements: a promoter sufficient for low-level expression in the pharyngeal muscle-specific expression is located near the transcriptional start site, and a

separable pharyngeal muscle-specific enhancer, 395 bp, located 300 bp upstream of the start site. This enhancer, which can induce pharyngeal muscle expression from a *myo-3::lacZ* fusion, involves at least three sub-elements that cooperate to activate transcription, two of which display distinct cell-type specificity (one for the whole pharynx, and two for a subset of pharyngeal cells). While individually, each of these subelements is inactive, any combination of two can drive transcription. Additionally, duplication of any of these elements is also sufficient to drive pharyngeal expression. Therefore, each of the subelements contains sufficient information to confer tissue-specific expression. Each subelement appears to contain multiple sites, as demonstrated by mutational analysis of each of these regions. Using a cDNA library, a *ceh-22* cDNA, which specifically binds one of the subelements, was identified (Okkema and Fire, 1994). Again, in this analysis, the transcriptional regulation of this gene appears to be regulated by multiple, discrete positive-acting elements. Subsequent studies have revealed that the organ-specific enhancer region contains a binding site for PHA-4 (Kalb *et al.*, 1998), a forkhead factor essential for pharyngeal development (Horner *et al.*, 1998; Kalb *et al.*, 1998; Mango *et al.*, 1994), and a binding site for DAF-3, which is a SMAD factor (Thatcher *et al.*, 1999). DAF-3, a negative regulator, is unlikely to modulate the organ specificity of this enhancer since a *daf-3* mutation does not affect the pharyngeal-specific expression pattern, or result in any pharyngeal defects, and may act to downregulate *myo-2* expression under as yet undescribed circumstances (Thatcher *et al.* 1999).

Similar experiments on the *ceh-24* upstream sequence revealed three distinct, separable tissue-specific enhancers for head neurons (57 bp), vulva muscles (48 bp) and the pharyngeal m8 cell (117 bp; Harfe and Fire, 1998).

The three previous examples demonstrate the relative simplicity of a handful of upstream *cis*-regulatory elements, which all act in a positive fashion to confer tissue-specific regulation. The following examples will show that not all promoters are as straightforward, and that, indeed, regulatory regions in *C. elegans* may contain both activator and repressor elements. Upon analysis of the carboxyesterase gene *ges-1*, it was shown that in particular deletions, it was expressed not in the gut (the E lineage, where normal expression is seen), but rather in muscle cells of the pharynx (which belong to a sister lineage of the gut, the MS lineage) and in body wall muscle and hypodermal cells (which belong to a cousin lineage of the gut). This 200-bp region responsible for the switch of expression from the E lineage to other lineages contains two binding sites for GATA factors, which have been subsequently shown to bind this sequence. Interestingly, when either of the two GATA sites or an adjacent sequence is eliminated, expression remains in the E lineage, but is restricted to a subset of cells, indicating that both of these sites are required for full expression in the gut. When any two of these three regions are eliminated, the switch to the MS lineage occurs and, when all three are eliminated, the vast majority of expression in all tissues is lost. These observations suggest that gut-specific gene expression in *C. elegans* involves not only gut-specific activators, but also multiple repressors that are present in particular non-gut lineages (Aamodt *et al.*, 1991; Egan *et al.*, 1995; Kennedy *et al.*, 1993). Subsequent studies have proposed a model in which the normal E lineage gut expression of *ges-1* is controlled by the gut-specific GATA factor such as ELT-2, while the pharynx and rectum (MS lineage) expression is controlled by PHA-4, which is normally bound to the *ges-1* 3' enhancer sequences. The

activation of PHA-4 is kept repressed by an unknown factor binding in the vicinity of the GATA factor binding sites (Marshall and McGhee, 2001).

The 10 neurons involved in mechanosensation in *C. elegans* express *mec-3*. The expression is maintained by autoregulation. Four conserved regions, each of 24-70 bp, were identified by intraspecies comparisons to *C. vulgarensis*. The downstream region (528 bp), which includes conserved blocks I, II and III, appear to mediate establishment of the expression pattern. An additional, more distal element (917 bp), also appears sufficient to establish *mec-3* expression. Mutations in region I, III and IV can all cause transient ectopic expression of the *mec-3::lacZ* fusions in some sister cells of the normal *mec-3* expressing cells. UNC-86 binding sites have been identified in conserved regions I, II and III of the 5' flanking sequence. (In an *unc-86* background, the cells that normally express *mec-3* are not specified to the correct terminal fate). However, it seems unlikely that the binding sites for UNC-86 are the sole players in this very complex upstream region (Wang and Way, 1996a,b; Way and Chalfie, 1988; Way *et al.*, 1991; Xue *et al.*, 1992, 1993).

Although *cis*-regulatory analysis has been performed on only a handful of upstream regions in *C. elegans*, it has been suggested that the complex regulation, particularly involving repressor elements, might be a general feature of transcriptional control in those genes expressed prior to cellular differentiation (Krause *et al.*, 1994). Genes that encode abundant structural proteins may be regulated in a simpler manner (Gilleard *et al.*, 1997). This simplicity may be an important feature of the transcription of large multigene families, or of genes that are transcribed following cellular differentiation. However, the DAF-3 binding studies on the *myo-2* enhancer serve as a

cautionary reminder that expression studies examine only one set of conditions. Under different conditions, repressor or activator activity may be utilized. They also demonstrated that in *C. elegans*, there are enhancers that function in all cell types of a tissue, and that these elements are not mutually exclusive from those that act in a distinct subtype of cells in this same tissue (Thatcher *et al.*, 1999). The *ceh-24* studies delineate that multiple modules, all apparently positive acting, may regulate tissue specificity in a variety of tissues that are not related by lineage (Harfe and Fire, 1998). These are just some of the complexities of transcriptional regulation in *C. elegans* that have been revealed to us so far. In other model organisms, such simplicity is almost unheard of. Which begs the question, “Is transcriptional regulation in *C. elegans* just that much simpler, or are we just not in deep enough to reveal all the layers of complexity that are seen in these other systems?”

An example of the complexity seen in other systems is the regulation of CD4 gene silencing expression during T-cell development. When three copies of the murine silencer were linked to a CAT reporter vector regulated by one of the CD4 enhancers and the CD4 promoter, expression of CAT was specifically repressed in CD4-CD8+, but not in CD4+CD8+ T cells. Using this system as an assay, a core 134 bp fragment was defined, which in triplicate reduced transcription 10-to 20-fold. This core silencer worked better than the larger fragment defined in transfection studies, but it had no silencing activity in transgenic mice. When flanking 5' or 3' sequences were added back to this core fragment, silencer activity was restored in the transgenic constructs. This functional redundancy of the flanking sequences in animals, and their dispensability in transient transfection studies, suggest that these flanking sequences contain elements needed for organizing the

chromatin structure to allow access of *trans*-acting factors to the silencing elements. When internal deletions were made in the core region, one of three outcomes was observed: (1) silencing, (2) no silencing, or (3) a variegation of silencing. The variegation suggested that, in many cases, the loss of a single nuclear factor binding site would not completely inactivate the silencer, but would decrease the probability of the establishment of silencing. A conclusion from these studies is that what may appear to be crucial, the 134-bp core fragment, may not be the whole story of elements involved in a gene's native transcriptional regulation. In addition, this is just one region that plays a role in CD4 gene transcription: two enhancers, a core promoter, and at least one other element in an intron have been implicated in the fidelity of the expression pattern (review in Ellmeier *et al.*, 1999).

Dissection of co-regulated genes

A common assumption in the modeling of genetic regulatory networks is that the cell-specific genes expressed in a given terminally differentiated cell type are likely to be subject to coordinate control, and hence possess similar upstream *cis*-acting sequences (Davidson, 2001). While some attempts to validate this assumption in *C. elegans* have failed, other studies have succeeded. A comparison of the cuticle gene *dpy-7*'s 5' flanking sequences with other *C. elegans* cuticle genes did not reveal any striking regions of similarity (Gilleard *et al.*, 1997). A dot-matrix comparison of two acetylcholinesterase genes, *ace-1* and *ace-2*, failed to show any similarities between the two promoters (Culetto *et al.*, 1999). And the comparison of *C. elegans* MyoD family member *hll-1* to

mouse myogenic regulatory factors presented no striking similarities between these promoters (Krause *et al.*, 1994).

One success story is that of the vitellogenin genes. There are six *C. elegans* vitellogenin genes that are subject to sex-, stage-, and tissue-specific regulation: they are expressed solely in the adult hermaphrodite intestine. Comparative sequence analysis of upstream sequences of these genes and their *C. briggsae* homologs revealed the presence of two repeated heptameric elements, vit promoter element 1 (VPE1) and VPE2. A functional analysis of the VPEs within the 5'-flanking region of the *vit-2* gene revealed that a 247 bp element containing the VPEs was sufficient for high-level, regulated expression. Furthermore, none of the four deletion mutations resulted in inappropriate expression (Blumenthal *et al.*, 1984; Spieth *et al.*, 1985, 1991a; Zucker-Aprison and Blumenthal, 1989).

Since every cell in the worm may have a unique identity at the molecular level, the use of a battery of cell type-specific markers might allow the identification of any common upstream element(s) responsible for driving expression in a specific cell or cell type. Indications that this type of analysis might work in *C. elegans* have started to appear. A comparison of the minimal promoters of *mtl-1* and *mtl-2* to other *C. elegans* intestinal cell-specific genes identified repeats of GATA transcription factor-binding sites. Mutation analyses determined that GATA elements are required for transcription, while electrophoretic mobility shift assays showed that ELT-2, a *C. elegans* GATA transcription factor, specifically binds these element. Furthermore, when *elt-2* is disrupted in *C. elegans*, *mtl-2* is not expressed. It was also shown that ectopic expression of ELT-2 can activate transcription of *mtl-2* in non-intestinal cells of *C. elegans*. These

results suggest that the binding of ELT-2 to GATA elements in these promoters regulates tissue-specific transcription of the *C. elegans* metallothionein genes (Moilanen *et al.*, 1999).

Another success story was the *C. elegans* gene *daf-19*, which encodes an RFX-type transcription factor that is expressed specifically in all ciliated sensory neurons (Swoboda *et al.*, 2000). Loss of *daf-19* function causes the absence of cilia, resulting in sensory defects. Twenty *C. elegans* promoters of genes that are expressed in ciliated sensory neurons were searched for X boxes. (X boxes are the mammalian targets for RFX-type transcription factors.) Target sites were found within the promoters of four of these genes, *che-2*, *daf-19*, *osm-1* and *osm-6*, which are expressed in most or all ciliated sensory neurons. Target sites were not found in the promoter regions of any of the genes that are expressed in only a subset of ciliated sensory neurons, e.g., *gcy-5*, *gcy-8* and *gcy-32*. Using an *in vivo* assay, it was shown that expression of the X box-containing genes was dependent on both *daf-19* function and the presence of the promoter X box. In a genome-wide search for X-box-containing genes, a novel gene was examined and found to be expressed in ciliated sensory neurons in a *daf-19*-dependent manner. These data suggest that *daf-19* is a transcriptional regulator of gene products that function broadly in sensory cilia (Swoboda *et al.*, 2000). To date, there are no studies that have looked at the co-regulation of genes at the cell-specific, rather than tissue-specific, level.

One of the fallbacks of this type of analysis is that assumptions have to be made on what genes may constitute a group of co-regulated genes. Groupings of co-regulated genes based on family function are not necessarily going to lead to the identification of a common element(s). The advent of microarray analysis and SAGE techniques will make

the determination of cohorts of co-regulated genes easier to identify. In a recent study, the expression pattern of 11,917 genes from *C. elegans* were monitored using microarrays to determine which of these genes was upregulated in response to heat-shock treatment. The upstream regions of the 28 genes that appeared to be upregulated by greater than four fold in response to heat-shock were examined using several computational and statistical methods. The resulting two heat-shock elements (HSE) were conserved in the upstream regions of the *C. briggsae* orthologs of the *C. elegans* genes. Upon mutational analysis of the *hsp-16-2::GFP*, these elements were found to be neither necessary nor sufficient, but did have an effect on the strength of the GFP expression, indicating that this type of element may be hard to isolate using the traditional experimental methods such as systematic deletion (GuhaThakurta *et al.*, 2002). In another recent study, *C. elegans* touch-receptor cells were cultured and used for microarray analysis. The culturing of these cells enabled the sensitivity of the microarray data to be increased, so that *mec-3*-dependent genes could be identified (there are only six touch-receptor cells in the worm). Using the 5' regions of genes that were significantly enriched in this analysis, Zhang *et al.* were able to determine that a heptanucleotide element was over-represented in this population (Zhang *et al.*, 2002). However, the functional significance of this element has not been shown. These are the first steps in a very promising future of experiments. The isolation of subpopulations of cells and microarray analysis will allow the identification of overrepresented upstream elements that are specific to a cellular function, or a specific cell type. However, what this technology does not ensure is the identification of all the important sequences involved in the fidelity of the expression pattern.

Phylogenetic footprinting

With whole genome sequences becoming readily available, and with the failure of de novo computational programs to recognize functional motifs in *cis*-regulatory regions (Loots *et al.*, 2000; Pennacchio and Rubin, 2001), there is a growing interest in comparing genome sequences to identify regulatory regions (Stojanovic *et al.*, 1999). Phylogenetic footprinting is a method for the identification of regulatory elements in a set of orthologous regulatory regions from multiple species; it does so by identifying the best-conserved motifs in those orthologous regions (Tagle *et al.*, 1988).

To see the real power of this technique, examine the studies performed on the human *epsilon-globin* gene, which undergoes dramatic changes in transcriptional activity during development. Elucidation of the mechanisms that govern these interactions could suggest strategies to reactivate fetal (*gamma*) or embryonic (*epsilon*) genes in individuals with severe hemoglobinopathies. The expression pattern of the *epsilon-globin* gene is conserved in all placental mammals. The *epsilon-globin* sequences from seven mammalian species- human, orangutan, gibbon, capuchin, monkey, galago, and rabbit- were used to compare the upstream regulatory regions of this gene. The total number of evolutionary years included in such an alignment is additive. Since the evolutionary time of these species is greater than 270 million years, nucleotide sequences have had ample time to accumulate changes. Twenty-one conserved elements were identified in the 2 kb of sequence immediately upstream of the coding region of the *epsilon* gene. Probes spanning each of these footprints bound proteins in gel-shift assays. Among the 47 binding interactions characterized were: eight sites for the yin and yang 1 (a protein

shown to have both activator and repressor properties); five binding sites for a putative stage-selective protein SSP; and seven sites for an as-yet-unidentified protein (Gumucio *et al.*, 1993). Such studies allow for an unbiased selection of factors involved in the transcriptional regulation of this gene, which speaks neither to the sufficiency nor the necessity of the individual factors, but rather to a more global picture of the milieu of the elements and factors involved.

For this type of analysis to be fruitful, the genomes that are used must be selected carefully. Comparison with too-closely related genome will reveal shared conservation in non-functional areas. However, if the comparison is performed on a species that is too-distantly related, the genomes will likely lack the conservation needed to be informative. Studies in bacteria and animals have suggested that a slightly less-diverged species is a better choice when looking for the conservation of *cis*-regulatory elements (Cargill *et al.*, 1999; Huynen and Bork, 1998).

Despite having diverged from each other an estimated 50-120 million years ago (Coghlan, 2002), both *C. elegans* and *C. briggsae* share almost identical development and morphology (Nigon and Dougherty, 1949). Cross-species rescue of mutant phenotypes has demonstrated that there is functional conservation between the two species (Culetto *et al.*, 1999; de Bono and Hodgkin, 1996; Kennedy *et al.*, 1993; Krause *et al.*, 1994; Kuwabara, 1996; Maduro and Pilgrim, 1996). This should not be taken to mean that all homologs will function and be expressed in a similar fashion between the two species. For instance, at least one aspect of the *hlh-1* gene's regulation, a homolog of the MyoD family of myogenic regulatory factors, differs between the two species. The *C. elegans* *hlh-1* is expressed in the MS-granddaughter cells during embryogenesis, while this

expression is not detected by *lacZ* reporter constructs and antibody staining in *C. briggsae* (Krause *et al.*, 1994). Despite this, the two almost completely sequenced genomes make *C. briggsae* an obvious choice for genome comparisons to *C. elegans*. The analysis of similarity within 142 pairs of orthologous intergenic regions shows regions of high similarity interspersed with non-alignable sequence (Webb *et al.*, 2002). The high degree of similarity in some of these regions suggests that they have undergone selective pressure. Such intergenic conservation between *C. elegans* and *C. briggsae* has been utilized in a handful of studies to isolate putative binding sites for *trans*-acting regulatory factors.

Upstream sequences from *ace-1* were compared to the orthologous *C. briggsae* gene by dot -matrix comparison. This analysis revealed four blocks (35, 58, 140 and 409 bp) of conserved sequence. These blocks were between 70-80% identical between species. The first block contained splicing site sequences and alternative splice-sites, indicating that this region was probably part of the minimal promoter. (Interestingly, it is devoid of TATA and CAAT boxes.) To test whether the other conserved sequences could qualitatively modulate the basal activity of the promoter, a CAT reporter gene expression system in mammalian cell lines was used. Two of the conserved blocks did not affect transcriptional activity, whereas one block in this system acted as a transcriptional repressor. However, in expression studies, the block that was found to repress CAT reporter gene expression was involved in driving expression in the body wall and anal muscle cells, and the two blocks that did not effect expression levels were also required for expression in other areas of the animal. Additionally, the conserved region that appeared to be a repressor in the CAT system, when combined with the minimal

promoter element, was sufficient to drive expression in body wall and anal muscles. These data suggest that *cis*-regulatory sequences of *C. elegans* are not recognized in the same way as in the transcriptional apparatus of the mouse cells. Intra-species comparisons with *C. briggsae* were able to identify the important *cis*-regulatory regions of this gene, but were unable to isolate distinct factor binding sites (Culetto *et al.*, 1999)

In *ceh-24* upstream sequences, *C. briggsae* was used in a species comparison to confirm the importance of a pair of NdE-boxes and the m8 pharyngeal cell enhancer. Intra-species comparison did not reveal any additional binding sites (Harfe and Fire, 1998).

Studies of the gut esterase gene, *ges-1* (discussed above), illuminate the benefits and risks of intra-species comparison studies between *C. elegans* and *C. briggsae*. A 17-bp region of conservation between the *C. elegans* and *C. briggsae* 5' flanking sequences was found, but deletion of this element had no effect on the expression pattern of the reporter transgene (Egan *et al.*, 1995). It is likely that not all conserved sequences between these two species will have a functional significance. On the other hand, an important binding site located in the 3' flanking regions of the coding sequence of this gene was identified using the comparison between these two species. This binding site, critical to the regulation of the *ges-1* gene in the pharynx and rectum, had not been found by conventional deletion analysis (Marshall and McGhee, 2001).

Thesis overview

In chapter one of this thesis, I analyze the *cis*-regulatory sequence regions sufficient to confer vulva cell- and anchor cell- specific expression of three putatively co-regulated

genes: *zmp-1*, *egl-17* and *cdh-3*. These genes are expressed in a restricted and overlapping expression pattern in specific vulva cell types and the uterine anchor cell within *C. elegans*. We chose these genes because their function is not required for the normal development of the cells in which they are expressed, and hence they lie downstream of the cell-fate-specification pathways.

In chapter two, I used an orthogonal approach to isolate vulva- and anchor cell-specific elements. I have identified the *C. briggsae* homologs of these three genes and used phylogenetic footprinting to identify the predicted control regions corresponding to the sufficiency regions identified in *C. elegans*. Together, these two approaches elucidate similar elements that are sufficient to confer expression to a subset of vulval cells and the uterine anchor cell.

REFERENCES

- Aamodt, E., Chung, M., and McGhee, J. (1991). Spatial control of gut-specific gene expression during *Caenorhabditis elegans* development. *Science* **252**, 579-582.
- Ambros, V. (1999). Cell cycle-dependent sequencing of cell fate decisions in *Caenorhabditis elegans* vulva precursor cells. *Development* **126**, 1947-1956.
- Aroian, R., Koga, M., Mendel, J., Ohshima, Y., and Sternberg, P. (1990). The *let-23* gene necessary for *Caenorhabditis elegans* vulval induction encodes a tyrosine kinase of the EGF receptor subfamily. *Nature* **348**, 693-699.
- Baron, M., Aslam, H., Flaszka, M., Fostier, M., Higgs, J., Mazaleyrat, S., and Wilkin, M. (2002). Multiple levels of Notch signal regulation (review). *Mol Membr Biol* **19**, 27-38.
- Beitel, G., Clark, S., and Horvitz, H. (1990). *Caenorhabditis elegans* ras gene *let-60* acts as a switch in the pathway of vulval induction. *Nature* **348**, 503-509.

- Beitel, G., Tuck, S., Greenwald, I., and Horvitz, H. (1995). The *Caenorhabditis elegans* gene *lin-1* encodes an ETS-domain protein and defines a branch of the vulval induction pathway. *Genes & Development* **9**, 3149-3162.
- Berset, T., Hoier, E., Battu, G., Canevascini, S., and Hajnal, A. (2001). Notch inhibition of Ras signaling through MAP kinase phosphatase LIP-1 during *C. elegans* vulval development. *Science* **291**, 1055-1058.
- Bettinger, J., Euling, S., and Rougvie, A. (1997). The terminal differentiation factor LIN-29 is required for proper vulval morphogenesis and egg laying in *Caenorhabditis elegans*. *Development* **124**, 4333-4342.
- Bird, D., and Riddle, D. (1989). Molecular cloning and sequencing of *ama-1*, the gene encoding the largest subunit of *Caenorhabditis elegans* RNA polymerase II. *Molecular and Cellular Biology* **9**, 4119-4130.
- Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, M., Spieth, J., and Sharrock, W. (1984). Cloning of a yolk protein gene family from *C. elegans*. *Journal of Molecular Biology* **174**, 1-18.
- Borland, C., Schutzman, J., and Stern, M. (2001). Fibroblast growth factor signaling in *Caenorhabditis elegans*. *Bioessays* **23**, 1120-1130.
- Bowerman, B., Eaton, B., and Priess, J. (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.
- Branda, C., and Stern, M. (2000). Mechanisms controlling sex myoblast migration in *Caenorhabditis elegans* hermaphrodites. *Developmental Biology* **226**, 137-151.
- Burdine, R., Branda, C., and Stern, M. (1998). EGL-17(FGF) expression coordinates the attraction of the migrating sex myoblasts with vulval induction in *C. elegans*. *Development* **125**, 1083-1093.
- Burdine, R., Chen, E., Kwok, S., and Stern, M. (1997). *egl-17* encodes an invertebrate fibroblast growth factor family member required specifically for sex myoblast migration in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences USA* **94**, 2433-2437.
- Cadigan, K., and Nusse, R. (1997). Wnt signaling: a common theme in animal development. *Genes & Development* **11**, 3286-3305.

- Camps, M., Nichols, A., and Arkinstall, S. (2000). Dual specificity phosphatases: a gene family for control of MAP kinase function. *FASEB J* **14**, 6-16.
- Carey, M. (1998). The enhanceosome and transcriptional synergy. *Cell* **92**, 5-8.
- Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C., Lim, E., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G., and Lander, E. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**, 231-238.
- Chang, C., Newman, A., and Sternberg, P. (1999). Reciprocal EGF signaling back to the uterus from the induced *C. elegans* vulva coordinates morphogenesis of epithelia. *Current Biology* **9**, 237-246.
- Chuang, P.-T., Albertson, D., and Meyer, B. (1994). DPY-27: A chromosome condensation protein homolog that regulates *C. elegans* dosage compensation through association with the X chromosome. *Cell* **79**, 459-474.
- Church, D., Guan, K., and Lambie, E. (1995). Three genes of the MAP kinase cascade, *mek-2*, *mpk-1/sur-1* and *let-60* ras, are required for meiotic cell cycle progression in *Caenorhabditis elegans*. *Development* **121**, 2525-2535.
- Clandinin, T., Katz, W., and Sternberg, P. (1997). *Caenorhabditis elegans* HOM-C genes regulate the response of vulval precursor cells to inductive signal. *Developmental Biology* **182**, 150-161.
- Clark, S., Chisholm, A., and Horvitz, H. (1993). Control of cell fates in the central body region of *C. elegans* by the homeobox gene *lin-39*. *Cell* **74**, 43-55.
- Clark, S., Lu, W., and Horvitz, H. (1994). The *Caenorhabditis elegans* locus *lin-15*, a negative regulator of a tyrosine kinase signaling pathway, encodes two different proteins. *Genetics* **137**, 987-997.
- Coghlan, A. W. K. (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genomic Research* **12**, 857-867.
- Culetto, E., Combes, D., Fedon, Y., Roig, A., Toutant, J., and Arpagaus, M. (1999). Structure and promoter activity of the 5' flanking region of *ace-1*, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *Journal of Molecular Biology* **290**, 951-966.

- Dantanel, J.C., Quintin, S., Lakatos, L., Labouesse, M., Tora, L. (2000). TBP-like factor is required for embryonic RNA polymerase II transcription in *C. elegans*. *Molecular Cell* **6**, 715-722.
- Davidson, E. H. (2001). "*Genomic Regulatory Systems: Development and Evolution*." Academic Press,
- de Bono, M., and Hodgkin, J. (1996). Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. *Genetics* **144**, 587-595.
- Egan, C., Chung, M., Allen, F., Heschl, M., Vanbuski, C., and McGhee, J. (1995). A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans ges-1* gene centers on two GATA sequences. *Developmental Biology* **170**, 397-419.
- Eisenmann, D., and Kim, S. (2000). Protruding vulva mutants identify novel loci and Wnt signaling factors that function during *Caenorhabditis elegans* vulva development. *Genetics* **156**, 1097-1116.
- Eisenmann, D., Maloof, J., Simske, J., Kenyon, C., and Kim, S. (1998). The B-catenin homolog BAR-1 and LET-60 Ras coordinately regulate the Hox gene *lin-39* during *Caenorhabditis elegans* vulval development. *Development* **125**, 3667-3680.
- Ellmeier, W., Sawada, S., and Littman, D. (1999). The regulation of CD4 and CD8 coreceptor gene expression during T cell development. *Annual Review of Immunology* **17**, 523-554.
- Euling, S., Bettinger, J., and Rougvie, A. (1999). The LIN-29 transcription factor is required for proper morphogenesis of the *Caenorhabditis elegans* male tail. *Developmental Biology* **206**, 142-156.
- Felsenfeld, G. (1996). Chromatin unfolds. *Cell* **86**, 13-19.
- Ferguson, E., and Horvitz, H. (1985). Identification and characterization of 22 genes that affect the vulval cell lineages of the nematode *C. elegans*. *Genetics* **110**, 17-72.
- Ferguson, E., and Horvitz, H. (1989). The multivulva phenotype of certain *Caenorhabditis elegans* mutants results from defects in two functionally redundant pathways. *Genetics* **123**, 109-121.

- Ferguson, E., Sternberg, P., and Horvitz, H. (1987). A genetic pathway for the specification of the vulval cell lineages of *C. elegans*. *Nature* **326**, 259-267.
- Gilleard, J., Barry, J., and Johnstone, I. (1997). *cis* regulatory requirements for hypodermal cell-specific expression of the *Caenorhabditis elegans* cuticle collagen gene *dpy-7*. *Molecular and Cellular Biology* **17**, 2301-2311.
- Gleason, J., Korswagen, H., and Eisenmann, D. (2002). Activation of Wnt signaling bypasses the requirement for RTK/Ras signaling during *C. elegans* vulval induction. *Genes Dev* **16**, 1281-1290.
- Grant, B., and Greenwald, I. (1996). The *Caenorhabditis elegans sel-1* gene, a negative regulator of *lin-12* and *glp-1*, encodes a predicted extracellular protein. *Genetics* **143**, 237-247.
- Greenwald, I. (1997). Development of the Vulva in: "*C. elegans II.*" DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press. **II**, 519-541.
- Grosschedl, R., and Birnstiel, M. (1980). Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc Natl Acad Sci U S A* **77**, 1432-1436.
- Gu, T., Orita, S., and Han, M. (1998). *Caenorhabditis elegans* SUR-5, a novel but conserved protein, negatively regulates LET-60 ras activity during vulval induction. *Molecular and Cellular Biology* **18**, 4556-4564.
- GuhaThakurta, D., Palomar, L., Stormo, G., Tedesco, P., Johnson, T., Walker, D., Lithgow, G., Kim, S., and Link, C. (2002). Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res* **12**, 701-712.
- Gumucio, D., Shelton, D., Bailey, W., Slightom, J., and Goodman, M. (1993). Phylogenetic footprinting reveals unexpected complexity in *trans* factor binding upstream from the *epsilon-globin* gene. *Proc Natl Acad Sci U S A* **90**, 6018-6022.
- Gutch, M., Flint, A., Keller, J., Tonks, N., and Hengartner, M. (1998). The *Caenorhabditis elegans* SH2 domain-containing protein tyrosine phosphatase

- PTP-2 participates in signal transduction during oogenesis and vulval development. *Genes & Development* **12**, 571-585.
- Hajnal, A., Whitfield, C., and Kim, S. (1997). Inhibition of *Caenorhabditis elegans* vulval induction by *gap-1* and by *let-23* receptor tyrosine kinase. *Genes & Development* **11**, 2715-2728.
- Han, M., Golden, A., Han, Y., and Sternberg, P. (1993). *C. elegans lin-45* raf gene participates in *let-60* RAS-stimulated vulval differentiation. *Nature* **363**, 133-140.
- Harfe, B., and Fire, A. (1998). Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in *Caenorhabditis elegans*. *Development* **125**, 421-429.
- Harris, J., Honigberg, L., Robinson, N., and Kenyon, C. (1996). Neuronal cell migration in *C. elegans*: regulation of Hox gene expression and cell position. *Development* **122**, 3117-3131.
- Hill, E., Broadbent, I., Chothia, C., and Pettitt, J. (2001). Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *Journal of Molecular Biology* **305**, 1011-1024.
- Hill, R., and Sternberg, P. (1992). The gene *lin-3* encodes an inductive signal for vulval development in *C. elegans*. *Nature* **358**, 470-476.
- Hoier, E., Mohler, W., Kim, S., and Hajnal, A. (2000). The *Caenorhabditis elegans* APC-related gene *apr-1* is required for epithelial cell migration and Hox gene expression. *Genes & Development* **14**, 874-886.
- Horner, M., Quintin, S., Domeier, M., Kimble, J., Labouesse, M., and Mango, S. (1998). *pha-4*, an HNF-3 homolog, specifies pharyngeal organ identity in *Caenorhabditis elegans*. *Genes & Development* **12**, 1947-1952.
- Horvitz, H., and Sternberg, P. (1991). Multiple intercellular signaling systems control the development of the *Caenorhabditis elegans* vulva. *Nature* **351**, 535-541.
- Horvitz, H., and Sulston, J. (1980). Isolation and genetic characterization of cell-lineage mutants of the nematode *C. elegans*. *Genetics* **96**, 435-454.
- Hsieh, J., Liu, J., Kostas, S., Chang, C., Sternberg, P., and Fire, A. (1999). The RING finger/B-box factor TAM-1 and a retinoblastoma-like protein LIN-35 modulate context-dependent gene silencing in *Caenorhabditis elegans*. *Genes & Development* **13**, 2958-2970.

- Huang, L., Tzou, P., and Sternberg, P. (1994). The *lin-15* locus encodes two negative regulators of *Caenorhabditis elegans* vulval development. *Molecular Biology of the Cell* **5**, 395-411.
- Hubbard, E., Wu, G., Kitajewski, J., and Greenwald, I. (1997). *sel-10*, a negative regulator of *lin-12* activity in *Caenorhabditis elegans*, encodes a member of the CDC4 family of proteins. *Genes & Development* **11**, 3182-3193.
- Huynen, M., and Bork, P. (1998). Measuring genome evolution. *Proc Natl Acad Sci USA* **95**, 5849-5856.
- Jongeward, G., Clandinin, T., and Sternberg, P. (1995). *sli-1*, a negative regulator of *let-23*-mediated signaling in *C. elegans*. *Genetics* **139**, 1553-1566.
- Kalb, J., Lau, K., Goszczynski, B., Fukushige, T., Moons, D., Okkema, P., and McGhee, J. (1998). *pha-4* is Ce-fkh-1, a fork head/HNF-3a,B,y homolog that functions in organogenesis of the *C. elegans* pharynx. *Development* **125**, 2171-2180.
- Katz, W., Hill, R., Clandinin, T., and Sternberg, P. (1995). Different levels of the *C. elegans* growth factor LIN-3 promote distinct vulval precursor fates. *Cell* **82**, 297-307.
- Kennedy, B., Aamodt, E., Allen, F., Chung, M., Heschl, M., and McGhee, J. (1993). The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Journal of Molecular Biology* **229**, 890-908.
- Kimble, J. (1981). Alterations in cell lineage following laser ablation of cells in the somatic gonad of *C. elegans*. *Developmental Biology* **87**, 286-300.
- Kimble, J., Sulston, J., and White, J. (1979). Regulative development in the post-embryonic lineages of *C. elegans* in: "*Cell Lineage, Stem Cells and Cell Determinations.*" *Le Douarin N (ed), Elsevier, NY.*, 59-68.
- Kornfeld, K., Hom, D., and Horvitz, H. (1995). The *ksr-1* gene encodes a novel protein kinase involved in Ras-mediated signaling in *C. elegans*. *Cell* **83**, 903-913.
- Korswagen, H., Coudreuse, D., Betist, M., van de Water, S., Zivkovic, D., and Clevers, H. (2002). The Axin-like protein PRY-1 is a negative regulator of a canonical Wnt pathway in *C. elegans*. *Genes Dev* **16**, 1291-12302.

- Kostrouch, Z., Kostrouchova, M., and Rall, J. (1995). Steroid/thyroid hormone receptor genes in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences USA* **92**, 156-159.
- Krause, M., Fire, A., White, H. S., Priess, J., and Weintraub, H. (1990). CeMyoD accumulation defines the body wall muscle cell fate during *C. elegans* embryogenesis. *Cell* **63**, 907-919.
- Krause, M., Harrison, S., Xu, S., Chen, L., and Fire, A. (1994). Elements regulating cell- and stage-specific expression of the *C. elegans* myoD family homolog *hllh-1*. *Developmental Biology* **166**, 133-148.
- Krause, M., and Hirsh, D. (1987). A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* **49**, 753-761.
- Kuwabara, P. (1996). Interspecies comparison reveals evolution of control regions in the nematode sex-determining gene *tra-2*. *Genetics* **144**, 597-607.
- Kwon, J., and Lee, J. (2001). Biological significance of a universally conserved transcription mediator in metazoan developmental signaling pathways. *Development* **128**, 3095-3104.
- Labouesse, M., Sookhare, S., and Horvitz, H. (1994). The *Caenorhabditis elegans* gene *lin-26* is required to specify the fates of hypodermal cells and encodes a presumptive zinc-finger transcription factor. *Development* **120**, 2359-2368.
- Lackner, M., Kornfeld, K., Miller, L., Horvitz, H., and Kim, S. (1994). A MAP kinase homolog, *mpk-1*, is involved in ras-mediated induction of vulval cell fates in *Caenorhabditis elegans*. *Genes & Development* **8**, 160-173.
- Latchman, D. S. (1998). "Eukaryotic Transcription Factors." Academic Press.
- Lee, J., Jongeward, G., and Sternberg, P. (1994). *unc-101*, a gene required for many aspects of *Caenorhabditis elegans* development and behavior, encodes a clathrin-associated protein. *Genes & Development* **8**, 60-73.
- Levitan, D., and Greenwald, I. (1995). Facilitation of *lin-12*-mediated signalling by *sel-12*, a *Caenorhabditis elegans* S182 Alzheimer's disease gene. *Nature* **377**, 351-354.
- Matrisian, LM. (2000). Quick guide Matrix Metalloproteinases. *Curr Biol* **10**, R692.

- Loots, G., Locksley, R., Blankespoor, C., Wang, Z., Miller, W., Rubin, E., and Frazer, K. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-140.
- Lu, X., and Horvitz, H. (1998). *lin-35* and *lin-53*, Two genes that antagonize a *C. elegans* Ras pathway, encode proteins similar to Rb and its binding protein RbAp48. *Cell* **95**, 981-991.
- MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T., and Spieth, J. (1992). Regulation of vitellogenin gene expression in transgenic *Caenorhabditis elegans* - Short sequences required for activation of the *vit-2* promoter. *Molecular and Cellular Biology* **12**, 1652-1662.
- Maduro, M., and Pilgrim, D. (1996). Conservation of function and expression of *unc-119* from two *Caenorhabditis* species despite divergence of non-coding DNA. *Gene* **183**, 77-85.
- Maloof, J., and Kenyon, C. (1998). The Hox gene *lin-39* is required during *C. elegans* vulval induction to select the outcome of Ras signaling. *Development* **125**, 181-190.
- Mango, S., Lambie, E., and Kimble, J. (1994). The *pha-4* gene is required to generate the pharyngeal primordium of *Caenorhabditis elegans*. *Development* **120**, 3019-3031.
- Marshall, S., and McGhee, J. (2001). Coordination of *ges-1* expression between the *Caenorhabditis* pharynx and intestine. *Developmental Biology* **239**, 350-363.
- McGhee, J., and Krause, M. (1997). Transcription Factors and Transcriptional Regulation in: "*C. elegans II*." DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press. **II**, 147-184.
- McKnight, S., and Tjian, R. (1986). Transcriptional selectivity of viral genes in mammalian cells. *Cell* 1986 **46**, 795-805.
- Meyer, B. (2000). Sex in the worm-counting and compensating X-chromosome dose. *Trends in Genetics* **16**, 247-253.
- Miller, L., Gallegos, M., Morisseau, B., and Kim, S. (1993). *lin-31*, a *Caenorhabditis elegans* HNF-3/fork head transcription factor homolog, specifies three alternative cell fates in vulval development. *Genes & Development* **7**, 933-947.

- Moilanen, L., Fukushima, T., and Freedman, J. (1999). Regulation of metallothionein gene transcription-Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from *C. elegans*. *Journal of Biological Chemistry* **274**, 29655-29665.
- Muller, M., Gerster, T., and Schaffner, W. (1988). Enhancer sequences and the regulation of gene transcription. *Eur J Biochem* **176**, 485-495.
- Narlikar, G., Fan, H., and Kingston, R. (2002). Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475-487.
- Nigon, V., and Dougherty, E. (1949). Reproduction patterns and attempts at reciprocal crossing of *Rhabditis elegans* Maupas, 1900, and *Rhabditis briggsae* Dougherty & Nigon, 1949 (Nematoda: Rhabditidae). *Journal of Experimental Zoology* **112**, 485-503.
- Okkema, P., and Fire, A. (1994). The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175-2186.
- Pazin, M., and Kadonaga, J. (1997). What's up and down with histone deacetylation and transcription? *Cell* **89**, 325-328.
- Pelham, H. (1982). A regulatory upstream promoter element in the *Drosophila hsp 70* heat-shock gene. *Cell* **30**, 517-528.
- Pennacchio, L., and Rubin, E. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**, 100-109.
- Pettitt, J., Wood, W., and Plasterk, R. (1996). *cdh-3*, a gene encoding a member of the cadherin superfamily, functions in epithelial cell morphogenesis in *Caenorhabditis elegans*. *Development* **122**, 4149-4157.
- Rachez, C., and Freedman, L. (2001). Mediator complexes and transcription. *Curr Opin Cell Biol* **13**, 274-280.
- Riddle, D., Blumenthal, T., Meyer, B., and Priess, J. (1997). Introduction to *C. elegans* in: "*C. elegans II*" DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press. **II**, 1-22.

- Roberts, S., Emmons, S., and Childs, G. (1989). Nucleotide sequences of *Caenorhabditis elegans* core histone genes. Genes for different histone classes share common flanking sequence elements. *Journal of Molecular Biology* **206**, 567-577.
- Roberts, S., Sanicola, M., Emmons, S., and Childs, G. (1987). Molecular characterization of the histone gene family of *C. elegans*. *Journal of Molecular Biology* **196**, 27-38.
- Rocheleau, C., Downs, W., Lin, R., Wittmann, C., Bei, Y., Cha, Y.-H., Ali, M., Priess, J., and Mello, C. (1997). Wnt signaling and an APC-related gene specify endoderm in early *C. elegans* embryos. *Cell* **90**, 707-716.
- Sanford, T., Golomb, M., and Riddle, D. (1983). RNA polymerase II from wild type and alpha-amanitin-resistant strains of *C. elegans*. *Journal of Biological Chemistry* **258**, 12804-12809.
- Sanford, T., Prenger, J., and Golomb, M. (1985). Purification and immunological analysis of RNA polymerase II from *C. elegans*. *Journal of Biological Chemistry* **260**, 8064-8069.
- Sanicola, M., Ward, S., Childs, G., and Emmons, S. (1990). Identification of a *Caenorhabditis elegans* histone H1 gene family. Characterization of a family member containing an intron and encoding a poly(A)⁺ mRNA. *Journal of Molecular Biology* **212**, 259-268.
- Sen, R., and Baltimore, D. (1986). Multiple nuclear factors interact with the immunoglobulin enhancer sequences. *Cell* **46**, 705-716.
- Sentenac, A. (1985). Eukaryotic RNA polymerases. *CRC Crit Rev Biochem* **18**, 31-90.
- Sewack, G., Ellis, T., and Hansen, U. (2001). Binding of TATA binding protein to a naturally positioned nucleosome is facilitated by histone acetylation. *Mol Cell Biol* **21**, 1404-1415.
- Sharma-Kishore, R., White, J., Southgate, E., and Podbilewicz, B. (1999). Formation of the vulva in *Caenorhabditis elegans*: a paradigm for organogenesis. *Development* **126**, 691-699.
- Sieburth, D., Sun, Q., and Han, M. (1998). SUR-8, a conserved Ras-binding protein with leucine-rich repeats, positively regulates Ras-mediated signaling in *C. elegans*. *Cell* **94**, 119-130.

- Sieburth, D., Sundaram, M., Howard, R., and Han, M. (1999). A PP2A regulatory subunit positively regulates Ras-mediated signaling during *Caenorhabditis elegans* vulval induction. *Genes & Development* **13**, 2562-2569.
- Solari, F., and Ahringer, J. (2000). NURD-complex genes antagonise Ras-induced vulval development in *Caenorhabditis elegans*. *Current Biology* **10**, 223-226.
- Spieth, J., Denison, K., Kirtland, S., Cane, J., and Blumenthal, T. (1985). The *C. elegans* vitellogenin genes: short sequence repeats in the promoter regions and homology to the vertebrate genes. *Nucleic Acids Research* **13**, 5283-5295.
- Spieth, J., Nettleton, M., Zucker-Aprison, E., Lea, K., and Blumenthal, T. (1991a). Vitellogenin motifs conserved in nematodes and vertebrates. *Journal of Molecular Evolution* **32**, 429-438.
- Spieth, J., Shim, Y., Lea, K., Conrad, R., and Blumenthal, T. (1991b). *elt-1*, an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. *Molecular and Cellular Biology* **11**, 4651-4659.
- Sternberg, P. (1988). Lateral inhibition during vulval induction in *Caenorhabditis elegans*. *Nature* **335**, 551-554.
- Sternberg, P., and Alberola-Ila, J. (1998). Conspiracy theory: RAS and RAF do not act alone. *Cell* **95**, 447-450.
- Sternberg, P., and Han, M. (1998). Genetics of RAS signaling in *C. elegans*. *Trends in Genetics* **14**, 466-472.
- Sternberg, P., and Horvitz, H. (1986). Pattern formation during vulval development in *C. elegans*. *Cell* **44**, 761-772.
- Sternberg, P., and Horvitz, H. (1989). The combined action of two intercellular signaling pathways specifies three cell fates during vulval induction in *C. elegans*. *Cell* **58**, 679-693.
- Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R. (1999). Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res* **27**, 3899-38910.
- Strahl, B., and Allis, C. (2000). The language of covalent histone modifications. *Nature* **403**, 41-45.

- Struhl, G., Fitzgerald, K., and Greenwald, I. (1993). Intrinsic activity of the *lin-12* and notch intracellular domains in vivo. *Cell* **74**, 331-345.
- Sulston, J., and Horvitz, H. (1977). Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology* **56**, 110-156.
- Sulston, J., and White, J. (1980). Regulation and cell autonomy during postembryonic development of *C. elegans*. *Developmental Biology* **78**, 577-597.
- Sundaram, M., and Greenwald, I. (1993). Genetic and phenotypic studies of hypomorphic *lin-12* mutants in *Caenorhabditis elegans*. *Genetics* **135**, 755-763.
- Sundaram, M., and Han, M. (1995). The *C. elegans ksr-1* gene encodes a novel Raf-related kinase involved in Ras-mediated signal transduction. *Cell* **83**, 889-901.
- Swoboda, P., Adler, H., and Thomas, J. (2000). The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in *C. elegans*. *Molecular Cell* **5**, 411-421.
- Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D., and Jones, R. (1988). Embryonic *epsilon* and *gamma* globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal Molecular Biology* **3**, 439-455.
- Tan, P., Lackner, M., and Kim, S. (1998). MAP kinase signaling specificity mediated by the LIN-1 Ets/LIN-31 WH transcription factor complex during *C. elegans* vulval induction. *Cell* **93**, 569-580.
- Tax, F., Thomas, J., Ferguson, E., and Horvitz, H. (1997). Identification and characterization of genes that interact with *lin-12* in *Caenorhabditis elegans*. *Genetics* **147**, 1675-1695.
- Tepass, U. (1999). Genetic analysis of cadherin function in animal morphogenesis. *Curr Opin Cell Biol* **11**, 540-548.
- Thatcher, J., Haun, C., and Okkema, P. (1999). The DAF-3 Smad binds DNA and represses gene expression in the *Caenorhabditis elegans* pharynx. *Development* **126**, 97-107.
- Thomas, J., and Horvitz, H. (1999). The *C. elegans* gene *lin-36* acts cell autonomously in the *lin-35* Rb pathway. *Development* **126**, 3449-3459.

- Thorpe, C., Schlesinger, A., Carter, J., and Bowerman, B. (1997). Wnt signaling polarizes an early *C. elegans* blastomere to distinguish endoderm from mesoderm. *Cell* **90**, 695-705.
- Tse, C., Sera, T., Wolffe, A., and Hansen, J. (1998). Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. *Mol Cell Biol* **18**, 4629-4638.
- Tsukiyama, T., and Wu, C. (1997). Chromatin remodeling and transcription. *Curr Opin Genet Dev* **7**, 182-191.
- Tuck, S., and Greenwald, I. (1995). *lin-25*, a gene required for vulval induction in *Caenorhabditis elegans*. *Genes & Development* **9**, 341-357.
- Vanfleteren, J., and Van, B. J. (1983). Nematode chromosomal proteins-III. Some structural properties of the histones of *C. elegans*. *Comparative Biochemistry & Physiology* **76B**, 179-184.
- Vanfleteren, J., Van, B. S., and Van, B. J. (1989). The histones of *Caenorhabditis elegans*: no evidence of stage-specific isoforms. *FEBS Letters* **257**, 233-237.
- Wada, K., Sato, H., Kinoh, H., Kajita, M., Yamamoto, H., and Seiki, M. (1998). Cloning of three *Caenorhabditis elegans* genes potentially encoding novel matrix metalloproteinases. *Gene* **211**, 57-62.
- Wang, B., Muller-Immergluck, M., Austin, J., Robinson, N., Chisholm, A., and Kenyon, C. (1993). A homeotic gene-cluster patterns the anteroposterior body axis of *C. elegans*. *Cell* **74**, 29-42.
- Wang, L., and Way, J. (1996a). Activation of the *mec-3* promoter in two classes of stereotyped lineages in *Caenorhabditis elegans*. *Mechanisms of Development* **56**, 165-181.
- Wang, L., and Way, J. (1996b). Promoter sequences for the establishment of *mec-3* expression in the nematode *Caenorhabditis elegans*. *Mechanisms of Development* **56**, 183-196.
- Way, J., and Chalfie, M. (1988). *mec-3*, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*. *Cell* **54**, 5-16.
- Way, J., Wang, L., Run, J., and Wang, A. (1991). The *mec-3* gene contains *cis*-acting elements mediating positive and negative regulation in cells produced by

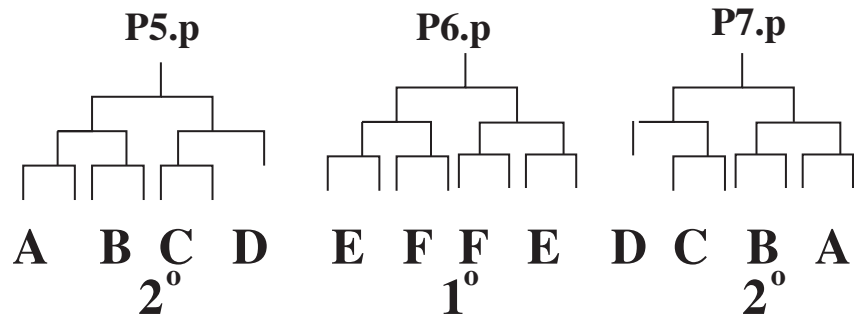
- asymmetric cell division in *Caenorhabditis elegans*. *Genes & Development* **5**, 2199-2211.
- Webb, C., Shabalina, S., Ogurtsov, A., and Kondrashov, A. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res* **30**, 1233-1239.
- Wen, C., Metzstein, M., and Greenwald, I. (1997). SUP-17, a *Caenorhabditis elegans* ADAM protein related to *Drosophila* KUZBANIAN, and its role in LIN-12/NOTCH signaling. *Development* **124**, 4759-4767.
- Williams-Masson, E., Heid, P., Lavin, C., and Hardin, J. (1998). The cellular mechanism of epithelial rearrangement during morphogenesis of the *Caenorhabditis elegans* dorsal hypodermis. *Developmental Biology* **204**, 263-276.
- Wu, Y., and Han, M. (1994). Suppression of activated Let-60 Ras protein defines a role of *Caenorhabditis elegans* Sur-1 MAP kinase in vulval differentiation. *Genes & Development* **8**, 147-159.
- Wu, Y., Han, M., and Guan, K. (1995). MEK-2, a *Caenorhabditis elegans* MAP kinase kinase, functions in Ras-mediated vulval induction and other developmental events. *Genes & Development* **9**, 742-755.
- Xue, D., Finney, M., Ruvkun, G., and Chalfie, M. (1992). Regulation of the *mec-3* gene by the *C. elegans* homeoproteins UNC-86 and MEC-3. *EMBO Journal* **11**, 4969-4979.
- Xue, D., Tu, Y., and Chalfie, M. (1993). Cooperative interactions between the *Caenorhabditis elegans* homeoproteins UNC-86 and MEC-3. *Science* **261**, 1324-1328.
- Yoon, C., Lee, J., Jongeward, G., and Sternberg, P. (1995). Similarity of *sli-1*, a regulator of vulval development in *C. elegans*, to the mammalian proto-oncogene c-cbl. *Science* **269**, 1102-1105.
- Zhang, H., and Emmons, S. (2000). A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes & Development* **14**, 2161-2172.

- Zhang, Y., Ma, C., Delohery, T., Nasipak, B., Foat, B., Bounoutas, A., Bussemaker, H., Kim, S., and Chalfie, M. (2002). Identification of genes expressed in *C. elegans* touch receptor neurons. *Nature* **418**, 331-335.
- Zorio, D., Cheng, N., Blumenthal, T., and Spieth, J. (1994). Operons as a common form of chromosomal organization in *C. elegans*. *Nature* **372**, 270-272.
- Zucker-Aprison, E., and Blumenthal, T. (1989). Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. *Journal of Molecular Evolution* **28**, 487-496.

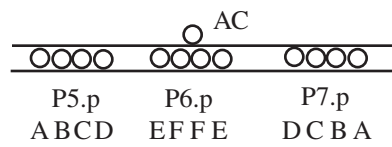
Figure 1: Vulva formation in *C. elegans*

The top panel shows the lineage relationship of P5, 6 and 7.p descendents that give rise to the vulva. In the bottom panels, nuclei are indicated by circles, and prominent cell boundaries are indicated by thin lines. The ventral surface is down in all panels, and the dark horizontal line represents the ventral cuticle. Since animals were typically observed from the side, different focal planes correspond to the midline (top panel), the sublateral plane (middle panel) and the lateral plane (bottom panel). A three-dimensional schematic is shown to the right. "A, B1, B2..." correspond to "vulA, vulB1, vulB2...".

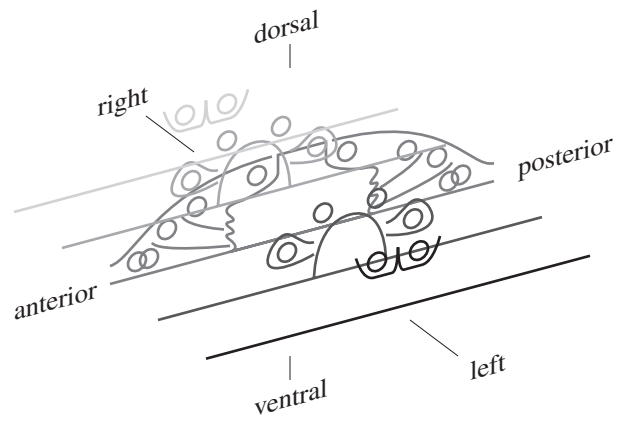
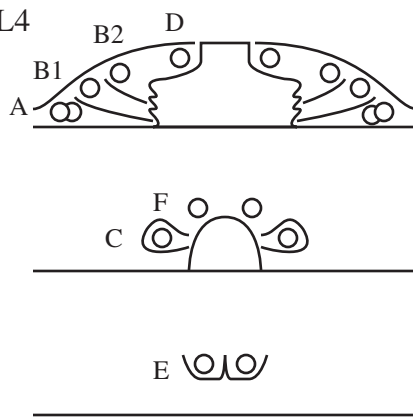
Figure 1: Vulva formation in *C. elegans*



late L3



mid L4



adult

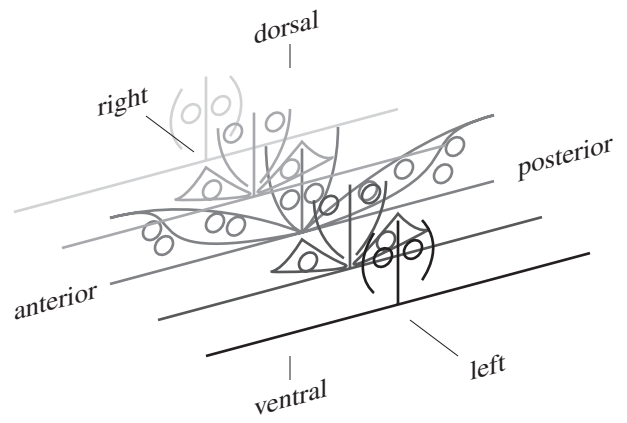
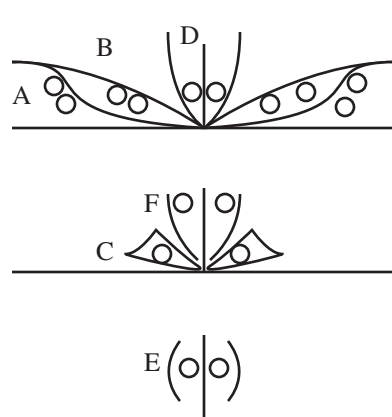


Figure 2: Available vulval marker gene's expression pattern in *C. elegans*

Filled bars indicate consistent expression observed in all animals, gray bars indicate expression observed in some but not all animals. The last round of cell division in the vulva takes place within the first one or two hours of the L4 stage. *egl-17::gfp* is also expressed earlier in the parents and grandparents of vulE and vulF cells, P6.p progeny, (Burdine *et al.*, 1998) (not shown). This expression occasionally persists into the L4 stage in some lines. The expression of *T04B2.6::gfp* is observed in old adults (animals with a significant number of eggs in the gonad) but not in young adults (animals without eggs in the gonad immediately after the L4 molt). The last panel is a side-by-side comparison of markers disregarding the temporal aspect, demonstrating that six different cell types can be distinguished based on the expression pattern.

Figure 3: *egl-17::GFP*

A schematic diagram of cell positions at various stages of development, late L3, mid-L4 and adult are shown (Sharma-Kishore *et al.*, 1999; Sulston and Horvitz, 1977). Nuclei are indicated by circles, and prominent cell boundaries are indicated by thin lines. The green filled-in circles depict the GFP expressing cells. The ventral is down in all panels, and the dark horizontal line represents the ventral cuticle. Each set of Nomarski images to the left have corresponding epifluorescence images to the right. The top panels are from the midline, the middle panels are from the sub-lateral plane, and the bottom panels are from the lateral plane (L3 animals were only photographed in midline plane). The strain and array photographed is MT2466 *ayIs4[egl-17::gfp]*.

Figure 4: *zmp-1::GFP*

A schematic diagram of cell positions at various stages of development, late L3, and adult stages are shown (Sharma-Kishore *et al.*, 1999; Sulston and Horvitz, 1977). Nuclei are indicated by circles, and prominent cell boundaries are indicated by thin lines. The green filled-in circles depict the GFP expressing cells. The ventral is down in all panels, and the dark horizontal line represents the ventral cuticle. Each set of Nomarski images to the left have corresponding epifluorescence images to the right. In the photomicrographs of the adult animals, the top panels are from the midline, the middle panels are from the sub-lateral plane, and the bottom panels are from the lateral plane. The strain and array photographed is PS3239 *syIs49[zmp-1::gfp]*.

Figure 5: *cdh-3::GFP*

A schematic diagram of cell positions at various stages of development, late L3, mid-L4 and adult stages are shown (Sharma-Kishore *et al.*, 1999; Sulston and Horvitz, 1977). Nuclei are indicated by circles, and prominent cell boundaries are indicated by thin lines. The green filled-in circles depict the GFP expressing cells. The ventral is down in all panels, and the dark horizontal line represents the ventral cuticle. Each set of Nomarski images to the left have corresponding epifluorescence images to the right. In the photomicrographs of the mid-L4 and adult animals, the top panels are from the midline, the middle panels are from the sub-lateral plane, and the bottom panels are from the lateral plane. In the photomicrographs of the mid-L4 and adult animals, the *cdh-3* is also expressed along with *ceh-3* in vulC cells (looks yellow in epifluorescence photomicrographs). The strain and array photographed is PS3528 *syIs51[cdh-3::cfp]; syIs55[ceh-2::gfp]* (for the mid-L4 and adult animals) and NL1008 *pkEx246[cdh-3::gfp]* (for anchor cell expression).

cis*-Regulatory Control of Cell Fate-Specific Genes in *Caenorhabditis elegans

Vulval Organogenesis

Martha Kirouac and Paul W. Sternberg

(submitted for publication)

ABSTRACT

The great-grandprogeny of the *Caenorhabditis elegans* vulval precursor cells (VPCs) adopt one of the final vulA, B1, B2, C, D, E, and F cell types in a precise spatial pattern; *egl-17*, *zmp-1*, and *cdh-3* are differentially expressed in the developing vulva lineages and provide a potential readout for different signaling pathways. We have identified upstream *cis*-regulatory regions of these three genes sufficient for their ability to confer vulval cell type specific regulation. A 143-bp region of *egl-17* is sufficient to drive vulC and vulD expression, while a separate 102-bp region drives the early expression in presumptive vulE and vulF cells. A 300-bp region of *zmp-1* is sufficient to drive expression in vulE, vulA, and the anchor cell. A 689-bp region of *cdh-3* is sufficient to drive expression in the anchor cell and vulE, vulF, vulD and vulC; a 155-bp region is sufficient to drive anchor cell expression; and a separate 563-bp region is also sufficient to drive expression in these vulval cells. We have found no evidence of repressor elements in any of these genes with respect to vulval and anchor cell expression.

INTRODUCTION

In the *C. elegans* vulval ectoderm, three intercellular signaling pathways, EGF, NOTCH, and WNT, induce six multipotential Vulval Precursor Cells (VPCs) to generate an invariant spatial pattern of cell fates. These signaling pathways stimulate both the division of the VPC cells, and the emergence of a precise pattern (reviewed in Greenwald, 1997; Sternberg and Han, 1998). The VPCs are of three types: 1° and 2° VPCs, which can be distinguished by their division pattern and differential expression of marker genes, and 3° VPCs, which generate non-vulval epidermis (Burdine *et al.*, 1997; Greenwald, 1997; Sternberg and Horvitz, 1986; Sulston and Horvitz, 1977). Once the VPCs terminally differentiate into one of the final vulval fates, vulF, E, D, C, B2, B1, and A (Figure 1), their morphogenetic interactions lead to the development of seven toroidal cells that connect the endothelium of the uterus to the external epithelium (Figure 1; Sharma-Kishore *et al.*, 1999). Little is known about the individual roles of these vulval cells following their terminal differentiation, and what cell-specific functions they possess. However, the differentiation of vulval cell types is likely to depend upon the *cis*-regulatory regions of the transcriptional targets of these intercellular signals in vulval development; the outcome of such differential activation will result in individual cell types. While a number of transcription factors are known to be involved in vulval development (e.g., *lin-1*, *lin-29*, *egl-38*, *lin-31*, *lin-39*, *lin-11*), their targets are not known (Beitel *et al.*, 1995; Bettinger *et al.*, 1997; Chang *et al.*, 1999; Clark *et al.*, 1993; Euling *et al.*, 1999; Freyd *et al.*, 1990; Tan *et al.*, 1998).

The gonadal anchor cell (AC) serves as the source of the inductive signal, LIN-3, which promotes vulval fates in the VPCs (Hill and Sternberg, 1992; Katz *et al.*, 1995; Kimble, 1981). The anchor cell also helps establish a functional connection between the

vulva and the uterus (Newman and Sternberg, 1996; Newman *et al.*, 1996). As in vulval development, we know few of the transcriptional regulators that control anchor cell gene expression. The isolation of response elements used by the anchor cell will facilitate identification of major transcriptional factors that control cell-specific gene expression.

Here we focus on three genes that are differentially regulated in these vulva cell types and in the anchor cell: *egl-17*, which encodes a fibroblast growth factor family member (Burdine *et al.*, 1997, 1998); *cdh-3*, which encodes a FAT-like cadherin (Pettitt *et al.*, 1996); and *zmp-1*, which encodes a zinc metalloproteinase, (J. Butler and J. Kramer, personal communication; Wada *et al.*, 1998). These genes offer the opportunity to find *cis*-regulatory elements for multiple vulval cell types as well as the anchor cell. The identification of sequences that direct expression in these cell types will lead to a deeper understanding of the regulatory networks that pattern the vulva. We have analyzed the *cis*-regulatory sequences of these genes in *C. elegans* and report here on the different regulatory regions that drive expression of these three genes in the vulva cells and the anchor cell.

MATERIALS AND METHODS

Generation of *C. elegans* promoter GFP constructs

Using PCR (supplemental material, Table 1), the regions of interest were amplified, with TaKaRa LA Taq (Takara Shuzo), and cloned into the minimal promoter *pes-10*, pPD107.94, (a gift from the Fire lab) using restriction sites engineered into the primers. The PCR protocol used was: 94.0 °C for 4 minutes; 30 cycles 94.0 °C for 30 seconds, 60.0°C for 30 seconds, 68.0°C for 45 seconds; and 68.0 °C for 7 minutes.

As a template for PCR, the following constructs were used: the *egl-17* promoter NH#293 (Burdine *et al.*, 1998); the *zmp-1* promoter pJB100 (J. Butler and J. Kramer, personal communication); and the *cdh-3* promoter jp#38 (Pettitt *et al.*, 1996).

The nomenclature of the constructs generated in this study is derived from the primers used to amplify the region. In all cases, the first one to three digits represent the 5' primer, and the digits after the hyphen represent the 3' primer. Although we performed a systematic dissection of these three upstream sequences, not all constructs made are shown in this paper because of space limitations. For a comprehensive list, see the supplemental material in figures 1, 2 and 3.

The *egl-17* genomic region of NH#293 contains 3819 bp of sequence upstream of the translational start site. The first exon of the transcript starts at nucleotide 4610, and translation starts at nucleotide 4708. Nucleotide 790 of the *egl-17* upstream region corresponds with nucleotide 17648 in Genbank cosmid F38G1 (Accession # AC006635). The *zmp-1* genomic region in pJB100 contains 3472 bp of sequence upstream of the translational start site. The translational start site of ZMP-1 is at nucleotide 3473. Nucleotide 1 of this *zmp-1* upstream region corresponds with nucleotide 7630 in Genbank cosmid EGAP1 (Accession # U41266). The jp#38 genomic region of *cdh-3* contains 5928 bp of sequence upstream of the translational start site, whose start codon occurs at nucleotide 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nucleotide 37343 in Genbank cosmid ZK112 (Accession # L14324).

Generation of *C. elegans* promoter deletion GFP constructs

An internal deletion was made using PCR primers (Supplemental Material, Table 1) that are homologous to 20 bp on either side of the region of the deletion. In primary PCR reaction, the deletion was generated using internal primers that span the deletion region, with outside primers mk151/mk50 and mk152/mk51 for construct $\Delta 3/4$. This generated two fragments with homologous ends containing the deletion. In a second round of amplification, just the outside primers mk50 and mk51 were used on the combined gel-purified products from the first PCR reaction that served as the template. The PCR protocol used for both the primary and secondary PCR reactions was: 94.0 °C 4 minutes; 30 cycles 94.0 °C for 30 seconds, 58.0 °C for 30 seconds, 65.0 °C for 40 seconds; and 65.0 °C for 7 minutes.

Sequencing of constructs

The following constructs were sequenced to confirm these sequences: mk158-159, mk66-156, mk155-67, mk64-65, mk66-67, mk96-63, mk135-119, mk96-145, mk146-144, mk135-134, mk135-147, mk96-143, mk135-143, mk102-56, mk102-104, mk80-104, mk103-148, mk50-111, mk50-115, mk52-51, mk52-74, mk36-74, mk76-51, mk107-51, mk121-51, mk50-124, mk50-74, mk50-123, mk $\Delta 3/4$, mk153-148, mk153-154, mk103-56, mk36-51, mk106-51, mk50-75, mk118-143, mk135-147, mk125-132, mk96-143.

Microinjection of promoter GFP constructs into *C. elegans*

The constructs were microinjected into the gonads of animals of genotype *pha-1(e2123ts)*; *him-5(e1490)* line using a standard protocol (Mello *et al.*, 1991). The constructs were injected at a concentration of 100 ng/ μ l, with 20 ng/ μ l pBluescript SKII (Stratagene), and 82

ng/μl *pha-1(+)*, pBX. Transgenic animals that stably transmitted the extrachromosomal arrays were isolated by selecting viable F1 animals at 22 °C to new plates, and examining their progeny for GFP expression in the anchor cell and the vulval cells.

Microscopy of transgenic animals

Animals were mounted on 5% noble agar pads and scored at 20°C for GFP expression under Nomarski optics using a Zeiss Axioplan microscope with a 200-watt HBO UV source, and a Chroma High Q GFP LP filter set (450 nm excitation/505 nm emission). At least two lines for each construct were examined.

egl-17 early expression in the granddaughters of P6.p, the precursor to vulE and vulF cells, was scored at the four-cell stage. *egl-17* vulC and vulD GFP expression was scored between the late L4 to young adult stages (Burdine *et al.*, 1998). *zmp-1* anchor cell GFP expression was scored between the L3 and the early L4 stage. VulE and vulD expression was scored between the late L4 and young adult stages. *zmp-1* vulA expression was scored between the young adult and adult stages (Wang and Sternberg, 2000). *cdh-3* AC GFP expression was scored between the L3 and the early L4 stage. *cdh-3* vulE, vulF, vulC, and vulD expression was scored between the late L3 stage through late L4 stages (Figure 1; Pettitt *et al.*, 1996).

Prediction of binding sites using Transfac database

Possible binding sites for known transcription factors in the regions defined by deletion analysis in the *egl-17*, *zmp-1* and *cdh-3* upstream regions were determined using the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al.*, 1995).

AlignACE predictions of over-represented sequences

AlignACE is based on a Gibbs sampling algorithm that computes a series of motifs that are over-represented in the input sequence(s) (<http://atlas.med.harvard.edu/cgi-bin/alignace.pl>; Roth *et al.*, 1998). This algorithm assigns a score to each motif; the MAP score (maximum a priori log likelihood) is the functional readout of the degree to which a motif is over-represented relative to the expectation for the random occurrence of such a motif in the sequence under consideration (Roth *et al.*, 1998). In this analysis, we chose a MAP cut-off of 10. When this cut-off was applied in a search of motifs in the genome of *Saccharomyces cerevisiae*, Hughes and colleagues found that this threshold did not lead to the rejection of any of the best examples of known *cis*-regulatory elements (Hughes *et al.*, 2000). We used a GC content setting of 0.35, and searched for motifs of eight and 10 nucleotides; these nucleotides do not have to be contiguous, but will receive higher MAP score if they are. A higher MAP score reflects: (1) a greater number of aligned sites; (2) a more tightly conserved motif; (3) less total input sequence; (4) more tightly packed information-rich positions; and (5) enrichment of the motif with nucleotides that are less prevalent in the genome (Hughes *et al.*, 2000).

RESULTS

To identify regulatory sequences sufficient to drive cell-specific expression, genomic fragments were tested for their ability to drive GFP expression from the heterologous *pes-10* basal promoter. This promoter does not drive expression of GFP in any of these tissues on its own. However, it contains the basic sites for the transcriptional machinery, which when

combined with an enhancer region, can drive GFP expression in a cell type-specific manner (Seydoux and Fire, 1994; G. Seydoux, personal communication).

Vulval specificity in the *egl-17* cis-regulatory region in *C. elegans*

The *egl-17::GFP* translational fusion NH#293 is detectable in P6.p, P6.p daughters and granddaughters (the presumptive vulE and vulF cells), turns off in early L4, and turns on again in vulC and vulD cells in mid L4. This *egl-17::GFP* construct contains 3.9 kb of sequence upstream of the translational start (Burdine *et al.*, 1998). We divided the 3.9 kb upstream region into four sub-fragments. Of the initial constructs, mk27-49 (3502-4586) showed expression in vulC and vulD cells, and mk15-20 (1716-3690) showed weak variable expression in the presumptive vulE and vulF cells, while mk153-154 (4565-4667) showed weak expression in the presumptive vulE and vulF cells (Figure 2).

We next sub-divided the mk27-49 (3502-4586) region to identify the minimal region sufficient for vulC and vulD specificity. A 143-bp region, mk125-132 (4331-4474), is sufficient to drive strong expression in vulC and vulD (Figure 3A). A comparison of mk125-132 (4331-4474), which drives expression clearly in both vulC and vulD cells, to mk102-56 (4359-4516), which drives expression weakly in vulD cells and not at all in vulC cells, suggests that the 5' end of this region is involved in vulC expression. Likewise, when we compare mk80-104 (4316-4466), which drives expression weakly in vulC cells but not at all in vulD cells, to mk80-132 (4316-4484), which drives expression in both vulC and vulD cells, it appears likely that the 3' end of this region is necessary for vulD expression (Figure 3A). However, the expression levels of both constructs are severely compromised when compared to the full-length construct. When both of these sites are removed, mk102-104

(4354-4466), no GFP expression is seen in either vulC or vulD. The nucleotide sequence of this entire region is shown in Figure 4A.

In addition to defining the region sufficient to drive GFP expression in vulC and vulD cells, we examined the regions defined by the initial constructs mk15-20 (4565-4667) and mk153-154 (4565-4667) for the minimal region sufficient to confer specificity of expression in the presumptive vulE and vulF cells. We have defined two regions that together confer strong expression in the presumptive vulE and vulF cells. One of these regions, 4565-4667, (mk153-154, Figure 3B), is located in the 5' UTR of the *egl-17* gene; 4565-4667 is sufficient to confer expression in these cells, but the expression is slightly variable and weaker than the full-length reporter construct. Additionally, a second element plays a role in conferring specificity to these cells. While constructs mk82-100 (2888-3611) and mk84-20 (3182-3640) show faint, inconsistent expression in the presumptive vulE and vulF cells, constructs mk82-85 (2088-3203) and mk27-20 (3502-3690) do not. This observation suggests that either the element responsible for this weak expression lies within the region 3203-3502, or that multiple sites are required that are dispersed throughout the larger region 2888-3690. When both of these regions are present, as in the much larger constructs mk15-148 (1716-4732) or mk84-148 (3182-4732), the expression is comparable to the level of early expression seen in the full-length reporter construct. The nucleotide sequence for mk84-148 (3182-4732) is shown in Figure 4A. In mk153-148 (4565-4732), despite containing the sequence that is sufficient to drive GFP expression in line mk153-154 (4565-4667), we see no GFP expression. This observation suggests the presence of a repressor of early expression in the presumptive vulE and F cells between 4667-4732. Another possibility is that the variability

of these lines (including mk103-148, 4427-4732) might be due to differences in transgene copy number.

Vulva and anchor cell specificity in the *zmp-1* cis-regulatory region in *C. elegans*

The *zmp-1::GFP* marker strain containing pJB100 has 3.8 kb of *zmp-1* upstream regulatory sequence; the GFP in this strain is expressed in the anchor cells of L3 larvae, and in late L4 and young adult animals it is also expressed in vulE, D and vulA cells (J. Butler & J. Kramer, personal communication; Wang and Sternberg, 2000). We divided this 3.8 kb upstream region into four fragments (Figure 2). Of the first constructs made, mk29-32 (791-1618) showed expression in the anchor cell, and in vulE and vulA cells. No construct was found to drive the expression in vulD cells, which is seen in the full-length reporter construct.

We next sub-divided the sequences defined by the initial construct mk29-32 (791-1618) to define a 380 bp region (mk50-51; 1052-1438) that is sufficient to confer anchor cell, vulE and vulA cell specificity on the *pes-10* promoter (Figure 5). This region also confers uterine cell expression; however, we chose not to analyze this expression pattern further. When this minimal region was further sub-divided, we were able to identify regions of the *zmp-1* promoter that drive expression in just the anchor cell, for example mk36-51 (1180-1438), mk50-124 (1052-1268), and mk76-74 (1147-1378), but we were not able to identify fragments active only in vulA or vulE. This failure is in spite of the fact that when successive deletions are made on either end, the expression pattern is lost in a reproducible manner, so that first vulA expression is lost, then vulE expression, and finally anchor cell expression (Figure 5). For instance, consider successive 5' deletions in which the 3' end is maintained at nucleotide 1438. In mk76-51 (1147-1438), vulA expression is lost first. Then, in mk107-51

(1165-1438), vulE expression becomes variable and, finally, in mk121-51 (1191-1438), anchor cell GFP expression is lost (Figure 5, constructs mk50-51 through mk121-51). Similarly, successive 3' deletions, in which the 5' end is maintained at nucleotide 1052, show the same pattern of expression loss (Figure 5, constructs mk50-75 through mk50-11). We did not observe vulA expression without expression in both vulE and the anchor cell, nor did we observe vulE expression without anchor cell expression. This hierarchy suggests that it is the number of binding sites, rather than just the qualitative aspects of these sites, that determines the expression pattern, and hence that the different cell types have different levels of the factor that binds these sites. The nucleotide sequence of mk50-51 (1052-1438) is shown in Figure 4B.

Further support for the hypothesis that the quantity of sites dictates the *zmp-1* expression pattern is our observation that when the end points are changed, the regions that were necessary for expression in a given cell type become important for other expression patterns. For instance, as seen in Figure 5, in constructs mk76-51 (1147-1438), mk107-51 (1165-1438), and mk36-51 (1180-1438), which have the same 3' end, 1438, but are successively smaller on the 5' end, the region 1147-1165 appears to play an important role in the expression in vulE cells, and the region 1165-1180 seems to play a role in both vulA and vulE expression. However, when we tested mk76-74 (1147-1378) and mk36-74 (1180-1378), which end at 1378 instead of 1438 on the 3' end, but whose 5' end is either at 1147 or 1180, respectively, we observe that the region 1147-1180 appears to be necessary for anchor cell expression. When we compare constructs mk52-74 (1119-1378) and mk76-74 (1147-1378) to mk50-51 (1052-1438) and mk76-51 (1147-1438), we find that the region 1119-1147 can be important for either vulE expression in the first constructs, or for vulA expression in the

latter two constructs. Also, when we compare mk52-51 (1119-1378) and mk50-74 (1052-1375) to mk52-74 (1119-1375), we see that either 1052-1119 or 1378-1438 is sufficient to drive GFP in vulA cells.

To test whether the level of expression is determined by the quantity of the sites alone, or whether qualitative aspects of the sites are also crucial, we made an internal deletion, 1262-1269 (mk Δ 3/4) (Figure 5). This region was shown to be important for anchor cell expression in our deletion analysis. If only anchor cell expression were lost, then the experiment would suggest that it is qualitative aspects of this site that are most important in determining the expression pattern. If vulA expression, or both vulA and vulE expression, is lost instead, it would suggest that it is the number of sites bound that determines the expression pattern. The resulting deletion had a more complex effect. Construct mk Δ 3/4 (Δ 1262-1269) showed expression in the anchor cell, and in vulA cells, but showed no expression in vulE cells. In this case, anchor cell expression was not lost, indicating that indeed there seem to be multiple sites that can drive expression in a given cell type. We also saw the loss of vulE expression preferentially over the loss of vulA expression. This pattern was not observed until the internal deletion was made, and it suggests that, while the quantity of bound sites is an important determining event in the expression pattern, there are also qualitative aspects of sites important for expression in a given cell type.

Vulva and anchor cell specificity in the *cdh-3* cis-regulatory region in *C. elegans*

A *cdh-3::GFP* fusion containing 6.0 kb of upstream sequence is expressed from the L2 stage in the anchor cell; during the L3 stage it is also expressed in vulE and vulF cells, and it is also expressed in the vulC and vulD cells of L4 larvae (Pettitt *et al.*, 1996). We divided this

6.0 kb upstream region into seven subfragments (Figure 2). Of the initial constructs, mk62-63 (1478-3008) showed anchor cell expression; mk66-67 (4434-4997) showed vulva expression; and mk135-134 (2412-3419) showed both anchor cell and vulva expression.

Since both mk62-63 (1478-3008) and mk135-134 (2412-3419) were sufficient to drive expression in the anchor cell, we focused our search for the minimal anchor cell element in the overlapping region (~2300-3200). This region is also sufficient to confer uterine cell expression, however, we chose to focus our attention on the vulva and anchor cell elements. The minimal *cis*-regulatory region we observed to drive anchor cell expression is 155 bp (mk146-144; 2367-2522) (Figure 6A). This construct displays variable expression, but the 232 bp construct mk96-144 (2290-2522) expresses in all animals observed. The details of the specific sequences driving anchor cell expression are more complicated. There appear to be at least three regions (α , 2290-2431; β , 2431-2522; and γ , 2989-4363) that play a role in anchor cell expression. While any one of these regions is insufficient to drive anchor cell GFP expression on its own, as demonstrated by mk96-145 (2290-2431) (α), mk135-119 (2412-2713) (β), and mk64-65 (2989-4363) (γ), any two of these regions are sufficient to drive this expression, as demonstrated by mk146-144 (2367-2522) (α and β), mk135-143 (2412-3164) (β and γ), and the co-injection of mk64-65 (2989-4363) with mk96-145 (2290-2431) (α and γ). However, in the case of the co-injection of mk64-65 and mk96-145, less than 10% of the animals show GFP expression in the anchor cell. The nucleotide sequence of mk96-134 (2290-3419), which contains the α , β , and γ sites, is shown in Figure 4C.

This anchor cell expression pattern has at least one additional layer of complexity. The expression from some constructs comes on at the VPC 2-cell stage, while from other constructs it does not express until the VPC 4-cell stage. When mk135-143 (2412-3164) and

mk135-147 (2412-3101) are compared, it appears that the region 3101-3164 can confer early anchor cell expression. This region is encompassed in the γ region. There appears to be another region, 1478-2290, that is sufficient to drive early anchor cell expression when mk62-63 (1478-3008), which expresses GFP at the VPC- 2-cell stage, is compared to mk96-63 (2290-3008), which does not express GFP at the VPC 2-cell stage (Figure 6B). This region is separate from the α , β and γ regions. It is possible that this is a separate temporal element that drives expression at the VPC 2-cell stage, or that there are general enhancers in these regions, and that without these enhancers the expression is not bright enough to see at the VPC two-cell stage. We chose not to analyse the early expression, but rather to focus on elements that drive expression at the VPC four-cell stage.

In the initial set of constructs, two separate regions were sufficient to confer vulval cell expression on the *pes-10* promoter: mk66-67 (4434-4997) and mk135-134 (2412-3419). We examined both of these regions to define the minimal sequence sufficient to confer vulval cell specificity. Vulval expression appears to be independent of anchor cell expression; the 689 bp region 2412-3101 (mk135-147), which is insufficient for AC expression, is sufficient to drive expression in vulE, C, D, and occasional expression in vulF (Figure 6C). Since mk96-143 (2290-31664) shows vulval expression while mk96-63 (2290-3008) does not, the 156 bp region 3008-3164 must be necessary for vulval expression. However, since construct mk64-65 (2989-4363) shows no expression and contains this region, it cannot be sufficient to confer vulval expression. The region 2412-2692 also appears to play a critical role in vulval expression, as demonstrated by comparing mk135-143 (2412-3164) to mk118-143 (2692-3164). Construct mk118-143 (2692-3146) shows weak expression in vulE and D, with occasional vulC expression; but there is no discernable expression in vulF, while mk135-143

(2412-3164) shows expression in vulE, F, C and D. There appear to be multiple sites involved in conferring vulva cell expression in the 2290-3164 region. The nucleotide sequence of mk96-134 (2290-3419) is shown in Figure 4C.

A second region, 4434-4997 (mk66-67, 563 bp) also appears to be sufficient to drive vulval expression (Figure 6C). There are some qualitative differences in the expression pattern when this region is compared to the other region (2412-3101) sufficient to drive vulval expression. The region 4434-4997 confers very bright vulF expression, while vulF expression in the region 2412-3101 is much weaker relative to other cell types. In addition, the vulC and vulD expression in the region 4434-4997 is weaker than the vulval expression in the 2412-3101 region. Thus, while multiple regions are sufficient to drive GFP expression in the vulva, they qualitatively differ in their detailed activity. The second region, 4434-4997, was subdivided into three overlapping regions: mk66-156 (4434-4729), mk155-67 (4719-4997), and mk158-159 (4680-4883) (Figure 6C). Of these constructs, mk66-156 (4434-4729), in a single line, showed very weak sporadic expression in vulC and vulD cells, and mk158-159 (4680-4883) drove very weak expression on rare occasion in vulE, or F. As with the other vulva cell sufficiency region, multiple sites important to all vulva cell expression must lie in this region. The nucleotide sequence of mk66-67 is shown in Figure 4D.

Transfac putative binding site predictions in upstream sequences

To find potential binding sites for transcription factors, we used the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al.*, 1995). We set the core matrix similarity to a minimum of 0.90 to maximize the specificity of the binding sites. We compared binding

sites in mk50-51 (1052-1438; *zmp-1* upstream region sufficient to drive expression in the anchor cell, vulE and vulA) to those in mk96-134 (2290-3419), the *cdh-3* region that is sufficient to drive expression in the anchor cell, as well as vulE, F, C and D) (Table 1). In this comparison, 21 shared binding sites are predicted. When we analyzed the two regions sufficient to drive *cdh-3* vulva expression, mk66-67 (4434-4962) and mk96-134 (2290-3419), we found 39 distinct binding sites that are shared between these regions. Finally, this process was utilized to compare mk84-148 (3182-4732; the *egl-17* region that is sufficient to drive vulC and D as well as early expression in the presumptive vulE and vulF cells), to mk50-51, mk96-134, and mk66-67. In this case, any putative binding site that is shared in three of these might indicate a factor involved in conferring cell specificity, since these genes express in overlapping cell types. In 12 cases, the same binding site showed up in all four regions. These might be candidates for a more general factor that drives tissue specific expression in all vulva cells. Some families are well represented in these analyses: the homeodomain family, the forkhead family, the cAMP-responsive element family, the octamer family, and the zinc finger family. However, the candidates are numerous and we have chosen to define the regions further using phylogenetic footprinting (Kirouac and Sternberg, in prep.) before attempting to distinguish between these candidates.

AlignACE predictions of over-represented sequences

The Transfac database (Quandt *et al.*, 1995) is used to identify binding sites of known transcription factors, but it is likely that motifs might exist that are uncharacterized, or that have altered binding specificities in *C. elegans* from the binding sites of known transcription factors in other systems. To determine if the apparent coordinate regulation of these genes

might indicate common DNA sequences, we used the AlignACE program (Roth *et al*, 1998), which computes motifs based on sequences that are over-represented in the input sequence. We were able to identify motifs over-represented in the sufficiency regions of *egl-17* mk84-148, *zmp-1* mk50-51, and *cdh-3* mk96-134 and mk66-67 individually. In addition, we identified motifs common to mk96-134 to mk50-51, which are each sufficient to confer expression in the anchor cell, and mk84-148, mk50-51, mk96-134, and mk66-67, which all drive expression in the vulva. These motifs may represent candidate transcription factor binding sequences that are critical for either anchor cell or general vulval expression respectively. We also compared mk96-134 to mk66-67 for reasons similar to those of the other vulval expression comparison, with the additional benefit that these two regions are located in the same upstream sequence, and might identify candidate motifs that are specific to *cdh-3* vulval expression. Analysis of these motifs should help us identify candidate sequences, known or unknown, for which to search for in upstream regulatory regions of genomic sequences for potentially co-regulated genes. One caveat of this approach to keep in mind is that some motifs that occur ubiquitously in a genome may be given a high MAP score, but have little relevance to the particular set of genes being examined.

In our analysis of *egl-17* region mk84-148, we found 14 eight-bp motifs, of which only five scored above our MAP score cut-off limit of 10, and we found an additional three motifs of six 10-bp motifs that were also above this threshold MAP score (Table 2A). Several of these candidate motifs showed multiple overlapping motifs (e.g. motif 4.8 and 2.8 share five sites); this overlap is indicative of the fact that either these motifs are really the same, or that these sites co-localize, which may identify binding sites of trans-acting factors that bind cooperatively. Some of these sites, as seen in Table 2B, are in regions that we defined in our

sufficiency analysis as being important for the fidelity of *egl-17* expression; for the early expression in the presumptive vulE and vulF cells, as well as later expression in the terminally differentiated vulC and vulD cells. For instance, motif 1.8 site 1158 is located between primers mk125 and mk102; this region (4331-4359, Figure 3A) is an important one for conferring expression in vulC and vulD. One motif, 3.8, is notable for its location in between primers mk154 and mk148. This region, 4667-4732 (Figure 3) may have repressor elements that play a role in controlling the early expression of *egl-17* seen in the presumptive vulE and vulF cells.

The analysis of the *zmp-1* sufficiency region mk50-51 yielded only one 8-bp motif (Table 2A). However, this motif has multiple sites, of which two are in regions important for the fidelity of *zmp-1* expression in vulA and vulE cells (Table 2B). Site 316 lies between the boundaries of primers mk74 and mk115. This region (1367-1378; Figure 5) is critical for conferring vulA expression. Site 100 is located between primers mk107 and mk36; this region (1165-1180, Figure 5) plays an important role in driving expression in both vulE and vulA.

We analyzed two *cdh-3* regions. The first region, mk96-134, is sufficient to drive expression both in the anchor cell and in the vulva cells. The second region, mk66-67, is able to confer vulval expression in vulE, F, C and D. In our analysis of mk96-134, we identified two 8-bp motifs and one 10-bp motif (Table 2A). The 10-bp motif, 3.10, overlaps almost entirely with motif 1.8 (Table 2B). All three of these motifs are located in multiple sites throughout the alpha, beta and gamma regions (Figure 6), which play a role in conferring anchor cell expression (Table 2B). Motif 1.8 shows a paucity of sites between primers mk136 and mk164 (Figure 4). This region may be important for conferring vulval expression.

Construct mk96-143 is able to confer vulval expression, but constructs mk96-63 and mk64-65, which divide this region in two cannot drive this expression. These results indicate that either the sites are located toward the center of this region where the break between constructs mk96-63 and mk64-65 occurs, or that sites from either end of the larger region 2290-3164 (mk96-143) are required for this expression pattern. If it is the latter case of multiple sites on either end of this region, then motif 1.8 may be a good candidate sequence, since it has few sites in the center, but multiple sites on either end of this region. We found no motifs above a MAP threshold of 10 in the region mk66-67.

To identify motifs that may be important in conferring anchor cell specificity, we compared the region mk96-134 to mk50-51; both of these regions are sufficient to drive anchor cell expression from a naïve promoter. We identified six candidate motifs (8-bp motifs and four 10-bp motifs, Table 2A). All of these motifs had sites that were located in regions important for conferring expression in the anchor cell of *zmp-1*, and all were present in multiple copies in the alpha, beta and gamma regions that are critical for anchor cell expression in *cdh-3* (Table 2B).

We identified no candidate motifs present in both mk96-134 and mk66-67. However, in our analysis of all the sufficiency regions that express in the vulva we found 13 candidates for motifs that might bind trans-acting factors that play a more general role in conferring vulva tissue specificity (Table 2A). Of these 13 candidates, all but one, motif 4.8, had at least one, and usually multiple, sites in all four regions, mk84-148, mk50-51, mk96-134 and mk66-67. Motif 4.8 was not found in mk66-67 (Table 2B).

DISCUSSION

A common assumption in the modeling of genetic regulatory networks is that the cell-specific genes expressed in a given terminally differentiated cell type are likely to be subject to coordinate control, and hence possess similar upstream *cis*-acting sequences (Davidson, 2001). While some attempts to validate this assumption in *C. elegans* have failed, other studies have succeeded. A comparison of the 5' flanking sequences of the cuticle gene *dpy-7* with other *C. elegans* cuticle genes did not reveal any striking regions of similarity (Gilleard *et al.*, 1997); a dot-matrix comparison of two acetylcholinesterase genes, *ace-1* and *ace-2*, failed to show any similarities between the two promoters (Culetto *et al.*, 1999); and the comparison of *C. elegans* MyoD family member, *hlh-1*, to mouse myogenic regulatory factors presented no striking similarities between these promoters (Krause *et al.*, 1994). The success stories lie in the studies of the inducible expression of the *C. elegans* metallothionein genes, *mtl-1* and *mtl-2*, which occur in intestinal cells (Moilanen *et al.*, 1999), and in the study of *daf-19* -regulated expression of genes expressed broadly in the sensory cilia (Swoboda *et al.*, 2000).

A comparison of the minimal promoters of *mtl-1* and *mtl-2* to other *C. elegans* intestinal cell-specific genes identified repeats of GATA transcription factor-binding sites. Mutation analyses determined that GATA elements are required for transcription, while electrophoretic mobility shift assays showed that ELT-2, a *C. elegans* GATA transcription factor, specifically binds these element. Furthermore, when *elt-2* is disrupted in *C. elegans*, *mtl-2* is not expressed, and it was also shown that ectopic expression of ELT-2 can activate transcription of *mtl-2* in non-intestinal cells of *C. elegans*. These results suggest that the binding of ELT-2 to GATA elements in these promoters regulates tissue-specific

transcription of the *C. elegans* metallothionein genes (Moilanen *et al.*, 1999). Another success story was the *C. elegans* gene *daf-19*, which encodes an RFX-type transcription factor that is expressed specifically in all ciliated sensory neurons (Swoboda *et al.*, 2000). Loss of *daf-19* function causes the absence of cilia and results in sensory defects. Twenty *C. elegans* promoters of genes expressed in ciliated sensory neurons were searched for X boxes. (X boxes are the mammalian targets for RFX-type transcription factors.) Target sites were found within the promoters of four of these genes: *che-2*, *daf-19*, *osm-1*, and *osm-6*, which are expressed in most or all ciliated sensory neurons. Target sites were not found in the promoter regions of any of the genes that are expressed in only a subset of ciliated sensory neurons, e.g. *gcy-5*, *gcy-8*, and *gcy-32*. Using an in vivo assay, it was shown that expression of the X box-containing genes was dependent on both *daf-19* function and the presence of the promoter X box. In a genome-wide search for X box-containing genes, a novel gene was examined and found to be expressed in ciliated sensory neurons in a *daf-19*-dependent manner. These data suggest that *daf-19* is a transcriptional regulator of gene products that function broadly in sensory cilia (Swoboda *et al.*, 2000).

We have attempted to address this assumption by analyzing the *cis*-regulatory sequences of three genes that have overlapping expression patterns in particular cell types within the *C. elegans* vulva and anchor cell. We chose three genes, *egl-17*, *zmp-1*, and *cdh-3*, whose function is not required for the normal development of the vulva and anchor cells, and hence lie downstream of the cell-fate specification pathways (Branda and Stern, 2000; Burdine *et al.*, 1997, 1998; Pettitt *et al.*, 1996; Wada *et al.*, 1998; J. Butler and J. Kramer, personal communication). While the roles of these three genes in the vulva and anchor cell

have yet to be determined, all three are members of families that have been shown to be involved in morphogenesis and extracellular matrix remodeling.

From our analysis of the upstream sequences of these genes, we were able to identify a number of cell-specific regulatory elements. Within these sufficiency pieces we have been able to identify multiple motifs, using the program AlignACE (Roth *et al.*, 1998), that may play a role in conferring anchor cell-specific expression, and general vulva tissue-specific expression.

A variety of regulatory mechanisms are used in *C. elegans* to control gene expression throughout development. Some genes, such as the acetylcholinesterase gene, *ace-1* (Culetto *et al.*, 1999), and the cuticle gene *dpy-7* (Gilleard *et al.*, 1997), are regulated in a relatively simple fashion by tissue-specific promoters. Other genes such as *ges-1* (Egan *et al.*, 1995), and *mec-3* (Wang and Way, 1996), are regulated in a more complex manner and require both activator and repressor elements to establish proper expression. Krause *et al.* (1994) has suggested that a more complex mechanism of control may be used in *C. elegans* to regulate genes that are expressed prior to terminal differentiation. Genes that are involved in controlling differentiation and cell fate would most likely be responsive to multiple inputs at many stages and cell interactions, as well as possessing the ability to regulate multiple gene regulatory networks, to dictate and shape these cell-fate decisions. In terminally differentiated cells, tissue identity is already established, and the need for such complex response mechanisms may be logistically unfavorable, especially in large families of genes that are likely to have partially redundant functions. In the analysis of the upstream sequences of *cdh-3*, *zmp-1*, and the late expression in vulC and vulD cells in *egl-17*, we indeed find discrete regions that direct tissue-specific expression, although each of these

regions appears to have multiple sub-elements. Moreover, we have found no evidence of repressor elements in these regions. Although, when this analysis is conducted in the context of the native promoter, negative regulatory elements may yet be found to play a role in regulating and establishing the cell specificity of these genes. We discuss below the regions directing expression of each of the cell markers that we examined in this study.

egl-17

The *egl-17* expression pattern is unique in our analysis: while it is expressed in the terminally differentiated cells vulC and vulD, it also shows expression in the presumptive vulE and vulF cells (Burdine et al., 1998). The early *egl-17* GFP expression in the presumptive vulE and vulF cells appears to be separable from the later expression in vulC and vulD cells. This early expression of *egl-17* in the presumptive vulE and vulF cells is the first marker indicating that the progeny of P6.p are specified to become primary cells (Ambros, 1999). Therefore, this expression may respond directly to the RAS-signaling pathway involved in the specification of these cells. There appear to be at least two regions directing this early expression pattern. One element that lies within 281 bp of the transcriptional start is sufficient for this pattern, but is not as strong as the full-length reporter. This expression is enhanced by an element that lies 1 kb further upstream. There also appears to be a region, 4667-4732, that may be involved in the negative regulation of early expression, as exhibited by mk153-148. This region inhibits early expression but its removal does not drive ectopic expression in *C. elegans*.

The minimal region that is sufficient for vulC and vulD cell-specific expression is 143 bp (mk125-132). There is some separability of the regions that drive vulC expression

from those that drive expression in vulD. The 5' end of this region plays a critical role for vulC expression. Likewise, the 3' end of this region is important for vulD expression. However, removal of either the 5' or 3' ends of this region substantially reduces expression levels when compared to the full-length reporter. While it is clear that there are sites that are required for expression that reside on either end of this region, it is unclear what role the remaining portion of this sufficient fragment plays in controlling the expression. Further systematic dissection of this region may elucidate other sites required for the fidelity of the expression pattern in vulC and vulD. *egl-17* expression conferred in the terminally differentiated vulC and vulD cells does not appear to be under negative regulation. Using the Transfac database (Quandt *et al.*, 1995), we found no evidence of convergence of signaling pathways, in particular ETS or WNT target sites in conserved regions at the level of this promoter. However, the alignACE program (Roth *et al.*, 1998) did identify several candidate motifs that might identify binding sites of components of these pathways, either new or as yet uncharacterized.

zmp-1

Using deletion analysis, we have defined a 380 bp region, mk50-51 (1052-1438), that is sufficient to confer vulE, vulA, and anchor cell specificity on the *pes-10* promoter. In our analysis of the *zmp-1* expression pattern, we did not identify any region that drives the weak vulD expression found in the full length reporter marker.

Multiple sites within the small 386 bp region confer expression in a reproducible, predictable fashion. When successive deletions are made on either end of this region, the expression pattern is lost in a reproducible manner: vulA expression is lost first, then vulE

expression, and finally anchor cell expression. We found segments of the *zmp-1* 5' region that drive expression in only the anchor cell, but we were not able to identify regions that confer expression only in vulA or vulE. The AlignACE program (Roth *et al.*, 1998) was able to identify a motif, 1.8, that is present in multiple copies throughout this region, and may serve as binding sites for such a factor. Yet, when sites necessary for the expression in the anchor cell are internally deleted, vulE expression, rather than anchor cell expression is lost. This observation suggests that while sequential deletions generate a reproducible pattern of expression loss, there are also some sites that are cell-type specific. Although this region appears to be part of a more complex regulatory mechanism, we saw no ectopic expression, suggesting that there are no repressor elements involved in the coordinated expression of this gene.

cdh-3

The complex *cdh-3* 5' regulatory region contains discernable tissue-specific *cis*-regulatory elements. *cdh-3* expression was examined in two tissues: the vulva (vulE, F, C, and D) and the gonad (anchor cell). The DNA elements that are sufficient to drive anchor cell expression are separable from the elements that drive expression in the vulva.

There appear to be at least three regions (α , β , and γ) that are important in anchor cell expression; two of these three elements must be present for expression. This mechanism of transcriptional control is reminiscent of the regulation of the *myo-2* gene in *C. elegans*, where three separable elements with pharyngeal enhancer activity were identified. Any combination of two of these elements, or duplication of a single element, was sufficient to confer expression in the pharynx, while singly they were inadequate to drive expression

(Okkema and Fire, 1994). When these regions are compared using the AlignACE program (Roth *et al.*, 1998), three motifs, 1.8, 2.8 and 3.10, were identified that each had multiple sites in the alpha, beta and Gamma regions. There are also two separable regions that are each sufficient to drive expression in the vulva cells. When these regions were compared to identify over-represented sequences with the AlignACE program (Roth *et al.*, 1998), no common motifs were found. Despite the fact that both regions were sufficient for vulval expression, there were qualitative differences in the strength of expression in the individual cell types. In the second region, the expression in vulF was stronger than in the first vulval region. GFP expression in vulC is weaker when driven by the second region than by the first. Although we found limited evidence that there are individual elements responsible for expression in each of the vulva cell types, we did find evidence that multiple binding sites within both of these regions are responsible for the fidelity of the expression pattern. It is possible that the loss of expression in all the vulva cells is the result of a more general regulatory mechanism on all vulva cells. This all-or-none mechanism of conferring tissue specificity is also reminiscent of the *C. elegans myo-2*, analysis in which one of the enhancer elements described above was responsible for conferring expression in all pharyngeal cells, while the other two elements identified conferred specificity for specific pharyngeal subtypes (Okkema and Fire, 1994). In the analysis of *cdh-3::GFP* expression, we never saw expansion or ectopic expression of this marker, suggesting that there are no repressor elements.

Distance of elements from translational start sites in *egl-17*, *zmp-1*, and *cdh-3*

The distances of elements that confer cell specificity do not seem to lie within a fixed distance from the translational start sites of their respective genes. In the case of *egl-17*, the transcriptional start site of *egl-17* is less than 400 bp from the elements that are sufficient to drive expression in vulC and vulD cells, and is less than 281 bp from the element that is sufficient to drive early expression in the presumptive vulE and vulF cells. However, the 386 bp *zmp-1* regulatory region that confers tissue specificity lies over 2.0 kb upstream of the translational start site of ZMP-1. Finally, the *cdh-3* regulatory regions that are responsible for anchor cell expression lie almost 3.8 kb upstream of the translational start site of the gene, while the elements that control vulva expression lie 3.6 kb and 1.6 kb from the translational start of the *cdh-3* gene.

Analysis of putative *trans*-acting factors

While the focus of this project was to isolate cell-specific *cis*-regulatory response elements rather than identifying *trans*-acting factors, we were also looking forward to the more distant goal of determining the integration of the signaling pathways in the downstream genes of these pathways. Our deletion analysis defined small regions that are critical for the fidelity of the expression pattern of these three genes; however, these regions are still broad enough to obscure the resolution of distinct binding sites. We used the Transfac database (Quandt *et al.*, 1995) to look for common putative *trans*-acting factor binding sites, as well as to indicate the integration of the known signaling pathways in the *cis*-acting regions. We found no obvious candidate sites based on location within analyzed regions to known transcription factors involved in *C. elegans* vulval patterning.

While a number of transcription factors (for example *egl-38*, *lin-26*, *lin-29*, *cog-1* and *lin-11*) (Bettinger *et al.*, 1997; Chamberlin *et al.*, 1997; Freyd *et al.*, 1990; Labouesse *et al.*, 1994; Rougvie and Ambros, 1995; Palmer *et al.*, in prep) are known to affect the marker-gene expression patterns in the vulva, it is unclear at this time whether they are acting directly in the regulation of the genes, or more proximally in the specification of these cell types (M. Wang and P. Sternberg, unpublished data). Biochemical studies using the sufficiency pieces defined in this study will determine which of these factors has a direct effect on the transcriptional regulation of these genes.

An anchor cell element that drives transcription of LIN-3 has been identified, and involves two trans-acting factors (B. Hwang and P. Sternberg, in prep). Removal of these factors does not disrupt the expression of *cdh-3* or *zmp-1::GFP* in the anchor cell (B. Hwang and P. Sternberg, in prep). This observation suggests that there are at least two different mechanisms and/or factors that are used to establish the anchor cell expression.

Analysis of overrepresented sequences in regions of sufficiency

While the Transfac database (Quandt *et al.*, 1995) identifies binding sites of known transcription factors, AlignACE (Roth *et al.*, 1998) identifies sequences that are over-represented in a given sequence. This approach allows the isolation of candidate motifs either within a gene, or between genes. As discussed above, we were able to identify a number of candidate motifs that bound in mk84-148, mk50-51, and mk96-134. Additionally, we compared sufficiency regions that are expressed in the same tissue to identify common motifs that may play a role in conferring cell-type-specific expression in co-regulated genes. Through these inter-regulatory region comparisons, we have identified candidate motifs that

may play a role in anchor cell expression, and a more general vulva tissue-specific expression.

Conclusions

By analyzing the functional anatomy of tissue-specific and cell-specific patterns of three reporter genes, *zmp-1*, *egl-17* and *cdh-3*, we have narrowed a 3.9 kb upstream region of *egl-17* to a 143 bp region of *egl-17* that confers vulC and vulD expression, and a separate 102 bp region sufficient to drive the early expression in presumptive vulE and vulF cells. A 3.5 kb *zmp-1* upstream region has been narrowed to a 300 bp region that is sufficient to confer expression in vulE, vulA, and the anchor cell. Moreover, a 6.0 kb upstream region of *cdh-3* upstream sequence has been delimited to: a 689 bp region sufficient to drive expression in the anchor cell and vulE, vulF, vulD and vulC; a 155 bp region sufficient to drive anchor cell expression; and a separate 563 bp region also sufficient to drive expression in these vulval cells. One theme that remained the same in all the analyses is that we failed to identify any repressor elements involved in conferring expression in terminally differentiated cell types. However, we identified regions of similarity between these three *cis*-regulatory sequences, and provided evidence for several different mechanisms through which *C. elegans* regulates transcription. These mechanisms include the use of discrete, separable elements that confer cell-type specific expression (*cdh-3* in the anchor cell and *egl-17* in sister cells), the use of complex patterns of binding sites that act combinatorially to establish the fidelity of expression in a variety of cell types from different lineages (*zmp-1*), and tissue-specific elements responsible for driving expression in an entire tissue rather than in sub-domains of its constituent cells (*cdh-3* vulval expression). Finally, we have been able to isolate candidate

motifs for each of these regions, and between anchor cell specific, and vulva specific regions that may be important for the fidelity of these expression patterns. We can now use these candidate motifs as targets for mutational analysis and for searching the genome for other candidate genes that have these motifs, to test if they are expressed in a similar fashion.

Acknowledgments

We are grateful to Jim Butler and Jim Kramer for *zmp-1::GFP* and the sharing of unpublished data; Rebecca Burdine and Micheal Stern for *egl-17::GFP*; and Jonathan Pettitt, William B. Wood and Ronald Plasterk for the *cdh-3::GFP*. We would like to thank Cheryl Van Buskirk, Takao Inoue, Erich Schwarz, David Sherwood, Bhagwati Gupta and Rene Garcia for the critical reading of this manuscript. P.W.S is an investigator with the Howard Hughes Medical Institute, which supported this work.

REFERENCES

- Ambros, V. (1999). Cell cycle-dependent sequencing of cell fate decisions in *Caenorhabditis elegans* vulva precursor cells. *Development* **126**, 1947-1956.
- Beitel, G., Tuck, S., Greenwald, I., and Horvitz, H. (1995). The *Caenorhabditis elegans* gene *lin-1* encodes an ETS-domain protein and defines a branch of the vulval induction pathway. *Genes & Development* **9**, 3149-3162.
- Bettinger, J., Euling, S., and Rougvie, A. (1997). The terminal differentiation factor LIN-29 is required for proper vulval morphogenesis and egg laying in *Caenorhabditis elegans*. *Development* **124**, 4333-4342.
- Branda, C., and Stern, M. (2000). Mechanisms controlling sex myoblast migration in *Caenorhabditis elegans* hermaphrodites. *Developmental Biology* **226**, 137-151.
- Burdine, R., Branda, C., and Stern, M. (1998). EGL-17(FGF) expression coordinates the attraction of the migrating sex myoblasts with vulval induction in *C. elegans*. *Development* **125**, 1083-1093.
- Burdine, R., Chen, E., Kwok, S., and Stern, M. (1997). *egl-17* encodes an invertebrate fibroblast growth factor family member required specifically for sex myoblast migration in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences USA* **94**, 2433-2437.
- Chamberlin, H., Palmer, R., Newman, A., Sternberg, P., Baillie, D., and Thomas, J. (1997). The PAX gene *egl-38* mediates developmental patterning in *Caenorhabditis elegans*. *Development* **124**, 3919-3928.
- Chang, C., Newman, A., and Sternberg, P. (1999). Reciprocal EGF signaling back to the uterus from the induced *C. elegans* vulva coordinates morphogenesis of epithelia. *Current Biology* **9**, 237-246.
- Clark, S., Chisholm, A., and Horvitz, H. (1993). Control of cell fates in the central body region of *C. elegans* by the homeobox gene *lin-39*. *Cell* **74**, 43-55.
- Culetto, E., Combes, D., Fedon, Y., Roig, A., Toutant, J., and Arpagaus, M. (1999). Structure and promoter activity of the 5' flanking region of *ace-1*, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *Journal of Molecular Biology* **290**, 951-966.

- Davidson, E. H. (2001). Regulatory Hardwiring: A Brief Overview of the Genomic Control Apparatus and its Causal Role in Development and Evolution in: "*Genomic Regulatory Systems: Development and Evolution.*" Academic Press, 1-24.
- Egan, C., Chung, M., Allen, F., Heschl, M., Vanbuski, C., and McGhee, J. (1995). A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans ges-1* gene centers on two GATA sequences. *Developmental Biology* **170**, 397-419.
- Euling, S., Bettinger, J., and Rougvie, A. (1999). The LIN-29 transcription factor is required for proper morphogenesis of the *Caenorhabditis elegans* male tail. *Developmental Biology* **206**, 142-156.
- Freyd, G., Kim, S., and Horvitz, H. (1990). Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-11*. *Nature* **344**, 876-879.
- Gilleard, J., Barry, J., and Johnstone, I. (1997). *cis*-regulatory requirements for hypodermal cell-specific expression of the *Caenorhabditis elegans* cuticle collagen gene *dpy-7*. *Molecular and Cellular Biology* **17**, 2301-2311.
- Greenwald, I. (1997). Development of the Vulva in: "*C. elegans II.*" DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press. **II**, 519-541.
- Hill, R., and Sternberg, P. (1992). The gene *lin-3* encodes an inductive signal for vulval development in *C. elegans*. *Nature* **358**, 470-476.
- Hughes, JD., Estep, PW., Tavazoie, S., and Church, GM. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205-1214.
- Katz, W., Hill, R., Clandinin, T., and Sternberg, P. (1995). Different levels of the *C. elegans* growth factor LIN-3 promote distinct vulval precursor fates. *Cell* **82**, 297-307.
- Kimble, J. (1981). Alterations in cell lineage following laser ablation of cells in the somatic gonad of *C. elegans*. *Developmental Biology* **87**, 286-300.

- Krause, M., Harrison, S., Xu, S., Chen, L., and Fire, A. (1994). Elements regulating cell- and stage-specific expression of the *C. elegans* myoD family homolog *hlh-1*. *Developmental Biology* **166**, 133-148.
- Labouesse, M., Sookhare, S., and Horvitz, H. (1994). The *Caenorhabditis elegans* gene *lin-26* is required to specify the fates of hypodermal cells and encodes a presumptive zinc-finger transcription factor. *Development* **120**, 2359-2368.
- Mello, C., Kramer, J., Stinchcomb, D., and Ambros, V. (1991). Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO Journal* **10**, 3959-3970.
- Moilanen, L., Fukushige, T., and Freedman, J. (1999). Regulation of metallothionein gene transcription-Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from *C. elegans*. *Journal of Biological Chemistry* **274**, 29655-29665.
- Newman, A., and Sternberg, P. (1996). Coordinated morphogenesis of epithelia during development of the *Caenorhabditis elegans* uterine-vulva connection. *Proceedings of the National Academy of Sciences USA* **93**, 9329-9333.
- Newman, A., White, J., and Sternberg, P. (1996). Morphogenesis of the *C. elegans* hermaphrodite uterus. *Development* **122**, 3617-3626.
- Okkema, P., and Fire, A. (1994). The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175-2186.
- Pettitt, J., Wood, W., and Plasterk, R. (1996). *cdh-3*, a gene encoding a member of the cadherin superfamily, functions in epithelial cell morphogenesis in *Caenorhabditis elegans*. *Development* **122**, 4149-4157.
- Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**, 4878-4884.
- Roth, FP., Hughes, JD., Estep, PW., and Church, GM. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **10**, 939-945.

- Rougvie, A., and Ambros, V. (1995). The heterochronic gene *lin-29* encodes a zinc finger protein that controls a terminal differentiation event in *Caenorhabditis elegans*. *Development* **121**, 2491-2500.
- Seydoux, G., and Fire, A. (1994). Soma-germline asymmetry in the distributions of embryonic RNAs in *Caenorhabditis elegans*. *Development* **120**, 2823-2834.
- Sharma-Kishore, R., White, J., Southgate, E., and Podbilewicz, B. (1999). Formation of the vulva in *Caenorhabditis elegans*: a paradigm for organogenesis. *Development* **126**, 691-699.
- Sternberg, P., and Han, M. (1998). Genetics of RAS signaling in *C. elegans*. *Trends in Genetics* **14**, 466-472.
- Sternberg, P., and Horvitz, H. (1986). Pattern formation during vulval development in *C. elegans*. *Cell* **44**, 761-772.
- Sulston, J., and Horvitz, H. (1977). Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology* **56**, 110-156.
- Swoboda, P., Adler, H., and Thomas, J. (2000). The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in *C. elegans*. *Molecular Cell* **5**, 411-421.
- Tan, P., Lackner, M., and Kim, S. (1998). MAP kinase signaling specificity mediated by the LIN-1 Ets/LIN-31 WH transcription factor complex during *C. elegans* vulval induction. *Cell* **93**, 569-580.
- Wada, K., Sato, H., Kinoh, H., Kajita, M., Yamamoto, H., and Seiki, M. (1998). Cloning of three *Caenorhabditis elegans* genes potentially encoding novel matrix metalloproteinases. *Gene* **211**, 57-62.
- Wang, L., and Way, J. (1996). Promoter sequences for the establishment of *mec-3* expression in the nematode *Caenorhabditis elegans*. *Mechanisms of Development* **56**, 183-196.
- Wang, M., and Sternberg, P. (2000). Patterning of the *C. elegans* primary vulval lineage by RAS and Wnt pathways. *Development* **127**, 5047-5058.

Figure 1: Marker gene expression summary

Nomarski image of the vulva of an L4 animal showing the anchor cell as well as vulA, B1, B2 and D cells (vulC, E, and F are not in this focal plane). The three markers *egl-17::GFP*, *zmp-1::GFP* and *cdh-3::GFP* are expressed in different cell types of the vulva at various developmental stages. This figure shows the stages at which the expression of three marker genes was scored.

Figure 2: Initial dissection of *egl-17*, *zmp-1* and *cdh-3* regulatory regions

For each of the three genes, the genomic region was divided into large sub-pieces of approximately 1kb each. **(A)** The genomic region of *egl-17* contains 3819 bp of upstream sequence. The first exon of the transcript starts at 4609, with the translational start at 4708. Nucleotide(nt) 790 of the *egl-17* upstream region corresponds with nt 17,648 in Genbank cosmid F38G1 (Accession # AC006635). **(B)** The *zmp-1* genomic region contains 3472 bp of upstream sequence. The start site of ZMP-1 is at nt 3,473. Nucleotide 1 of the *zmp-1* upstream region corresponds with nt 7630 in Genbank cosmid EGAP1 (Accession # U41266). AC stands for anchor cell. **(C)** The genomic region of *cdh-3* contains 5,928 nucleotides of upstream sequence. The translational start site occurs at nt 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nt 37,343 in Genbank cosmid ZK112 (Accession # L14324).

Figure 3: Upstream regions that direct *egl-17* expression

The list of constructs does not encompass all constructs that were made. (For a comprehensive list see Supplemental Figure 2.) (A) The constructs that were informative in determining two regions, 4331-4359 and 4466-4474, shown in pink, important in driving *vulC* and *vulD* expression. (B) The constructs that were most informative in determining the regions that drive the early expression in the presumptive *vulE* and *vulF* cells are depicted. The first region, 3182-3611, highlighted in orange, shows that, while not sufficient alone to drive the early *egl-17::GFP* expression, when combined with the region shown in blue, 4565-4667, drives GFP expression at a level comparable to the full-length reporter construct. The region highlighted in blue depicts the region that alone is sufficient to drive expression in the presumptive *vulE* and *vulF* cells. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. Mk80-104 showed very weak *vulC* expression in 1/2 lines. 102-56 showed weak expression in *vulD* in 3/3 lines. On rare occasion, expression in *vulC* and *vulD* was seen in mk103-148. The early expression for this construct was variable from line to line. mk153-154 shows variable expression in the presumptive *vulE* and *vulA* cells, although this expression is neither as weak nor as variable as that seen in mk103-148, mk84-20, mk82-100 and mk15-20.

Figure 4: Upstream sequences of mk84-148, mk50-51, mk96-134 and mk66-67

Nucleotide sequences are shown for: the *egl-17* region mk84-148, with the translational start site in bold (A); *zmp-1* region mk50-51 (B); *cdh-3* region mk96-134 (C) ; and *cdh-3* region mk66-67 (D). Arrows show the end points and direction of primers in the regions.

Figure 4: Upstream sequences of mk84-148, mk50-51, mk96-134 and mk66-67

(A) *C. elegans egl-17* mk84-148

3181 ^{mk84} TCACGTCTCTCCCCCGTCACCTCCCTTCTTTCACGTCCTTGGTAATTTTCATATGT
 3241 ATGTTTGCTTGGCCACACATGGCGAAAAAGACAGTTTCATAACCGAAAGCGTAGCCCAA
 3301 TTTCTTAAACTACTTTTCCAAATGACGTTTTTAAAGACATGAGAAGCCAGGAAAAACCGGG
 3361 TAAAGTTGTTGCGTAATCTATACCAAACGTTTTTTTTTTCTGCTTGTCTCTGTTTAC
 3421 TTGTCACCGTTCAGTTTTTCATGTGATGTTAATAAATTTTCTGAGGTTAAAGTTTTT
 3481 CAATGTTTTTTTTGTTTAAAGTGACTATCTCTGTTGGAGATTGCTTTAAAGATT
 3541 CCTATGGGTCACAATGACCGAATATCATGATATAAAAAATTCAAAAAAATTCAGATTT
 3601 TATATGATTTTTGGGAATTTGGAATAAATCTCAGTTTTCCCTAATTCCTATTGAATTAC
 3661 CGCCTATTGAACTCGTTCTGTTGGAGCGCTTGAATTAATTTTCATTAATGTTTTATTGG
 3721 TTCTCATTTTTCACGTTGTTAGTGAATAATGAGAACATAAAAAATTAATGAAAAATAAT
 3781 GCAATCGCGCTCCAACGAACGAGTTCAATTTGGCGGTAATTCAAATAGGAATTAGGGGAAA
 3841 ACTGAGATTTTTCGAATTTCAAAAAAATAATTTAAAATCTAGAAATGTTTGAATTT
 3901 TTTATCATGATATTCGGTCATTTGACCCCATAGGCAAGTTCCGTATAGGTGTGATAAG
 3961 TAGCTTCGAGAAAACAATTTAGACTAAATCTCATCGTTTGAATTAATTTGGTTCATGTA
 4021 CAGATCTTTCATATATAACTACTTTTTATGCTCTTTCGATTACTTTCAAATCTGTGCA
 4081 TTACTCCAGAAGGGGATTTTGCAAATTTCTGAAGATTGTAGTAGCATTTAAGGGTATAG
 4141 CTCTCCGCTAAATTTTGGCGATACCTACTTTCAAAAAAACGAAAACATGTTTCTTGTA
 4201 AGCTTTAAACCTACTCACCACAAGTTATATTTTGTGTGTACCACATGTATGAAAA
 4261 TGTCATCTTAATATGATGTCAGTCAATAGTTTTCTCAGTTTTCTAGTTTCCCCCTCA
 4321 TCTCTTATATCGTCTGTCTTTACCACTTTCCCGCTCTCGAACAATAATGTCGGACAAC
 4381 TCAAGTTGTAATTACAATGTGTTTTGAAAGAAAAAGTGACAAAAAGTTGATTAATTC
 4441 TTGTTTCTGATCTGATTTCTTCCAACGAACACCGCGCTTCTTCTACGTGGCGTCTCAGC
 4501 CGCTCGATATGTTACTTTTGAATATGTTTTTCAATTGCATTTTTAGTTTCCGTTTTTGT
 4561 TTTTACCAATGTGTGTCGCCGCTGTGAAAAATCGTTTTACAGGCATCCATCTTGTATTCC
 4621 GACTCTAATTTATAAAATTTCAAGTTGGTCCACTTGTTCATGTCACAATTAATAACAAT
 4681 GATTTTTCAGGTGCCGAAATGTGAGCTATGCTCAAAGTCTACTACCCCTGATG

(B) *C. elegans zmp-1* mk50-51

992 TTTTATGTAAGTTTATGCGCCCTCGAGAGAAAGATGATTTTCGTAACCCATTTCAAAA
 1052 GAAGGACGGCTCGTTGAACAGAAATACACAGATTTCTGTTCCAATGGAGATTTTCCCTTT
 1112 TCTGATTTGATCATCAAAGTATTCGAGTACGTTTACACTGGTTTCTGTTCTTCCGTTTT
 1172 TAATTTCTCTGCTGCCAGATGCAAACTGATTCATGTGTACGTATTGCTTGAAAAAAGAGTA
 1232 ACAAGAAAAAGTAGAAGGGTATTAGTCTAGTAGTAGTATTTCAGTTGTAGTAATATAT
 1292 TTCTACTAATTTGTTAGTTTTCGCCACTTAAGATGGTCAATCGCAATTTTCAATTAATTT
 1352 TTGGTGGACTTTTCAGAAGAGAAACCTCGAAATATTTTATGAATGGAAATGTGACAGT
 1412 TTTTGTTCATATGGCCATTTCTAG

(C) *C. elegans cdh-3* mk96-134

2290 ^{mk96} CCGCATTTTTCATCAAGATTCCACAAAAGTTCAGATTTCCAAAACAGAAAAAAACAAATAAAAGGCA
 2357 CCTGACAAATCTCAGAAATCGGAGAATGATTGAGAAGGAGCAGGTGCACACAGTTCTCTGCCACTT
 2424 GCCCCATTTCTTTTAAAGCAGTTGAAATAAGAACACCTGCTTCTCGGAGATTGACACAAAACCGGAA
 2491 CGGTAGCCAATGTTTATGTGTATGAATAATGAATGTTTGGATTCTTCTATAAATTTAGATTTTT
 2558 TGTCTTTTTAGTGATAGGTTACTGCAGAGTTTTGTTTACATTGATTAAGTCAATTTGAAATCTGATT
 2625 TTTAATTTTTGAAATGAGTTTTTAATTAATCTTCTGCATTTCAAATATTTCTGTTAATTTTATT
 2692 GACGACAACTTAATGAAATTTGAAATGTAGCTACCAAAAAATGGCTTGTCTGAAAAAATCTCT
 2759 TACTTCTTGGCAAACTTTTACAACCTTCTATGTATCTGTCAACATATTTAAGGGGTTTTAGTAAAT
 2826 TGTAGTGTGATACACTACCACAGCCTTAAGCCTATATCTTTGATAACTCGTATTCAGATTTT
 2893 TCACATCTTTTCAATTTTCATTTTCATATTTTATTCCTGATACCGTTTTGCGTATTGTCA
 2960 AACACCGAGACGATGGTCACCTCCCTATACAAAACGAGCCGACCGTCCCAAAAAAAGTTGTGAAACA
 3027 ATTAGAGTCTCGAGGCGGTTGTTGTTGTCGTCATCCCGCTTCCAATCCATTTTCGAGCCCTATGAC
 3094 TACACTACCACCTGCCTTTTGTGTGTCGTTCCGCGGTGCCGCCCTGTTCAACTTCGACCAATGCA
 3161 TGCTCAATTTTGTTCATCTAGGACCGATTTTTGGGATGAAGAACCTTGTGTTATGTTACTCTTAAT
 3228 GATTTGGGTTATTTCTACTTTTTTAAATTTTAAATATTTTCATGAAATGGTAGCGATTCCGTACCTTAT
 3295 ATTTTGTACACAAGCATAATTTTCTTATATTTCTGTCATTTTGTCTCAAAAACGAGTAAAAAA
 3362 TTTTCTAGTAAAAAATTTTGATATAAAAGTTAAATAACAAAGCCGGCAGTTTTATG

(D) *C. elegans cdh-3* mk66-67

4434 ^{mk66} GTGAAAGCTCCAGGAGCTGAAACCAATAGTTTTTTTTTCAATTTGAATTTTTCATCTTATTATTTC
 4500 TAACCTCTTTGAATTTAATGAATAAACCTTTCACATTACAATCCTGTTTTTATTCACCGAATTTTC
 4566 AGCCTGTAATAATTTGTATCCCAAGTCAAAGATTTCTATAAAAAGTATTTTCCAACTGTTCCGAT
 4632 GTTGCCGAAACTCATGTAACCTTTGAAAAGTCTGTTCAAACCTTATTACCTTGATTTCTTGTATA
 4698 TCCAATTTTCGAGATTGCTTTCACACCACACAGTGCCAAATTTGCTTTCCACTTAGATCGGAAGGC
 4764 GGTCTCTTCTGTCTCTCATAGTTTCACACCTTTTCCCTTCCGTCAGTCACAGTCCCTTTTCTT
 4830 CCAATCTCCAATCCAATATGTCCTTTTGTATGCTAATTTGCATTTCTGTCGCCGCGCCCAAT
 4896 TCAACCTAATCTAACACATTTTCTGTTGATTTCCGGCCCTGTATCTATTTGTTGAAATACCG
 4962 CATCGTCTTCTTTAGCGTTTCTTGGGACCACT

Figure 5: Multiple regions direct *zmp-1* expression

The construct that confers GFP expression in vulE, vulA, and anchor cell (AC) at a similar level as the full-length reporter construct is shown at the top. The expression pattern of each construct shown is located at the right. The colored zones represent regions that confer GFP expression in a particular cell type: orange regions are those areas that contribute to vulA cell expression; yellow regions contribute to both vulE and vulA expression; blue regions important for driving expression in vulE; and purple regions are those regions that contribute to anchor cell expression. mk106-51 is the smallest construct that drives expression in the three cell types (depicted in green). The small boxes depict these regions graphically. If there is more than one box in a region, then that region is important for driving GFP expression in multiple cell types, depending on the surrounding DNA context. The sizes of each of these regions are listed in the appropriate box. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. A gap in the graphical depiction in construct mk Δ 3/4 indicates an internal deletion.

Figure 6: Regions that direct *cdh-3* expression

The name of the constructs, a graphical depiction of their location with respect to the full-length upstream region, and a summary of the expression of each construct is shown for each panel. These lists of constructs do not encompass all constructs that were made. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. (A) *cdh-3* constructs that illustrate the importance of two regions that direct the expression of the GFP in the anchor cell from the VPC 4-cell stage. The alpha, beta and gamma sub-regions are also shown. (B) The constructs that are listed illustrate the importance of two regions that confer expression of GFP in the anchor cell from the 2-cell stage of the VPC. (C) The constructs that are listed illustrate the importance of these two regions in directing the expression of the GFP in vulE, F, C and D. The first region, in blue, is bounded by nucleotides 2412-3101, and the second region, in yellow, is bounded by nucleotides 4434-4997. Construct mk66-156 shows variable weak expression in the occasional animal in vulC and vulD, while mk158-159 shows variable weak expression in the occasional animal in vulC, E and F, but never vulD.

Table 1: Transfac binding site predictions for regions that confer cell-specific expression

The program MatInspector was used to make Transfac Database binding site predictions. Construct mk50-51 was compared to mk96-134, both of which can drive expression in the anchor cell. Construct mk96-134 was compared to mk66-67 since both of these *cdh-3* regions can confer vulva expression. Finally, *egl-17* mk84-148, *zmp-1* mk50-51, *cdh-3* constructs mk96-134 and mk66-67 were compared, since all of these constructs overlap in the vulva cells. Transfac prediction binding sites were listed that meet the following criteria: (1) the minimum core binding specificity had to be at least 0.90, and (2) the maximum Random Expectation Value, "re", which is the number of times this site would appear in a random 1000 bp, was not exceed 0.51. The number of sites in the first site is followed by a slash, and then the number of sites in second region. * These factors were not necessarily found in both, but were included because they are part of potentially interesting transcription families.

Table 1: Transfac binding site predictions for regions that confer cell-specific expression

		Anchor Cell	vulF, E, C and D	vulF, E, C and D
FACTOR FAMILY	FACTOR	mk50-51/ mk96-134	mk96134/ mk66-67	mk84-148/mk50-51/ mk96-134/mk66-67
ARS binding factor	ABF1.01	1/2		
cAMP-Responsive Element	ATF.02		1/1	1/0/1/1
cAMP -Responsive Element	CREB.01		1/1	1/0/1/1
cAMP -Responsive Element	CREB.03		1/1	
cAMP -Responsive Element	CREB.04		1/1	1/0/1/1
cAMP -Responsive Element	CREBP1.02		1/1	1/0/1/1
Cart-1 (cartilage homeoprotein)	CART1.01	1/1	1/1	4/1/1/1
CLOX FAMILY	CDP.01	1/1		1/1/1/0
CLOX FAMILY	CDPCR3.01		1/1	3/1/1/1
Enhancer-CcAaT binding factors	NFY.01			2/3/2/0
Enhancer-CcAaT binding factors	NFY.02			2/1/2/1
ETS	PU.1ETS *		2/1	2/0/2/1
EVI myleoid transforming protein	EVI1.02	1/2	2/3	0/1/3/2
Floral determination	MADSA.01	1/4		2/1/5/0
Fork Head and Related	FREAC2.01			2/1/4/0
Fork Head and Related	FREAC4.01			1/1/2/0
Fork Head and Related	HFH1.01			1/1/2/0
Fork Head and Related	HNF1.01	1/1		1/1/1/0
Glucocorticoid Responsive	GRE.01		2/1	
Glucocorticoid Responsive	PRE.01	1/1		
Homeodomain Factor	FTZ.01		3/4	4/0/3/1
Homeodomain Factor	NKX25.02	1/7	8/1	6/1/8/1
Homeodomain Factor	NKX31.01		1/3	
Homeodomain Factor	PBX1.01	1/3		
Homeodomain Factor myeloid leukemia	MEIS1.01	1/1		3/1/1/0
Homeodomain Pancreatic /Intestinal (LIM domain family)	ISLI1.01	1/2	2/2	5/1/2/2
Homeoprotein Caudel	CDX2.01	1/4	4/2	5/1/4/2
HSF family	FHSF.03	1/2		2/1/2/0
HSF family	FHSF.04			2/1/0/1
Human muscle-specific Mt binding site	MTBF.01		2/2	2/0/2/2
Interferon Regulated Factor	IRF1.01	1/2	2/1	2/1/2/1
Interferon Regulated Factor	IRF2.01		1/1	2/1/1/1
Interferon Regulated Factor	ISRE.01			1/1/1/0
MYB-Like protein (Petunia)	MYBPH3.01		1/1	
Octamer Family	OCT.01			2/1/2/2
Octamer Family	OCT1.01		3/2	3/1/1/2
Octamer Family	OCT1.02			2/2/0/2
Octamer Family	OCT1.03		3/3	7/0/3/3

Octamer Family	OCT1.04		1/2	2/1/1/2
Octamer Family	OCT1.05			4/0/1/2
Octamer Family	OCT1.06		3/3	10/0/3/3
Paired homeodomain factors	PAX6 HD.01	1/1		6/1/2/0
<i>Phaseolus vulg.</i> SILEncer reg. of chalcone synth. prom.	SBF1.01	1/4		3/1/4/0
Plant P-Box binding sites	PBOX.01	1/1	1/1	0/1/1/1
Poly A	APOLYA.01		2/2	3/0/2/2
Poly A	POLYA.01		1/2	1/0/1/2
Promoter-CcAaT binding factors	ACAAT.01		1/2	1/0/1/2
Repr. of RXR-mediated & retinoic acid responses	COUP.01		1/1	
signal transducers and activators of txn	STAT1.01		1/1	
signal transducers and activators of txn	STAT3.01		1/1	1/0/1/1
SMAD Family TGF-B	FAST1.01	1/2		
Special AT rich binding Sequence	SATB1.01	1/1	1/1	0/1/1/1
TATA FAMILY	TATA.02		1/2	9/0/1/2
Tata-Binding Protein Factor	ATATA.01			
TCF/LEF	LEF1.01 *	2/1	2/1	2/1/2/1
Vertebrate steroidogenic factor	SF1.01		1/1	
Yeast CCAAT binding factors	HAP234.01			4/0/1/1
<i>Zea mays</i> Transcriptional activator OPAQue-2	O2.03		1/1	
zinc finger POZ domain B-Cells	BCL6.02	1/1	1/1	0/1/1/1
zinc finger W Box family	WRKY.01		2/1	1/0/2/1
zinc finger <i>Xenopus</i> MYT1 C2HC	MYT1.01		4/1	
zinc finger <i>Xenopus</i> MYT1 C2HC	MYT1.02			5/0/4/1
<i>C. elegans</i> maternal gene	SKN1.01		1/1	2/0/1/1

Table 2: AlignACE predictions of overrepresented sequences

(A) A summary table of the number of motifs found in each of the listed regions. The total number of motifs identified by AlignACE is shown in parentheses, while the number of motifs that scored above the MAP score threshold of 10 is shown outside the parentheses for both the eight- and 10-bp motifs. The left-hand column identifies the cell-type expression of interest when two or more regions were compared. * Indicates that all four regions, mk84-148, mk50-51, mk96-134, and mk66-67 were compared. (B) This table summarizes the data for each of the motifs listed in Table 2A that had a MAP score over 10. The region is listed in the left-hand column. The motif numbers are consecutive and are followed by the size of the motif. The MAP score for each motif is shown under the column head MAP. The sites for each motif are listed. If more than one region was compared, the sites for the first, as indicated by the left-hand column, are in parentheses, followed by the second set of sites in parentheses, and so on. Abbreviations are as follows: expr., expression; imp., importance and; elem. element. The pictograms were generated using the Pictogram program (<http://genes.mit.edu/pictogram.html>).

Supplemental Table 1: PCR primers

The PCR primers used to generate the constructs that were analyzed in this study are listed.

Supplemental Table 1: PCR Primers

PRIMER	GENE	SITE	SEQUENCE OF PRIMER
mk01	egl-17	SphI	5' CCC CCG CAT GCC ACT ATA GAA TAC ATA GGA TC 3'
mk02	egl-17	Sall	5' CCC CCG TCG ACT TTT CAC AGC GGG GAC ACA CAT TGG 3'
mk09	zmp-1	SphI	5' CCC CCG CAT GCG TGT TTA ATT TTG ACC CAA AGA TGC 3'
mk15	egl-17	SphI	5' CCC CCG CAT GCC CAT CTT ACG GTT ATA TTC 3'
mk16	egl-17	StuI	5' CCC CCA GGC CTG GAA TAT AAC CGT AAG ATG G 3'
mk20	egl-17	StuI	5' CCC CCA GGC CTG CGC GCT CCA ACG AAC GAG 3'
mk27	egl-17	SphI	5' CCC CCG CAT GCG TGG ACT ATA CTC TGT GGG 3'
mk29	zmp-1	SphI	5' CCC CCG CAT GCC TTG AAT CTA GCT ATA TGT AG 3'
mk30	zmp-1	XbaI	5' CCC CCT CTA GAC TAC ATA TAG CTA GAT TCA AG 3'
mk31	zmp-1	SphI	5' CCC CCG CAT GCC AGT AAC CAA GCA CTC GTT ATC 3'
mk32	zmp-1	XbaI	5' CCC CCT CTA GAG ATA ACG AGT GCT TGG TTA CTG 3'
mk33	zmp-1	SphI	5' CCC CCG CAT GCC ATA TGC TAC CTT CAC CAG C 3'
mk34	zmp-1	XbaI	5' CCC CCT CTA GAG CTG GTG AAG GTA GCA TAT G 3'
mk35	zmp-1	XbaI	5' CCC CCT CTA GAG CTG ACT CAT TAG CAC AAG AC 3'
mk36	zmp-1	SphI	5' CCC CCG CAT GCC TGC CAG ATG CAA ACT GAT TC 3'
mk37	zmp-1	XbaI	5' CCC CCT CTA GAG AAT CAG TTT GCA TCT GGC AG 3'
mk45	egl-17	HindIII	5' CCC CCA AGC TTC GCG CTC CAA CGA ACG AGT TC 3'
mk48	egl-17	SphI	5' CCC CCG CAT GCG CTT ACA AGA AAC ATG TTT TC 3'
mk49	egl-17	SphI	5' CCC CCG CAT GCC ACA GCG GGG ACA CAC ATT GG 3'
mk50	zmp-1	SphI	5' CCC CCG CAT GCG AAG GAC GGC TCG TTG AAC AG 3'
mk51	zmp-1	XbaI	5' CCC CCT CTA GAC TAG AAA ATG GCC AAT ATG C 3'
mk52	zmp-1	SphI	5' CCC CCG CAT GCG ATC ATC AAA GTA TTC GAG 3'
mk53	zmp-1	XbaI	5' CCC CCT CTA GAC TAC AAC TGA ATA CTA CTA CGA C 3'
mk54	zmp-1	XbaI	5' CCC CCT CTA GAC AAG CAA TAC GTA CAC ATG 3'
mk55	zmp-1	XbaI	5' CCC CCT CTA GAG CGA TGA CCA TCT TAA GTG GCG 3'
mk56	egl-17	XbaI	5' CCC CCT CTA GAG TAA CAT AAT CGA GCG GCT GAG 3'
mk57	egl-17	SphI	5' CCC CCG CAT GCG CAT TTA AGG GTA TAG CTC TTC CC 3'
mk58	cdh-3	SphI	5' CCC CCG CAT GCG GAG GGT ACC ATG GCC ATC CC 3'
mk59	cdh-3	XbaI	5' CCC CCT CTA GAG CGG AAC ATC GAT TCT ATG G 3'
mk60	cdh-3	SphI	5' CCC CCG CAT GCC CAT AGA ATC GAT GTT CCG C 3'
mk62	cdh-3	SphI	5' CCC CCG CAT GCC TAG AGC ATG ATG TCC TTA CC 3'
mk63	cdh-3	XbaI	5' CCC CCT CTA GAG GGA CGG TCG GTC CGT TTT G 3'
mk64	cdh-3	SphI	5' CCC CCG CAT GCC AAA ACG GAC CGA CCG TCC C 3'
mk65	cdh-3	XbaI	5' CCC CCT CTA GAC ACT AGT TAC TCC AAC TGA TC 3'
mk66	cdh-3	SphI	5' CCC CCG CAT GCG TGA AAG CTC CAG GGA GCT G 3'
mk67	cdh-3	XbaI	5' CCC CCT CTA GAC AGA TGG TCC CAA GAA ACG C 3'
mk68	cdh-3	SphI	5' CCC CCG CAT GCG CGT TTC TTG GGA CCA TCT G 3'
mk69	cdh-3	XbaI	5' CCC CCT CTA GAG TCA TCT ATT CAG CAT TGA TC 3'
mk70	zmp-1	SphI	5' CCC CCG CAT GCC GCC ACT TAA GAT GGT CAT CGC 3'
mk71	zmp-1	SphI	5' CCC CCG CAT GCC ATG TGT ACG TAT TGC TTG 3'
mk72	zmp-1	SphI	5' CCC CCG CAT GCG TAG AAG GGT ATT AGT CGT AG 3'
mk73	zmp-1	SphI	5' CCC CCG CAT GCC AGT TGT AGT AAT ATA TAT TTC 3'
mk74	zmp-1	XbaI	5' CCC CCT CTA GAC GTT TTC TCT TCT GAA AAG TCC 3'
mk75	zmp-1	XbaI	5' CCC CCT CTA GAC TGT CAC ATT TTC CAT TC 3'
mk76	zmp-1	SphI	5' CCC CCG CAT GCC ACT GGT TTC TGT TCT TTC CG 3'

mk77	egl-17	SphI	5' CCC CCG CAT GCG TCT GCT GCC TCG CCT CAT CG 3'
mk78	egl-17	XbaI	5' CCC CCT CTA GAC TAT GTT TCT AGA GAA TTT TG 3'
mk79	zmp-1	none	5' CTA CGA CTA ATA CCC TTC TAC GAG AAA TTA AAA ACG GAA AG 3'
mk80	egl-17	SphI	5' CCC CCG CAT GCC CTC ATC TCT TAT ATC GTC TG 3'
mk81	egl-17	XbaI	5' CCC CCT CTA GAC AGA CGA TAT AAG AGA TGA GG 3'
mk82	egl-17	SphI	5' CCC CCG CAT GCG TAT TAC ATT CCC TAT CAG TC 3'
mk84	egl-17	SphI	5' CCC CCG CAT GCC ACT GTC TCC TCC CCC GTC ACC 3'
mk85	egl-17	XbaI	5' CCC CCT CTA GAG GTG ACG GGG GAG GAG ACA GTG 3'
mk87	zmp-1	none	5' CAT TCA TAA AAT ATT TCG ACC TTT TTC TTG TTA CTC TTT TTT TC 3'
mk89	zmp-1	none	5' GAA AAA AAG AGT AAC AAG AAA AAG GTC GAA ATA TTT TAT GAA TG 3'
mk92	zmp-1	none	5' GCA TGC GTA GAA GGG TAT TAG TCG TAG TAG TAG TAT TCA GTT GTA GTC TAG A 3'
mk93	zmp-1	none	5' TCT AGA CTA CAA CTG AAT ACT ACT ACT ACG ACT AAT ACC CTT CTA CGC ATG C 3'
mk96	cdh-3	SphI	5' CCC CCG CAT GCC CGC ATT TTC ATC AAG ATT CC 3'
mk97	cdh-3	XbaI	5' CCC CCT CTA GAG GAA TCT TGA TGA AAA TGC GG 3'
mk98	cdh-3	StuI	5' CCC CCA GGC CTC AGC TCC CTG GAG CTT TCA C 3'
mk100	egl-17	XbaI	5' CCC CCT CTA GAC GGT CAT TGT GAC CCC ATA GG 3'
mk102	egl-17	SphI	5' CCC CCG CAT GCC GAT ACA ATT GTC CGA CAA C 3'
mk103	egl-17	SphI	5' CCC CCG CAT GCG TTG ATT AAA TTC TTG TTT C 3'
mk104	egl-17	XbaI	5' CCC CCT CTA GAG TTG GAA GAA ATC AGA TCA G 3'
mk105	zmp-1	SphI	5' CCC CCG CAT GCC AAA GTA TTC GAG TAC GTT TAC 3'
mk106	zmp-1	SphI	5' CCC CCG CAT GCG TAC GTT TAC ACT GGT TTC TG 3'
mk107	zmp-1	SphI	5' CCC CCG CAT GCC CGT TTT TAA TTT CTC CTG CC 3'
mk108	zmp-1	SphI	5' CCC CCG CAT GCG AAA AAA AGA GTA ACA AG 3'
mk109	zmp-1	SphI	5' CCC CCG CAT GCG TAT TAG TCG TAG TAG TAG 3'
mk110	zmp-1	XbaI	5' CCC CCT CTA GAC CCT TCT ACT TTT TCT TGT TAC 3'
mk111	zmp-1	XbaI	5' CCC CCT CTA GAC TAC GAC TAA TAC CCT TCT AC 3'
mk112	zmp-1	SphI	5' CCC CCG CAT GCG TAA CAA GAA AAA GTA GAA G 3'
mk113	zmp-1	XbaI	5' CCC CCT CTA GAG CAA AAA AAA ACT GTC ACA TTT TCC 3'
mk114	zmp-1	XbaI	5' CCC CCT CTA GAG TAA GTA TTT TAT AAA GCT G 3'
mk115	zmp-1	XbaI	5' CCC CCT CTA GAC TGA AAA GTC CAC CAA AAA ATT 3'
mk116	zmp-1	XbaI	5' CCC CCT CTA GAC AAA AAA TTA ATT GAA AAT TGC G 3'
mk117	zmp-1	XbaI	5' CCC CCT CTA GAC TCT TTT TTT CAA GCA ATA C 3'
mk118	cdh-3	SphI	5' CCC CCG CAT GCG ACG ACA ACT TAA TGA AAT TTG 3'
mk119	cdh-3	XbaI	5' CCC CCT CTA GAC AAA TTT CAT TAA GTT GTC GTC 3'
mk120	zmp-1	SphI	5' CCC CCC GTA CGC TGT TCT TTC CGT TTT TTA ATT TC 3'
mk121	zmp-1	SphI	5' CCC CCC GTA CGC AAA CTG ATT CAT TGT GTA CG 3'
mk122	zmp-1	SphI	5' CCC CCC GTA CGA GTA GAA GGG TAT TAG TCG TAG 3'
mk123	zmp-1	XbaI	5' CCC CCT CTA GAA ATA CTA CTA CTA CGA CTA ATA C 3'
mk124	zmp-1	XbaI	5' CCC CCT CTA GAC TAC TAC TAC GAC TAA TAC CC 3'
mk125	egl-17	SphI	5' CCC CCC GTA CGC GTC TGT CTT TAC CAA CTT TC 3'
mk129	egl-17	XbaI	5' CCC CCT CTA GAC GAG CGG CTG AGA CGC CAC G 3'
mk130	egl-17	XbaI	5' CCC CCT CTA GAG ACG CCA CGT AGA AGA AGC GG 3'
mk131	egl-17	XbaI	5' CCC CCT CTA GAG AAG AAG CGG CGG TGT TCG TTG 3'
mk132	egl-17	XbaI	5' CCC CCT CTA GAC GGT GTT CGT TGG AAG AAA TC 3'
mk133	cdh-3	XbaI	5' CCC CCT CTA GAG AAG CAA GAC TGT TGA CAG C3'
mk134	cdh-3	XbaI	5' CCC CCT CTA GAC ATA AAA CTG CCC GGC TTT G3'
mk135	cdh-3	SphI	5' CCC CCG CAT GCC CTG TCC CAC TTG CCC ATT C3'




























II-58

mk136	cdh-3	SphI	5' CCC CCG CAT GCG TTT ATG TGT CAT GAA TAA TG3'
mk137	cdh-3	XbaI	5' CCC CCT CTA GAC ATT TCA AAA ATT AAA AAT CAG3'
mk138	zmp-1	SphI	5' CCC CCG CAT GCA ATA TTT CGA CGT TTT CTC TTC 3'
mk141	zmp-1	none	5' GTA GAA GGG TAT TAG TCG TAT ATT CAG TTG TAG TAA TAT ATA TTT C 3'
mk142	zmp-1	none	5' GAA ATA TAT ATT ACT ACA ACT GAA TAT ACG ACT AAT ACC CTT CTA C 3'
mk143	cdh-3	XbaI	5' CCC CCT CTA GAG ACA TGC ATT GGT GCA AGT TG 3'
mk144	cdh-3	XbaI	5' CCC CCT CTA GAC ATT ATT CAT GAC ACA TAA AC 3'
mk145	cdh-3	XbaI	5' CCC CCT CTA GAG AAT GGG CAA GTG GGA CAG G 3'
mk146	cdh-3	SphI	5' CCC CCG CAT GCC TCA GAA ATC GGA GAA TGA TTG 3'
mk147	cdh-3	XbaI	5' CCC CCT CTA GAG TAG TGT AGT CAT AGA GGT CCG 3'
mk148	egl-17	StuI	5' CCC CCA GGC CTC ATC AGG GTG AGT AGG ACT TTG 3'
mk151	zmp-1	none	5' CGT ACA CAT GAA TCA GTT TGG AAA TTA AAA ACG GAA AGA AC 3'
mk152	zmp-1	none	5' GTT CTT TCC GTT TTT AAT TTC CAA ACT GAT TCA TGT GTA CG 3'
mk153	egl-17	SphI	5' CCC CCG CAT GCC CCA ATG TGT GTC CCC GCT G 3'
mk154	egl-17	XbaI	5' CCC CCT CTA GAG TGA CAT GAA CAA GTG GAC C 3'
mk155	cdh-3	SphI	5' CCC CCG CAT GCC ACA CCA CAC AGT GCC AAT TG 3'
mk156	cdh-3	XbaI	5' CCC CCT CTA GAC AAT TGG CAC TGT GTG GTG TG 3'
mk158	cdh-3	SphI	5' CCC CCG CAT GCC CTT GAT TCTCTT GTA TAT CC 3'
mk159	cdh-3	XbaI	5' CCC CCT CTA GAG GAC AGA GAATGC AAA TTA GC 3'

Supplemental Figure 1: *egl-17* cis-regulatory deletion analysis

The upstream region of *egl-17* is depicted at the top of this figure. The translational start occurs at nucleotide 4610. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. Mk80-104 showed very weak vulC expression in 1/2 lines. 102-56 showed weak expression in vulD in 3/3 lines. On rare occasion expression in vulC and vulD was seen in mk103-148. The early expression for this construct was variable from line to line. mk153-154 shows variable expression in the presumptive vulE and vulA cells, although this expression is neither as weak nor as variable as that seen in mk103-148, mk84-20, mk82-100 and mk15-20.

Supplemental Figure 1: *egl-17* cis-regulatory deletion analysis

		EXPRESSION		
		Early	vulC	vulD
0	 4610	+	+	+
0	 1726 mk01-16 (1726bp.)	-	-	-
	mk15-20 (1974 bp.) 1716  3690	+/-	-	-
	mk82-100 (723 bp.) 2888  3611	+/-	-	-
	mk84-20 (508 bp.) 3182  3690	+/-	-	-
	mk45-48 (417 bp.) 3786  4203	-	-	-
	mk77-78 (404 bp.) 2484  2888	-	-	-
	mk82-85 (315 bp.) 2888  3203	-	-	-
	mk57-81 (211bp.) 4125  4336	-	-	-
	mk27-20 (188 bp.) 3502  3690	-	-	-
	mk15-148 (3016 bp.) 1716  4732	+	+	+
	mk84-148 (1550 bp.) 3182  4732	+	+	+
	mk103-148 (305 bp.) 4427  4732	+/-	-	-
	mk153-148 (167 bp.) 4565  4732	-	-	-
	mk153-154 (102 bp.) 4565  4667	+	-	-
	mk27-49(1084bp.) 3502  4586	-	+	+
	mk80-56 (200 bp.) 4316  4516	-	+	+
	mk80-129 (190bp.) 4316  4506	-	+	+
	mk80-130 (179bp.) 4316  4495	-	+	+
	mk80-131 (168bp.) 4316  4484	-	+	+
	mk80-132 (158bp.) 4316  4474	-	+	+
	mk80-104 (150 bp.) 4316  4466	-	+/-	-
	mk125-132 (143 bp.) 4331  4474	-	+	+
	mk57-56 (381bp.) 4125  4516	-	+	+
	mk102-56 (157 bp.) 4359  4516	-	-	+/-
	mk103-56 (89 bp.) 4427  4516	-	-	-
	mk102-104 (107 bp.) 4359  4466	-	-	-

Supplemental Figure 2: *zmp-1* cis-regulatory deletion analysis

The *zmp-1* upstream region is depicted at the top of the figure. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. AC stands for anchor cell.

		EXPRESSION		
		AC	vuIE	vuIA
1052	mk50-53 (229 bp.)	+	+/-	-
1052	mk50-123 (220bp.)	+	+/-	-
1052	mk50-124 (216bp.)	+	-	-
1052	mk50-111 (210 bp.)	-	-	-
1052	mk50-110 (198 bp.)	-	-	-
1052	mk50-117 (177 bp.)	-	-	-
1052	mk50-54 (167 bp.)	-	-	-
1119	mk52-51 (319 bp.)	+	+	+
1119	mk52-74 (259 bp.)	+	+	-
1119	mk52-55 (215 bp.)	+	+	-
1119	mk52-53 (162 bp.)	+	+	-
1119	mk52-54 (100 bp.)	-	-	-
1119	mk52-74 (259 bp.)	+	+	-
1147	mk76-74 (231 bp.)	+	-	-
1180	mk36-74 (198 bp.)	-	-	-
1201	mk71-74 (177 bp.)	-	-	-
	mk92-93 (39 bp.)	-	-	-
1180	mk36-53 (101 bp.)	-	-	-
1180	mk36-31 (438 bp.)	+	-	-
1147	mk76-75 (263 bp.)	+	-	-
791	mk29-37 (410 bp.)	-	-	-
1119	Δ 3 deletion (163bp.)	-	-	-
1052	Δ 3/4 deletion (-7 bp.)	+	-	+

Supplemental Figure 3: *cdh-3* cis-regulatory deletion analysis

The *cdh-3* upstream region is depicted at the top of the figure. A +/- indicates that either the expression level was reduced with respect to other constructs, or that not all animals showed consistent expression in the cell. Mk66-156 shows variable weak expression in the occasional animal in vulC and vulD, while mk158-159 shows variable weak expression in the occasional animal in vulC, E and F, but never in vulD.

.

Supplemental Figure 3: *cdh-3* cis-regulatory deletion analysis

		EXPRESSION				
		AC	vuIE	vuIF	vuIC	vuID
0	7000	+	+	+	+	+
113	996 mk58-59 (883 bp.)	-	-	-	-	-
977	1478 mk60-64 (501 bp.)	-	-	-	-	-
mk62-63 (1530 bp.)	1478 3008	+	-	-	-	-
mk62-97 (831 bp.)	1478 2310	-	-	-	-	-
mk64-65 (1374 bp.)	2989 4363	-	-	-	-	-
mk66-67 (563 bp.)	4434 4997	-	+	+	+	+
mk66-156 (295 bp.)	4434 4729	-	-	-	+/-	+/-
mk155-67 (278 bp.)	4719 4997	-	-	-	-	-
mk158-159 (203 bp.)	4680 4883	-	+/-	+/-	-	-
mk68-69 (1067 bp.)	4978 6045	-	-	-	-	-
mk96-98 (2163 bp.)	2290 4453	+	+	+	+	+
mk96-133 (1629 bp.)	2290 3919	+	+	+	+	+
mk96-134 (1129 bp.)	2290 3419	+	+	+	+	+
mk96-143 (874bp.)	2290 3164	+	+	+	+	+
mk96-63 (708 bp.)	2290 3008	+	-	-	-	-
mk96-119 (423 bp.)	2290 2713	+	-	-	-	-
mk96-137 (350 bp.)	2290 2640	+	-	-	-	-
mk96-144 (222 bp.)	2290 2522	+	-	-	-	-
mk96-145 (141 bp.)	2290 2431	-	-	-	-	-
mk96-145 (141 bp.)	2290 ²⁴³¹ ²⁹⁸⁹ 4363 mk64-65 (1374 bp.)	+/-	-	-	-	-
mk146-144 (155bp.)	2367 2522	+/-	-	-	-	-
mk135-134 (1008 bp.)	2412 3419	+	+	+	+	+
mk135-143 (752 bp.)	2412 3164	+	+	+	+	+
mk135-119 (301 bp.)	2412 2713	-	-	-	-	-
mk135-147 (689 bp.)	2412 3101	+/-	+	+/-	+	+
mk136-147 (599 bp.)	2502 3101	+/-	+	-	+	+
mk136-119 (211 bp.)	2502 2713	-	-	-	-	-
mk118-143 (472 bp.)	2692 3164	-	+/-	-	-	+/-
mk118-63 (306 bp.)	2692 3008	-	-	-	-	-

**Three Genes, Two Species: A Comparative Analysis of Upstream Regulatory Sequences
Sufficient to Direct Vulval Expression in *C. elegans* and *C. briggsae***

Martha Kirouac and Paul W. Sternberg

(submitted for publication)

ABSTRACT

We have identified the *Caenorhabditis briggsae* homologs of three *C. elegans* genes, *egl-17*, *zmp-1* and *cdh-3*, that are differentially expressed in subsets of vulval cells and the anchor cell. Upstream cis-regulatory regions of the *C. elegans* genes sufficient to confer vulval and anchor cell specific regulation are known (Kirouac and Sternberg, accompanying manuscript). We have identified the corresponding *C. briggsae* control regions and tested these regions for activity in *C. elegans*. We find that a 748-bp region of *C. briggsae egl-17* confers expression in *C. elegans* in the primary lineage, occasional secondary lineage expression and late expression in vulC and D. We have identified a 755-bp upstream region of *C. briggsae zmp-1* that confers expression in vulE, vulA, and the anchor cell in *C. elegans*. Finally, we have identified a 1.4-kb region of *C. briggsae cdh-3* that drives expression in vulE, F, C, and D cells in *C. elegans*, and a separate 277-bp region of *C. briggsae cdh-3* that confers expression to *C. elegans* vulC, E and F, but not vulD. We conclude that these phylogenetic footprints promote vulval cell expression in both species. Lastly, we compare the efficacy of phylogenetic footprinting with respect to deletion analysis in transgenic animals.

INTRODUCTION

One of the hallmarks of metazoan development is the transition of an undifferentiated population of cells into unique terminal-cell types. Intercellular signaling plays a major role in the differentiation of cell populations compared to the number of cell types, but, there are relatively few signaling pathways that specify a broad range of terminal fates. The mechanisms by which unique populations of cells are generated from these general signaling components are not well understood.

In the development of the *C. elegans* vulva, at least three intercellular signaling pathways, the EGF, NOTCH, and WNT pathways, induce six multipotential Vulval Precursor Cells (VPCs; reviewed in Greenwald, 1997; Kornfeld, 1997; Sternberg and Han, 1998) to generate an invariant spatial pattern of seven cell fates; vulA-F (Sharma-Kishore *et al.*, 1999). This patterning is likely to depend upon the cis-regulatory regions of the transcriptional targets of these intercellular signals. The isolation of response elements in their transcriptional targets will facilitate biochemical and bioinformatic identification of major transcriptional factors that control cell specific gene expression downstream of these canonical signaling pathways.

Regulatory regions sufficient for vulva and anchor cell expression of three target genes have been described (Kirouac and Sternberg, in prep.): *egl-17*, a fibroblast growth factor family member; *zmp-1*, which encodes a zinc metalloproteinase gene; and *cdh-3*, which encodes a FAT-like cadherin gene. These sufficiency regions probably encode multiple binding sites spread over an extended area. To delimit what regions might be the most important in determining vulva and anchor cell specificity, we have identified the *C. briggsae* homologs of these three genes, and then used phylogenetic footprinting to

identify the control regions predicted to correspond to the sufficiency regions in *C. elegans*. Phylogenetic footprinting is a method for the identification of regulatory elements in a set of orthologous regulatory regions from multiple species by identifying the best-conserved motifs in those regions (Tagle *et al.*, 1988).

Despite having diverged from one another an estimated 50-120 million years ago (Coghlan, 2002), both *C. elegans* and *C. briggsae* share almost identical development and morphology (Nigon and Dougherty, 1949), and the sequences of both species are now known. Rescue of *C. elegans* mutant phenotypes with *C. briggsae* has demonstrated that there is functional conservation between the two species (e.g. de Bono and Hodgkin, 1996; Kennedy *et al.*, 1993; Krause *et al.*, 1994; Kuwabara, 1996; Maduro and Pilgrim, 1996). In addition, analysis of similarity within 142 pairs of orthologous intergenic regions shows regions of high similarity interspersed with non-alignable sequence (Webb *et al.*, 2002). The high degree of similarity in some of these regions suggests that they are under selective pressure. Such intergenic conservation between *C. elegans* and *C. briggsae* has been utilized in various studies to isolate putative binding sites for trans-acting regulatory factors (e.g., Culetto *et al.*, 1999; Gilleard *et al.*, 1997; Gower *et al.*, 2001; Krause *et al.*, 1994; Xue *et al.*, 1992).

In this paper, we test intergenic conserved regions from *C. briggsae* for their ability to drive GFP expression in the vulva cells and anchor cell from the basal *pes-10* promoter for expression in both *C. elegans* and *C. briggsae*.

MATERIALS AND METHODS

Protein prediction of EGL-17, ZMP-1, and CDH-3 homologs in *C. briggsae*

The sequence of the *C. elegans* translated protein used for the TBLASTX was obtained either through Wormbase (<http://www.wormbase.org/>; Stein et al., 2001), as was the case for EGL-17 and CDH-3, or from personal communication in the case of ZMP-1 (J. Butler and J. Kramer, personal communication). For each of these three predicted genes, the corresponding *C. briggsae* cDNA was partially sequenced from an RT-PCR product made from poly (A)⁺ RNA that was isolated from mixed-staged *C. briggsae* worms. The following primers were used for RT-PCR: mk166 5' AGGCGAAACCCACTGGCAAC 3' and mk167 5' TTTGGCGGAGCAGAACACAC 3' for *egl-17*; mk168 5' ATGGGTATT TGCCCCGTGGC 3' and mk169 5' GATTCCTTCTCATAGGTGAACGC 3' for *zmp-1*; and mk170 5' CCTCTCCAACCTCGACATGAATCTC 3' and mk171 5' ACAGTCAAGT TTTCGATTGCGG 3' for *cdh-3*.

Analysis of homologous upstream sequences in *C. elegans* and *C. briggsae*

The Seqcomp and Family Relations programs (Brown *et al.*, 2002) were used to identify homologous upstream sequences conserved between *C. elegans* and *C. briggsae*. The Seqcomp algorithm compares a window of fixed size from one sequence against a same sized window in the second sequence. All 20-bp windows were compared between the two species, at an 80-85% threshold level. This threshold level allows three to four mismatches in a 20-bp window. The upstream sequences of *egl-17*, *zmp-1* and *cdh-3* lie on *C. briggsae* contigs c000300114, c010400937, and c01090600, respectively.

Generation of *egl-17*, *zmp-1* and *cdh-3* *C. briggsae* promoter GFP constructs

Using PCR primers designed from the predicted conserved regions between the upstream regions of *C. elegans* and *C. briggsae egl-17*, *zmp-1* and *cdh-3*, the regions of interest were amplified, with TaKaRa LA Taq (Takara Shuzo), and cloned into the minimal promoter *pes-10*, pPD107.94 (a gift from the Fire lab) using Sph I (5') and Xba I (3') restriction sites engineered into the primers. The sequence of these primers were as follows: mk160, 5' CCCCCGCATGCCACGACCTCCTGGTGTGAGG 3', and mk161, 5' CCCCTCTAGACTAACAA ATGACAAGCGGAAG 3', for *egl-17*; mk172, 5' CCCCCGCATGCGAGTTTCTGGAG GATTCTG 3', and mk173, 5' CCCCTCTAGACGGAA TACTTTAGAATCTC 3', for *zmp-1*; mk162, 5' CCCCCGCATGCCTGACTATGGGGC AGGTGGCC 3', and mk163, 5' CCCCTCTAGAGGTGCGGGAAGAGCCGAGC 3', for the *cdh-3* region containing elements A-F; mk164, 5' CCCCCGCATGCGTCTGTTT GTCCCGATGTCGA 3', and mk165, 5' CCCCTCTAGAGTAGATGGCTGGGATGA CAGG 3', for the *cdh-3* region containing elements H-K. The following PCR protocol was used: 94.0 °C for 4 minutes, followed by 30 cycles 94.0 °C for 30 seconds, 58.0-60.0°C for 30 seconds, 68.0 °C for 7 minutes, followed by 7 minutes at 68.0 °C. *C. briggsae* genomic DNA served as a template for the PCR reaction.

The nomenclature of the constructs generated in this study is derived from the primers used to amplify the region. In all cases, the first 1-3 digits represent the 5' primer and the digits after the hyphen represent the 3' primer.

Microinjection of promoter GFP constructs into *C. elegans*

The constructs were microinjected into the gonads of animals of genotype *pha-1(e2123ts); him-5(e1490)* line using a standard protocol (Mello et al., 1991). The constructs were injected at a concentration of 100 ng/μl, with 20 ng/μl pBluescript SKII (Stratagene), and 82 ng/μl *pha-1(+)*, pBX. Transgenic animals that stably transmit the extrachromosomal arrays were isolated by selecting viable F1 animals at 22.0 °C to new plates and examining their progeny for GFP expression in the anchor cell, and the vulval cells.

Microinjection of promoter GFP constructs into *C. briggsae*

The constructs were microinjected into the gonads of AF16, a wild-type *C. briggsae* line (Fodor et al., 1983), using a standard protocol (Mello et al., 1991). Constructs were injected at a concentration of 100 ng/μl, with 110 ng/μl pBluescript- SKII, and 10 ng/μl *myo-2::GFP*. Transgenic animals stably transmitting the extra-chromosomal arrays were isolated by selecting for *myo-2::GFP* expression in the pharynx of F2 animals. These animals were transferred to new plates, and lines that stably transmitted the array were examined for vulva GFP expression in their progeny.

Microscopy of transgenic animals

Animals were mounted on 5% noble agar pads and scored at 20.0°C for GFP expression under Nomarski optics using a Zeiss Axioplan microscope with a 200-watt HBO UV source, and a Chroma High Q GFP LP filter set (450 nm excitation/505 nm emission). At least two lines for each construct were examined.

egl-17 early expression in the granddaughters of P6.p, the precursors of vulE and vulF cells, was scored at the four-cell stage. *egl-17* vulC and vulD GFP expression was scored between the late L4 to young adult stages (Burdine *et al.*, 1998). *zmp-1* anchor cell GFP expression was scored between the L3 and the early L4 stage. VulE and vulD expression was scored between late L4 and young adult stages. *zmp-1* vulA expression was scored between young adult and adult stages (Wang and Sternberg, 2000). *cdh-3* AC GFP expression was scored between the L3 and the early L4 stage. *cdh-3* vulE, vulF, vulC and vulD expression was scored between the late L3 stage through late L4 stages (Pettitt *et al.*, 1996).

Prediction of binding sites using Transfac database

Putative binding sites for known transcription factors in the conserved regions defined by comparative analysis between *C. elegans* and *C. briggsae* in the *egl-17*, *zmp-1* and *cdh-3* upstream regions were determined using the Transfac database and the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al.*, 1995). Particular emphasis was placed on the regions that were sufficient to confer expression in transgenic *C. elegans* on *pes-10* (Kirouac and Sternberg, in prep.).

AlignACE predictions of overrepresented sequences

AlignACE is based on a Gibbs sampling algorithm that computes a series of motifs that are over-represented in the input sequence(s) (<http://atlas.med.harvard.edu/cgi-bin/alignace.pl>; Roth *et al.*, 1998). The MAP score (maximum a priori log likelihood) is the functional readout of the degree to which a motif is over-represented relative to the

expectation for the random occurrence of such a motif in the sequence under consideration (Roth *et al.*, 1998). We chose a MAP cut-off of 10, which has been shown to be adequate to identify the best-studied examples of known transcription factor binding sites in yeast (Hughes *et al.*, 2000). We used a GC content setting of 0.35, and we searched for motifs of eight and 10 nucleotides. A greater number of aligned sites that are more tightly conserved with information-rich positions, and with nucleotides that are less prevalent in the genome, will lead to higher MAP scores (Hughes *et al.*, 2000).

RESULTS

C. briggsae homologs of *egl-17*, *zmp-1* and *cdh-3*

Because genomic regions that have a biological function are often conserved through evolution, non-coding regions conserved between species are more likely to contain regulatory sequences (Stern, 2000). Therefore, we examined *egl-17*, *zmp-1* and *cdh-3* in the related nematode species, *C. briggsae*.

To identify conserved upstream regulatory regions, we first identified the homologs of ZMP-1, EGL-17 and CDH-3 in *C. briggsae*. Predictions of the *C. briggsae* cDNAs were based on TBLASTX searches of Jim Mullikin's PHUSION assembler data (11/11/2001) at Washington University (<http://genome.wustl.edu/gsc/>), combined with prediction of splice-site donor and acceptor sites using the NetGene2 (<http://www.cbs.dtu.dk/services/NetGene2/>) program. The *C. briggsae* cDNAs were isolated from mixed-staged poly (A)⁺ RNA and sequenced using primers based on these predictions.

The predicted *C. briggsae* EGL-17 cDNA lies on contig c000300114. As seen in the ClustalW alignment (Figure 1), the EGL-17 proteins in both species consist of five translated exons. The *C. elegans* protein has 216 amino acids, and the predicted *C. briggsae* protein has 218 amino acids. The *C. briggsae* exons three and four were sequenced, as were most of exons two and five (Genbank accession #AF529234). The six beta strands and three hairpin structural domains that make up the beta trefoil-fold structural element of the FGF ligand family is conserved in this prediction.

The predicted ZMP-1 *C. briggsae* cDNA lies on two non-overlapping contigs, c010400937 and c000100134. As seen in the ClustalW alignment (Figure 2), the *C. elegans* ZMP-1 protein consists of eight translated exons, as does the *C. briggsae* protein. The *C. elegans* protein has 521 amino acids, and the predicted *C. briggsae* protein has 517 amino acids. There are several interesting features of the sequences. First, the length of the large third intron of approximately 3 kb is conserved in both species. Second, the *C. briggsae* genomic sequence has a large intron of ≥ 5 kb after exon six, where the sequence jumps between non-overlapping contigs. The cDNA from *C. briggsae* was sequenced and the prediction was confirmed for the entirety of exons four, five and six, and most of exons three and seven (Genbank accession #AF529235). Additionally, the conserved matrix metalloproteinase motif, HEXXH, was sequenced and found to be conserved in the sixth exon, and the predicted PRCGXPD motif of the matrix metalloproteinase family located in the second exon is conserved in the prediction.

The predicted *C. briggsae* CDH-3 cDNA lies on two overlapping contigs, c014100642 and c01090600. As seen in the ClustalW alignment (Figure 3), the *C. elegans* protein consists of 23 translated exons, while the *C. briggsae* protein consists of

21 exons. Exons three and four in *C. elegans* are present in a single exon in *C. briggsae*. Similarly, the exons corresponding to *C. elegans* exons nine and ten are present in exon eight in *C. briggsae*, and exons 18 and 19 in the *C. elegans* transcript are represented by exon 16 in *C. briggsae*. Finally, exon 21 from *C. elegans* is split into exons 18 and 19 in *C. briggsae*. Overall, the *C. elegans* protein has 3343 amino acids, and the predicted *C. briggsae* protein has 3221 amino acids. The cDNA from *C. briggsae* was sequenced, and the prediction was confirmed for exons three through five, and parts of exons two and six (Genbank accession #AF529236). The eleven predicted cadherin domains, and the lamin G domain in the *C. elegans* protein (wormPD report CDH-3 at <http://www.incyte.com/proteome/WormPD>; Costanzo *et al.*, 2000) are conserved in the *C. briggsae* prediction.

Comparative sequence analysis

Previous comparisons of intergenic regions have relied on gross alignment of these sequences to find regions of similarity using ClustalW (Higgins *et al.*, 1996) or other alignment programs. In our analysis, we used the Seqcomp and Family Relations programs that perform a comparison of two genomic sequences (Brown *et al.*, 2002). This algorithm allows the isolation of possible conserved regions regardless of location or orientation (i.e., this allows the isolation of similarities from the reverse complement of the sequence). Regions of high similarity between two species such as *C. elegans* and *C. briggsae* are termed phylogenetic footprints (Tagle *et al.*, 1988). The footprints between these two species are, on average, 80% similar, while whole intergenic regions are, on average, 47% similar in *C. elegans* and 50% similar in *C. briggsae* (Webb *et al.*, 2002).

Therefore, a comparison of these regions at a threshold value of 85-90% identity should allow selection of the most similar non-coding regions.

For the *egl-17* comparison, we used the entire 3.9 kb genomic region upstream of the translational start site in *C. elegans* as a basis for comparison against the *C. briggsae* sequence upstream of the predicted *egl-17* translational start site. At the 90% threshold level, four regions of similarity are found (Figure 4A). These elements (A, B, C and D) were located in the same orientation and order with respect to each other in the two species (Figure 5). Elements B, C and D all appear at a 100% threshold level, and at lower thresholds, these regions expand. Element A shares 90% identity between the two species. Two of these four elements, B and D, are in regions of the *C. elegans* sequence that were shown by our sufficiency analysis to be important for either early expression in the presumptive vulE and vulF cells, or in vulC and vulD cells, respectively (Kirouac and Sternberg, in prep.). Element B resides within a region in *C. elegans* that is important for early expression in the presumptive vulE and vulF cells (Kirouac and Sternberg, in prep.). However, this region alone in *C. elegans* was not sufficient to drive this expression pattern consistently. Element D is in a region in *C. elegans* that was shown by sufficiency analysis to be important for driving vulC and vulD expression (Kirouac and Sternberg, in prep.). Element A and C lie in regions that are not needed to drive vulC and vulD expression in *C. elegans*.

When this analysis was performed at a lower threshold of 85% identity, with *C. elegans* sequence mk80-132 (4316-4474) (Figure 5A) needed to drive expression in vulC and vulD, another region, element E, is identified (Figure 5B).

For the *zmp-1* comparison, we used the *C. elegans* genomic sequence from the region mk50-51 (Figure 6A), which we have shown through sufficiency analysis to be important for vulva expression in vulA, vulE, and the anchor cell, as a basis for comparison against the *C. briggsae* sequence upstream of the predicted *zmp-1* translational start site. This comparison was performed in the same manner as for *egl-17*. At this threshold level, four regions of similarity were found (Figure 4B). The order of these four elements (A, B, C and D) is conserved (Figure 6). However, element D is in the reverse orientation with respect to the other elements and the coding region; element D lies within a region in *C. elegans* that is crucial for anchor cell and vulE cell expression (Kirouac and Sternberg, in prep.). Part of this region was deleted in the $\Delta 3/4$ *zmp-1* internal deletion, which shows loss of expression in vulE. The B element is located in a region in *C. elegans* that was shown by deletion analysis to be important for vulA expression (Kirouac and Sternberg, in prep.). Element A appears at the 90% threshold level, while the rest of these elements appear at the 85% level.

For the *cdh-3* comparison, we performed two separate analyses. The first analysis was performed using the upstream region from *C. elegans*, 2290-3419 (mk96-134) (Figure 7A) that was shown to drive both anchor cell expression and vulva cell expression (the first vulval region; Kirouac and Sternberg, in prep.). This sequence was analyzed using the Family Relations and Seqcomp programs to identify regions of similarity when compared to the sequence upstream of the predicted translational start site of *C. briggsae* *cdh-3*. At a threshold level of 85% identity, six elements were found (Figure 4C). These elements, A-F, are scrambled with respect to each other between the two species, both in location and orientation (Figure 7). Element A resides within the α

region, element B resides within the β region, and element F resides within the γ region defined by the sufficiency analysis in *C. elegans* (Kirouac and Sternberg, in prep.). These three sites are important for anchor cell expression, and may also help drive expression in vulE, F, C, and D (Figure 7). Element F shares 100% identity between the two species, while the rest of these elements share 85% homology. All three of the remaining elements D, E, and F, as well as part of C, are contained in the *C. elegans* region mk118-143 that drives variable expression in the vulD, vulE and occasional vulC cells (Kirouac and Sternberg, in prep.).

The second analysis of *cdh-3* was performed with the *C. elegans* genomic sequence corresponding to the mk66-67 (4434-4997)(Figure 8A), which contains the second region that was sufficient to drive expression in the vulva cells (Figure 4D; Kirouac and Sternberg, in prep.). When this region was compared at an 85% threshold level with the sequence upstream of the predicted translational start of *C. briggsae cdh-3*, four elements were found: H, I, J and K. Again, the order of these elements were scrambled between the two species, and these elements partially overlap (Figure 8). Element K shares 100% identity, elements J and H share 95% identity, and element I shares 85% identity between the two species.

Analysis of *C. briggsae* upstream regions

To assess the role of these conserved elements in the cell-specific regulation of these genes, we made constructs containing the elements found in the upstream region of *egl-17*, *zmp-1* and *cdh-3* in *C. briggsae* (Table 1).

Construct mk160-161 (a 748-bp fragment containing the *C. briggsae egl-17* elements B, C, D and E) (Figure 5B), when injected into *C. elegans*, drives expression in both vulC and vulD cells, as well as early expression in the presumptive vulE and vulF cells (Table 1, and Figure 9A). In all lines examined, animals showed variable early expression. Not only was GFP expressed in the presumptive vulE and vulF cells, but GFP was also expressed in the presumptive vulA, B, C and D cells; this latter expression perdured into later stages of invagination (through L3 in some cases, but never in L4) than in *C. elegans*. Furthermore, GFP was sometimes not expressed in the presumptive vulE and vulF cells, while it was expressed in presumptive A, B, C, and D cells. It is possible that in this construct, a negative regulatory element is missing, thereby giving rise to the expanded expression pattern and extending the duration of expression. It is also possible that this expression pattern is the result of species differences either in regulatory control, or in protein function. Element B, which plays a role in expression in the presumptive vulE and vulF cells, is located ~200 bp upstream of the region that correlates with vulC and vulD expression. However, in *C. elegans* this potential enhancer element is located over 1 kb away from the elements that are driving the vulC and vulD expression (Kirouac and Sternberg, in prep.). This observation suggests that the spacing of these elements may not be critical for their functionality.

The *C. elegans egl-17::GFP* reporter, containing 3.9 kb of upstream sequence, shows the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). An occasional animal does not express GFP in vulC and vulD cells at the L4 stage. However, when the 748 bp construct mk160-161 was injected into *C. briggsae*, expression was not seen in the presumptive vulE and vulF cells at the VPC 4-cell stage, although an

occasional animal that was starting to invaginate did show expression in P5.p (Table 1). This observation suggests that either all the elements required for the fidelity of the early expression in *C. briggsae* are not contained in this construct, or that the native gene in *C. briggsae* is not expressed in these cells. In L4 animals, GFP was expressed in the vulva in about 50% of the animals. Of this 50%, GFP was consistently expressed in vulC, and sometimes vulD cells. We infer that an element that is necessary for the fidelity of the expression in *C. briggsae* in vulC and vulD may be missing. Furthermore, this missing element plays a proportionally larger role in regulating the expression in vulD than in vulC cells.

Construct mk172-173 (5138-5892), a 755 bp fragment containing the *C. briggsae* *zmp-1* elements A, B, C and D (Table 1 and Figure 6B), when injected into *C. elegans*, drives expression in the anchor cell, vulE and vulA (data not shown). The only apparent difference between the expression pattern in *C. elegans* and *C. briggsae* is that the vulA expression is variable, and seems to occur at slightly later time points. This difference suggests that there may be an additional element(s) not present in mk172-173 that ensures the fidelity of the vulA expression. In *C. elegans*, vulA expression can be seen in the young adult, but mk172-173 drives vulA expression slightly later than its *C. elegans* counterpart; the majority of animals do not express GFP in vulA cells until eggs are present in the uterus.

The *C. elegans* *zmp-1::GFP* reporter, containing 3.5 kb of upstream sequence, shows the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). Consistent expression was seen in the anchor cell, vulA and vulE. Expression in vulD cells in *C. briggsae* was not determined because of its weak expression in *C. elegans*.

A 1.4-kb fragment containing the *C. briggsae cdh-3* elements A, B, D, E and F (mk162-163) (Figure 7B), when injected into *C. elegans*, drives expression in the vulE, F, C and D cells, but less than 10% of the animals showed any expression in the anchor cell (Table 1 and Figure 9B). A similar fragment, mk96-134 (2290-3419) (Figure 7A) from *C. elegans*, drives expression in vulE, F, C, D and anchor cell (Kirouac and Sternberg, in prep.).

A 277 bp fragment containing the *C. briggsae cdh-3* elements H, I, J, and K (mk164-165) (Table 1 and Figure 8B), when injected into *C. elegans*, drives expression in vulC, E, and F, but not in vulD (data not shown). This expression pattern varies from animal to animal, with vulF showing the strongest and the most penetrant expression. A similar fragment, mk66-67 (4434-4997) (Figure 8A), from *C. elegans*, drives expression in vulE, F, C and D cells (Kirouac and Sternberg, in prep.).

The *C. elegans cdh-3::GFP* reporter, containing 6.0 kb of upstream sequence, does not show the same expression pattern in *C. briggsae* as it does in *C. elegans* (Table 1). Although the expression in the anchor cell is present consistently, only rarely is there expression in vulC, D, E, or F. When there is expression in the vulva cells, it is usually not present in more than a single cell in any given animal. This is in spite of the fact that when the *cdh-3 C. briggsae* sequences are placed in the context of *C. elegans*, there is some expression in vulval cells. We infer that the factor(s) that drive expression in *C. elegans* might be absent in the corresponding *C. briggsae* cells, or the factors have altered binding specificity in *C. briggsae*. It is possible that this gene may have different functions in these two species. Alternatively, *C. briggsae cdh-3* may use binding sites not present in the 6.0 kb of the *C. elegans* sequence to drive expression in the vulva cells.

Transfac binding site prediction in conserved regions

As one approach to finding potential binding sites for known transcription factors in the conserved region, we used the MatInspector program (http://www.genomatix.de/mat_fam; Quandt *et al.*, 1995). We set the core matrix similarity to a minimum of 0.90 to maximize the specificity of the binding sites. We then compared the output from the program of *C. elegans* mk84-148 (3182-4732) to the output for the *C. briggsae* mk160-161 (17543-18289). Only binding sites that appear in both these sequences and had a maximum Random Expectation Value (re-value; the "re" value is the number of times the sequence would appear by chance in 1000 bp of sequence) of ≤ 0.51 were considered for further analysis (Table 2). This process was repeated to compare *C. elegans* sequences from mk96-134 (2290-3419) to *C. briggsae* sequences for mk162-163 (22710-21306), and *C. elegans* sequences for mk66-67 (4434-4962) to sequences for *C. briggsae* construct mk164-165 (18143-17867). Finally, this analysis was done for *C. elegans* construct mk50-51 (1052-1438), and to sequences for *C. briggsae* construct mk172-173 (5138-5892). A total of four potential binding sites were found in the conserved regions of *egl-17* (Table 2). All four of these sites were located in element D. *zmp-1* contained eight factor binding sites in conserved regions (all located in conserved region B or D). The first *cdh-3* region containing conserved elements A-F had three factor binding sites in conserved regions (located in elements B and F; Table 2), and the second *cdh-3* region containing elements H-K also had three conserved binding sites (all located in element K; Table 2). Although this program predicted putative binding sites for families thought to play a role in the specification or terminal differentiation of

these cells (e.g. ETS family members, TCF/LEF-1), we found only two putative binding sites for factors from these families whose site is located in one of the conserved regions of *C. elegans*, and whose corresponding element in *C. briggsae* also contains the same site. The first family was the LIM homeodomain family; *lin-11* is a LIM domain family member and is known to play a role in the specification of secondary cells (Freyd *et al.*, 1990). LIM domain family member sites are found in conserved regions of *egl-17* and *cdh-3* (mk66-67/ mk164-165 region). The second family is the HOX homeodomain family (Kenyon *et al.*, 1998). There is a conserved site in *cdh-3* (mk96-134/ mk162-163) and *zmp-1* (mk50-51/ mk172-173). However, the consensus for the homeodomain families is very weak outside the TAAT core. Given the low specificity, we did not mutate these sites.

AlignACE predictions of overrepresented sequences

We used the AlignACE program (Roth *et al.*, 1998), which computes motifs based on sequences that are over-represented in the input sequence, to identify motifs in the upstream sequences of the *C. briggsae* *egl-17*, *zmp-1* and *cdh-3* (Table 3). We then looked to see which of those motifs were localized in conserved elements. We chose this approach instead of searching for common motifs between homologous upstream regions, because homologous upstream regions, by definition, are likely to be more similar. While looking for regions of similarity was an effective approach to identifying important regulatory sequences within a large upstream sequence, the Seqcomp and Family Relations programs (Brown *et al.*, 2002) recognizes matches based on 85%-100% percent identity over a window of 20 base pairs. The AlignACE program identifies motifs

based on a consensus of eight to ten base pairs. These matches will likely occur much more frequently between two homologous upstream regions than those in two coregulated genes, and may not be functionally meaningful. We also searched for motifs that were common to *C. briggsae zmp-1* region mk172-173 and *C. elegans cdh-3* region mk96-134, each of which are sufficient to drive expression of a naïve promoter in the anchor cell.

In our analysis of *C. briggsae egl-17* sufficiency region mk160-161, AlignACE identified three 8-bp motifs and two 10-bp motifs above the threshold MAP cut-off of 10 (Table 3A). Several of these motifs have common sites, which suggests that they are either variants of the same motif or that they might represent binding sites of trans-acting factors that cooperatively bind DNA. Motifs 1.8, 2.8, 3.8 and 5.10 all have roughly the same site in conserved element B, which was implicated in a sufficiency analysis to be important for the fidelity of the early expression in the presumptive vulE and vulF cells (Table 3B; Kirouac and Sternberg, in prep.). In addition, all of the motifs except 5.10 had sites that resided within conserved element D; element D is located in a region that is critical for conferring expression in vulC and vulD cells (Table 3B; Kirouac and Sternberg, in prep.).

The analysis of the *C. briggsae zmp-1* region mk172-173 identified three 8-bp motifs and two 10-bp motifs (Table 3A). While motifs 1.8, 3.8, and 4.10 all contained sites in conserved element D, only motif 1.8 was found within the part of this element that is contained in the sufficiency region mk50-51 in *C. elegans* (Table 3B; Kirouac and Sternberg, in prep.). It is possible that conserved element D plays a role in conferring expression in vulA cells. Motif 5.10 has one site that is found in conserved element A;

conserved element A is a region that was shown by sufficiency analysis to be critical for anchor cell expression in *C. elegans* (Table 3B; Kirouac and Sternberg, in prep.).

In *C. briggsae cdh-3* construct mk162-163 nine 8-bp motifs and five 10-bp motifs were identified (Table 3A). Of these motifs, 4.8, 5.8, 7.8, 8.8, 10.10, and 12.10 each had one site in conserved element F. This element is in a region that by sufficiency analysis in *C. elegans* was important for both vulval and anchor cell expression (gamma region, Kirouac and Sternberg, in prep.). Motifs 8.8, 12.10 and 13.10 all contain a site in conserved element D, and a site in conserved element A (Table 3B). It is unclear at this time what role conserved element D might be playing in regulating *cdh-3* expression. Conserved element A is located in the alpha region that is important for anchor cell expression in *C. elegans*, but mk162-163 was not able to drive expression in the anchor cell except in few rare cases. Element A's role, if any, in driving expression in vulval cells is not evident.

Mk164-165 was examined using the AlinACE program and was found to contain one 8-bp and two 10-bp motifs (Table 3A). Taken together, these motifs have sites in conserved elements H, J K and I. Mk164-165 drives vulE, F, C, but not D cell expression in *C. elegans* vulval cells (Table 3B, Kirouac and Sternberg, in prep.). The conservation through this region is extensive, suggesting that these regions of conservation and, as an extension of this, these motifs may be important in conferring this expression.

We also compared the *C. briggsae zmp-1* mk172-173 to the *C. elegans cdh-3* mk96-134; both of these regions are sufficient to confer anchor cell expression on a naïve promoter. AlignACE was able to identify one 8-bp motif and two 10-bp motifs that scored above the MAP score cut-off of 10 (Table 3A). An ideal candidate motif would

have sites in conserved regions of both *cdh-3* and *zmp-1* (in essence giving a four-way comparison). Unfortunately in this case, while all three motifs have at least one site that is located in conserved element A of the *cdh-3* region, no sites fall in the conserved elements identified in *zmp-1* (Table 3B). We did not do the reciprocal comparison since the *C. briggsae* construct, which contains the conserved elements that appear to be important in conferring anchor cell specificity in *C. elegans*, does not drive expression in the anchor cell in *C. elegans*.

DISCUSSION

Experiments testing the sufficiency of genomic fragments to direct expression of a heterologous promoter defined small regions that are critical for the fidelity of the expression pattern of *C. elegans egl-17*, *zmp-1* and *cdh-3* (Kirouac and Sternberg, in prep.). However, these regions were still too large to identify specific putative binding sites for known transcription factors. In order to further experimentally define possible binding sites for transcription factors, we used phylogenetic footprinting of the cis-regulatory regions between two species of *Caenorhabditis*, *C. elegans* and *C. briggsae*: *C. briggsae*, by molecular criteria, is 50-120 million years diverged from *C. elegans* (Coghlan, 2002). The Seqcomp program (Brown *et al.*, 2002) was crucial in identifying conserved elements between *C. elegans* and *C. briggsae* in upstream regions. By using phylogenetic footprinting in homologous genes in addition to correlating putative binding sites in potentially co-regulated genes (Kirouac and Sternberg, in prep.), we have maximized the likelihood of identifying regulatory elements responsible for cell-type specific expression.

Phylogenetic footprinting

When phylogenetic footprinting is carried out on a whole-genome scale, it identifies the most highly conserved elements in the regulatory regions; these are promising candidates for binding trans-acting factors (reviewed in Blanchette and Tompa, 2002). In our analysis, we already had in our hands relatively small regions from the homologous *C. elegans* genes that were sufficient to direct vulva and anchor cell expression (Kirouac and Sternberg, in prep.). In the case of *egl-17*, there was a coincidence of the conserved region with the functionally defined sequences at the 95-90% identity level; there were only four elements that were conserved in the 3.9 kb of the original reporter construct. However, for both *cdh-3* and *zmp-1*, there were many conserved elements that did not necessarily fall in the realm of the previously defined sufficiency pieces (Kirouac and Sternberg, in prep.). In *zmp-1*, at a threshold level of 85% identity, there are two to four blocks of conservation in the upstream regions. One of these blocks is the region around mk50-51. In *cdh-3* at a threshold level of 100% identity, three conserved regions appear; elements K and F are two of these three regions. At a threshold level of 90%, element K expands as does the third site, and one additional region appears. Finally, at the 85% threshold level, we see multiple sites spread out throughout the upstream region. This fact made the sufficiency data invaluable for determining which of these conserved elements may play a role in directing vulva and anchor cell specificity. It seems likely that these other conserved regions may be conserved elements involved in the regulation of this gene in other tissues. *egl-17* ::GFP is expressed in a limited number of other tissues: in two large unidentified cells in the head at the three-fold stage of embryogenesis, in the

M4 pharyngeal neuron, and occasionally in the ventral hypodermis of late first-stage larvae (Burdine *et al.*, 1998). In *C. elegans*, *zmp-1::GFP* is expressed in a variety of other cell types, from multiple lineages, including uterine and tail cells, and body muscle and subsets of neurons (J. Butler and J. Kramer, unpublished data). In hermaphrodites, *cdh-3::GFP* is expressed in the seam cells, the buccal and rectal epithelia, the excretory cell, two hypodermal cells in the tail, the uterine epithelium closest to the invaginating vulval cells followed by the multinucleate uterine seam cell (utse), the vulva and associated neurons (Pettitt *et al.*, 1996). The complexity of the expression patterns, and the variety of tissues in which both *zmp-1* and *cdh-3* expression are expressed contrasts with the relatively simple expression pattern of *egl-17::GFP*, thus these other conserved regions in *zmp-1* and *cdh-3* may be other cis-regulatory regions driving transcription in other tissues. It may also be the case that some genes have undergone a faster rate of divergence than others have, and may be under less selective pressure.

Potential for specific isolation of trans-acting factors binding sites by phylogenetic footprinting between *C. elegans* and *C. briggsae*

By comparing the phylogenetic footprints in the upstream regions of homologous sequences from *C. elegans* and *C. briggsae*, we were able to narrow down regions that were responsible for the vulva and anchor cell specific expression of these genes.

However, we could not determine distinct binding sites. Cis-regulatory binding sites can be eight to 10 bp long and they are often highly variable; since DNA has only four-fold variation instead of the 20-fold seen in protein, its level of random variation can be quite high. Comparison to *C. briggsae* will be helpful in locating a phylogenetic footprint of

conserved regulatory regions and confirming the presence of a putative binding site(s). However, when there are no obvious trans-acting candidates, it may be necessary to compare co-regulated or homologous genes from several other species to detect signal above background.

Analysis of putative trans-acting factors using the Transfac database

The focus of these studies was to isolate cell-specific cis-regulatory response elements. However, we also used the Transfac database to look for putative trans-acting factors in the conserved regions that drive expression in the anchor and vulva cells (Table 2), and to compare these data to the putative binding sites in found in the sufficiency analyses (Kirouac and Sternberg, in prep.). Putative binding sites in the conserved elements between *C. elegans* and *C. briggsae* upstream sequences overlap with only a few putative sites defined by the sufficiency analysis of these potentially co-regulated genes (Kirouac and Sternberg, in prep.). Among the overlap were: the CLOX family members, CDP and CDPCR3; the glucocorticoid response family member, GRE; the octamer family member, Oct1; and the homeodomain proteins ISLI and MEIS-1. It is likely that the expression is driven in these cells by different combinations of factors, and that we will not be able to isolate a factor(s) responsible for driving the expression in a single cell type across a panel of coregulated genes, or in orthologous genes in different species.

While a number of genes (for example, *egl-38*, *lin-26*, *lin-29*, *cog-1* and *lin-11*) (Freyd *et al.*, 1990; Labouesse *et al.*, 1994; Rougvie and Ambros, 1995; Bettinger *et al.*, 1997; Chamberlin *et al.*, 1997; Palmer *et al.*, in press) are known to effect the marker gene expression patterns in the vulva, it is not yet known whether they act directly in the

regulation of these genes, or more proximally in the specification of these cell types (M. Wang, T. Inoue, and P. Sternberg, unpublished data). Of these genes, only a site potentially bound by *lin-11* showed up in our Transfac analysis. Biochemical studies using the sufficiency pieces and the conserved regions defined in these studies might help determine which of these transcription factors has a direct effect on the transcriptional regulation of these genes.

Analysis of over-represented sequences in regions of sufficiency

While the Transfac database (Quandt *et al.*, 1995) identifies binding sites of known transcription factors, AlignACE (Roth *et al.*, 1998) identifies sequences that are over-represented in a given sequence. This approach allows the isolation of candidate motifs either within a gene, or between genes. We were able to use this program to identify motifs in our *C. briggsae* constructs, and evaluate whether these motif sites resided in any of the conserved regions that were found using the Seqcomp and Family Relations programs. When we compared *C. briggsae* mk172-173 and *C. elegans* 96-134, each of which are expressed in the anchor cell, we were able to isolate several motifs that may be binding sites of factors that play a role in conferring this cell-specific expression.

Implications of cross-species comparison of *egl-17*, *zmp-1* and *cdh-3*

By comparing the expression patterns of the full-length *C. elegans* GFP reporter constructs in *C. elegans* and *C. briggsae*, it appears that there might be inter-species differences in gene regulation and function. Both *egl-17* and *cdh-3* show differences in expression patterns in the vulva and anchor cell in *C. briggsae*.

The *C. elegans egl-17::GFP* reporter, containing 3.9 kb of upstream sequence, shows expression in the same vulval cells in *C. briggsae* as it does in *C. elegans*. However, there are some differences. Occasionally, *C. briggsae* animals do not express *egl-17::GFP* in vulC and vulD cells at the L4 stage. It is unknown whether this is a result of DNA-mediated transformation differences between *C. elegans* and *C. briggsae*, or if it reflects differences in gene regulation. Early expression is grossly the same between the two species when we examined the full-length *C. elegans* construct in *C. briggsae*. However, when the *C. briggsae egl-17* conserved upstream sequence mk160-161 was injected into *C. elegans*, early expression was highly variable, and was driven in P5.p and P7.p and their descendants as often as it was driven in P6.p. This same region, when injected into *C. briggsae*, does not show consistent expression in the primary lineage, but does show occasional expression in the secondary lineage, P5.p. This difference suggests that there may be a repressor site in *C. elegans* that inhibits expression in vulval cells outside of the primary lineage. However, occasionally, in *C. elegans*, the *C. elegans egl-17::GFP* expression is observed in the secondary lineages at the VPC four-cell stage, but this expression is always in addition to expression in P6.p (M. Wang, D. Sherwood and M. Kirouac, unpublished observations).

While, the differences in the *egl-17::GFP* expression pattern may only be the result of quantitative differences in binding specificity of one or more transcription factors, the differences in *cdh-3::GFP* expression are more substantial. These differences indicate that *cdh-3* may be playing a different role in the vulval cells in *C. briggsae*. In *C. elegans* it is clear that CDH-3 is required for the morphogenesis of a single cell that forms the tip of the tail in the hermaphrodite, while the other cells that express the *cdh-3*

reporter appear to be unaffected by a null allele (Pettitt *et al.*, 1996). However, the genesis of the egg-laying system requires several sets of cell-cell recognition events, all of which occur during the expression of *cdh-3::GFP*. The vulval epidermal cells invaginate and form a connection with the uterus, and the utse cell makes contacts with the seam cells. In addition, during the formation of the seven toroidal rings of the vulva, the vulva cells are involved in complex interactions (Pettitt *et al.*, 1996; Sharma-Kishore *et al.*, 1999). It is possible that in *C. elegans*, other genes can compensate for the loss of CDH-3. There are 12 predicted cadherin superfamily members in *C. elegans*. Of these 12, two, *hmr-1* and *cdh-3*, have been defined by experimental work on their structure and function (Tepass, 1999). Since it appears that in *C. elegans*, *cdh-3* is not required in the vulva cells, it is even less clear what is going on in *C. briggsae*. Perhaps, in *C. briggsae* other members of the cadherin family are active in the vulva cells, or perhaps this gene family is not active at all in the *C. briggsae* vulva.

Conclusions

Independent analysis by phylogenetic footprinting and sufficiency testing (Kirouac and Sternberg, in prep.) can define similar control regions for conferring cell-type specific expression (e.g., regions that drive *egl-17* expression in the vulval cells can be found independently by both methods). However, the success of *de novo* analysis using phylogenetic footprinting techniques will likely depend on the complexity of the cis-regulatory control region. The more complex the control region, the more one must rely on other data, such as sufficiency testing, in establishing the appropriate region for any given cell-type specific expression. In our study, both the *zmp-1* and *cdh-3* upstream

regions had multiple regions of similarity, and it was only through the use of our sufficiency data that we were able to correctly identify regions that conferred vulval cell and anchor cell expression. While these modules may not be narrow enough to resolve discrete binding sites, the addition of other species may allow sub-domains of these phylogenetic footprints to be identified and tested for their ability to confer cell-type specific expression. Also, we found evidence of differences in the expression of both *egl-17* and *cdh-3* full-length *C. elegans* reporter constructs in *C. briggsae*; such differences suggest that either the regulation, or the function, or both, of these proteins has changed in the last 50-120 million years. The convergence of cross-species sufficiency studies and phylogenetic footprinting studies is an efficient way to identify candidate factor binding sites.

Acknowledgments

We are grateful to Jim Butler and Jim Kramer for *zmp-1::GFP* and the sharing of unpublished data; Rebecca Burdine and Micheal Stern for *egl-17::GFP*; and Jonathan Pettitt, William B. Wood and Ronald Plasterk for the *cdh-3::GFP*. We would like to thank Cheryl Van Buskirk, Takao Inoue, Erich Schwarz, Gary Schindelman, and Bhagwati Gupta for the critical reading of this manuscript. P.W.S is an investigator with the Howard Hughes Medical Institute, which supported this research.

REFERENCES

- Bettinger, J.C., Euling, S., and Rougvie, A.E. (1997). The terminal differentiation factor LIN-29 is required for proper vulval morphogenesis and egg laying in *Caenorhabditis elegans*. *Development* **124**, 4333-4342.
- Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* **12**, 739-748.
- Brown, C., Rust, A., Clarke, P., Pan, Z., Schilstra, M., De Buysscher, T., Griffin, G., Wold, B., Cameron, R., Davidson, E., and Bolouri, H. (2002). New Computational Approaches for Analysis of cis-Regulatory Networks. *Developmental Biology* **246**, 86-102.
- Burdine, R., Branda, C., and Stern, M. (1998). EGL-17(FGF) expression coordinates the attraction of the migrating sex myoblasts with vulval induction in *C. elegans*. *Development* **125**, 1083-1093.
- Chamberlin, H.M., Palmer, R.E., Newman, A.P., Sternberg, P.W., Ballie, D.L. and Thomas, J.H. (1997). The PAX gene *egl-38* mediates developmental patterning in *Caenorhabditis elegans*. *Development* **124**, 3919-3928.
- Coghlan, A. W. K. (2002). Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Research* **12**, 857-867.
- Costanzo, M., Hogan, J., Cusick, M., Davis, B., Fancher, A., Hodges, P., Kondu, P., Lengieza, C., Lew-Smith, J., Lingner, C., Roberg-Perez, K., Tillberg, M., Brooks, J., and Garrels, J. (2000). The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research* **28**, 73-76.
- Culetto, E., Combes, D., Fedon, Y., Roig, A., Toutant, J., and Arpagaus, M. (1999). Structure and promoter activity of the 5' flanking region of *ace-1*, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *Journal of Molecular Biology* **290**, 951-966.

- de Bono, M., and Hodgkin, J. (1996). Evolution of sex determination in *Caenorhabditis*: Unusually high divergence of *tra-1* and its functional consequences. *Genetics* **144**, 587-595.
- Fodor, A., Riddle, D., Nelson, F., and Golden, J. (1983). Comparison of a new wild-type *Caenorhabditis briggsae* with laboratory strains of *C. briggsae* and *C. elegans*. *Nematologica* **29**, 203-217.
- Freyd, G., Kim, S., and Horvitz, H. (1990). Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-11*. *Nature* **344**, 876-879.
- Gilleard, J., Henderson, D., and Ulla, N. (1997). Conservation of the *Caenorhabditis elegans* cuticle collagen gene *col-12* in *Caenorhabditis briggsae*. *Gene* **193**, 181-186.
- Gower, N., Temple, G., Schein, J., Marra, M., Walker, D., and Baylis, H. (2001). Dissection of the promoter region of the inositol 1,4,5-triphosphate receptor gene, *itr-1*, in *C. elegans*: A molecular basis for cell-specific expression of IP3R isoforms. *Journal of Molecular Biology* **306**, 145-157.
- Greenwald, I. (1997). Development of the Vulva in: "*C. elegans II*." *DL Riddle, T Blumenthal, BJ Meyer and JR Priess (eds), Cold Spring Harbor Laboratory Press. II*, 519-541.
- Higgins, D.G., Thompson, J.D., and Gibson, T.J. (1988). Using ClustalW for multiple alignments. *Methods Enzymol.* **266**, 387-402.
- Hughes, JD., Estep, PW., Tavazoie, S., and Church, GM. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205-1214.
- Kennedy, B., Aamodt, E., Allen, F., Chung, M., Heschl, M., and McGhee, J. (1993). The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Journal of Molecular Biology* **229**, 890-908.
- Kenyon, C., Austin, J., Costa, M., Cowing, D., Harris, J., Honigberg, L., Hunter, C., Maloof, J., Muller-Immergluck, M., Salser, S., Waring, D., Wang, B., and Wrischnik, L. (1998). The dance of the Hox genes - Patterning the anteroposterior

- body axis of *Caenorhabditis elegans*. *Cold Spring Harbor Symposia on Quantitative Biology* **62**, 293-305.
- Kornfeld, K. (1997). Vulval development in *Caenorhabditis elegans*. *Trends in Genetics* **13**, 55-61.
- Krause, M., Harrison, S., Xu, S., Chen, L., and Fire, A. (1994). Elements regulating cell- and stage-specific expression of the *C. elegans* myoD family homolog *hlh-1*. *Developmental Biology* **166**, 133-148.
- Kuwabara, P. (1996). Interspecies comparison reveals evolution of control regions in the nematode sex-determining gene *tra-2*. *Genetics* **144**, 597-607.
- Labouesse, M., Sookhare, S., and Horvitz, HR. (1994). The *Caenorhabditis elegans* gene *lin-26* is required to specify the fates of hypodermal cells and encodes a presumptive zinc-finger transcription factor. *Development* **120**, 2359-2368.
- Maduro, M., and Pilgrim, D. (1996). Conservation of function and expression of *unc-119* from two *Caenorhabditis* species despite divergence of non-coding DNA. *Gene* **183**, 77-85.
- Mello, C., Kramer, J., Stinchcomb, D., and Ambros, V. (1991). Efficient gene transfer in *C. elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO Journal* **10**, 3959-3970.
- Nigon, V., and Dougherty, E. (1949). Reproduct patterns and attempts at reciprocal crossing of *Rhabditis elegans* Maupas, 1900, and *Rhabditis briggsae* Dougherty & Nigon, 1949 (Nematoda: Rhabditidae). *Journal of Experimental Zoology* **112**, 485-503.
- Pettitt, J., Wood, W., and Plasterk, R. (1996). *cdh-3*, a gene encoding a member of the cadherin superfamily, functions in epithelial cell morphogenesis in *Caenorhabditis elegans*. *Development* **122**, 4149-4157.
- Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**, 4878-4884.
- Roth, FP., Hughes, JD., Estep, PW., and Church, GM. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **10**, 939-945.

- Rougvie, AE., and Ambros, V. (1995). The heterochronic gene *lin-29* encodes a zinc finger protein that controls a terminal differentiation event in *Caenorhabditis elegans*. *Development* **121**, 2491-2500.
- Sharma-Kishore, R., White, J., Southgate, E., and Podbilewicz, B. (1999). Formation of the vulva in *Caenorhabditis elegans*: a paradigm for organogenesis. *Development* **126**, 691-699.
- Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research* **29**, 82-86.
- Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution Int J Org Evolution* **54**, 1079-91.
- Sternberg, P., and Han, M. (1998). Genetics of RAS signaling in *C. elegans*. *Trends in Genetics* **14**, 466-472.
- Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D., and Jones, R. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology* **3**, 439-455.
- Tepass, U. (1999). Genetic analysis of cadherin function in animal morphogenesis. *Current Opinion in Cell Biology* **11**, 540-548.
- Wang, M., and Sternberg, P. (2000). Patterning of the *C. elegans* primary vulval lineage by RAS and Wnt pathways. *Development* **127**, 5047-5058.
- Webb, C., Shabalina, S., Ogurtsov, A., and Kondrashov, A. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucl. Acids Res* **30**, 1233-1239.
- Xue, D., Finney, M., Ruvkun, G., and Chalfie, M. (1992). Regulation of the *mec-3* gene by the *C. elegans* homeoproteins UNC-86 and MEC-3. *EMBO Journal* **11**, 4969-4979.

Figure 1: EGL-17 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The *C. elegans egl-17* has one untranslated exon that is not shown in the exon structure. The exon that starts with the translational start is labeled exon 1. Exon boundaries are indicated by an inverted triangle. The *C. briggsae* cDNA corresponding to the amino acids highlighted in blue was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : identifies a highly conserved amino acid substitution, and . indicates there is a semi-conserved amino acid substitution. The red boxes show the location of the six beta strands, and the green boxes show the location of the three hairpin regions that together make up the beta-trefoil fold, which is conserved in the FGF ligand family.

Figure 2: ZMP-1 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The exon begins with the translational start is labeled exon 1. Exon boundaries are indicated by an inverted triangle. The *C. briggsae* cDNA corresponding to the amino acids highlighted in purple was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : indicates a highly conserved amino acid substitution, and . indicates a semi-conserved amino acid substitution. The location of the conserved PRCGXPD and HEXXH domains of the matrix metalloproteinase family is shown in black boxes.

Figure 3: CDH-3 clustalW alignment in *C. elegans* and *C. briggsae*

The exon structures are shown at the top of the figure. The exon begins with the translational start is labeled exon 1. The lower panel shows the alignment of the first few exons of CDH-3. Exon boundaries are indicated by an inverted triangle, and an inverted triangle with an apostrophe means that an exon boundary was found only in the *C. elegans* protein. The *C. briggsae* cDNA corresponding to the amino acids highlighted in green was sequenced from a RT-PCR. In this alignment, * indicates amino acid identity, : indicates a highly conserved amino acid substitution, and . indicates a semi-conserved amino acid substitution. The conserved cadherin domains of the cadherin family located in this part of CDH-3 are located in the black boxes.

Figure 4: Seqcomp and Family Relations predictions for *egl-17*, *zmp-1* and *cdh-3* upstream sequences

In these analyses the window size is 20 bp. After the Seqcomp program found a region of similarity, this region was examined by eye for other conservation near by. These regions are shown in red. In all four analyses, the translational start site is located on the far right and side of the schematics. (A) In the EGL-17 upstream comparison, we used a threshold value of 90% similarity. Elements A, B, C and D are shown on the schematic of the upstream sequence. The four smaller panels below show the nucleotide conservation of these four elements between the two species. (B) For the ZMP-1 upstream comparison, we used a 85% threshold level. (C) In the first *cdh-3* comparison, we used sequences that corresponded to sequences that resided within *C. elegans* construct mk96-134. We used a threshold of 85% identity. (D) In the second *cdh-3* comparison, we used sequences that corresponded to sequences residing within *C. elegans* construct mk96-134. We used a threshold of 85% identity.

Table 1: Summary of construct expression patterns

This table lists the origin of the upstream region. The names of the construct, features of this construct (e.g., conserved elements (elem.) contained within the region), and the promoter from which expression is driven are listed, as well as which species was injected, and the resulting expression pattern. * This construct showed variable expression in the presumptive vulE and vulF cells, as well as variable expression in the secondary lineages, the presumptive vulA-D. Constructs mk84-148, mk50-51, mk96-134 and mk66-67 were generated in a sufficiency analysis of these three genes in *C. elegans* (Kirouac and Sternberg, in prep.).

Table 1: Summary of construct expression patterns

Origin	Construct Name	Features	Promoter	Species injected	Expression
<i>Ce-egl-17</i>	NH#293	Full length	native	<i>C. elegans</i>	Early, vulC and vulD
<i>Ce-egl-17</i>	NH#293	Full length	native	<i>C. briggsae</i>	Early, vulC and vulD (vulC/D slightly variable)
<i>Cb-egl-17</i>	mk160-161	Elem. B-E	<i>pes-10</i>	<i>C. elegans</i>	Variable early*, vulC and vulD
<i>Cb-egl-17</i>	mk160-161	Elem. B-E	<i>pes-10</i>	<i>C. briggsae</i>	No early, variable vulC and vulD
<i>Ce-egl-17</i>	mk84-148	Elem. B-E	<i>pes-10</i>	<i>C. elegans</i>	Early, vulC and vulD
<i>Ce-zmp-1</i>	pJB100	Full length	native	<i>C. elegans</i>	vulE, vulA and anchor cell
<i>Ce-zmp-1</i>	pJB100	Full length	native	<i>C. briggsae</i>	vulE, vulA and anchor cell
<i>Cb-zmp-1</i>	mk172-173	Elem. A-D	<i>pes-10</i>	<i>C. elegans</i>	vulE, vulA and anchor cell
<i>Ce-zmp-1</i>	mk50-51	Elem. A-D	<i>pes-10</i>	<i>C. elegans</i>	vulE, vulA and anchor cell
<i>Ce-cdh-3</i>	jp#38	Full length	native	<i>C. elegans</i>	vulE, F, C and D and anchor cell
<i>Ce-cdh-3</i>	jp#38	Full length	native	<i>C. briggsae</i>	anchor cell, rare vulval cell expresses
<i>Cb-cdh-3</i>	mk162-163	Elem. A, B, and D-F	<i>pes-10</i>	<i>C. elegans</i>	vulE, F, C, and D
<i>Ce-cdh-3</i>	mk96-134	Elem. A-F	<i>pes-10</i>	<i>C. elegans</i>	vulE, F, C and D and anchor cell
<i>Cb-cdh-3</i>	mk164-165	Elem. H-K	<i>pes-10</i>	<i>C. elegans</i>	vulE, F, C (variable) not vulD
<i>Ce-cdh-3</i>	mk66-67	Elem. H-K	<i>pes-10</i>	<i>C. elegans</i>	vulE, F, C and D

Figure 5: *egl-17* nucleotide sequences of important regions

(A) The nucleotide sequence of *C. elegans egl-17* mk84-148 is shown. The *egl-17* genomic region of NH#293 contains 3819 bp of upstream sequence. The first exon of the transcript starts at nucleotide 4610, and translation starts at nucleotide 4708. Nucleotide 790 of the *egl-17* upstream region corresponds with nucleotide 17648 in Genbank cosmid F38G1 (Accession # AC006635). (B) The nucleotide sequence of *C. briggsae egl-17* upstream region mk160-translational start site is shown. The *C. briggsae egl-17* upstream region lies on contig c000300114 (nucleotides 17543-18504). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that neither of these sequences shows conserved element A.

Figure 6: *zmp-1* nucleotide sequences of important regions

(A) The nucleotide sequence of *C. elegans zmp-1* mk50-51 is shown. The *zmp-1* genomic region in pJB100 contains 3472 bp of upstream sequence. The translational start site of ZMP-1 is at nucleotide 3473. Nucleotide 1 of this *zmp-1* upstream region corresponds with nucleotide 7630 in Genbank cosmid EGAP1 (Accession # U41266). In this panel, nucleotides 992-1438 are shown. (B) The *C. briggsae zmp-1* upstream region mk172-173 that contains the conserved elements predicted by Seqcomp program lies on contig c010400937. Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors.

Figure 7: *cdh-3* nucleotide sequences of mk96-134 and mk162-163

(A) The nucleotide sequence of *C. elegans cdh-3* mk96-134 is shown. The jp#38 genomic region of *cdh-3* contains 5928 bp of upstream sequence, whose start codon occurs at nucleotide 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nucleotide 37343 in Genbank cosmid ZK112 (Accession # L14324). In this panel, nucleotides 2290-3419 are shown. (B) The *C. briggsae cdh-3* upstream region mk162-163 that contains conserved elements predicted by Seqcomp program lies on contig c014100642 (20582-22703). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that elements C and E that are found in *C. elegans* mk96-134 are not in mk162-163.

Figure 8: *cdh-3* nucleotide sequences of mk66-67 and mk164-165

(A) The nucleotide sequence of *C. elegans cdh-3* mk66-67 is shown. The jp#38 genomic region of *cdh-3* contains 5928 bp of upstream sequence, whose start codon occurs at nucleotide 6041. Nucleotide 113 of the *cdh-3* upstream region corresponds with nucleotide 37343 in Genbank cosmid ZK112 (Accession # L14324). In this panel, nucleotides 4434-4997 are shown. (B) The *C. briggsae cdh-3* upstream region, mk164-165, which contains conserved elements predicted by the Seqcomp program lies on contig c014100642 (nucleotides 17869-18145). Arrows show the end points and direction of primers in the region. The conserved elements found by the Seqcomp and Family relations programs are depicted in different colors. Note that elements H and J overlap in mk164-165, and elements H and I overlap in mk66-67.

Figure 9: *C. briggsae* upstream regions injected in *C. elegans*

Panel A shows the expression pattern of *C. briggsae* mk160-161 when it is injected into *C. elegans*. mk160-161 (A) Nomarski DIC photomicrograph of an animal as the vulva has started to invaginate is shown. mk160-161 (B) All of P5.p, P6.p, and P7.p are GFP positive. Another example of this variable expression pattern is seen images C and D. mk160-161 (C) Nomarski DIC photomicrograph of a slightly older animal. mk160-161 (D) The fluorescent image of this same animal is seen; clear expression is seen in the descendants of P5.p and P7.p, but not in P6.p (not in this focal plane and not expressing).

mk160-161 (E) Nomarski DIC photomicrograph of an L4 animal with vulD cells labeled. The vulC cells are not in this plane of focus. mk160-161 (F) This is the same animal and the fluorescence is clearly visible in vulD cells. mk160-161 (G) The same animal is shown again in a slightly different focal plane to see the GFP expression in the vulC cells. In panel (B), are shown some representative pictures from *C. elegans* animals that were injected with *C. briggsae* mk162-163. mk162-163 (A) Nomarski DIC photomicrograph of an animal that has just start to invaginate. The P6.p, the presumptive vulE and vulF, cells are labeled. mk162-163 (B) Shows the fluorescence image of the same animal and GFP is clearly seen in both vulE and vulF cells. mk162-163 (C) Nomarski DIC photomicrograph of an L4 animal, with vulD cells labeled. The vulC cells are not in this plane of focus. mk162-163 (D) Same animal; fluorescence is clearly visible in vulD cells. mk162-163 (E) Same animal again in a slightly different focal plane. The GFP in vulC cells is evident. All photomicrographs are lateral views of the animals.

Table 2: Transfac binding site predictions in regions of similarity between *C. elegans* and *C. briggsae*

Transfac prediction binding sites were listed that meet the following criteria: (1) the minimum core binding specificity had to be at least 0.90, (2) the maximum Random Expectation Value, "re", which is the number of times this site would appear in a random 1000 bp, was not exceed 0.51, and (3) the sites had to appear in both the *C. elegans* region and the homologous *C. briggsae* region. The number of sites in the *C. elegans* region is followed by a slash, and then the number of sites in the *C. briggsae* region is listed. In addition, if the site was in a conserved region, inside the parentheses is denoted how many sites are conserved and in what element. There are several factors marked by *: these factors where not necessarily found in both *C. elegans* and *C. briggsae*, but were included because they are part of some potentially interesting transcription families. The letters B, C, D, F and K refer to the conserved elements in these regions.

Table 2: Transfac database prediction in conserved regions

		<i>egl-17</i>	<i>zmp-1</i>	<i>cdh-3</i>	<i>cdh-3</i>
FAMILY OF FACTORS	FACTOR	mk84-148/ mk160-161	mk50-51/ mk172-173	mk96-134/ mk162-63	mk66-67/ mk164-165
AP1 and related factors	NFE2.01		1/1 (1, B)		
<i>Arabidopsis</i> HomeoBox Protein	ATHB1.01	6/1			
ARS binding factor	ABF1.01		1/1	1/2	
ARS binding factor	ABF1.01		1/		
ARS binding factor	ABF1.02		2/2		
<i>Aspergillus</i> Spore/Developmental regulator	ABAA.01		1/2		
Brn POU domain factors	BRN3.01	5/1			
<i>C. elegans</i> maternal gene product SKN-1	SKN1.01	2/1			1/1
cAMP-Responsive Element Binding proteins	E4BP4.01		2/2 (1, D)		
Ccaat/Enhancer Binding Protein	CEBP.02		1/1 (1, D)		
Cell-death specification 2	CES2.01		2/2 (1, D)		
CLOX FAMILY	CDP.01	1/1	1/1	1/3 (1, B)	
CLOX FAMILY	CDPCR3.01	3/2			1/2 (1, K)
CRP binding Site	CRP.01			1/2	
BRoad-Complex ecdysone steroid response	BRCZ4.01			1/1	
<i>Drosophila</i> gap gene hunchback	HB.02	4/2 (1, D)		1/3	
E2F-myc activator/cell cycle regulator	E2F.01			2/3	
E2F-myc activator/cell cycle regulator	E2F.03			2/2 (1, F)	
ETS	c-ETS-1 (p54) *	0/1			0/1
ETS	ETS1.01			2/1	
ETS	PU.1ETS *	2/1		0/1	1/1
EVI myleoid transforming protein	EVI1.01			1/2	
EVI myleoid transforming protein	EVI1.02		1/2	2/2	
Floral determination	MADSA.01		1/1	4/7	
Fork Head and Related	FREAC2.01	1/3		2/2	
Fork Head and Related	FREAC4.01		1/1		
Fork Head and Related	XFD2.02			1/1	
GATA FAMILY	GATA1.04			1/1	
Glucocorticoid Responsive	ARE.01			2/1	
Glucocorticoid Responsive	GRE.01				1/1 (1, K)
Glucocorticoid Responsive	PRE.01		1/1	1/2	
Homeodomain Factor	FTZ.01	4/1		3/6	
Homeodomain Factor	NKX25.02		1/2		
Homeodomain Factor	NKX31.01		1/1		

Homeodomain Factor	PBX1.01			3/2	
Homeodomain Factor myeloid leukemia	MEIS1.01	3/2	1/2 (1, B)		
Homeodomain Pancreatic /Intestinal LIM domain	ISLI1.01	5/2 (1, D)	1/1	2/3	2/1 (1, K)
Homeodomain Pancreatic /Intestinal	PDX.01	1/1 (1, D)			
Homeoprotein Caudal	CDX2.01	5/2	1/4	4/3	
HOXF	HOX1-3.01	5/2		1/4 (1, B)	
HOXF	HOXA9.01		1/1 (1, B)		
HSF family	FHSF.01		1/3		2/1
HSF family	FHSF.02		1/1		
HSF family	FHSF.03	2/2	1/5	2/1	
HSF family	FHSF.04		1/2		
HSF family	IHSF.01		1/2		
HSF family	IHSF.03	2/1			
HSF family	IHSF.04		1/3		
Interferon Regulated Factor	IRF1.01		1/2	2/2	1/1
Interferon Regulated Factor	IRF2.01		1/2		
Interferon Regulated Factor	ISRE.01		1/1		
MEF2-myocyte-specific enhancer-binding	AMEF2.01	1/1		2/6	
MEF2-myocyte-specific enhancer-binding	HMEF2.01			1/2	
MYB-Like protein (Petunia)	MYBPH3.01		1/2 (1, D)	1/7	
Octamer Family	OCT1.01	1/1	1/1		
Octamer Family	OCT1.06	6/2		1/2	
Octamer Family	OCT1.06	4/2 (1, D)			
papilloma virus E2 Txn activator	E2.02			1/1	
PAX3 FAMILY	PAX3.01		1/1 (1, D)		
<i>Phaseolus vulg.</i> SiLencer reg. of chalcone	SBF1.01	3/3	1/3	3/6	
Plant I-Box sites	IBOX.01	1/1			
Plant P-Box binding sites	PBOX.01			1/2	
Poly A	APOLYA.01	3/3			
Poly A	POLYA.01	1/1		1/1	
Promoter-CcAaT binding	ACAAT.01		1/1	1/1	2/2
Repr. of RXR-mediated activ. & retinoic	COUP.01			1/1	
signal transducers and activators of txn	ISTAT.01			1/1	2/2
signal transducers and activators of txn	STAT6.01	1/1			
SMAD Family TGF-B	FAST1.01		2/2	1/2	
Special AT rich binding Sequence	SATB1.01		1/1		
TATA FAMILY	TATA.02	6/2			
Tata-Binding Protein Factor	ATATA.01	2/1		2/5	
TCF/LEF	LEF1.01 *	1/1	1/3	2/2	
TCF/LEF T-cell Homolog	TCF/LEF *			0/1	

III-57

TCF/LEF	TCF/LEF *	2/1		0/1	1/1
TCF/LEF	TCF11/KCR-F1/NRF1 *	2/1		1/1	
Vertebrate steroidogenic	SF1.01			1/1	
XhoI site-binding protein I	XBP1.01	1/1			
Yeast CCAAT binding	HAP234.01	4/3		1/2	
Yeast GC-Box Proteins	MIG1.01			1/1	
Yeast MADS-Box factors	RLM1.01			1/1	
zinc finger W Box family	WRKY.01	1/3 (1, C)		2/1	1/1
zinc finger <i>Xenopus</i> MYT1 C2HC	MYT1.01	1/1			
zinc finger <i>Xenopus</i> MYT1 C2HC	MYT1.02	5/6 (1, D)	1/9	4/8	

Table 3: AlignACE predictions of overrepresented sequences

(A) A summary of the number of motifs found in each of the listed regions. The total number of motifs identified by AlignACE is shown in parentheses, while the number of motifs that scored above the MAP score threshold of ten is shown outside the parentheses for both the eight- and 10-bp motifs. The last entry on this table is a comparison of *C. elegans cdh-3* to *C. briggsae zmp-1*, each of which drives expression in the anchor cell. As indicated in the left-hand column, this comparison was performed to isolate motifs that might be important in conferring anchor cell expression on a naïve promoter. (B) This table summarizes the data for each of the motifs listed in Table 3A that had a MAP score over 10. The region is listed in the left-hand column. The motif numbers are consecutive and are followed by the size of the motif. The MAP score for each motif is shown under the column head MAP. The sites for each motif are listed. If more than one region was being compared, the sites for the first as indicated by the left-hand column are in parentheses, followed by the second set of parentheses, and so on. Abbreviations are as follows: expr. stands for expression; imp. stands for importance and elem. stands for element. The pictograms were generated using the Pictogram program (<http://genes.mit.edu/pictogram.html>).

Summary

Thesis summary

I have taken two complementary approaches to isolating cell-type specific *cis*-regulatory regions upstream of three genes, *egl-17*, *zmp-1* and *cdh-3*. In the first approach (Chapter 2), I used a sufficiency analysis to test genomic regions of DNA upstream of three genes for their ability to confer cell-specific expression on a naïve promoter, *pes-10*. In a second, orthogonal, approach (Chapter 3), I compared homologous upstream regions (phylogenetic footprints) to identify regions of similarity responsible for conferring cell type-specific patterns of expression.

The selection of these three genes stemmed from the fact that they are expressed in a restricted number of overlapping cell types at similar times. Genes that are specifically expressed in the same tissue at the same time might have common regulatory programs and might be recognized by common *trans* factors. Therefore, conserved motifs in genes showing common expression profiles are likely to be involved in spatial/temporal expression. Additionally, with the exception of the early expression of *egl-17* in the presumptive vulE and vulF cells, all vulval and anchor cell expression occurs after terminal differentiation. The isolation of elements that drive post-terminal differentiation expression allows us to determine what makes each of these cell types unique, and to try to make connections between the known signaling pathways involved in these cell's specification and terminal fates decisions.

While it seems that no single approach is going to identify and define all the *cis*-acting regulatory elements responsible for conferring cell type-specific expression, the corroboration of approaches allows for significant progress to be made.

Sufficiency analysis

The goals of this study was to define the minimal sequences responsible for conferring specificity off a naïve promoter to several vulval cells and the anchor cell in order to search the genome for similar elements. I have narrowed down a 3.9 kb region to: a 143 bp region of *egl-17* that drives vulC and vulD expression, and a separate 102 bp region that is sufficient to drive the early expression in presumptive vulE and vulF cells. I have narrowed a 3.5 kb region to a 300bp region of *zmp-1* that is sufficient to confer expression in vulE, vulA and the anchor cell. And finally, I have examined a 6.0 kb region to define a 689 bp region of *cdh-3* that is sufficient to drive expression in the anchor cell and vulE, vulF, vulD and vulC; a 155 bp region that is sufficient to drive anchor cell expression; and a separate 563 bp region that is also sufficient to drive expression in these vulval cells. One theme that remains the same in all three analyses is that I failed to identify any repressor elements involved in conferring expression in terminally differentiated cell types. Furthermore, it became clear from this study that there are multiple mechanisms used to ensure fidelity of expression patterns even between genes that are expressed in the same cell. These mechanisms include: the use of discrete separable elements that confer cell-type specific expression (*cdh-3* anchor cell expression and *egl-17* expression in sister cells vulC and vulD); the use of complex patterns of binding sites that combinatorially act to establish the fidelity of expression in a variety of cell types from different lineages (*zmp-1*); and the use of tissue-specific elements responsible for driving expression in an entire tissue rather than in sub-domains of its constituent cells (*cdh-3*).

Determining the necessity of regions defined by sufficiency analysis

In one sense, the necessity of these elements was irrelevant to our immediate goal of determining sequences that possess the ability to confer cell type-specific expression of these genes. In our case, the genes themselves are somewhat superfluous compared to the elements, which are sufficient to confer this specificity. What the necessity testing will be invaluable for is putting the results of these analyses back into the context of the native promoters. It will be especially interesting to observe the relative importance of the two non-overlapping regions in upstream sequences of *cdh-3*, both of which, despite qualitative differences, appear sufficient to confer expression in the same cells.

Additionally, mutation analysis of the individual elements defined in the sufficiency and phylogenetic footprint studies will allow us to further delimit the boundaries of these regions. If conducted in the context of the native promoter, the significance of these mutations may be weighed in the natural milieu of the gene.

Phylogenetic footprinting studies of *cis*-regulatory sequences

Since continuously occurring mutational events accumulate at neutral positions but are eliminated in functional regions, it is argued that conserved motifs in diverse orthologous promoter sequences are more likely to have a functional role (Tagle *et al.*, 1988). In this study, I used two species of *Caenorhabditis*, *C. elegans* and *C. briggsae*, for sequence comparisons. With a two-species comparison, I was able to identify several blocks of homology. In the cases of *zmp-1* and *cdh-3*, these blocks were located throughout the upstream region, and only by using the sufficiency data was I able to hone in on a single

block in each as conferring expression in the anchor cell and/or the vulva cells.

Presumably, these other blocks of similarities throughout the upstream regions confer expression in other cell types, as these markers are expressed in a variety of tissues. In the case of *egl-17*, the only elements found, by our sequence comparison were in a region that was found to drive expression in the vulval cells. This is not surprising since the expression of this marker is restricted to very few tissues.

The regions of similarity that did direct vulval and anchor cell specific expression are still broad enough to obscure the resolution of distinct binding sites; furthermore, multiple *trans*-acting sites may be needed to confer a specific expression pattern. In order to get a more defined picture of the regions I have found, it will be helpful to compare co-regulated or homologous genes from several other species in order to distinguish signal from the background noise. With the addition of other species, it may be possible to define this region in greater detail. The present nematode tree gives two additional siblings, CB5161 and PS1010, that may be very useful for such comparisons (Figure 1) (Fitch et al., 1995). I am currently trying to isolate the upstream regions of the *egl-17*, *zmp-1* and *cdh-3* genes from these species for use in a four-way comparison. As one adds more species to the analysis, the distinction between conserved motif and diverged background should become clearer. One risk with this type of analysis is that when including many sequences, particularly distantly related ones, there is an increased chance that some of them may have lost, or completely altered, some regulatory elements over the course of evolution (reviewed in Blanchette and Tompa, 2002). This makes the selection of species imperative to the successful outcome of the analysis. One advantage of this type of approach over others is that while other approaches will distinguish a

single site as necessary and/or sufficient, this approach may help delimit multiple elements in the *cis*-acting regions to give a broader view of the *cis*-acting sequences.

Practical considerations when identifying phylogenetic footprints

ClustalW (Higgins *et al.*, 1994) alignments do not always work for identifying such footprints. Regulatory elements tend to be short (8-10 bp) relative to the entire regulatory region. If the species are more diverged, the noise of the diverged nonfunctional background will overcome the short conserved signal. The result is that the alignment will not align the short regulatory elements well; the regulatory elements would go undetected. This failure was the case for the *zmp-1* and *cdh-3* upstream regulatory regions. There is enough divergence in the sequence that elements picked up by the Seqcomp and Family Relations programs were completely obscured in the clustalW alignment (data not shown). The *egl-17* clustalW alignments were able to identify regions of similarity (data not shown). However, there are large blocks of similarity in the upstream sequences of this gene, making this method, while still fruitful, less helpful than in the case of the other two genes. Additionally, many alignment tools and comparisons do not allow the identification of reverse complement similarities, which can be functionally significant in the context of enhancers that may operate in either direction.

Combining the results of sufficiency testing and phylogenetic footprinting studies

By combining the results of my sufficiency testing with the results of the phylogenetic footprinting, it was satisfying to find that both methods were able to hone in on similar regions as those that were important for conferring tissue specific expression. As can be

seen in figures 2 (*egl-17*), 3 (*zmp-1*), and 4 (*cdh-3*), there are conserved elements that fall in the regions of sufficiency in each of these three genes. In the case of *egl-17*, the location of element D, which falls in the middle of the minimal region defined by sufficiency, is very encouraging; putative binding sites or over-represented sequences in this region should provide good candidates for cell-specific elements. The location of element B in *egl-17* is in a region that plays a role in conferring GFP expression in vulE and vulF. In *zmp-1*, the locations of all conserved elements appear to fall in regions that were important for vulE, vulA and anchor cell expression. Multiple conserved elements in *cdh-3* are found in the regions defined by sufficiency analysis to be important for vulval and anchor cell expression.

Analysis of putative *trans*-acting factors

The sufficiency analysis and phylogenetic footprinting experiments defined overlapping regions of importance in conferring cell-type specific expression of several vulva cells and the uterine anchor cell. However, these regions are still broad enough to obscure the resolution of distinct binding sites. To identify putative *trans*-acting factors that drive expression in these cells, I turned to the Transfac database (see Transfac analysis in Chapters 2 and 3) and our knowledge of genes that are likely to be involved in the specification of these cells (Table 1).

In *lin-29* animals (Horvitz et al., 1983), a gene involved in the heterochronic pathway (Arasu et al., 1991; Bettinger et al., 1997), *egl-17* expression in the presumptive vulE and vulF cells persists, and vulC and vulD expression does not ensue. In the case of *zmp-1*, there is no vulE expression in the young adult, though this background does not

affect the *cdh-3* expression during the L3 and early L4 stages. Since this mutation causes the reiteration of earlier developmental stages, it is not surprising that early expression persists at the expense of the later expression pattern (M. Wang and T. Inoue, unpublished observations).

In the PAX family member *egl-38* (Chamberlin et al., 1997), there is no *egl-17* expression in the presumptive vulE and vulF cells, and no *zmp-1* expression in vulE. The HOM-C family member *lin-39* also decreases the *egl-17* expression in the presumptive vulE and vulF cells, suggesting that these genes may play a role in regulating expression in vulE (M. Wang, unpublished observations)

In animals mutant in the *lin-1* gene (Beitel et al., 1995), which encodes an ETS family member, there is no *egl-17* expression in the presumptive vulE and vulF cells. However, *zmp-1* expression in vulE is normal in the *lin-1* background. *lin-1* also effects vulC and vulD expression in the *egl-17* background. This altered expression suggests that *lin-1* may play a specific role in *egl-17* regulation (M. Wang, unpublished observations).

In the *lin-26* animals, a predicted zinc finger transcription factor that plays a role in the generation of Pn.p cells (Labouesse et al., 1994), *egl-17* vulC expression is lost and the vulD expression is dramatically reduced. Additionally, *cdh-3* expression is dramatically reduced in vulC, D and E (T. Inoue, unpublished observations). The *lin-26* gene may play an important role in the specification of these cells.

In animals carrying one allele of the gene encoding a GTX NKx6.2 family member *cog-1* (R. Palmer et al., in press), *sy275*, *egl-17* vulE expression is seen in addition to vulC and vulD expression in the L4 stage. This expression is separate from

the early expression in this cell. This same allele shows no vulE *zmp-1* expression. Perhaps, *cog-1* (*sy275*) plays a role regulating late *egl-17* expression in vulE cells (M. Wang and T. Inoue, unpublished observations). However, no GTX binding sites were found using the MatInspector program. A second allele of *cog-1*, *sy607*, does not effect vulE expression. However, this allele shows no *cdh-3* expression in vulC and vulD cells, and a dramatic reduction in vulE cells (M. Wang and T. Inoue, unpublished observations).

In the LIM domain protein, *lin-11* (Freyd et al., 1990), there is no *egl-17* expression in vulC or vulD cells, but there is no effect on the early expression in the presumptive vulE and vulF cells. In *lin-11*, there is also no *zmp-1* expression in either vulA or vulE cells, yet it also alters *cdh-3* expression levels in vulF, vulE, vulC and vulD. This result is surprising because of the *lin-11* effect on *zmp-1* and *cdh-3* expression in the primary lineage. Although we know that *lin-11* animals have altered secondary cell lineage, we have no evidence of it having any effects on the analysis of primary fate (B. Gupta, unpublished observations). Our analysis using the MatInspector program did identify binding sites for the putative LIM homolog, ISLI-1, in conserved regions responsible for driving *egl-17* expression in vulC and vulD. The significance of this finding is not known. This site came up in all the analyses, and has a very loose consensus sequence with a core matrix sequence of TAAT similar to that of other homeodomains.

A loss of function mutation in *lin-17* (Sternberg and Horvitz, 1988), which encodes a WNT-family receptor, causes variable *cdh-3* expression in vulC and vulD and ectopic variable expression in vulA and vulB (T. Inoue, unpublished observations). This

result suggests that this gene probably plays an intimate role in mediated secondary cell fate or transcriptional regulation.

An anchor cell element that drives transcription of LIN-3 has been isolated, and involves *trans*-acting factors that bind to a nuclear hormone receptor site and E-box protein-binding sites (B. Hwang and P. Sternberg, unpublished results). Disruption of these elements does not disrupt the expression of *cdh-3* or *zmp-1::gfp* in the anchor cell. A different mechanism and/or factors must be used to establish the anchor cell expression of these late markers. We have few candidate factors that may be involved in the regulation in this cell.

While the focus of this project was to isolate cell-specific response *cis*-regulatory elements rather than identifying *trans*-acting factors, I was also looking forward to the more distant goal of determining the integration of signaling pathways in the downstream targets of these pathways. The integration, in the upstream sequences, of members of the RAS, NOTCH and WNT pathways, whose signaling is intimately bound with the establishment of these fates, would help establish the hierarchy of action of these pathways and their interactions. In the case of the early expression of the *egl-17* gene (expression in the presumptive vulE and vulF cells), it is still a matter of debate regarding the determination status of these cells at the time of this expression. *egl-17* is expressed at a time when crucial signaling events that result in an invariant cell fate pattern are still occurring, which makes this particular gene, and the elements responsible for conferring its early expression, of special interest. There are several approaches to the identification of the *trans*-acting factors involved in conferring the cell type-specific expression patterns. The preceding section has talked about various genetic backgrounds that have

been examined in the context of the full-length reporter constructs. Some of these genetic backgrounds have a dramatic effect on the ability of these reporters to confer expression. One approach is to use the minimal sufficiency regions defined in this thesis to look at the genetic backgrounds that had an effect on expression patterns, to establish that they are working through these elements, and also to extend this to a greater diversity of genetic backgrounds. This, however, will not get to the crux of the matter of whether these factors are directly binding these sequences, or are regulating something in turn that is directly binding them. It will, however, tell you which genes appear to be involved in establishing the differential gene expression in these cells.

To categorically establish which *trans*-acting factors are binding these sites directly will require biochemical testing of the ability of a specific *trans*-acting factor to bind a particular sequence.

Genomic analysis

Once elements responsible for conferring cell-type specific expression have been defined as concisely as bench-work will allow us (through mutational analysis, or further phylogenetic analysis), it will be both feasible and exciting to search the genome of *C. elegans* and *C. briggsae* for other genes whose *cis*-regulatory sequences contain these elements.

When a single promoter sequence is searched, one often finds many putative elements conserved all over the sequence, making it difficult to choose for further experimental analysis. On the other hand, when multiple promoter sequences are searched simultaneously, the conserved motifs are more likely to be functionally

important. To this end, I used the AlignACE program to look for over-represented sequences in elements of intergenic regions found in our sufficiency analysis; I also looked for over-represented sequences between elements that conferred the same cell specificity (Chapter 2, Table 2 and Chapter 3, Table 3). One caveat of this approach is that its efficacy, while seemingly good in yeast (Hughes *et al.*, 2000), has not been tested on metazoans. The metazoans have much larger non-coding regions use a more combinatorial based system of regulation show long distance regulation via chromatin, and appear to have a vast number of transcription factors not present in yeast. These over-represented sequences that fall into regions which, by our other analysis, appear to be important in conferring cell/ tissue specificity make good candidates to search for in the genome, and also make good candidates for mutational analysis. In order to perform this search with a consensus sequence, we can modify the program ScanACE, which performs a similar search on the genome of *Saccharomyces cerevisiae*.

References

- Arasu, P., Wightman, B., and Ruvkun, G. (1991). Temporal regulation of *lin-14* by the antagonistic action of two other heterochronic genes, *lin-4* and *lin-28*. *Genes & Development* **5**, 1825-1833.
- Beitel, G., Tuck, S., Greenwald, I., and Horvitz, H. (1995). The *Caenorhabditis elegans* gene *lin-1* encodes an ETS-domain protein and defines a branch of the vulval induction pathway. *Genes & Development* **9**, 3149-3162.
- Bettinger, J., Euling, S., and Rougvie, A. (1997). The terminal differentiation factor LIN-29 is required for proper vulval morphogenesis and egg laying in *Caenorhabditis elegans*. *Development* **124**, 4333-4342.

- Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* **12**, 739-748.
- Chamberlin, H., Palmer, R., Newman, A., Sternberg, P., Baillie, D., and Thomas, J. (1997). The PAX gene *egl-38* mediates developmental patterning in *Caenorhabditis elegans*. *Development* **124**, 3919-3928.
- Fitch, D., Bugaj-Gaweda, B., and Emmons, S. (1995). 18S Ribosomal RNA gene phylogeny for some *Rhabditidae* related to *Caenorhabditis*. *Molecular Biology and Evolution* **12**, 346-358.
- Freyd, G., Kim, S., and Horvitz, H. (1990). Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-11*. *Nature* **344**, 876-879.
- Higgins, D.G., Thompson, J.D., and Gibson, T.J. (1988). Using ClustalW for multiple alignments. *Methods Enzymol.* **266**, 387-402.
- Horvitz, H., Sternberg, P., Greenwald, I., Fixsen, W., and Ellis, H. (1983). Mutations that affect neural cell lineages and cell fates during the development of the nematode *C. elegans*. *Cold Spring Harbor Symposia on Quantitative Biology* **48**, 453-463.
- Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205-1214.
- Labouesse, M., Sookhare, S., and Horvitz, H. (1994). The *Caenorhabditis elegans* gene *lin-26* is required to specify the fates of hypodermal cells and encodes a presumptive zinc-finger transcription factor. *Development* **120**, 2359-2368.
- Sternberg, P., and Horvitz, H. (1988). *lin-17* mutations of *Caenorhabditis elegans* disrupt certain asymmetric cell divisions. *Developmental Biology* **130**, 67-73.
- Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D., and Jones, R. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology* **3**, 439-455.

Figure 1: Selection of nematode species for comparative genomic analysis

Closest sibling species to *C. elegans* are listed in this tree diagram, adapted from Fitch *et al.*, 1995. Dates of divergence are hard to predict, but the current prediction of divergence between *C. elegans* and *C. briggsae* is 50-120 million years.

Figure 2: Combined results of the *egl-17* sufficiency and phylogenetic analyses

This figure depicts both the *egl-17* sufficiency data as seen in Chapter 2, Figure 3, and the conserved regions identified in the phylogenetic footprinting studies, which have been superimposed on this schematic. A, B, C, D represents element A, element B, and so forth. The boundaries of each element are listed in the top right-hand corner of the figure. The box in the upper right-hand corner depicts the expression pattern of each of the three markers used in these studies.

Figure 3: Combined results of the *zmp-1* sufficiency and phylogenetic analyses

This figure depicts both the *zmp-1* sufficiency data as seen in Chapter 2, Figure 5, and the conserved regions identified in the phylogenetic footprinting studies, which have been superimposed on this schematic. A, B, C, D represents element A, element B and so forth. The boundaries of each element are indicated at the bottom of each element.

The box in the upper right-hand corner depicts the expression pattern of each of the three markers used in these studies.

Figure 4: Combined results of the *cdh-3* sufficiency and phylogenetic analyses

This figure depicts both the *cdh-3* sufficiency data as seen in Chapter 2, Figure 6, and the conserved regions identified in the phylogenetic footprinting studies, which have been superimposed on this schematic. A, B, C, D represents element A, element B and so forth. Elements H, I, J, K are overlapping a consecutive, and so have been represented by a single box labeled “HIJK”. The boundaries of each element are listed in the top right-hand corner of the figure. The box in the upper right-hand corner depicts the expression pattern of each of the three markers used in these studies.

Table 1: Effect of genetic background on marker expression

For each marker gene listed in the first column, the expression pattern in a variety of different genetic backgrounds (listed in column two) is summarized for cells vulA-F. The expression pattern in the anchor cell was not determined. An “nd” means that the expression pattern was not determined. A “+/-“ indicates that expression was variable or weak. (These data summarize expression studies performed by M. Wang and T. Inoue, unpublished results.)

Table 1: Effect of genetic background on marker expression

marker	Genetic background	vulA	vulB	vulC	vulD	vulE	vulF
<i>egl-17::GFP</i>	wt			+	+	+	+
	<i>lin-29</i> (<i>sy292 /n333</i>)			-	-	+	+
	<i>lin-26 (ga91)</i>			-	+/-	nd	nd
	<i>cog-1 (sy275)</i>			+	+	++ (at time 2°)	+
	<i>cog-1 (sy607)</i>			+	+	+	+
	<i>lin-11 (n389)</i>			-	-	+	+
	<i>egl-38 (n578)</i>			nd	nd	-	-
	<i>lin-1 (sy254)</i>			+/-	+/-	-	-
	<i>lin-39 (n709)</i>			nd	nd	+/-	+/-
	<i>lin-17</i>			nd	nd	nd	nd
	<i>sqv-3 (n2842)</i>			+	+	+	+
	<i>evl-2 (ar101)</i>			+	+	+	+
	<i>evl-22 (ar104)</i>			+	+	+	+
	<i>cdh-3::GFP</i>	wt			+	+	+
<i>lin-29</i>				+	+	+	+
<i>lin-26 (ga91)</i>				+/-	+/-	+/-	+
<i>cog-1 (sy275)</i>				+	+	+	+
<i>cog-1 (sy607)</i>				-	-	+/-	+
<i>lin-11</i>				-	-	-	+/-
<i>egl-38 (n578)</i>				nd	nd	nd	nd
<i>lin-1</i>				nd	nd	nd	nd
<i>lin-39</i>				nd	nd	nd	nd
<i>lin-17</i>		+/-	+/-	+/-	+/-	+	+
<i>zmp-1::GFP</i>		wt	+				+
	<i>lin-29 (sy292)</i>	+				-	
	<i>lin-26 (ga91)</i>	nd				nd	
	<i>cog-1 (sy275)</i>	+				-	
	<i>cog-1 (sy607)</i>	nd				nd	
	<i>lin-11 (n389)</i>	-				-	
	<i>egl-38 (n578)</i>	nd				-	
	<i>lin-1 (sy254)</i>	nd				+	
	<i>lin-39</i>	nd				nd	
	<i>lin-17</i>	nd				nd	
<i>lin-31 (n301)</i>	+				nd		