GLOBAL TRANSCRIPTOME PROFILING OF SINGLE CELLS REVEALS KEY MOLECULES INVOLVED IN CELLULAR FUNCTION AND DEVELOPMENT IN *C. ELEGANS*

By

William Clayton Spencer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Cell and Developmental Biology

August, 2011

Nashville, Tennessee

Approved:

Professor David M. Miller, III

Associate Professor Guoqiang Gu

Professor James R. Goldenring

Professor Randy D. Blakely

Associate Professor Ethan Lee

To Kylee and Isaac

ACKNOWLEDGEMENTS

It has been a long haul to reach this point in my scientific career and I have numerous people to thank for their support, encouragement, and friendship. I started graduate school again after working for 1.5 years in David Miller's lab as a research assistant. I promptly joined David's lab as a graduate student. I couldn't imagine joining a different lab for my Ph.D. work. The combination of David's training and the lab environment was second to none. From learning how to give better presentations to thinking about alternative hypotheses (many at times!) I can't thank David enough for his effort in training me as a scientist.

When I joined the lab, there were three graduate students and a research assistant who have provided great advice and friendship through the years. Steve Von Stetina, Rebecca Fox, and Joseph Watson formed the core of "Team Slacker" and were extremely welcoming to me. Steve, thanks for your advice and being the model lab guru. Rebecca, I try to emulate your dogged determination to get things done (but mostly fail). Joseph, you have challenged me in life and science like no one else. Kathie Watkins was my fellow research assistant when I joined the lab, stuck it out in the lab with me, and will be departing with me. She has been the glue of the lab and will always be Mama turtle. Laurie and JJ, you all are a pair like no other.

Several new people joined the lab over the years and have continued the great Miller lab environment tradition. Jud, science + beer = awesome! Sarah and I entered grad school the same year and she has provided me a number of

interesting questions to think about. Rachel, the combination of your own unique personality with Lou and your dogs provided endless entertainment. Cody and Mallory, I have thoroughly enjoyed sharing in quality beer consumption and the scientific vigor you brought to the lab. Becky, I truly appreciate what you have done for the modENCODE project. Tim, always remember that you are the man! Tyne, I am happy you joined the lab and I can't wait to see Miller and Miller, 20XX.

Lastly, but most importantly, I thank Kylee for her love, support, and patience. Grad school is not an easy process for anyone, and I am truly grateful that we were able to support each other when experiments were not working and celebrate when they did. You have also been extremely patient with myself and other Miller lab members when we have bugged you with our statistics questions! I am excited about the next step in our life and careers as you start at Heidelberg University and I start my postdoc work. I can't wait to see how things unfold and what Isaac does next.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

GENE EXPRESSION: APPLICATIONS AND ANALYSIS

Introduction

Organism development and cellular function depend on precise control of gene expression. This outcome is achieved by the application of multiple levels of regulatory control. However, the extreme sensitivity of biological processes to perturbations in gene expression can result in significant developmental defects or disease when even one of these mechanisms is affected. Thus, understanding the intricacies of how genes are expressed and how they are regulated can impact all aspects of biology. There has been a tremendous effort over several decades to define when and where genes are expressed by assaying RNA and protein levels in whole animals, tissues, and cells. This approach not only provides evidence for where genes are expressed but can also infer biological function by noting the biochemical properties of gene products that are consistently co-regulated.

Methods for the quantification of gene expression levels have progressed from the simplest approach of assaying the product of a single gene to more recently developed technology with which it is possible to measure expression of all genes in a given genome in a single experiment. This introduction provides an overview of transcription, transcriptional regulation, post-transcriptional

regulation, and how gene expression can be measured by various methodologies, in particular, focusing on RNA transcript quantification.

Transcription

Genes represent a union of DNA sequences that produce a functional product. Historically, the central dogma dictated that one gene encoded one protein product. After decades of research, it is now apparent that genes can produce many forms of RNA as a final product as well as multiple protein isoforms. To produce a protein product, a gene must first be transcribed into RNA, and the RNA message is translated into a protein based on the genetic code (CRICK et al. 1961). As DNA is double-stranded, the coding strand carries the genetic information, while the opposite strand serves as the template for transcribing DNA to RNA. Transcription is performed by the RNA Polymerase II macromolecular complex. RNA Polymerase uses the sequence of the template DNA strand to polymerize a chain of ribonucleotides to form a strand of RNA. Since the template DNA strand is complementary to the coding DNA strand, the RNA strand represents the same sequence as the coding strand of DNA substituting uracil for thymine. This RNA strand is called a transcript, which in eukaryotes typically contains regions called introns that must be removed by the splicing machinery, leaving untranslated regions (UTRs) at the 5’ and 3’ ends of the molecule, and the exons that carry the genetic information. RNA binding proteins bind the messenger RNA (mRNA) to stabilize and protect the molecule from degradation and shuttle it outside of the nucleus to the cytoplasm where it can be translated by ribosomes.

There are many genes that produce RNAs that do not code for proteins (noncoding RNA, ncRNA) and function in many cellular processes. The most abundant and well-studied ncRNAs are the ribosomal RNAs (rRNA) and transfer RNAs (tRNA). These ncRNAs provide critical functions in translation of mRNA into protein. rRNAs are named based on their sedimentation rate, which depends on shape and mass. The 18S is a component of the small ribosome subunit and 28S is a component of the large ribosome subunit. Two smaller rRNAs, 5S and 5.8S are both located in the large ribosomal subunit. The rRNAs function with other proteins and tRNAs to use mRNA as a template for polypeptide synthesis.

There are many other classes of ncRNAs including small nucleolar RNAs (snoRNA), microRNAs (miRNA), small nuclear RNAs (snRNA), small interfering RNAs (siRNA), Piwi-interacting RNAs (piRNA), and long ncRNAs. These ncRNAs have various functions from template-guided processing of rRNAs (snoRNAs)(Kiss 2001) to suppression of retrotransposon expression in germ line cells (Vagin et al. 2006).

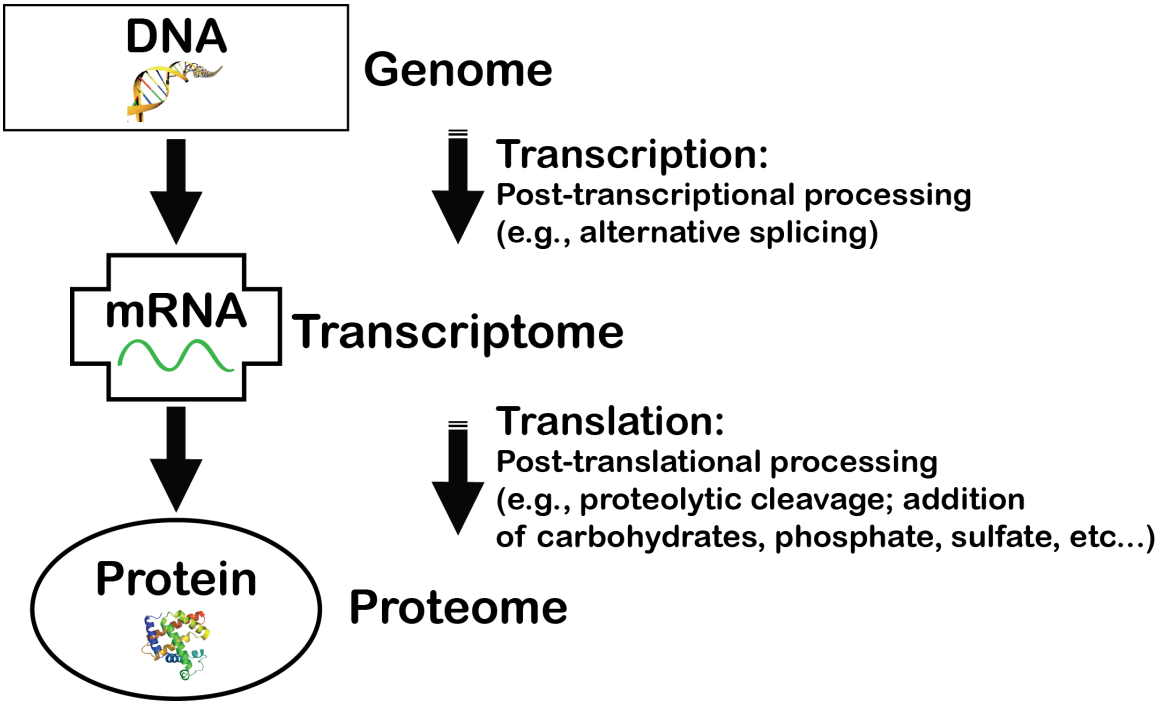


Figure 1.1 Representation of the “central dogma” of molecular biology showing the flow of genetic information from DNA to RNA to protein (inspired/adapted from (Walker et al. 2005)).

Transcriptional regulation

Gene expression depends not only on the RNA Polymerase complex (RNA Pol II) and general transcription factors (e.g. TFIID), but also on sequence-specific enhancer-binding transcription factors that contribute to the regulation of when and where a gene is expressed (Mitchell and Tjian 1989; Karin 1990; Orphanides et al. 1996; Latchman 1997; Lee and Young 2000). Transcription factors can promote or inhibit the activity of the RNA Polymerase complex and other factors to positively or negatively regulate transcription. In general, transcription factors can positively regulate transcription by promoting the binding and stabilization of the RNA Pol II and general transcription factors, recruiting histone acetyltransferases, and recruiting chromatin remodeling factors. Negative regulators can inhibit transcription through several mechanisms: compete for the same binding site as a positive regulator, mask the activation domain of a positive regulator, bind to the transcriptional machinery, recruit repressive chromatin remodeling factors, or recruit histone deacetylases (Lee and Young 2000). Through the combinatorial action of positive and negative transcription regulators, the expression pattern of a gene can be sculpted to restrict activity to specific cell-types and developmental stages.

Post-transcriptional regulation

Once transcripts are produced, several mechanisms act to modify the transcript prior to translation or degradation. The splicing machinery (spliceosome) binds to the nascent transcript to remove introns. The spliceosome recognizes sequences surrounding exon-intron boundaries to identify where the

transcript will be spliced. The spliceosome loops the intron and performs two sequential trans-esterification reactions to cleave the intron and bind the two flanking exons together (Rio 1993). This process takes place constitutively, but alternative splicing can occur through regulation of several splicing factors. Alternative splicing can produce many transcript isoforms from a single gene locus. Some alternative splicing events include exon skipping, mutually exclusive exon inclusion, intron retention, and alternate acceptor/donor usage (Blencowe 2006). Up to 95% of multi-exon human genes are alternatively spliced demonstrating how transcripts can be modified to potentially produce many more protein products than previously assumed by the central dogma (Pan et al. 2008).

MicroRNA (miRNA) regulation of transcript stability and translational-inhibition is another mechanism for regulating gene expression post-transcriptionally. MicroRNAs were originally thought to be a quirk of *C. elegans*, but have now been described and well-studied in many organisms including humans (Lee and Ambros 2001; Bentwich et al. 2005; Bartel 2009). MicroRNAs can be found as an independent genomic locus, in polycistronic clusters, or within introns of protein-coding genes (Moss 2002; Berezikov et al. 2007; Ruby et al. 2007). MicroRNAs are usually transcribed by Pol II and become polyadenylated and capped as with most mRNAs generating the primary miRNA transcript (Cai et al. 2004; Lee et al. 2004; Zhou et al. 2007). The primary miRNA transcript can contain multiple miRNAs that form hairpins. DGCR8/Pasha binds the primary miRNA and associates with the enzyme Drosha that cleaves the

primary miRNA into separate precursor miRNAs (Gregory et al. 2006). The precursor miRNAs are exported from the nucleus and the cleaved by the RNAse III enzyme Dicer (Lund and Dahlberg 2006; Ji 2008). Argonaute family proteins bind with Dicer and the mature miRNA to form the RNA induced silencing complex (RISC) (Rana 2007). RISC usually binds target mRNAs at the 3' UTR. If the miRNA base-pairs perfectly with the target, Argonaute2 degrades the target mRNA (Matranga et al. 2005). If there is imperfect base-pairing with the target, translational inhibition can occur (McManus et al. 2002; Seggerson et al. 2002). Additionally, miRNAs are able to destabilize target mRNAs likely by removal of the poly(A) tail, which is the initial step in mRNA decay (Lim et al. 2005; Roush and Slack 2006). Since miRNAs are transcriptional regulated by their own promoter or are transcribed with a host gene, miRNA function can be restricted to particular cell-types (Martinez et al. 2008; Isik et al. 2010). Thus, miRNA regulation of gene expression adds another layer of control on gene expression patterns. To study how complex patterns of gene expression arise, what factors contribute to the control of gene expression, and what genes are important for the function of a cell, gene expression profiling is frequently used to delineate those processes.

Gene expression profiling

Gene expression profiling is the systematic identification and characterization of genes expressed in a cell or tissue. This can involve measurement of RNA transcripts or proteins. Depending on the technique used, a few genes to all genes in the genome are assayed in a single experiment.

Gene expression profiling can be used for both hypothesis testing and hypothesis generation. For example, a novel gene is identified which encodes a protein that contains a DNA-binding domain. A reasonable hypothesis would suggest that the protein might regulate transcription. To test this hypothesis, an experiment could be performed to generate a gene expression profile of cells with a loss-of-function mutation in the gene to compare to a gene expression profile of wildtype cells. The profiles could be analyzed for gene expression that is higher or lower in the mutant cells vs. wildtype cells. Gene expression profiling is also useful for hypothesis generation. If the function of a cell is unknown, producing a gene expression profile of the cell can provide evidence for a particular function based on prior knowledge (*e.g.,* Gene Ontology terms) of the molecular function of genes detected as expressed. Gene expression profiling is a powerful methodological approach and many more complex experiments can be performed than the ones described above (Kim et al. 2001; Elemento et al. 2007; Ramakrishnan et al. 2010).

Methods for gene expression profiling of the transcriptome

In this section, I will review the benefits and drawbacks of methods used for measurement of RNA levels. There are many methods available to perform transcriptome profiling. The choice of method largely depends on the experiment being performed, available resources/reagents and organism under study. The methods also differ on whether they can measure gene expression *in vivo*, or require isolation of RNA from the sample for an *in vitro* measurement.

The methods can be divided into groups based on the approximate number genes that can be assayed in an experiment. Low to mid-plex methods include (1 – 800 targets): Northern blot, ribonuclease-protection assay, *in situ* hybridization, RT-PCR, and NanoString. High-plex methods include (>1,000 targets): SAGE, MPSS, microarrays, and RNA-Seq.

Low to Mid-plex techniques

Northern blotting is a traditional method to detect native transcripts. RNA is isolated from cells, tissues, or whole animals, size fractionated by denaturing agarose gel electrophoresis and transferred to nitrocellulose or nylon membrane by blotting. The transferred RNA is hybridized with labeled (radioactive, fluorescent, chemiluminescent) cDNA or antisense RNA probes specific to the transcript of interest. The blot is developed and image analysis is performed to determine the intensity of the signal. The levels of intensity can be normalized to an internal control transcript (e.g., actin, GAPDH, 28S rRNA) that does not vary in levels between conditions to obtain a quantitative expression value. Northern blots also provide information about the length of the transcript, since the probe hybridizes the native transcript. Thus, Northern blots can also be used to analyze alternative splicing of transcripts. Northern blots are not as sensitive as many modern approaches and the procedure is moderately complicated by the many steps involved, yet provides a very accurate measure of gene expression.

Ribonuclease protection assays are another method for gene expression analysis. Isolated RNA is hybridized to a labeled cDNA or antisense RNA probe specific to the transcript of interest in solution. A single-strand specific nuclease

is added to the solution to digest all unhybridized single-stranded RNA and leave the “protected” transcript intact. The protected double-stranded nucleic acids are electrophoresed on a polyacrylamide gel and imaged. The gel image is quantified similarly to Northern blots. Ribonuclease protection assays are more sensitive than Northern blots and are performed in solution, allowing quantification of transcripts expressed at low levels.

In situ hybridization (ISH) is a method that allows detection of RNA transcripts in cells, tissues, or whole animals. Labeled probes are incubated with a fixed sample to allow the probes to enter the cell and hybridize with the transcript. The sample can be exposed to film or imaged for fluorescence and quantified. ISH has the advantage of revealing the endogenous localization of a transcript in cells, but quantification of the signal can be difficult depending on the complexity of the sample.

One of the largest advances in the quantification of transcript levels occurred with the development of real-time monitoring of polymerase chain reaction (PCR) product accumulation. Previously, semi-quantitative RT-PCR could be used to measure transcript levels, but only the end-point product could be measured using DNA binding dyes, or radiolabeled nucleotides incorporated into the product. The advent of real-time monitoring of the reaction allows a cycle threshold (C_t) to be set while the reaction is proceeding in the linear phase before saturation. A standard curve is used to establish the cycle threshold for known amounts of cDNA for the target gene. Then, the C_t can be determined for the target gene in an experimental sample. As with other techniques, it is critical to

normalize levels with an internal control. Real-time PCR is one of the most sensitive methods for quantification of gene expression due to the exponential nature of PCR. By multiplexing samples, it is possible to measure more than a single target per reaction enabling higher throughput.

The nCounter Analysis system from NanoString is a new platform that offers medium throughput (~800 targets), high sensitivity (500 attomolar), and wide dynamic range (500X)(Geiss et al. 2008). This system uses solution hybridization with two probes (one “capture” probe and one “reporter” probe) per target. Each probe contains repeat sequences that are used for tandem affinity purification of the tri-molecular complex. After purification, the complex is immobilized on a solid support, and the color-coded tag on the reporter probe is imaged. The color code is specific for one target, so transcripts are quantified by counting the number of detected probes for each target transcript and normalized to multiple internal housekeeping genes. One advantage of this system is the ability to directly use 100 ng of total RNA without reverse transcription to cDNA and further amplification. This system will likely excel where transcription of several hundred genes is needed with the sensitivity of real-time PCR and low cost per sample.

High-plex techniques

To quantify gene expression for thousands to all genes in a genome, several techniques have been developed over the past two decades. SAGE and cDNA microarrays presented the first opportunity to assay many genes in parallel with one experiment (Schena et al. 1995; Velculescu et al. 1995). The SAGE

method generates small-specific fragments (“tags”) from cDNA that are concatamerized, cloned and sequenced. Transcription levels are quantified by the number of times a tag for a gene is counted and normalized to total tag counts (Velculescu et al. 1995). Improved versions of the SAGE method have been described to alleviate some of the issues involved with SAGE, such as, short tag length (14-15 bases), concatamer cloning efficiency, and small insert sizes (Saha et al. 2002; Gowda et al. 2004).

Microarray technology has been the most popular method for measuring expression of thousands of genes in a single experiment. The original method involved spotting individual known cDNA clones in arrays on a glass slide (Schena et al. 1995). Fluorescent RNA probes were generated from purified poly(A) RNA and hybridized to the array. To test for differential gene expression probes from two samples were synthesized using two different fluorescent markers and were simultaneously hybridized to the same microarray. The hybridized microarray is scanned with a laser and the fluorescence is imaged and quantified based on intensity. The first iteration of spotted cDNA arrays was limited to 48 genes, but the technology progressed rapidly and many companies adopted their own design for cDNA or oligonucleotide based arrays. A summary of microarray technologies is shown in Figure 1.2. One of the largest advances came when Affymetrix developed short oligonucleotide microarrays using photolithography processes adopted from the semiconductor industry. This

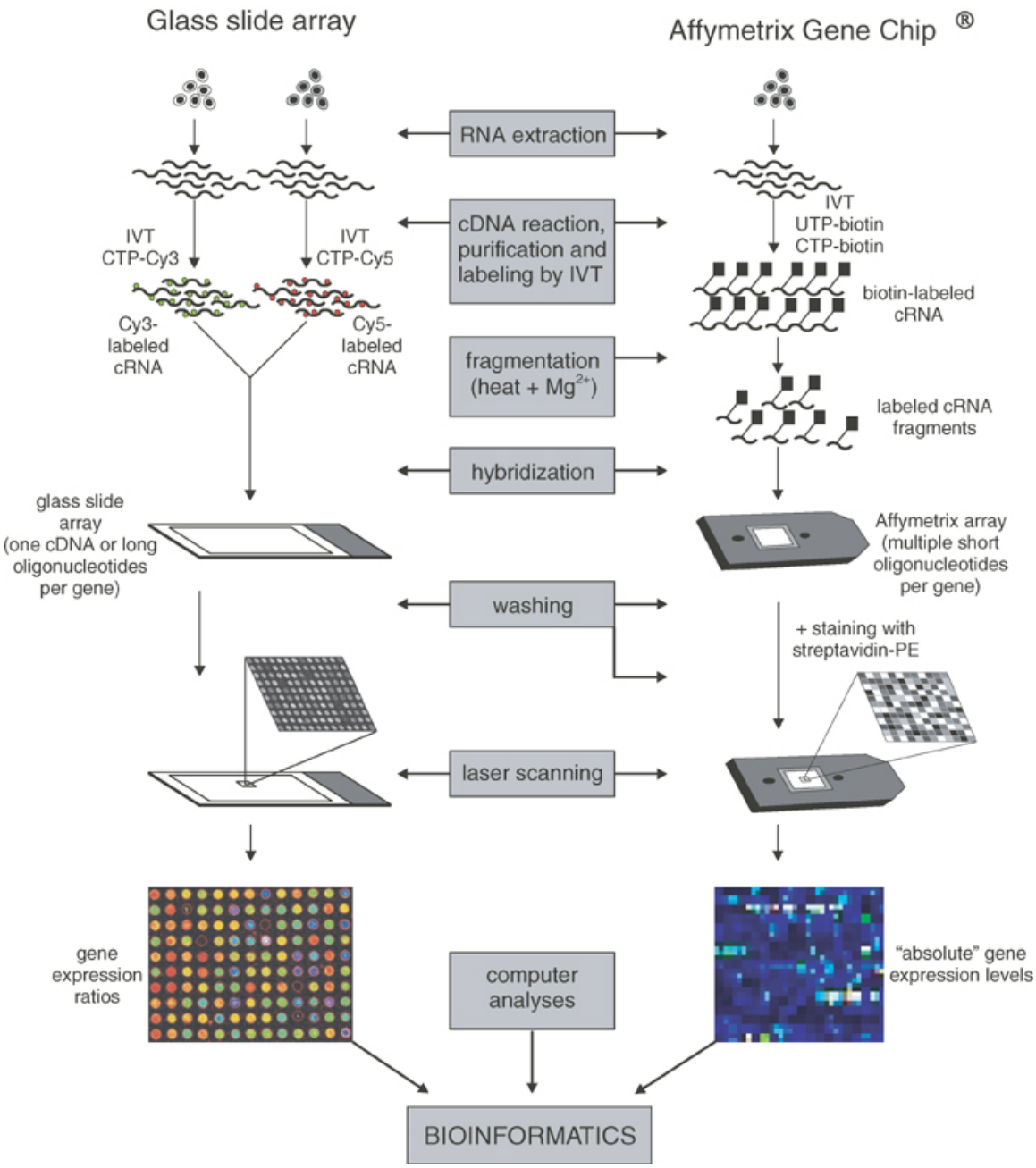


Figure 1.2 Typical procedures for spotted cDNA (left) and oligonucleotide microarrays (right) are shown.
In both procedures, labeled targets are prepared and hybridized to the array. For cDNA arrays, the label is fluorescent (Cy3 or Cy5) and after hybridization, the array is scanned and raw data is generated. For oligonucleotide arrays, after hybridization, the targets are stained with phycoerythrin-labeled streptavidin. Then the array is scanned and raw data is generated. Figure adapted from (Staal et al. 2003).

allowed Affymetrix to design microarrays that contained hundreds of thousands to millions of 25mer oligonucleotide probes on a single array to target nearly every gene in a genome (depending on genome complexity). Affymetrix microarrays are single-color arrays, where biotin-labeled RNA or cDNA is hybridized to the array, and subsequently stained using phycoerythrin-labeled streptavidin. The microarray is scanned with a laser and fluorescence is imaged in gray scale and quantified. Only a single sample can be hybridized to an oligonucleotide array, so treatment and control samples are hybridized to separate microarrays. The traditional GeneChip arrays were designed to measure expression of protein-coding genes. Probe sets were designed to most known gene models with most probes located within the 3'-most exons (Lipshutz et al. 1999). Affymetrix also designed arrays with stringent probe selection criteria requiring similar melting temperatures, uniqueness relative to family members, and lack of similarity of to other abundant RNAs in the sample (rRNA, tRNA, actin). The GeneChip arrays also limited downstream analyses by only targeting 3' ends of protein-coding genes. It is not possible to measure differences in transcript isoform expression or interrogate splice site usage. Later, Affymetrix released exon arrays that designed probe sets to most exon in protein-coding gene models to allow simultaneous analysis of gene expression and alternative splicing. Recently, tiling microarrays were developed to allow interrogation of the whole genome. Probes are designed to "tile" across all non-repetitive portions the genome, which provides a relatively unbiased platform for assaying gene expression irrespective of known gene annotation. Since tiling

arrays usually represent only one strand of the genome and it is not known *a priori* whether a potential novel transcript is transcribed from the plus or minus strand, it is necessary to synthesize ds cDNA to allow either strand of the target to bind the probe. Many reports using tiling microarrays for different species including humans, have suggested widespread transcription occurs throughout the genome (Bertone et al. 2004; David et al. 2006; Consortium et al. 2007). These novel transcriptionally active regions (TARs) have been controversial with some groups suggesting that most bona fide novel transcripts are associated with known protein-coding genes (van Bakel et al. 2010). Due to their utility and reasonable cost, microarrays have become the most routinely used method for measuring global gene expression.

Second-generation sequencing technologies have recently expanded the number of options for gene expression profiling (RNA-Seq). The first of the new high throughput sequencing technologies was provided by 454 Life Sciences (now part of Roche Applied Science). 454 sequencing technology is based on amplification of targets using emulsion PCR, immobilization on a solid substrate, and sequencing-by-synthesis. Emulsion PCR uses fragments of cDNA bound to microbeads that are suspended in droplets of a water-in-oil emulsion. This set up produces small amplification reactors that can produce 10^7 clonal copies of a template DNA (Margulies et al. 2005). The beads are attached to a solid support allowing hundreds of thousands of sequencing reactions to be performed in parallel. Sequencing is performed using a pyrosequencing technique (Nyrén et al. 1993; Ronaghi et al. 1996). Solutions of dNTPs are added to the reaction one

at a time and when DNA polymerase incorporates a nucleotide, inorganic pyrophosphate is released and can be measured by chemiluminescence. The 454 approach is currently able to produce > 1 million high-quality reads with an average length of 400 bases (http://www.454.com/products-solutions/system-features.asp). Studies have utilized 454 sequencing for gene expression profiling, which demonstrate accurate quantification of transcript levels (Bainbridge et al. 2006; Torres et al. 2008).

The Applied Biosystems SOLiD system provides a platform that uses a similar emulsion PCR step for amplification of cDNA fragments, but the sequencing reaction is performed using a hybridization-ligation reaction. Sixteen combinations of dinucleotides are labeled with one of four fluorescent dyes (one dye labels four different dinucleotides). A primer initiates the reaction by annealing at position 0. The dinucleotides probes compete for hybridization and ligation to the sequencing primer. The dye fluorescence is imaged and the dye is cleaved from the dinucleotide. The hybridization, ligation, detection, and cleavage represents one cycle and it is repeated a specified number of times to generate the read length. After a full set of cycles, the synthesized produced is removed and another primer is annealed to the -1 position. Then another round of cycling proceeds. This process is repeated a total of five times allowing every base to be sampled twice. By sampling every base twice, the error rate is extremely low (99.99% accuracy claimed). SOLiD generates much shorter reads than 454, between 35 and 75 bp, but produces many more reads (>100 million per sample, http://www.appliedbiosystems.com). Fewer studies have utilized

SOLiD sequencing for gene expression profiling, but report high accuracy and sensitivity of ~1 copy of transcript per cell (Tang et al. 2009; Bradford et al. 2010).

The most widely adopted second-generation sequencing platform is produced by Illumina. The Genome Analyzer platform differs from the 454 and SOLiD emulsion PCR approach, by ligating bridge PCR adapters to double-stranded cDNA, which immobilize both ends of the cDNA molecule on a solid support. The adapters serve as primers for PCR amplification of the template producing a cluster of cDNA. Sequencing is performed by using DNA polymerase and adding fluorescently-labeled dNTPs one at a time to the reaction. After nucleotides are incorporated, fluorescence is imaged (see Figure 1.3). The current Illumina system produces up to 100 base reads/reaction and ~100 million reads per sample (http://www.illumina.com/). The widespread adoption of the Illumina system is likely due to easier sample preparation using bridge PCR instead of emulsion PCR and higher number of reads generated. There have been numerous studies of gene expression using the Genome Analyzer including the large-scale genome annotation projects mod/ENCODE (Denoeud et al. 2008; Marioni et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Hillier et al. 2009; Mane et al. 2009; Gerstein et al. 2010; Trapnell et al. 2010).

The application of second-generation sequencing technology for gene expression profiling has provided a tremendous benefit and will continue to do so as the technology progresses. Not only can transcript levels can be accurately

and sensitively quantified, but additional information is also gained by validating splicing, alternative splice isoforms, and identifying novel transcripts and transcript fusions. There are still a number of issues to overcome in the areas of sample preparation and data analysis. Many cDNA libraries can be prepared in parallel, but sequencing the libraries presents a major limiting factor in analyzing many samples. A major hurdle for RNA-Seq is the vast abundance of ribosomal RNA in total RNA preparations. Most studies have focused on protein-coding genes and use a poly-adenylated RNA purification procedure to enrich for mRNA. However, to achieve the goal of measuring all RNA transcripts in a given sample, rRNA must be efficiently depleted to prevent an excessive number of sequence reads derived from rRNA. A variety of methods have been developed for depleting rRNA including: Invitrogen Ribominus, RNAse H, Illumina duplex-specific nuclease, and Epicentre Terminator exonuclease, which are discussed in Chapter V. Finally, the cost of sequencing has decreased significantly over the past several years and depending on the experimental design, can now approach the cost of a microarray experiment.

Analysis of RNA-Seq data involves several steps that are designed to exclude random errors. Quality scores are assigned to each base in a read that are logarithmically linked to error probabilities (Ewing and Green 1998). First, the sequence reads must first be filtered by quality of base calling to prevent erroneous reads from being used in downstream analyses. The filtered reads are mapped to the target genome using parameters that determine the level of

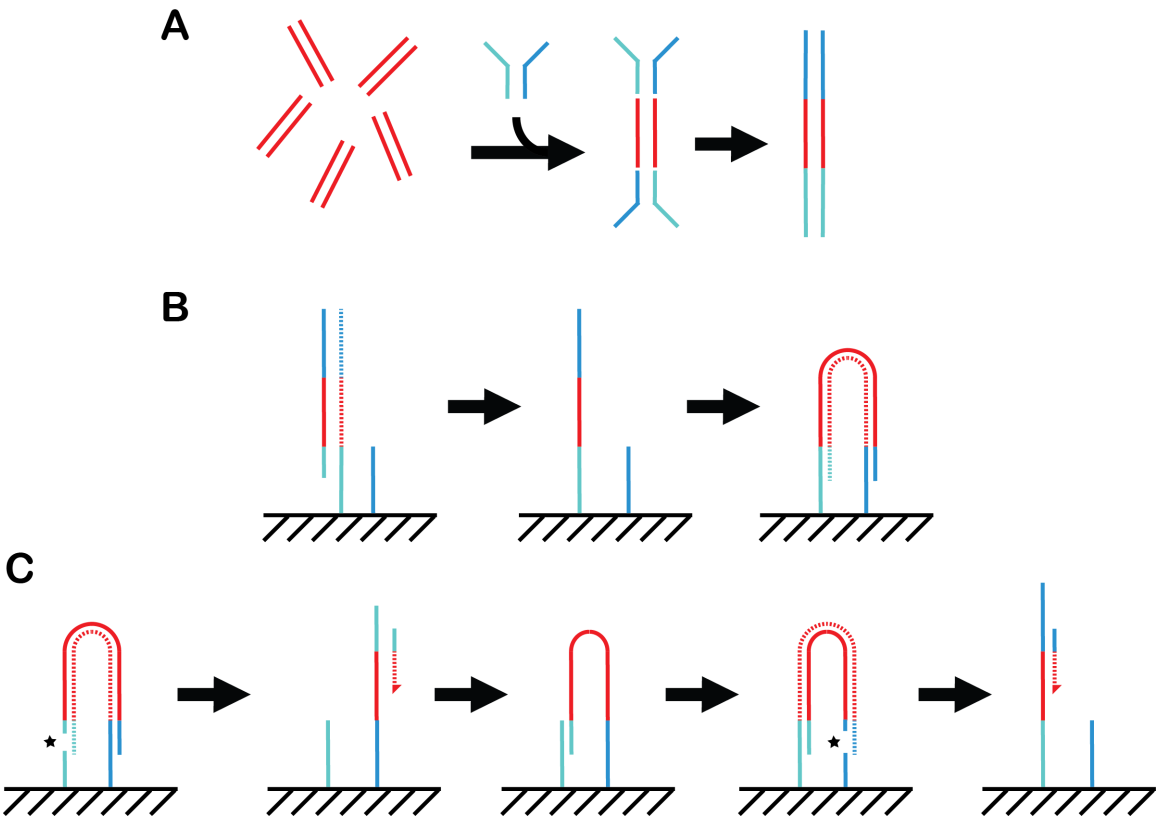


Figure 1.3 Illumina sample preparation procedure.
(A) Forked adapters are ligated to ds cDNA. The ligated product is amplified producing a double-stranded molecule with different sequences at each end. (B) Products are annealed to oligonucleotides attached to the surface of the flow cell (hatches). The oligos serve as primers to synthesize a new strand (dotted). The products are denatured and the newly synthesized strand binds complementarily to a different oligo attached to the surface of the flow cell, forming a bridge, and another product is synthesized. Multiple rounds of amplification produce a clonal cluster of template. (C) A restriction enzyme site in one oligo is used to linearize the cluster and sequencing-by-synthesis is performed (sequence produce shown dotted). To perform paired-end sequencing, the products are denatured and the other surface-bound oligo is cleaved to linearize the product, and the template is sequenced from the other end.

stringency for the alignment. Reads can map to multiple locations in the genome, contain a number of mismatches, or not map to the genome despite high sequence quality. When analyzing RNA-Seq data, reads that span a splice junction will not map to the genome unless a specific splice-aware alignment strategy is used (Trapnell et al. 2009). Normalization of RNA-Seq data also requires careful consideration. To accurately quantify gene expression levels within an experiment, normalization to the number of mapped reads is necessary as with SAGE (number of tags), but since whole transcripts are being analyzed, a large transcript will generate many more reads than a small transcript. Without a computational correction, this circumstance would artificially inflate the apparent expression level of a large versus small gene. A widely used normalization procedure, reads per kilobase of exon model per million mapped reads (RPKM) attempts to correct for these biases (Mortazavi et al. 2008). RPKM will normalize to the length of exons in a gene model and to the total number of reads mapped to the genome. A recent study has suggested that using a normalization method such as RPKM for differential expression analysis may inflate variance between replicates depending on differences in depth of coverage, thus reducing power and increasing Type II error (Anders and Huber 2010). Instead, Anders and Huber suggest using read counts for each gene model and then normalizing to total number of mapped reads for differential expression analysis. While RNA-Seq data analysis is still in its infancy, the power of the technique has motivated discussion and comparison of various methodologies (Zheng and Chen 2009;

Bullard et al. 2010; Oshlack et al. 2010; Garber et al. 2011; Malone and Oliver 2011; Roberts et al. 2011).

Cellular-enrichment strategies

The gene expression profiling strategies reviewed above can be applied to numerous experimental designs. Many experiments measure gene expression from whole animals or tissues, but due to the heterogonous environment of a tissue, many complex processes can be obscured by lack of cellular resolution. To overcome this hurdle, several approaches have been utilized to isolate specific cell-types for gene expression profiling. The major methods include: Laser Capture Microdissection (LCM)(Emmert-Buck et al. 1996), RNA immunoprecipitation (Roy et al. 2002; Keene et al. 2006; Doyle et al. 2008; Heiman et al. 2008), Immunopanning (PAN, immunoselection) (Antoine et al. 1978), Fluorescence-Activated Cell Sorting (FACS)(Bonner et al. 1972), microfluidics-based FACS (Fu et al. 1999; Hu et al. 2005), and manual isolation (Frohlich and Konig 2000; Smith et al. 2000). These methods require a cell-specific marker or distinct morphology to identify the cell for selection. Each method has distinct advantages and caveats, but they all enable specific cell-types to be purified for experimental analysis.

Manual dissection of single cells is the lowest throughput method, but can produce a highly purified sample of cells (Tietjen et al. 2003; Tietjen et al. 2005; Okaty et al. 2011). Frequently a fluorescent marker expressed in the cell-type is used to identify the cell, which can be aspirated with a mouth pipette and placed in a tube for downstream experiments. This approach has been successfully

adopted for real-time PCR analysis and microarray profiling (Smith et al. 2000; Tietjen et al. 2003; Tietjen et al. 2005; Cherry et al. 2009).

FACS is one of the most widely used methods for isolation of a purified cell-type (see Figure 1.4 for a schematic)(Bonner et al. 1972). A heterogenous cell sample is mixed with a carrier fluid and pushed through a nozzle to create a laminar flow stream. The laminar flow orders the cells, which pass through a detector one at a time and are pulsed with lasers to measure cell size and fluorescence properties. If a given cell-type expresses a marker to identify that cell, then it can be selected by establishing gating criteria determined empirically from analyzing cells with and without the marker. After detection, the stream of cells is broken into droplets by intense vibrations with only one cell per droplet. If a cell was selected for sorting, an electrical charge is added to the droplet, which passes between charged metal plates that deflect the cell droplet into a collection tube. Many thousands of cells can be analyzed per second allowing the collection of a large sample of cells as well as the isolation of rare cells from a large starting population of cells. Numerous studies have applied FACS-based purification of cell-types for gene expression profiling (Zhang et al. 2002; Colosimo et al. 2004; Fox et al. 2005; Lobo et al. 2006; Okaty et al. 2011; Pan et al. 2011). Methods based on FACS have been developed using microfluidics devices to analyze and sort cells, which provide a customized solution, but typically lack the throughput of traditional FACS (Hu et al. 2005; Ishii et al. 2010).

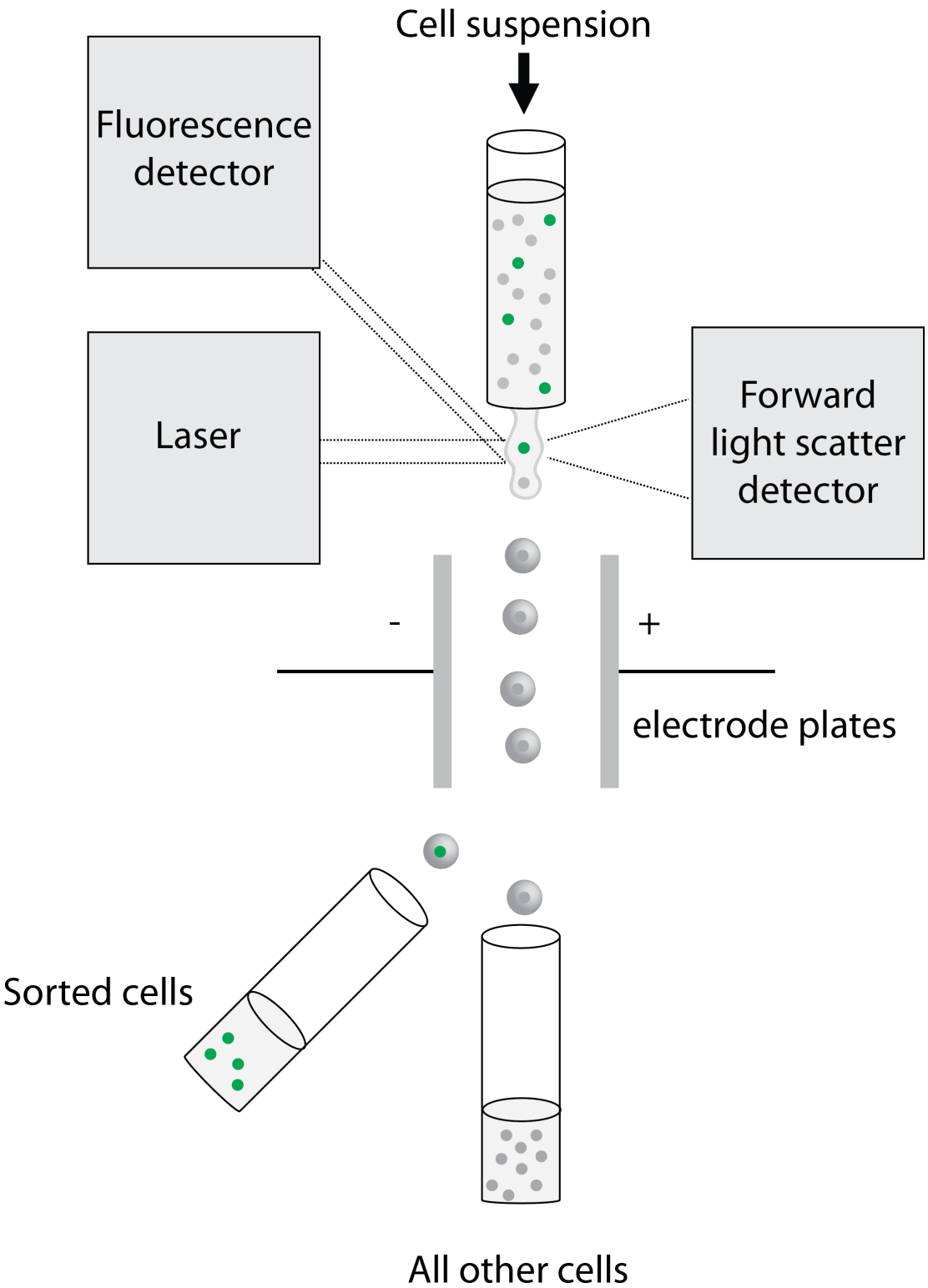


Figure 1.4 Schematic overview of fluorescence activated cell sorting (FACS). Cells are scanned with a laser and analyzed for size and fluorescence, droplets are formed and droplets with positive cells are charged and deflected by electrode plates into a collection tube.

Laser capture microdissection typically uses a microscope mounted UV laser to cut a region of cells or tissue out of a sample. The isolated cell sample is then removed from the remainder of the tissue using various methods such as catapulting the sample into a tube using a defocused UV laser. LCM typically has a resolution down to 5 μm, so cells with a diameter less than 5 μm cannot be isolated without risk of contamination from other cells. LCM is widely used for pathology samples, since fixed samples are primarily used with LCM and are more readily accessible for human samples. Gene expression profiling using LCM isolated cells is common, but can have contamination issues (Leethanakul et al. 2000; Chung et al. 2005; Rossner et al. 2006; Caretti et al. 2007; Emrich et al. 2007; Kube et al. 2007; Okaty et al. 2011).

Immunopanning or immunoselection uses antibodies directed against a particular cell-surface protein expressed in a cell-type of interest. The antibodies are covalently bound to a solid surface and target cells are bound to the surface. This method has been used for selective purification or depletion of cell-types for expression profiling resulting in an enriched population of cells and is particularly useful for cell-types that do not survive FACS (Farkas et al. 2004; Ivanov et al. 2006; Cahoy et al. 2008; Okaty et al. 2011).

One general method obviates the need to mechanically enriching the cell-type of interest by isolating RNA directly from that cell. Several variations of RNA immunoprecipitation have been developed, but two primary methods have aimed at isolating cell-specific mRNA. The first method, mRNA-tagging, utilized the poly(A)-binding protein (PABP) that binds to the poly(A) tails of mRNAs. The

PABP is expressed in specific cell-types by the control of a known promoter and is tagged for immunoprecipitation. Animals are fixed to crosslink the mRNA to the PABP and the complex is immunoprecipitated. The crosslinks are reversed and the mRNA is purified for expression analysis. This approach has been primarily used in *C. elegans*, but has also been used in *Drosophila melanogaster* (Roy et al. 2002; Kunitomo et al. 2005; Pauli et al. 2005; Yang et al. 2005; Jiao et al. 2007; Von Stetina et al. 2007; Watson et al. 2008; Spencer et al. 2011). A similar method, genetically targeted translating ribosome affinity purification (TRAP), has been implemented in mouse aimed at isolating mRNAs being actively translated (Doyle et al. 2008; Heiman et al. 2008). In this method, a BAC construct encoding a tagged L10a ribosomal protein is expressed in specific cell-types and the RNA:ribosomal protein complex is purified as with mRNA-tagging. This strategy may provide a more accurate representation of expressed genes than the mRNA tagging method since TRAP specifically detects mRNAs as they are translated whereas the mRNA tagging method should not have this bias. Each method serves a particular strategic approach and likely could be used in parallel to define all mRNA transcripts in a cell and those that are translated under particular conditions.

In summary, the combination of cell enrichment and global gene expression profiling provides a powerful approach to define cellular identity and provide clues for molecules that are critical for the function of that cell type. Methods for gene expression profiling in *C. elegans* have focused on the use of MAPCeL (MicroArray Profiling of *C. elegans* celLs) and mRNA-tagging in

combination with microarrays or SAGE (Roy et al. 2002; Zhang et al. 2002; Blacque et al. 2005; Fox et al. 2005; Fox et al. 2007; McGhee et al. 2007; Von Stetina et al. 2007; Watson et al. 2008; McGhee et al. 2009). To enhance our ability to isolate individual cells using FACS, I have implemented a multicolor FACS approach in which a single cell is marked by the unique overlap of two different fluorophores driven by promoters that co-express in only the cell of interest. This method facilitated the isolation of two command interneurons of the *C. elegans* motor circuit that otherwise would not be accessible to cell specific profiling (Chapter II). The expression profile of one of the command interneurons revealed an enriched transcript for an immunoglobulin domain cell adhesion molecule that I then showed is necessary for synaptic connectivity between the command neuron and its motor neuron targets (Chapter III). Thus, this profiling approach has identified a candidate gene for synaptic specificity which can now be systematically tested to establish its mechanism of action. For the broader purpose of systematically defining gene expression across the *C. elegans* anatomy, we used a combination of FACS and mRNA-tagging to isolate cell specific RNAs for hybridization to tiling microarrays that revealed the timing, location and expression levels of all *C. elegans* genes (Chapter IV). The advent of RNA-Seq provided an opportunity to enhance our gene expression analysis, but also required the implementation of new methods to allow sequencing from small quantities of RNA (< 10 ng) and to exclude rRNA templates. Chapter V describes an empirical analysis of rRNA depletion strategies and methods for quantification and differential expression analysis of genes. Our use of *C.*

elegans as a model organism allows the use of genetic screens for identifying genes that function in a specific pathway, but mapping the causal mutation from a screen can be laborious. By taking advantage of the compact nature of the *C. elegans* genome and second-generation sequencing, it is now possible to sequence the genome of individual genetic mutants to identify the lesion (Chapter V).

CHAPTER II

METHOD FOR ISOLATION OF SINGLE NEURONS

In this chapter, data production and analysis was a joint effort. I optimized 2-color FACS and profiled the AVA neuron and all embryonic cell reference. A research assistant, Rebecca McWhirter profiled the AVE neuron. Stefan Henz parsed WormBase gene annotation and mapped tiling array probes to gene models. Georg Zeller performed the microarray normalization and differential expression tests.

Introduction

The nematode *C. elegans* has a simple and well-defined nervous system with only 302 neurons, for which nearly all synaptic connections are described in a comprehensive wiring diagram (White et al. 1986; Chen et al. 2006). Therefore, *C. elegans* is an ideal model organism for establishing the relationships between neuron identity and connectivity. A catalog of genes neuron specific gene expression would facilitate this study by providing a link between the function of a neuron and its molecular signature. This knowledge also has the potential of shedding light on the biological mechanisms responsible for the communication between neurons. For example, an earlier work revealed specific genes with potential roles in synaptic connectivity by correlating single-neuron gene expression data with the wiring diagram (Varadan et al. 2006). Here I describe

the application of microarray profiling methods to L+R pairs of *C. elegans* embryonic neurons using in a method that exploits multicolor FACS to isolate specific neuron types that are uniquely marked with a unique combination of different colored fluorescent reporter genes.

Ideally, individual neuron-specific promoters can be used to tag specific neurons for isolation and RNA extraction. However, due to the dearth of neuron-specific single promoters, combinations of promoter elements are needed for this task. There are in total 82 L+R pairs of embryonic neurons and an additional 58 single embryonic neurons for a total of 222 embryonic neurons. Given the appropriate combination of promoters, many or all embryonic could potentially be uniquely marked for isolation by FACS and expression profiling (Varadan and Anastassiou ; Zhang et al. 2004).

Methods and Materials

C. elegans culture and strains used in this study

Nematodes were grown as described (Brenner 1974). Strains were maintained on nematode growth media (NGM) plates inoculated with the *E. coli* strain OP50. Strains used to isolate cells were N2 (wildtype Bristol strain), NC1749 [hdIs32(P*glr-1*::DsRed2), otEx239(P*rig-3*::GFP;*pha-1*+)] for AVA, and NC1750 [hdIs32(glr-1::DsRed2), gvEx173(opt-3::GFP + pRF4(*rol-6*))] for AVE.

Preparation of embryonic cells and primary cell culture

Methods used for generating preparations of embryonic cells and for primary cell culture have been previously described (Christensen et al. 2002) and are summarized here. Embryos were obtained by hypochlorite treatment of

synchronized populations of adult hermaphrodites and digested with chitinase to remove the egg shell. The resultant single-cell suspension of embryonic cells (in egg buffer) (Christensen et al. 2002) was resuspended in L-15 cell culture medium, supplemented with 10% FBS and penicillin/streptomycin and plated at a density of 1 x 10-6 ml-1 on 1-well chamber slides (Nunc) coated with poly-L-lysine (Sigma). Primary cultures were maintained overnight at 25 °C.

Isolation of fluorescently-labeled embryonic cells by FACS

FACS was used to isolate AVA and AVE, each labeled with GFP and DsRed2. Cells derived from freshly dissociated embryos were passed through a 5 m filter (Durapore - Millipore) to remove debris. Primary cultures were examined 24 hr after plating to confirm expression of fluorescent markers (GFP and DsRed2). Cultured cells were resuspended in egg buffer and prepared for FACS as previously described (Fox et al. 2005). Dead cells were labeled by staining with 7-AAD (Invitrogen) (~1-2 g/mL of cells). Viable cells were isolated using either a FACStar Plus (AVA) or FACSAria (AVE) flow cytometer (75 m nozzle, ~10,000-15,000 events/sec) (Becton Dickinson, San Jose, CA). FACS gates were empirically adjusted to achieve >80% purity for AVA and ~90% for AVE. The fraction of target cells (80-90%) for each cell type was determined by direct inspection in the fluorescence microscope 24 hr after plating on 4-well chamber slides coated with peanut lectin (Sigma) (Fox et al. 2005). Yields of target cells ranged from ~5,000 to ~20,000 for each FACS run. At least 3 independent samples were collected for each cell type. Reference samples for

cells obtained from primary cultures (late embryos, LE) were obtained by isolating all viable cells from the wildtype (N2) strain (Fox et al. 2005).

RNA extraction from embryonic cells isolated by FACS

Cells collected for RNA isolation were sorted directly into Trizol LS (Invitrogen) up to a final 1X concentration. The sample was extracted with chloroform, RNA precipitated with isopropanol, washed 2X with 75% EtOH and resuspended in RNAase-free H_2O. A DNA-free RNA purification kit (Zymo Research) was used to DNAase-treat and purify RNA according to the manufacturer's instructions. RNA quality and yield was determined using a Bioanalyzer (Agilent). Total amounts of RNA for each sample ranged from 600 pg to ~10 ng.

RNA amplification

The WT-Ovation Pico kit (NuGEN Technologies, Inc) was used to amplify RNA (0.6 ng to 5 ng starting material). 3 µg from each reaction was used to generate double stranded cDNA with the WT-Ovation Exon module (NuGEN Technologies, Inc). 4-5 µg of ds-cDNA was fragmented and labeled using the FL-Ovation Biotin V2 module (NuGEN Technologies, Inc).

Microarray hybridization

The *C. elegans* 1.0R tiling array (Affymetrix) contains > 3 million perfect match (PM)/mismatch (MM) probe pairs representing the *C. elegans* non-repetitive genome. Probes are 25 nt in length and tiled at an average distance of 25 nt as measured from the centers of adjacent probes. Double-stranded cDNA targets were used for hybridization because all probe sequences match a single

DNA strand whereas individual transcripts can be derived from either the plus or minus strands. At least 3 independent replicates were obtained for each cell type. *Interse* Pearson correlation coefficients were calculated between replicates to ensure consistent sample preparation and hybridization.

Mapping tiling probes to the *C. elegans* genome and its annotation

Perfect match (PM) 25mer tiling probe sequences were mapped to the *C. elegans* genome sequence (release WS200) (Rogers et al. 2008) using vmatch to detect all (direct and inverse) matches of length ≥17 with at most one mismatch or indel (Abouelhoda 2004). Only probes that perfectly aligned to a single genomic location were retained thereby discarding the most highly repetitive probes. Repeat information was kept for probes with multiple imperfect alignments as a filter for subsequent analyses. These included 70,189 PM tiling probes with exact matches and an additional 113,054 probes with inexact matches leaving a total of 2,758,587 non-repetitive probes according to the above criteria.

Probe set definition and estimation of expression for annotated genes

For each protein-coding gene model annotated in WS200 (Rogers et al. 2008), we constructed a probe set containing all PM tiling probes that could be perfectly aligned to corresponding constitutive exons. Repetitive probes (see above definition) were removed from gene probe sets and probe set information was converted into CDF. Subsequently, expression was estimated for genes with a minimal probe set size of three using RMA, which involves quantile normalization and summarization with median polish (RMA's default array-

background normalization was omitted (Bolstad et al. 2003; Irizarry et al. 2003; Gautier et al. 2004).

Testing genes for expression above background

To establish whether expression of a particular gene was significantly higher than the array background intensity, we compared its hybridization signal to an empirical null model. For each gene probe set we constructed a background probe set from an equally sized random sample of probes mapped to annotated intergenic regions. This sampling process was repeated until $\geq 10^6$ background probe sets had been collected. For a given biological sample, we established the null model from the median of the PM intensities of the background samples pooling replicate data. The empirical p-value of a gene's expression was estimated as the proportion of background probe sets with the same or higher median intensity than the median PM intensity of the gene probe set. Expression p-values were adjusted for multiple testing using the false discovery rate (FDR) method by Benjamini & Hochberg (as implemented in the R function p.adjust(x, method="fdr") (Benjamini and Hochberg 1995).

Determining differentially expressed genes

Differentially expressed genes were identified using a linear model and an empirical Bayes moderated t-statistic (Smyth 2004) implemented in the Bioconductor package Limma (Smyth 2005).

Microscopy

Isolated embryonic cells were imaged using differential interference contrast (DIC) and epifluorescence optics with a Zeiss Axiovert inverted

microscope equipped with an ORCA ER (Hamamatsu) high-resolution, cooled CCD camera. Intact animals were imaged with a Zeiss Axioplan compound microscope equipped with an ORCA ER camera, Leica TCS SP5 confocal microscope, or Zeiss LSM510 confocal microscope.

Results

To demonstrate the ability to use more than one promoter to isolate and profile individual neurons or L+R pairs, we selected the AVA (L+R) and AVE (L+R) neurons for two-color FACS. The well-characterized promoter of the *C. elegans* AMPA-type glutamate receptor subunit, *glr-1*, is expressed in both AVA and AVE as well as in other command interneurons (*i.e.,* AVB, AVD, PVC) (Maricq et al. 1995; Hutter 2004; Schmitz et al. 2007). Other promoters that drive expression in each of the individual command interneurons, but otherwise do not overlap with the P*glr-1* expression pattern can be selected to uniquely mark each interneuron (Table 2.1). We chose the combination of P*glr-1* and P*rig-3* to test for unique identification of AVA neurons. According to WormBase, 35 neurons express *glr-1*, nine neurons (along with 2 amphid sheath cells and intestine) express *rig-3*, but only one neuron, AVA (L+R) expresses both markers. We confirmed this prediction by direct inspection in the confocal microscope of a strain containing the P*glr-1*::DsRed2 and P*rig-3*::GFP transgenes (Figure 2.1).

The next set of experiments, were designed to optimize FACS sorting gates for these dual-color cells. First, we generated primary cultures of GFP and DsRed2 cells to set FACS gates for each fluorophore. We also established a sorting gate for the fluorescent dye (7-AAD) used for marking (and excluding)

Table 2.1. Promoter combinations for profiling the command interneurons.

Gene Combination	**Isolated L+R pair**	
rig-3 AND *glr-1*	AVA	Backward circuit
tol-1 AND *nmr-1*	AVD	
opt-3 AND *glr-1*	AVE	
sra-11 AND *glr-1*	AVB	Forward circuit
flp-11 AND *glr-1*	PVC	

dead cells in the preparation. With these parameters in place, we isolated GFP + DsRed2 neurons ("dual positive cells") from the transgenic line expressing both P*rig-3*::GFP and P*glr-1*::DsRed (Figure 2) at ≥80% purity (Figure 2.1). The same strategy was applied to the AVE command interneurons which are labeled with P*opt-3*::GFP and P*glr-1*::DsRed2 (see Table 2.1). AVE neurons were sorted to ≥88% purity.

Because our approach isolated only a single L+R pair of neurons from each animal, minute quantities of RNA were obtained from each FACS run. For further experimentation, it was therefore necessary to optimize RNA isolation from sorted cells. This goal was achieved in part by sorting cells directly into Trizol LS (Invitrogen) in order to limit RNA degradation and then purifying RNA with a DNA-free RNA kit (Epicentre).

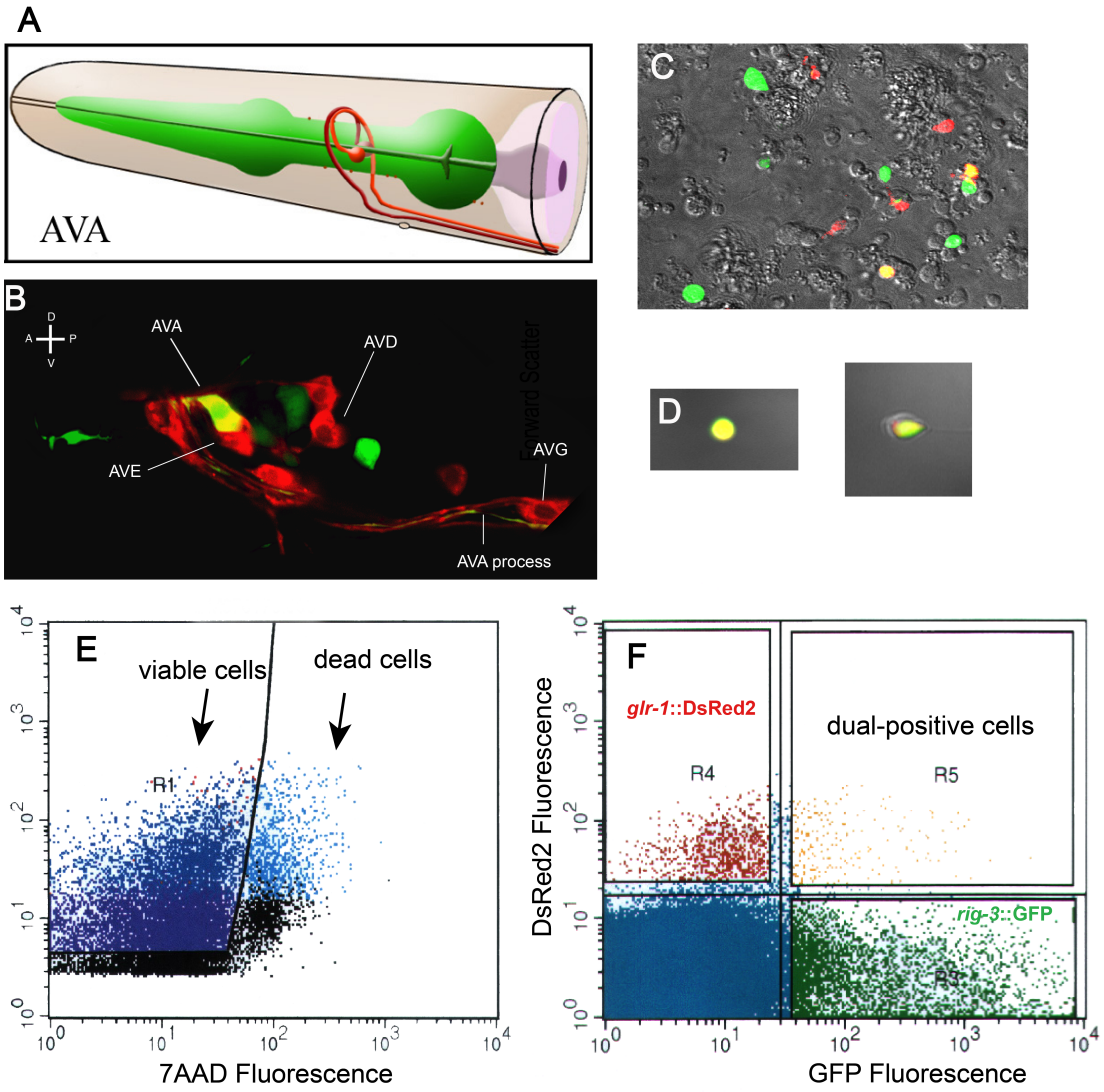


Figure 2.1. Isolation of AVA command interneuron by FACS.
(A) Cartoon showing bilateral pair of AVA (L+R) neurons in the head region extending single processes into the ventral nerve cord.
(B) Confocal projection of head region (left side, posterior to right, adult) showing co-expression (yellow) of *rig-3*::GFP (green) and *glr-1*::DsRed2 (red) in AVAL. Other command interneurons (AVD and AVE) and the single neuron, AVG, express *glr-1*::dsRed2, but not *rig-3*::GFP.
(C) Cultured cells from the *rig-3*::GFP;*glr-1*::DsRed2 strain before FACS. Arrows point to neurons (yellow) expressing both DsRed2 and GFP.
(D) GFP + DsRed2 neurons (AVA) after isolation by FACS.
(E) Scatter plot with sorting gate to exclude dead cells labeled with 7AAD.
(F) Fluorescence gate to isolate GFP + DsRed2 cells (R5, dual positive cells) from *rig-3*::GFP or *glr-1*::DsRed2 cells.

allowed amplification using the Ovation WT-Pico kit and Exon module (NuGEN, Inc.) to produce double-stranded cDNA. RNA quality was evaluated on the Agilent Bioanalyzer and amplified with the NuGen Pico-Ovation kit. Because we used tiling arrays (see chapter IV) for these samples, a double-stranded hybridization probe was generated for each sample with the NuGen Exon module. Fragmented and labeled ds cDNA was hybridized to *C. elegans* whole-genome tiling microarrays (Affymetrix). The tiling microarrays contain probes covering all non-repetitive sequences, allowing us to interrogate gene expression throughout the genome. Probes with redundant matches (>17 nt) to the genome were removed. Probes located entirely within exons of protein-coding gene models (WS200) were used to create a custom chip definition file (CDF) for analysis of gene expression. Three reproducible biological replicates were obtained for AVA, AVE, and or an all-embryonic cell reference profile.

To determine which genes are differentially expressed in specific cell-types, we first needed to compare normalization procedures and statistical tests to determine which combination identifies the most true-positives and the fewest false-positives. We have compared two normalization procedures (RMA and FARMS) and two statistical tests (a moderated t-test/limma and the RankProduct test) with the AVA and embryonic pan-neural data sets. RMA (robust-multiarray analysis) is a standard normalization procedure for background adjusting probe intensities (Irizarry et al. 2003). FARMS (Factorial analysis for robust microarray summarization) is a newer normalization method that uses a factorial analysis

model to estimate RNA concentration assuming Gaussian measurement noise (Hochreiter et al. 2006). These normalization procedures were applied to the raw array data and a moderated t-test (limma) or the non-parametric RankProduct test were used to detect significant changes in gene expression against the all-cell reference sample (Breitling et al. 2004; Wettenhall and Smyth 2004; Smyth 2005).

We first looked for known AVA-expressed genes in each of the 4 lists (see Table 2.2). The RMA-limma analysis detected the most known AVA-expressed genes with 15/26. RMA-RankProduct was next at 10/26, then FARMS-RankProduct (10/26) and FARMS-limma (5/26). The RMA-limma approach detects the most known expressed genes and contains the highest number of enriched genes (782), but could also contain the highest number of false positives. To approximate the number of false positives, we compared each of the 4 lists with an embryonic muscle expression profile (Fox et al. 2007). The RMA-limma and FARMS-RankProduct lists both detected 19 muscle-enriched genes, but the RMA-limma “false-positive” rate is lower at 2.4% vs. 4%. Because a significant number of genes are known to be expressed in both muscle and neurons, it is possible that at least some of the overlapping transcripts in these data sets represent true positives. Therefore, we used the RMA-limma approach for determining which genes are differentially expressed in each of our data sets (this work ; Gerstein et al. 2010; Spencer et al. 2011).

For purpose of detecting enrichment of proteins with particular functions in these neuron-specific expression profiles, I used a software tool, DAVID (The

Table 2.2 Known genes expressed in AVA (WormBase)

WBGene00000054	*acr-15*	acetylcholine receptor
WBGene00001400	*fax-1*	nuclear hormone receptor
WBGene00001444	*flp-1*	FMFRamide neuropeptide precursor
WBGene00001461	*flp-18*	FMFRamide neuropeptide precursor
WBGene00001587	*ggr-2*	GABA/Glycine receptor
WBGene00001612	*glr-1*	Glutamate receptor
WBGene00001613	*glr-2*	Glutamate receptor
WBGene00001615	*glr-4*	Glutamate receptor
WBGene00001616	*glr-5*	Glutamate receptor
WBGene00001676	*gpa-14*	G-protein alpha
WBGene00002129	*inx-7*	innexin
WBGene00003000	*lin-11*	LIM domain homeobox protein
WBGene00003774	*nmr-1*	Glutamate receptor
WBGene00003775	*nmr-2*	Glutamate receptor
WBGene00003969	*pef-1*	protein phosphatase
WBGene00004370	*rig-3*	IgCAM
WBGene00006747	*unc-7*	innexin
WBGene00006748	*unc-8*	degenerin
WBGene00006749	*unc-9*	innexin
WBGene00006778	*unc-42*	Homeobox domain
WBGene00006830	*unc-103*	ERG-like K+ channel
WBGene00006890	*vem-1*	cytochrome b5-like heme/steroid-binding domain
WBGene00009562	*flp-22*	FMFRamide neuropeptide precursor
WBGene00019641	K10G6.4	unknown
WBGene00020368	*ast-1*	ETS protein
WBGene00020952	*kel-8*	keltch-repeat protein

Database for Annotation, Visualization and Integrated Discovery) that detects potential over-representation of specific GO terms (Huang et al. 2007). The AVA expression profile was significantly enriched for 16 clusters of GO terms for molecular functions, protein domains, and pathways (see Table 2.3). The most significant cluster consisted of the GO terms helicase and ATPase activity. The fourth cluster consisted of neuron-associated categories including: synapse, postsynaptic cell membrane, glutamate receptor activity, etc. The AVE neuron expression profile was significantly enriched for 29 clusters of GO terms. The first annotation cluster had a strikingly high number of neuron-associated categories including: synapse, ion channel, neurotransmitter-gated ion channel, etc. Other annotation clusters included G-protein coupled receptors, PDZ domains, neuropeptide activity, and immunoglobulin domains. The enrichment for annotation categories related to neuron function underscores the validity of the expression profiles and provides evidence for novel functions of these neurons.

AVA and AVE are capable of responding to multiple neurotransmitter signals

AVA and AVE have been previously shown to receive signals from presynaptic neurons that secrete the neurotransmitter glutamate. The command interneurons express a variety of glutamate receptors (*glr-1*, *glr-2*, *glr-4*, *glr-6*, *nmr-1*, *nmr-2*) as evidenced by expression reporters, antibodies, and functional studies (Maricq et al. 1995; Zheng et al. 1999; Brockie et al. 2001a; Brockie et al. 2001b; Brockie and Maricq 2003). The command interneurons have also been shown to express acetylcholine receptors (*acr-15* and *acr-16*) and a GABA

receptor (*ggr-2*), although no functional experiments have been performed on the GABA receptor (Feng et al. 2006; Wormbase 2006).

To address whether these and additional neurotransmitter receptors are detected in the AVA/AVE expression profiles I mined each profile for known/predicted receptors. Each profile robustly detected many of the known receptors and identified additional receptors (Figure 2.2). All known AVA/AVE glutamate and acetylcholine receptors were identified in one or both profiles. The ionotropic glutamate receptors (iGluRs) *glr-1*, *glr-2*, *glr-4*, *glr-5*, *nmr-1*, and *nmr-2* have been studied and functionally well-described in the command interneurons (Maricq et al. 1995; Brockie et al. 2001a; Mellem et al. 2002; Zheng et al. 2004; Walker et al. 2006a; Walker et al. 2006b). For example, the activity of GLR-1 is required for the light nose touch response and the frequency of turning in response to certain stimuli, such as food (Hills et al. 2004; Chalasani et al. 2007). Both AVA and AVE were enriched for *glr-1*, *glr-4*, *glr-5*, and *nmr-2*, with *glr-2* enriched in AVA and *nmr-1* enriched in AVE. The metabotropic glutamate receptors, *mgl-1* and *mgl-2*, are enriched in AVA and AVE respectively, and are thought to have a neuromodulatory role by functioning at either presynaptic sites (*mgl-2*, Class I) or postsynaptic sites (*mgl-1*, Class II) to inhibit or promote NMDA receptor (iGluRs) activity, respectively. The muscarinic acetylcholine receptor, *gar-1*, is annotated as expressed in AVA in WormBase, but is only enriched in AVE. It is possible both neurons express *gar-1*, but is much more highly expressed in AVE. It is also possible that due to the close proximity of the AVA cell body to the AVE cell body, the cell expressing *gar-1* was incorrectly identified

Table 2.3 Enriched GO/KEGG/protein domain categories

AVA	Fold enr.	AVE	Fold enr.
helicase activity	4.3	synapse	8.6
ATPase activity	3.2	GPCR	6.4
helicase	2.8	glutamate receptor activity	3.9
synapse	2.6	PDZ	2.8
transcription factor	2.2	neuropeptide receptor	2.6
		ion transport	2.0
		calcium-dependent membrane targeting	1.9
		transmembrane	1.6
		oxygen binding	1.6
		immunoglobulin domain	1.5

as AVA instead of AVE. Previously the only GABA receptor annotated as expressed in AVA is *ggr-2* with no GABA receptor annotated as expressed in AVE. In the expression profiles, *ggr-2* is detected as enriched in the AVE profile, but not in the AVA profile. Two additional GABA receptors are also identified in the expression profiles with AVA enriched for *gbb-1* and AVE enriched for both *gbb-1* and *gbb-2*. These two GABA receptor subunits are orthologs of the GABBR1 and GABBR2 GABAB receptors, respectively, and function as a dimer, so both are likely to be expressed in AVA and AVE.

Previously, no biogenic amine receptor was annotated as expressed in either AVA or AVE. Interestingly; AVA is enriched for one dopamine/serotonin receptor, *dop-5*, and AVE is enriched for two dopamine receptors *dop-1* and *dop-2* and the dopamine/serotonin receptor *dop-5*. AVE is also enriched for the serotonin receptor *ser-4*, an ortholog of the mammalian 5-HT1 metabotropic serotonin receptor. The dopaminergic neurons located in the nerve ring of *C. elegans*, ADE and CEPs, make parallel connections to AVA and AVE, respectively, with the ADEs making 11 synapses onto AVA and the CEPs making 17 synapses onto AVE (Figure 2.3). The primary output of the ADEs is to another head interneuron RIG, but AVA has the next highest number of synapses with ADE. Likewise, the CEP neurons primarily synapse with the head interneuron RIC, but AVE has the next highest number synapses among other interneurons. These data suggest ADE to AVA and CEP to AVE signaling likely occurs using dopamine through the *dop-1*, *dop-2*, and *dop-5* receptors and could be necessary for dopamine-mediated behaviors in *C. elegans*. Dopamine signaling

mediates the slowing response when animals encounter food (Chase et al. 2004). It is possible this effect is partially controlled through dopamine neuron signaling to the command interneurons.

AVA and AVE could also be responsive to neuropeptide signaling. Accumulating evidence suggests neuropeptide signaling has a modulatory role in certain animal behaviors (Li et al. 1999a; Li et al. 1999b; Davis and Stretton 2001). No neuropeptide receptors were found enriched in AVA, but AVE is enriched for *npr-11*, *npr-13*, *npr-14*, and Y58G8A.4. *npr-11* has been shown to be responsive to the FMRFamide-like peptide *flp-1* in the olfactory system of *C. elegans* (Chalasani et al. 2010). Y58G8A.4 is responsive to the *flp-18* peptides, which are expressed in AVA and other neurons in the head (Kubiak et al. 2008). Therefore, neuropeptide signaling could provide a function to modulate the activity of AVE.

Evidence for neurotransmitters used by AVA/AVE to signal to postsynaptic targets

Neurotransmitters released by neurons to signal to downstream neurons, usually require enzymes necessary for synthesis and/or transporters used for uptake of specific molecules made available by other sources. These molecules can then be used directly as a neurotransmitter by packaging into vesicles for release at the synapse, or are further processed into another product that is packaged for synaptic release. The AVA/AVE enriched gene lists were analyzed for genes required for synthesis or uptake of common neurotransmitters (glutamate, acetylcholine, GABA, biogenic amines, and neuropeptides). The

vesicular glutamate transporter, *eat-4*, which is required for loading glutamate into synaptic vesicles, is enriched in AVE, but not AVA. No plasma membrane localized glutamate transporter, which is required for uptake of glutamate from the extracellular space, is enriched in either AVA or AVE, but the glutamate transporter, *glt-6*, is detected as an expressed gene (EG, see Methods) in AVA and *glt-1*, *glt-4*, and *glt-6* are detected as expressed genes in AVE. These data suggest it is likely that AVA and AVE use glutamate as a neurotransmitter to signal to postsynaptic neurons.

The use of acetylcholine as a neurotransmitter, requires uptake of choline into the cell via the choline transporter, *cho-1*, and the action of a choline acetyltransferase enzyme, *cha-1*, to synthesize acetylcholine. Acetylcholine is then loaded into vesicle by the vesicular acetylcholine transporter, *unc-17*. *cho-1* is enriched in the AVE profile and is detected as an EG in AVA. Yet, the biosynthetic enzyme *cha-1* and vesicular transporter *unc-17* are not detected as enriched or as expressed genes in either neuron. A previous study performed immunohistochemistry to detect endogenous CHA-1 and UNC-17 protein, but did not observe expression in AVA or AVE. Choline has a major role in all cells as an essential compound necessary in the synthesis of membrane phospholipid components of the plasma membrane (Michel et al. 2006). Choline is the head group of phosphatidylcholine, which is a major constituent of neuronal membranes and is required for normal axon outgrowth and neuron survival (Wurtman 1992; Yen et al. 2001). High expression of the choline transporter in AVA and AVE suggests that phosphatidylcholine could be required for survival

and/or axon outgrowth that is occurring in embryonic development when these neurons were profiled.

Four biogenic amine neurotransmitters are synthesized from the amino acids tryptophan and tyrosine. Serotonin (5-HT) synthesis requires tryptophan hydroxylase (*tph-1*) to generate 5-hydroxytryptophan (5-HTP), and the aromatic amino acid decarboxylase (*bas-1*) to generate serotonin from 5-HTP. *tph-1* and *bas-1* are detected as EGs in AVA, but not in AVE. Interestingly, both AVA and AVE are enriched for the vesicular monoamine transporter, *cat-1*, which loads the monoamines into synaptic vesicles. It is therefore possible that AVA could use serotonin as a neurotransmitter, but both neurons may use other biogenic amines as well. Dopamine synthesis requires the tyrosine hydroxylase, *cat-2*, to make L-DOPA from tyrosine, and the aromatic amino acid decarboxylase, *bas-1*, to convert L-DOPA to dopamine. *cat-2* is not detected in either neuron, but *bas-1* is detected as an EG in AVA. With these genes expressed at very low levels or not detectable, it is not likely that either neuron uses dopamine as a neurotransmitter.

Tyrosine can also be converted to the neurotransmitter, tyramine, by the tyrosine decarboxylase, *tdc-1*, and tyramine can then be converted to the neurotransmitter octopamine by the tyramine β-hydroxylase, *tbh-1*. *tdc-1* is enriched in both AVA and AVE, while *tbh-1* is only detected as an EG in AVE. These data suggest AVA may use tyramine as a neurotransmitter and AVE may use tyramine and/or octopamine.

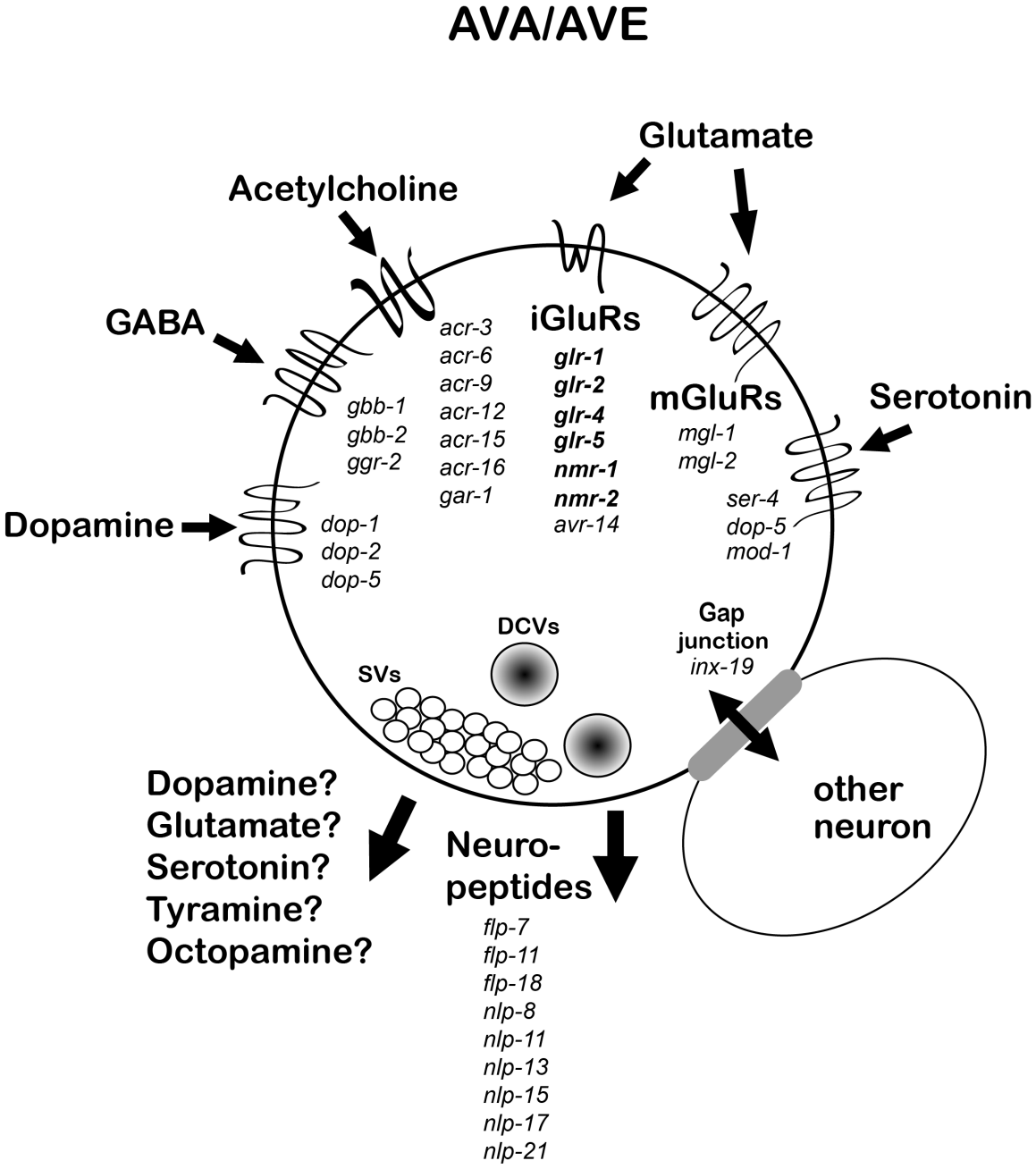


Figure 2.2. Signaling components detected in the AVA and AVE command interneurons.

Neuropeptides are small peptide neurotransmitters that are initially expressed as larger pro-proteins that are cleaved by the pro-protein convertase, *egl-3*, and carboxypeptidase E, *egl-21* in the endoplasmic reticulum for release as dense-core vesicles from the *trans*-golgi network. These vesicles have been visualized by electron microscopy vesicles containing a relatively electron dense region in comparison to chemical neurotransmitter vesicles, which are small and clear (Pysh and Wiley 1974). *egl-3* is detected as an EG in AVE and *egl-21* is detected as an EG in both AVA and AVE. The CAPS homolog *unc-31*, which is involved in post-docking calcium-regulated dense-core vesicle (DCV) fusion is enriched in both AVA and AVE. With the high-level expression of many neuropeptides in AVA and AVE (see Figure 2.2), peptidases and a protein necessary for DCV fusion, neuropeptide secretion is likely a major signaling pathway for AVA and AVE.

Transcription factors detected in AVA and AVE

Two transcription factors are known to control aspects of AVA and AVE development and function. The paired-like homeodomain gene *unc-42* is the primary transcription factor required for AVA and AVE cell-fate specification by controlling expression of glutamate receptors and axon outgrowth (Hart et al. 1995; Maricq et al. 1995; Wightman et al. 1997; Baran et al. 1999; Galliot et al. 1999; Brockie et al. 2001b). While the AVA profile does not detect *unc-42*, it is enriched in the AVE profile. Both AVA and AVE are enriched for the *fax-1* nuclear receptor that is required for expression of specific glutamate receptors and other transcripts in parallel to *unc-42* (Much et al. 2000; Wightman et al. 2005). *lin-11*,

encodes a LIM homeodomain protein that has been reported to be expressed in AVA and AVE, but *lin-11* mutants did not show a loss of *glr-1* or *unc-42* expression in AVA (Sarafi-Reinach et al. 2001). Additionally, *lin-11* expression was not dependent on *unc-42* expression, suggesting *lin-11* and *unc-42* act in parallel pathways in AVA. In our expression profiles, *lin-11* is detected as enriched in AVE and as expressed in AVA.

By comparing the AVA and AVE enriched gene lists to the worm transcription factor compendium (wTF2.1) (Reece-Hoyes et al. 2005), many new transcription factors are identified in each neuron. In the AVA enriched gene list, 28 transcription factors are identified (see Figure 2.4a), while in the AVE enriched gene list, 16 transcription factors are detected (see Figure 2.4b). Three transcription factors are enriched in both profiles: *fax-1*, F10B5.3, and ZK686.4. *fax-1* is previously known to be expressed as mentioned above, but F10B5.3 and ZK686.4 are both novel. F10B5.3 is an uncharacterized nematode-specific C_2H_2 class zinc finger transcription factor and ZK686.4 is an uncharacterized conserved C_2H_2 class zinc finger transcription factor. ZK686.4 is the homolog of the vertebrate Zinc finger matrin-type protein 2 (Zmat2), which has not been previously studied, but is highly expressed in the mouse brain (http://mouse.brain-map.org/brain/gene/69080733.html). In total, 41 transcription factors are detected in either AVA or AVE, with the zinc finger family being the most prevalent in each neuron, which is not surprising since most of the transcription factors encoded in the *C. elegans* genome are in the zinc finger family (575/931, 62%) (Reece-Hoyes et al. 2005).

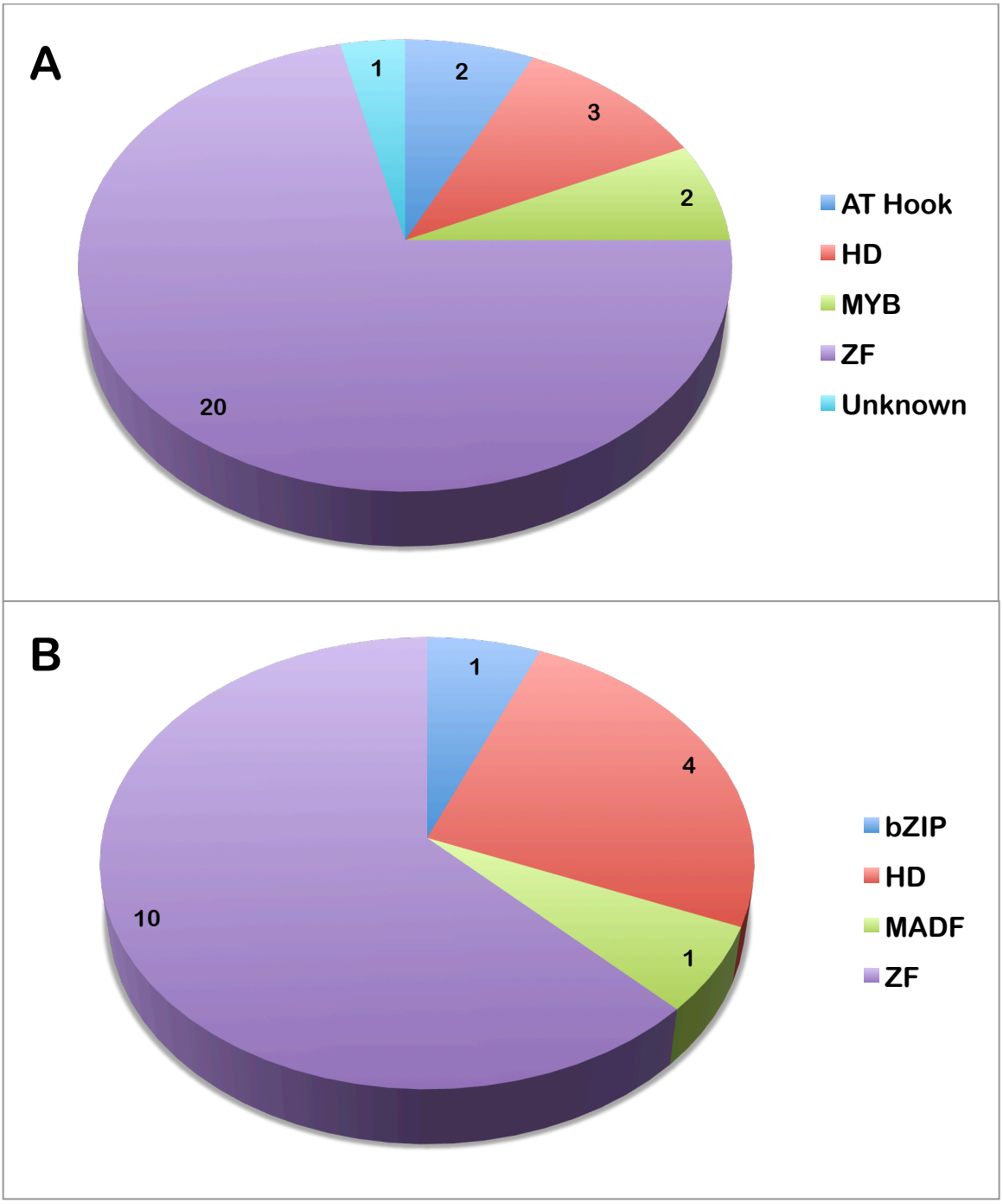


Figure 2.4. Transcription factor families identified in (A) AVA and (B) AVE expression profiles.

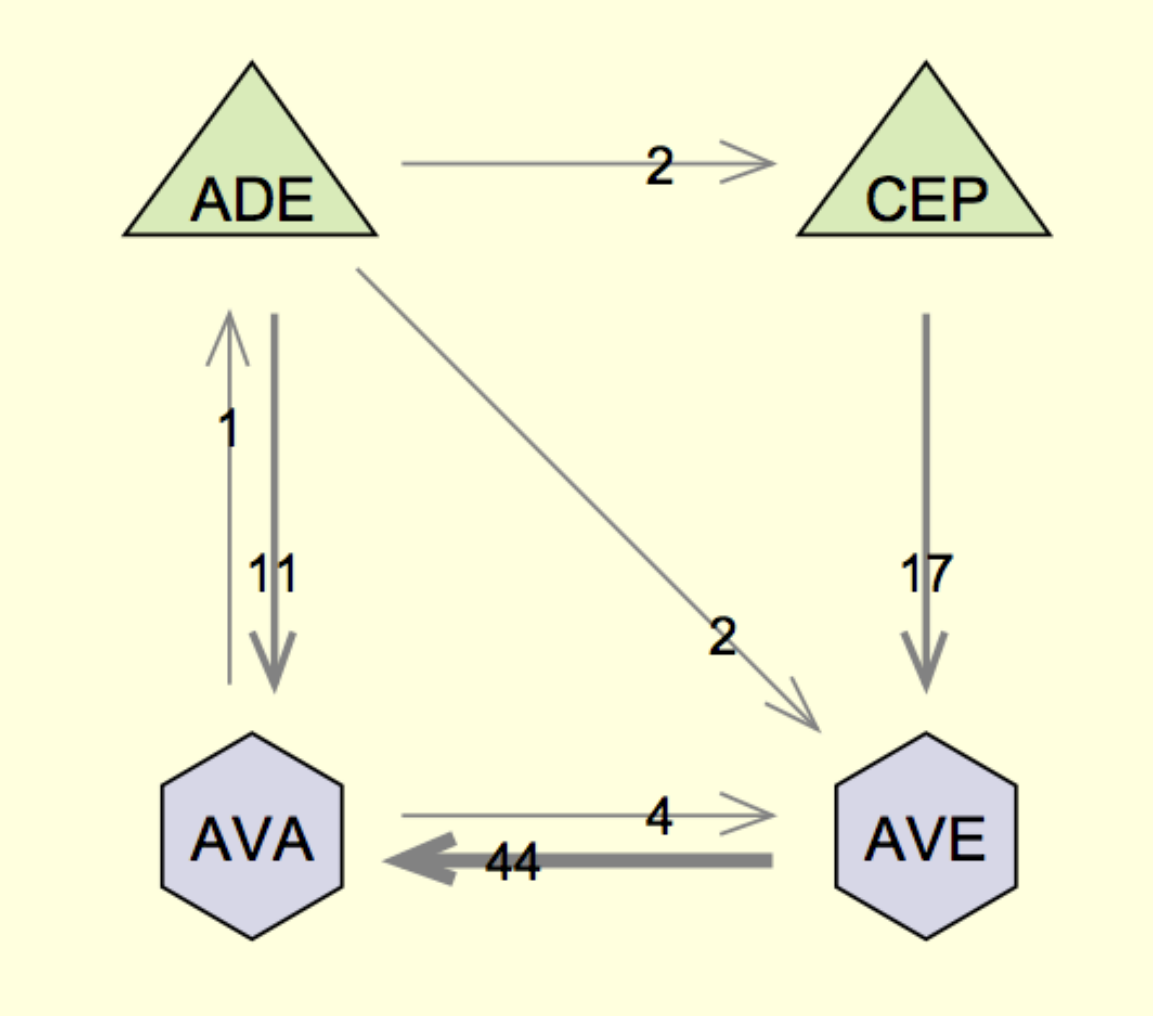


Figure 2.3. Synaptic connections between the dopaminergic neurons ADE and CEP and the command interneurons AVA and AVE.

Adhesion molecules enriched in AVA and AVE

Adhesion molecules contribute to the architecture and connectivity of the nervous system. To identify adhesion molecules that may play a role in AVA and AVE axon guidance and connectivity, I mined the AVA and AVE enriched gene lists for genes encoding proteins with either immunoglobulin (Ig) or cadherin (Cdh) protein domains. Ig domains have been previously shown to play an important role in axon guidance and connectivity (Biederer et al. 2002; Shen and Bargmann 2003; Shen et al. 2004; Washbourne et al. 2004). There are four genes each, enriched in AVA and AVE that encode Ig domain containing proteins. Three of the four genes are shared between AVA and AVE, *rig-3*, *lad-2*, and *zig-8*. AVA is also enriched for a peroxidasin, *pxn-1*, and AVE is enriched for *syg-1*. The *rig-3* Ig-domain cell adhesion molecule has been shown to have a minor contribution to proper axon guidance of the command interneurons (Schwarz et al. 2009). We used the promoter of *rig-3* to drive GFP expression as a marker for AVA in this study, which shows the *rig-3* transcript is highly expressed in AVA. *rig-3* is also enriched in AVE, which is not seen with the GFP reporter (*i.e.*, there is no overlap with P*glr-1*::DsRed2 expression, which does express in AVE, see Figure 2.1), suggesting the promoter fragment used to drive GFP expression does not contain all of the endogenous regulatory elements to control *rig-3* expression. *rig-3* has been previously detected as enriched in DA motor neurons, which are major postsynaptic targets of AVA and AVE. This circuit is responsible for driving backward locomotion. These results therefore identify *rig-3* as a candidate cell surface protein for promoting the formation of

synapses between AVA/AVE and A-class motor neurons (see Chapter III). *lad-2* encodes is a non-canonical member of the L1 cell-adhesion molecule family. Previously, *lad-2* was not detected as expressed in either AVA or AVE, but a *lad-2* mutant was shown to have axon guidance defects in the SDQL neuron (Wang et al. 2008). AVA and AVE are also enriched for the *zig-8* gene, which encodes a putative secreted Ig-domain cell adhesion molecule and loss of *zig-8* results in minor PVQ, PVP, and HSN axon guidance defects (Benard et al. 2009). The peroxidasin homolog, *pxn-1*, is enriched in AVA and has been suggested to negatively regulate basement membrane formation or function (Gotenstein et al. 2010). In addition to the peroxidase domain, *pxn-1*, encodes a leucine-rich repeat domain and 2 Ig domains. Due to its known role in basement membrane function, AVA may secrete PXN-1 to modify the surrounding basement membrane. AVE is enriched for a novel Ig domain adhesion molecule, *syg-1*, that is required for proper localization of HSN motor neuron synapses on vulval muscle (Shen and Bargmann 2003). *syg-1* is also enriched in A-class motor neurons suggesting a role in this circuit (Von Stetina et al. 2007).

Cadherins are single-pass transmembrane proteins that are involved in Ca^{2+}-dependent homotypic interactions that are necessary for tissue morphogenesis (Hill et al. 2001). The *C. elegans* genome contains 15 genes that encode proteins with cadherin-like extracellular domains (Hill et al. 2001). No cadherins are enriched in AVA, but AVE is enriched for two genes, *cdh-1* and *cdh-10*. Little is known about *cdh-1* and *cdh-10* except that in a genome-wide RNAi screen, knockdown of *cdh-1* in an RNAi-sensitized background [*rrf-*

3(*pk1426*)] resulted in locomotion defects (Simmer et al. 2002). The AVE axon is unique among the command interneurons since it stops short of the vulva, whereas the AVA axon extends along the entire length of the ventral nerve cord into the tail region. Adhesion molecules unique to AVE like *cdh-1* and *cdh-10* are therefore candidates for proteins that terminate axon outgrowth. This possibility could be readily determined by visualizing AVE morphology in *cdh-1* or *cdh-10* mutant backgrounds.

Discussion

The combination of cell culture and FACS has proven invaluable to numerous areas of biology from cell cycle research to high-throughput drug screening. Primary cell culture of *C. elegans* embryonic cells has been an accessible method for 10 years with the majority of publications focusing on isolating cell-types of interest for gene expression profiling (Christensen and Strange 2001; Zhang et al. 2002; Colosimo et al. 2004; Fox et al. 2005; Etchberger et al. 2007; Fox et al. 2007; Von Stetina et al. 2007; Spencer et al. 2011). Our focus has not only been to utilize this method for identification of genes involved in particular cellular pathways, but to extend the method allowing routine isolation of a single cell-type. In this work, we demonstrate that using available fluorescent protein-based expression reporters, a single cell can be marked using the combination of two different fluorophores expressed under the control of promoters that uniquely overlap in that cell. Cells expressing GFP and DsRed2 are easily detected by modern flow cytometers and dead cells can be discarded by use of the DNA-binding dye 7-aminoactinomycin D. The isolation of

a rare cell-type results in a pure, but relatively small number of cells and therefore limited amounts of RNA. To overcome this obstacle to expression analysis, I optimized isolation intact RNA from sorted cells used an RNA amplification protocol to generate sufficient material for microarray profiling (Fox et al. 2005; Fox et al. 2007; Von Stetina et al. 2007; Watson et al. 2008; Spencer et al. 2011).

AVA and AVE utilize multiple neurotransmitter receptors and signaling systems to control locomotion

We applied this method to individual left/right pairs of neurons that function in controlling backward locomotion in the *C. elegans* motor circuit. The AVA command interneuron is the most highly connected neuron in the *C. elegans* nervous system (White et al. 1986). The AVE command interneuron also displays are large number of synapses and gap junctions, but plays a lesser role in controlling locomotion (Chalfie et al. 1985; White et al. 1986). Integration of signals from diverse sensory neurons and interneurons would be expected to require a broad array of ligand-specific receptors. From our expression profiles, we detected receptors that for multiple types of neurotransmitters including glutamate, acetylcholine, GABA, dopamine, serotonin, and neuropeptides. These receptors are not solely activating receptors, but also have inhibitory and modulatory roles. These results are consistent with a model in which AVA and AVE command interneurons are responsive to a wide spectrum of neurotransmitters that are likely integrated to produce coherent output to motor neurons that control locomotion.

The neurotransmitters that interneurons release to regulate motor neuron activity have not been directly identified, however. Gene expression profiles of the principle AVA and AVE postsynaptic targets, the ventral cord DA and VA motor neurons (A-class motor neurons) (White et al. 1986) have provided some clues. These studies detected robust expression in A-class motor neurons of receptors for multiple classes of neurotransmitters including acetylcholine, GABA, dopamine, serotonin, and neuropeptides. Acetylcholine receptor subunits are the most prevalent (Fox et al. 2005; Von Stetina et al. 2007). The AVA and AVE expression profiles present moderate evidence for the packing and/or synthesis of several signaling molecules. Interestingly, both AVA and AVE are highly enriched for the vesicular monoamine transporter *cat-1*, which loads biogenic amines into synaptic vesicles. The A-class motor neurons express dopamine (*dop-1*) and serotonin receptors (*ser-4* and *ser-7*). Since AVA also expresses the serotonin synthetic enzymes *tph-1* and *bas-1*, it is possible AVA signals to A-class motor neurons using serotonin. AVA and AVE also express many neuropeptides and A-class motor neurons express several neuropeptide receptors, thus peptidergic signaling may play a crucial role in communication between these neurons.

Connectivity between AVA, AVE, A-class motor neurons

Due to the reproducible specificity of synaptic inputs of AVA and AVE, with A-class motor neurons (White et al. 1986; White et al. 1992), this motor circuit presents an attractive model for studying synaptic specificity. The *unc-4* homeodomain transcription factor functions in the A-class motor neurons to

inhibit B-class motor neuron fate and to maintain connections with AVA, AVE, and AVD (Miller et al. 1992; White et al. 1992; Miller et al. 1993; Miller and Niemeyer 1995; Pflugrad et al. 1997), but little is known of how the command interneurons interact with motor neurons targets to coordinate synaptogenesis. Studies in *C. elegans* and other systems have shown that adhesion molecules can contribute to synaptic specificity (Shen and Bargmann 2003; Shen et al. 2004) and laminar targeting (Yamagata et al. 2002). In this study, we have identified three Ig-domain cell adhesion molecules enriched in AVA and AVE (*lad-2*, *zig-8*, and *rig-3*). The embryonic A-class motor neurons are also enriched for *rig-3* (Fox et al. 2005). Some IgCAMs are able to participate in homophilic binding (Miller et al. 1995; Tian et al. 2000), suggesting the RIG-3 protein may bind to itself homophilically between AVA/AVE and A-class motor neurons to trigger synaptogenesis (see Chapter III).

The availability of gene expression data set for the AVA and AVE interneurons should facilitate future studies to identify specific gene products that are used by these command interneurons to integrate signals received from sensory neurons and propagate them to motor neurons. Additionally, due to the well-defined specificity synaptic connections involving these neurons, candidate signaling pathways and adhesion molecules suggested by these neuron-specific profiles can be tested for roles in defining this circuit diagram. It should also be possible to augment these transcriptome-profiling strategies by applying new methods of measuring gene expression (RNA-Seq) and by exploiting a new larval cell isolation procedure to generate companion profiles of these neurons

during the period in which they establish connections with postembryonically derived motor neurons (McWhirter RD, Spencer WC, Miller DM, unpublished results; Zhang et al. 2011).

CHAPTER III

THE CELL ADHESION MOLECULE, RIG-3, PROMOTES SYNAPSE FORMATION IN THE BACKWARD MOTOR CIRCUIT

Introduction

Development of the nervous system is a complex multi-step process. Synaptic choice appears to occur after cell-fate specification and axon outgrowth to the target region and thus is likely to depend on mechanisms whereby neurons recognize specific partners (Jontes and Phillips 2006). Molecules that promote general synaptogenesis, such as NCAM and some FGFs are known, but the mechanisms that allow neurons to discriminate between many potential synaptic targets are not well understood (Rutishauser et al. 1985; Tosney et al. 1986; Caday et al. 1990; Li et al. 2002; Salinas 2005; Waites et al. 2005). It has been hypothesized that cell-type specific expression of different isoforms of the adhesion molecules DSCAM, neurexin/neuroligin, or SynCAM would generate the needed specificity (Zipursky et al. 2006; Hattori et al. 2007). These molecules have been shown to be important for synaptic specificity in certain contexts, yet there are many circuits in the brain that do not rely on these proteins for connectivity (Caday et al. 1990; Salinas 2005; Gerrow and El-Husseini 2006). Therefore, additional novel mechanisms for the establishment of synaptic partners must exist.

Cell adhesion molecules and synaptic specificity.

It has become apparent that cell adhesion molecules (CAMs) are important for synaptic specificity. Specific CAMs localize to synaptic regions and show various effects on neuronal function and synapse formation (Biederer et al. 2002; Yamagata et al. 2003). These molecules include cadherins, proto-cadherins, NCAM, nectins, Fasciclin II, sidekick-1/-2, SynCAM, SYG-1/SYG-2, among others (Gerrow and El-Husseini 2006). The cadherins, sidekicks, and SYG-1/SYG-2 have provided the greatest insight into synaptic specificity thus far (Yamagata et al. 2002; Shen and Bargmann 2003; Shen et al. 2004; Takeichi and Abe 2005). These molecules have been shown to be critical for the establishment of specific circuits in the nervous system.

Cadherins are one class of classical Ca^{2+}-dependent adhesion molecules. Originally described as generic adhesion molecules for various tissues, evidence for their potential role in synaptic specificity is growing (Hatta and Takeichi 1986; Takeichi and Abe 2005). Recent studies indicate that cell- and lamina-specific expression of certain cadherins allow homophilic binding to promote synaptogenesis (Masai et al. 2003). Using a short intracellular domain, cadherins bind another family of proteins, the catenins, to mediate protein-protein interactions that are necessary for synaptogenesis (Benson and Tanaka 1998). β-catenin is known to recruit PDZ domain containing proteins through its PDZ binding motif (Bamji et al. 2003). These results suggest cadherins are involved in clustering of synaptic components, but additional work is needed to clarify the molecular components of this signaling pathway.

Sidekick, a transmembrane Immunoglobulin cell adhesion molecule (IgCAM) was originally identified in a screen for retinal patterning factors in *Drosophila* (Nguyen et al. 1997). An independent study identified the vertebrate (chick) homolog, sidekick-1 (sdk-1) and paralog sidekick-2 (sdk-2), in a screen for retinal ganglion cell-specific genes. The vertebrate sidekicks are expressed in specific retinal neurons and sub-lamina of the internal plexiform lamina, the target of retinal neurons. Ectopic expression of sdk-1 was sufficient to alter connectivity to the expressing target sub-lamina. Thus, the "sidekicks" appear to function as instructive cues for lamina specificity (Yamagata et al. 2002; Yamagata and Sanes 2008). The downstream mechanisms that trigger synapse formation in this location are unknown, however.

Recent studies in *C. elegans* have revealed a pair of IgCAM proteins that specify the creation of a particular set of synapses. In *C. elegans*, egg-laying is controlled by the HSNL motor neuron, which synapses with vulval muscle. SYG-1 is expressed in the HSNL motor neuron and functions as the receptor to the guidepost protein SYG-2, which is expressed in adjacent vulval epithelial cells (Shen and Bargmann 2003; Shen et al. 2004). Thus, SYG-2 expression effectively marks the location for synapse formation. Both molecules contain immunoglobulin and fibronectin type-III domains. Absence of SYG-2 results in reduced number of synapses on vulval muscle and ectopic expression of SYG-2 drives accumulation of SYG-1 clusters and ectopic synapse formation. SYG-1 and SYG-2 are expressed in other specific *C. elegans* cells, but it is not yet apparent if they also provide instructive cues for synapse formation in other

neural circuits or how they promote pre-synaptic organization. It has recently been shown that in addition to specifying the location of synapse formation, SYG-1 protects the adjacent synapse from degradation by the ubiquitin-proteasome system (UPS) (Ding et al. 2007).

Previous approaches for identifying synaptic specificity molecules have included candidate gene analysis and genetic screens for connectivity-defective mutants (Mann et al. 2002; Shen and Bargmann 2003). Cell-specific microarray profiling offers an alternative strategy for identifying synaptogenic components. One successful example of this approach used the *Drosophila* neuromuscular junction (NMJ) as a model for delineating synaptic specificity genes (Inaki et al. 2007). This study focused on two neighboring embryonic muscle cells with that are connected to separate sets of motor neurons. The muscle cell M12 is innervated by the motor neurons RP5 and V whereas the adjacent M13 muscle cell is innervated by the motor neurons RP1 and RP4. To identify genes encoding instructive cues for innervation, the authors generated single-cell expression profiles of the two muscle cells using the Affymetrix *Drosophila* expression microarray. Comparison of the M12 vs M13 profiles revealed differentially expressed transcripts. Of particular interest, Wnt4 is expressed at high levels in M13 and low levels in M12. Genetic analysis showed that Wnt4 acts as an inhibitory cue that prevents innervation of M13 by motor neurons (RP5, V) that normally make connections with M12. The microarray experiment also detected additional transcripts for a subset of membrane and secreted proteins with potential roles in this process. This study demonstrates that cell-

specific expression profiles can provide a wealth of candidate molecules for determining synaptic choice. This approach will likely be invaluable for identifying synaptic specificity molecules in other systems.

The *C. elegans* motor circuit provides a useful model for elucidating additional synaptic specificity mechanisms. EM reconstruction has provided a complete wiring diagram of the nervous system and the lineage of each cell is known (White et al. 1976; Sulston et al. 1983). This comprehensive atlas provides a clearly defined map of neuron-specific connections. For example, Figure 3.1 depicts connections for the excitatory neurons in the motor circuit. The motor circuit consists of command interneurons, which reside in head and tail ganglia and their motor neuron targets In the ventral nerve cord. Separate circuits drive forward or backward locomotion. The command interneurons AVA, AVD, and AVE synapse with the DA and VA motor neurons to mediate backward locomotion. Forward movement depends on the command interneurons AVB and PVC, which provide inputs to DB and VB motor neurons. Genetic experiments have shown that synaptic choice in the backward circuit is controlled by selective expression of the homeodomain transcription factor UNC-4 in DA and VA motor neurons (Miller et al. 1992). *unc-4* mutant animals can move forward but not backward. This behavioral defect arises from specific miswiring of VA motor neurons. In *unc-4* mutants, the synapses between the backward movement command interneurons (AVA, AVD, AVE) to VA motor neurons are replaced with gap junctions from AVB and chemical synapses with PVC (White et al. 1992).

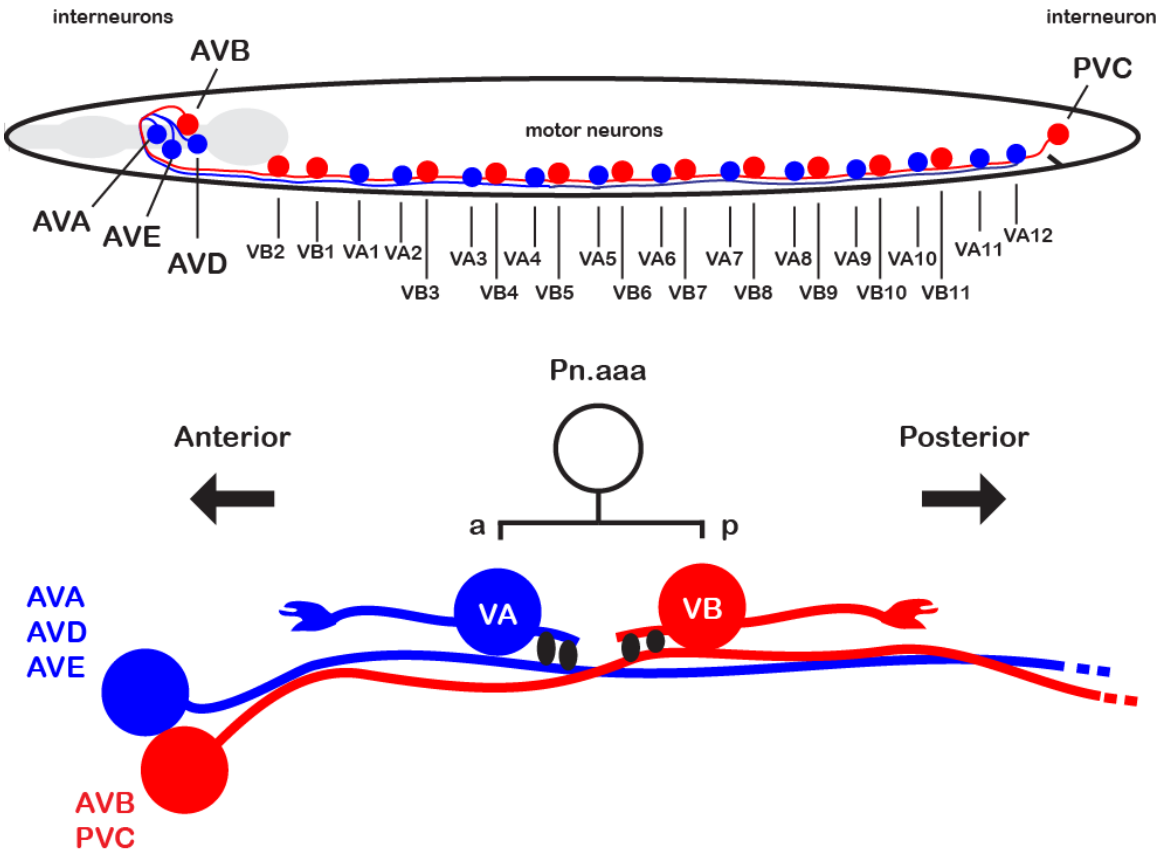


Figure 3.1. Schematic of the excitatory components of the motor circuit.
The backward movement circuit is depicted in blue and the forward movement circuit is depicted in red. The command interneurons form *en passant* synapses onto the motor neurons in the ventral nerve cord.

Thus, in *unc-4* mutants, VA motor neurons acquire synaptic inputs normally reserved for their sister VB motor neurons. Recent work has shown that VA miswiring in *unc-4* mutants depends on de-repression of a VB-specific homeodomain transcription factor, the HB9 homolog, CEH-12 (Von Stetina et al. 2007). Interestingly, genetic experiments indicate that only VA motor neurons in the posterior ventral cord are miswired due to ectopic *ceh-12*, but not VA motor neurons in the anterior cord. This finding suggests that UNC-4 is likely to control additional downstream pathways that function in parallel to *ceh-12* to preserve normal inputs to anterior VAs. These results show that the Unc-4 phenotype is a result of changes in gene expression in VA motor neurons. The downstream molecules which are regulated by these transcription factors (UNC-4 and CEH-12/HB9) and which allow the command interneurons to discriminate between A vs B-class motor neurons are unknown.

The mechanisms underlying synaptic specificity remain poorly understood. Previous work indicates that CAMs are likely candidates for molecules necessary for synaptic specificity (Abbas 2003). Several different classes of IgCAMs are expressed in the *C. elegans* nervous system, yet few have been investigated for functional roles in synaptic choice (Vogel et al. 2003). The *C. elegans* motor circuit provides a unique opportunity for detailed analyses of these molecules and for strategies to identify additional determinants of synaptic specificity (see Figure 3.1 for a diagram of the *C. elegans* motor circuit). The existence of a well-defined wiring diagram and predictable movement phenotypes arising from specific changes in this network can be exploited to evaluate the function of

candidate synaptic determinants. In the previous chapter, I described adhesion molecules expressed in AVA that could play a role in synaptic connectivity between AVA and the A-class motor neurons. One of those adhesion molecules, *rig-3*, encodes an immunoglobulin domain cell adhesion molecule. The RIG-3 protein is predicted to contain 2-3 immunoglobulin (Ig) domains and a diverged Fibronectin-type III (Fn-III) domain (see Figure 3.2). The RIG-3 IgCAM was an attractive candidate gene and here I describe my efforts to test the hypothesis that RIG-3 is required for normal connectivity between AVA and the A-class motor neurons. Phenotypic analysis reveals that ablation of the *rig-3* gene results in abnormal movement and defects in synaptic connectivity between AVA and the A-class motor neurons.

Materials and Methods

Nematode Strains

Nematodes were grown as described (Brenner 1974). The RB1712 strain containing the *rig-3(ok2156)* mutation was obtained from the *Caenorhabditis* Genetics Center (CGC). RB1712 was backcrossed 5 times to the standard N2 (Bristol) strain to remove potential background mutations. The *rig-3*::GFP strain OH4326 was obtained from the CGC. The spGFP strain containing wyEx1845[*unc-4*::*nlg-1*::spGFP1-10 (20ng/ml), *flp-18*::*nlg-1*::spGFP11 (30 ng/ml), *odr-1*::DsRed2 (50 ng/ml)] was obtained from Kang Shen's laboratory.

Microscopy

Videos of animal locomotion were recorded using an RGB video camera mounted to a Zeiss Stemi 2000 dissecting microscope. GFP expressing animals

were visualized by epifluorescence microscopy using a Zeiss Axiovert 200M compound microscope and confocal microscopy using a Zeiss LSM510. Images were captured with an ORCA ER CCD camera (Hamamatsu Corporation, Bridgewater, NJ).

Results

rig-3 mutants show a backward locomotion defect.

The *rig-3* gene encodes an IgCAM that is expressed in AVA and AVE, with no evidence for expression in other command interneurons. As shown in Figure 3.1, AVA is uniquely marked by co-expression of *rig-3*::GFP, which is expressed in AVA and a handful of additional head neurons and *glr-1*::DsRed2, which is expressed in all command interneurons. To explore the potential role of this cell adhesion molecule in AVA development or function, I obtained a deletion mutation, which ablates most of the *rig-3* coding region (Fig. 3.1). The *rig-3(ok2156)* mutants display an obvious defect in backward, but not forward locomotion (Figure 3.3). This observation is consistent with a defect in the ventral cord motor circuit that AVA regulates.

rig-3 mutants have a reduced number of synapses.

The ability of AVA to control backward locomotion depends on normal synaptic connections to the A-class motor neurons. By visualizing the chemical synapses formed between AVA and A-class neurons I can determine whether neuronal connectivity is normal. To test this idea, I used a synaptic marker designed to label specific synapses (Feinberg et al. 2008). This marker utilizes split GFP to localize specifically to AVA to A-class neuron synapses. In AVA, *nlg-*

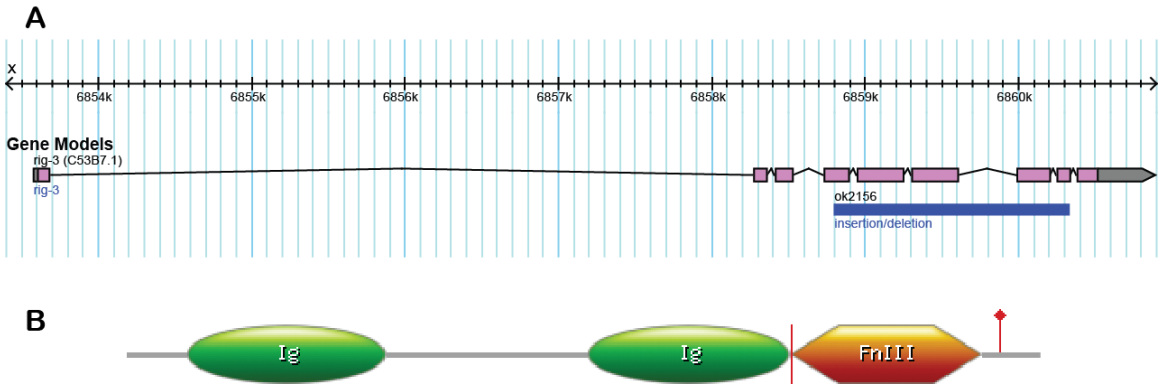


Figure 3.2. *rig-3* encodes an immunoglobulin-containing adhesion molecule. (A) Gene model of *rig-3*. The mutation of the *rig-3* gene is depicted with the blue bar representing the deletion spanning 5 exons. (B) Protein domain model of RIG-3. Two Ig domains are shown in green ellipses and a fibronectin type-III domain with the orange hexagon. RIG-3 is predicted to be GPI-anchored at amino acid 466 (red line with diamond marker).

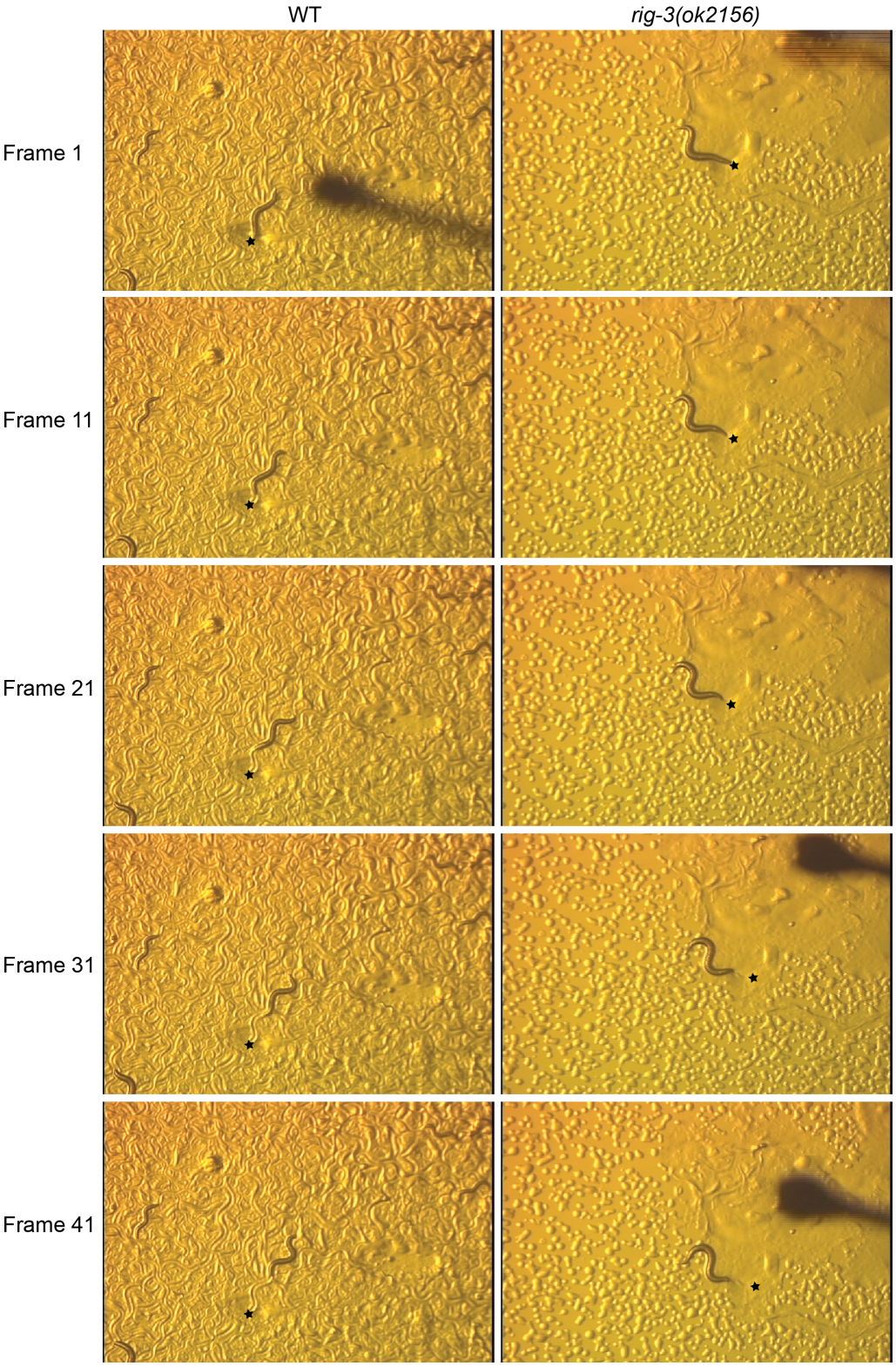


Figure 3.3. *rig-3(ok2156)* mutants display a backward movement phenotype. Over 40 frames, WT animals complete 2 backward body bends (left column). *rig-3* mutants fail to complete a backward body bend (right column). Beginning of backward locomotion is marked with a star.

1/Neuroligin is fused to one fragment of GFP and localizes to the presynaptic terminals of AVA. In the A-class motor neurons, *nlg-1* is fused to a complementary GFP fragment and is positioned at the postsynaptic membrane of the A-class neurons (NLG-1 localizes to both pre- and postsynaptic domains). The close proximity of the split GFP fragments at the AVA-A motor neuron synapses results in reconstitution of fluorescent GFP molecules. I crossed the split GFP (spGFP) marker into the *rig-3* deletion mutant and compared the number of fluorescent puncta to wildtype animals. As shown in Figure 3.4, *rig-3* mutants show significantly fewer fluorescent puncta than wildtype ($P < 0.05$, Wilcoxon Rank Sum test, $N > 20$). This result suggests that *rig-3* mutants have a synaptic connectivity defect, but leaves open the possibility for other explanations: AVA or A-class neurons could have a process placement defect, AVA or A-class neurons may have a synaptic organization defect, or *rig-3* could positively regulate the expression of either spGFP transgene driven by the *flp-18* and *unc-4* promoters.

rig-3 mutant AVA axons have minor guidance defects

To address the possibility of AVA axon guidance defects, I crossed the *rig-3*::GFP reporter into the *rig-3* mutant and then evaluated the GFP-marked AVA neuron for process placement defects in the ventral cord. AVA shows minor axon guidance defects in the *rig-3* mutant, which therefore suggests that the misplacement of AVA does not account for the loss of synapses (see Figure 3.5). My results corroborate earlier and an earlier finding that also detected a mild axon guidance phenotype for *rig-3* mutants that becomes more severe when

combined with other mutations in other adhesion molecules (Schwarz et al. 2009). This result suggests that the RIG-3 protein could exercise independent functions in synaptogenesis and axon guidance.

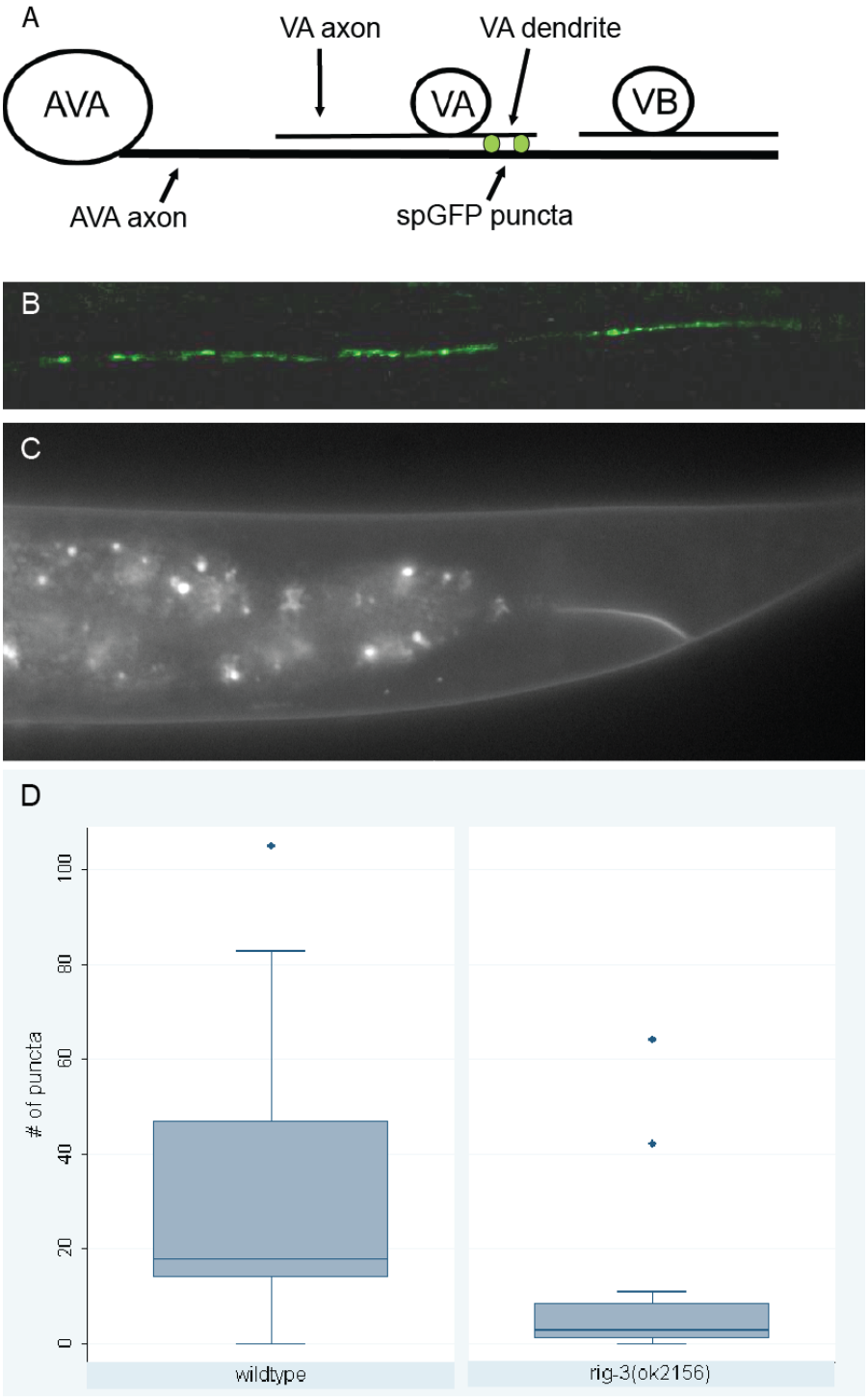


Figure 3.4 RIG-3 is required for AVA to A-class neuron synaptic connectivity. (A) Diagram of spGFP strain marking synapses between AVA and A-class neurons. (B) spGFP puncta in a wildtype animal (100x mag.) (C) spGFP puncta in a *rig-3(ok2156)* mutant (63x mag.). (D) Mean number of spGFP puncta for wildtype and *rig-3(ok2156)* mutant animals (Whiskers = lower/upper bounds of distribution, Box edges = 25th and 75th percentiles, *P* = 0.002, Wilcoxon Rank Sum test, N > 20 each group).

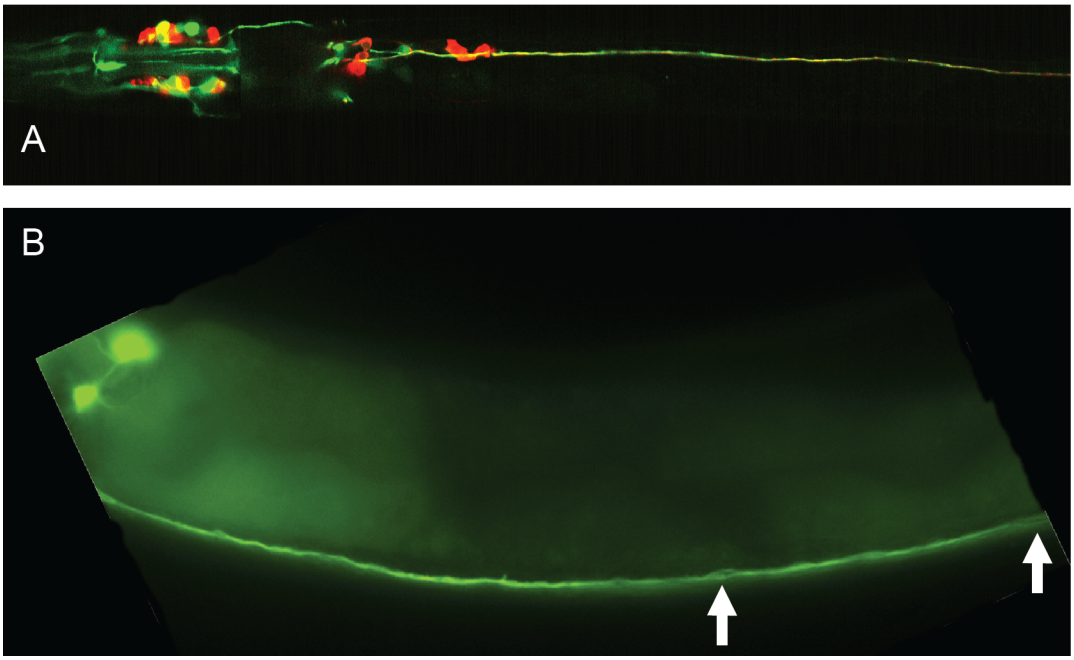


Figure 3.5 AVA axons show mild defects in *rig-3* mutants.
(A) P*rig-3*::GFP in WT (B) P*rig-3*::GFP in *rig-3(ok2156)*. Young adult animals are shown from head to near vulva, with anterior to the left, ventral view in A, and lateral view in B. White arrows mark minor AVA axon defasiculations in the ventral nerve cord.

Discussion

Studies on neuronal circuit formation have largely focused on axon guidance and synapse formation. Efforts to reveal how neurons identify the proper synaptic partners have been stymied by the need to reproducibly target synaptic partners for experimental analysis. In this study, I have taken advantage of the stereotyped connectivity in the *C. elegans* motor circuit and available molecular tools to test the role of the IgCAM, RIG-3, in synapse formation between the AVA command interneuron and the A-class motor neurons.

rig-3 mutants have an obvious backward movement phenotype implying a defect in the backward locomotory circuit. By using a fluorescent marker that specifically labels AVA to A-class motor neuron synapses, I have shown that *rig-3* is required for formation of these synapses. It is not yet clear whether RIG-3 is provides a permissive or an instructional cue. One experiment to address this question would be to cell-specifically express RIG-3 in a neuron that does not normally synapse with AVA and determine whether ectopic synapses are formed between the two neurons. It is also possible, that RIG-3 is necessary for assembly of all AVA synapses. This hypothesis can be tested by assaying synapses between AVA and other pre- and postsynaptic neurons, such as PVD and PVC.

A major prerequisite for experimentally testing synaptic specificity is the lack of an effect on axon guidance. If a molecule has dual roles in axon guidance and synaptic specificity, it would be difficult to determine whether the molecule is involved in specificity. Axons and dendrites must be adjacent for *en passant*

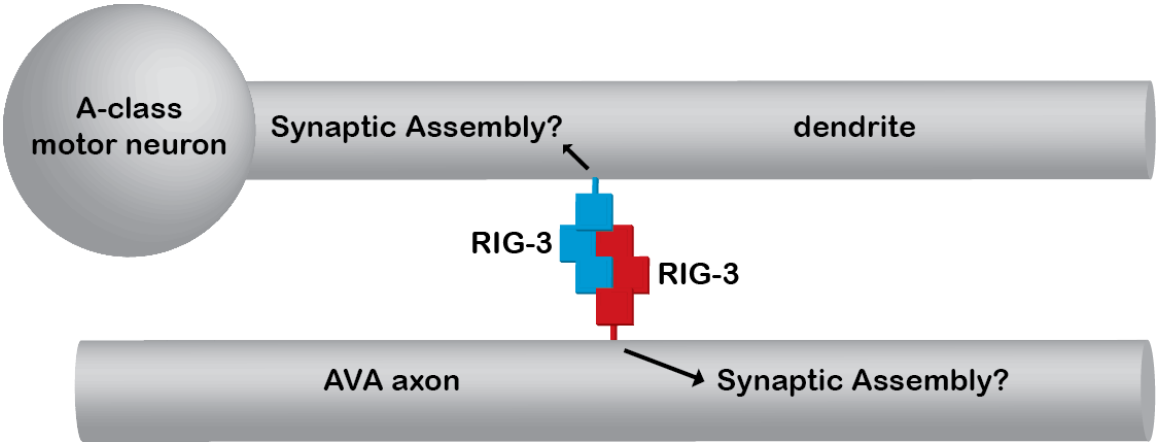


Figure 3.6 Model of RIG-3 mediated initiation of synaptogenesis between AVA and A-class motor neurons

synapse formation. Previous work and my results show that the axons of AVA are essentially wildtype with few deviations from the normal placement in the ventral nerve cord (Schwarz et al. 2009). Schwarz, et al., also show that *rig-3* mutants do not have significant DA motor neuron process defects. It is not yet known whether VA motor neuron dendrites are normal in *rig-3* mutant animals. By crossing an *unc-4*::mCherry expressing strain with the *rig-3*::GFP strain in the *rig-3* mutant background, would determine if AVA axons and A-class dendrites are directly adjacent.

rig-3 was known to be expressed in AVA and other microarray studies have shown that *rig-3* is also expressed in AVE and the A-class motor neurons (II ; Fox et al. 2005). This expression pattern is thus far exclusive to neurons that function in the backward-movement circuit. Cell-specific transcriptome profiling of AVB and PVC command interneurons and the B-class motor neurons will be necessary to determine whether *rig-3* is expressed in the forward motor circuit as well. If *rig-3* is exclusively expressed in the backward-movement circuit, then it is possible *rig-3* could provide the instructive cue for synapse formation between AVA, AVE, AVD and the A-class motor neurons. It would also be necessary to test whether synapses between AVB, PVC, and the B-class motor neurons are normal in *rig-3* mutant animals.

The closest homolog to RIG-3 is the *Drosophila* IgCAM *klingon* (Hutter et al. 2000; Teichmann and Chothia 2000). Both proteins are predicted to contain 2-3 Ig domains and a divergent Fibronectin type-III domain. This domain architecture is similar to mammalian NCAM (Teichmann and Chothia 2000).

NCAM has been extensively studied revealing a wide variety of roles in synapse formation, axon guidance, learning and memory (Rutishauser et al. 1985; Muller et al. 1996; Cremer et al. 1997; Dityatev et al. 2000; Eckhardt et al. 2000; Bukalo et al. 2004; Kleene and Schachner 2004). The roles of IgCAMs in the nervous system are likely conserved, thus future studies of RIG-3 function in the *C. elegans* motor circuit could provide clues on how IgCAMs contribute to generating the complex architecture of the brain.

CHAPTER IV

A SPATIAL AND TEMPORAL MAP OF *C. ELEGANS* GENE EXPRESSION

This chapter has been published by the journal ***Genome Research*** in the February 2011 issue (http://genome.cshlp.org/content/21/2/325.full). I am co-first author with Georg Zeller on the basis of my role in planning this project and paper and in producing and analyzing tiling array results. My work on the paper involved development of 2-color fluorescence-activated cell sorting for isolating single neurons, optimizing RNA purification from FACS isolated cells for NuGEN WT-PICO amplification, basic tiling microarray analysis, performing quality control on microarray data, microscopy of GFP reporters, RT-PCR analysis of novel transcripts, qRT-PCR analysis of novel transcripts, analyzing microarray results and advising Rebecca McWhirter and Kathie Watkins in cell culture, RNA isolation and RNA amplification procedures. Georg Zeller performed a majority of the data analysis assisted by Stefan Henz, Gunnar Rätsch, and Clay Spencer. Cell-specific plasmids and transgenic lines used for profiling were generated by Joseph Watson, Kathie Watkins, Steven Von Stetina, Sarah Anthony, and Clay Spencer. Rebecca McWhirter, Kathie Watkins and Joseph Watson produced most of the cell-specific data sets. Staged whole-animal tiling array data sets were produced by Jeanyoung Jo and Valerie Reinke. I worked closely with David Miller to plan and write this paper.

Introduction

The generation of specific cell types depends on spatial and temporal control of gene expression. The nematode *C. elegans* has been widely utilized to address this question because of its simple body plan and fully sequenced genome (Hillier et al. 2005). Although comprised of fewer than 1,000 somatic cells, the tissues of *C. elegans* adults include cell types characteristic of all metazoans such as muscle, nerve, intestine, skin, *etc* (Altun 2002-2010). Moreover, the developmental origin of each of these cells is fully described in a complete map of cell divisions from fertilized zygote to sexually mature adult (Sulston and Horvitz 1977; Sulston et al. 1983). The *C. elegans* genome sequence is also precisely defined, and at ~100 Mb is about 1/30 the size of the human genome (Hillier et al. 2005). However, with 20,168 predicted genes (http://wiki.wormbase.org/index.php/WS200), the *C. elegans* protein-coding genome is only slightly smaller than that of humans (http://www.sanger.ac.uk/PostGenomics/encode/stats.html). Major classes of non-coding RNAs (ncRNAs) such as microRNAs (miRNAs) are also represented in *C. elegans* (Ruby et al. 2006; Kato et al. 2009). Thus, *C. elegans* provides a simple but representative model of development that depends on differential expression of a compact, well-described genome. Although *C. elegans* is completely sequenced, some predicted genes lack direct evidence of transcription and other cryptic protein-coding genes and ncRNAs are likely to have been overlooked by gene prediction software (Hillier et al. 2009; Schweikert et al. 2009). In addition, the cell-specific expression patterns of the majority of *C.*

elegans genes are unknown. Thus, the anatomy and development of the animal is defined at the resolution of the single cell but a comparably precise atlas of gene expression is not currently available.

The goal of a comprehensive gene expression map has been achieved in part by analysis of promoter::GFP fusions for a broad array of protein coding genes (Dupuy et al. 2007; Hunt-Newbury et al. 2007; Murray et al. 2008; Liu et al. 2009). This methodology, however, is generally not quantitative and can be misleading if key regulatory elements are omitted from the reporter genes (Hunt-Newbury et al. 2007). We have adopted the alternative strategy of measuring native transcripts from a broad array of specific tissues and cell-types. In addition, we used whole genome tiling arrays in order to sample the entire non-repetitive genome and therefore achieve an unbiased approach to transcript discovery. In addition to assigning gene expression to identified tissues and stages, our approach of analyzing different cell types and developmental periods also ensures detection of RNAs that may be selectively expressed during discrete temporal intervals or in limited numbers of cells. We accomplished this goal by utilizing recently developed methods for obtaining RNA from specific *C. elegans* cells (Roy et al. 2002; Zhang et al. 2002; Fox et al. 2005). Altogether, we sampled 13 embryonic cell types and 12 larval and adult tissues. We also produced tiling array data sets for whole animal RNA isolated from seven different developmental stages. Additional profiling results were obtained from larval males and from the hermaphrodite gonad and soma. Thus, our datasets significantly enhance a growing body of tissue and stage specific gene

expression for *C. elegans* (McKay et al. 2003; Pauli et al. 2005; Von Stetina et al. 2007; Meissner et al. 2009). Our results indicate that most protein coding genes (~75%) are differentially expressed among the stages and cell types that we sampled. In addition to providing evidence of extensive gene regulation, these results should also greatly aid genetic analysis by suggesting cell types or developmental stages in which highly expressed transcripts are likely to function. For example, our results provide the first comprehensive description of gene expression in *C. elegans* primordial germ cells and led to the discovery that proteins encoded by a subset of these genes are expressed well before their established roles in meiosis and oogenesis. To identify novel transcripts, we utilized a recently developed computational method for recognizing transcribed regions irrespective of their annotation status (Laubinger et al. 2008). This approach revealed a large number of previously unannotated transcripts encoded by at least 10% of the *C. elegans* genome. These novel transcripts show striking cell specificity that may be indicative of tissue-specific functions. To facilitate the use of these data for future studies of gene function, we provide online resources for visualizing transcribed regions in a genome browser and for estimating relative gene expression levels across tissue types and developmental stages.

Methods and Materials

Detailed methods for cell culture, FACS, amplification, microarray hybridization, probe mapping and annotation, expression above background and differential expression analysis are described in Chapter II.

Construction of cell-specific 3XFLAG::PAB-1 plasmids

A Gateway (Invitrogen) compatible mRNA-tagging vector, *pSV41* (Pgateway::3XFLAG::PAB-1 + *unc-119* minigene) was constructed to provide a convenient method for inserting cell-specific promoters and for generating transgenic lines by bombardment. The *unc-119* minigene plasmid, MM051 (Maduro and Pilgrim 1995), was digested with HindIII, blunted with T4 DNA polymerase, digested with BamHI and the resulting fragment subcloned into plasmid pSV15 which contains the 3XFLAG::PAB-1 insert (Von Stetina et al. 2007), using BamHI and EcoRV restriction sites. The resulting plasmid was digested with KpnI then treated with T4 DNA polymerase for blunt end ligation with the Gateway vector conversion fragment A (Invitrogen). The resulting plasmid (pSV41) contains the *unc-119* minigene in the opposite orientation *vs*. promoter sequences inserted between the attR1 and attR2 sites upstream of the 3XFLAG::PAB-1 coding region.

Constructs generated using Gateway LR recombination with pSV41 plasmid

Cell-specific promoter fragments were generated from genomic DNA for *unc-122* (coelomocytes), *dpy-7* (hypodermis), *glr-1* (*glr-1*-expressing neurons) and subcloned into pCR8/GW-TOPO (Invitrogen). PCR amplicons and primer pairs were: *unc-122* (800 bp, *unc-122*_5prime/*unc-122*_3p); *dpy-7* (354 bp, *dpy-7*_5p/*dpy-7*_3p); *glr-1* (5.3 kb, *glr-1*_5p/*glr-1*_3p). The 716 bp *dat-1* (dopaminergic neurons) promoter was PCR amplified from plasmid pRN200, (a gift from R. Blakely), using primers *dat-1*p1 and *dat-1*p2 and cloned into pCR8/GW-TOPO. The *hlh-17* (CEP sheath cell) promoter was generated by

amplifying the 4 kb promoter sequence upstream of the first ATG start (McMiller and Johnson 2005) with primers containing flanking attB recombination sites. This fragment was subcloned into pDONOR221 (Invitrogen) by a BP recombination reaction. LR recombination reactions were performed using pSV41 as the destination vector to create the following expression plasmids: pJW7 (P*glr-1*::3XFLAG::PAB-1), pJW5 (P*unc-122*::3XFLAG::PAB-1), pJW8 (P*dpy-7*::3XFLAG::PAB-1), pKW63 (P*dat-1*::3XFLAG::PAB-1), pMK107L (P*hlh-17*::3XFLAG::PAB-1).

Generating cell-specific::3XFLAG::PAB-1 strains by microparticle bombardment

Microparticle bombardment was used as previously described (Fox 2005) to generate transgenic lines from plasmids containing the *unc-119* minigene. Additional modifications were used for plasmids pJW5, pJW7, pJW8 and pKW63, which were linearized by digesting with a unique ApaI restriction site upstream of the *unc-119* + minigene cassette. The reaction was then ethanol-precipitated and re-suspended in ddH2O. 8-10 µg of linearized plasmid was used to coat gold beads for bombardment. Animals were bombarded at 1800 psi, allowed to recover for 1 hr and washed to 7 x 100 mm NGM plates seeded with OP50-1 bacteria. Plates were allowed to starve for 2 weeks at 23-25 °C and viable animals showing wildtype movement were picked for selfing. Transgenic lines were screened by anti-FLAG immunostaining (Von Stetina et al. 2007) to confirm specific expression.

Other constructs generated for cell-specific profiling

The excretory cell-specific promoter P*clh-4* was amplified from genomic DNA using primers *clh-4* F and *clh-4* R. The 4 kb PCR product was then cloned into TOPO-2.1 (Invitrogen) to generate pDM1. pDM1 was used as a template to construct a Gateway donor vector by PCR amplification of the *clh-4* promoter using *clh-4* primers flanked with attB1 and attB2 sites. The promoter fragment was subcloned into pDONOR221 by performing a BP recombination reaction to create pDM2. pDM2 was combined with destination vector pSV41 in a LR recombination reaction creating the expression vector pJW6 (P*clh-4*::3XFLAG::PAB-1). The P*clh-4*::3XFLAG::PAB-1 cassette was then PCR amplified and 6ul of PCR product was coinjected with pRF4 [*rol-6 (su1006)*] at 25 ng into wild type animals. The transgenic line was integrated by gamma irradiation and outcrossed five times. The 861 bp putative promoter of *ttr-39* was amplified via PCR with primers pC04G21_5 and pC04G21_3 and inserted into pENTR-D-TOPO (Invitrogen) via TOPO TA reaction. *Pttr-39* was then inserted upstream of 3XFLAG::PAB-1 via Gateway LR reaction with pSV41 resulting in the expression vector pSA2. The P*ttr-39*::3XFLAG::PAB-1 cassette was then amplified via PCR from pSA2 with primers pC04G21_5 and PAB1UTR_3 (5' CAATAGCAGCCAAATGCA 3'). The PCR reaction (12 l) was co-injected with *dpy-5* rescuing plasmid pCes361 (25ng) into *dpy-5(e907)* animals. Gamma irradiation of the transgenic line yielded NC1645 *dpy-5(e907); wdIs31[Pttr-39::3xFLAG::PAB-1 dpy-5(+)]* IV. The integrant was outcrossed five times prior to microarray profiling. Expression of the epitope-tagged PAB-1 for both the

excretory cell and D-class motor neuron cell-specific lines was confirmed by immunostaining (Roy et al. 2002) with monoclonal mouse anti-FLAG antibodies (Sigma).

Isolation of cell-specific RNA by the mRNA tagging method

The mRNA-tagging method was used to isolate RNA from 12 different cell types in either larvae or young adults. Methods for obtaining RNA from L2 stage larvae were as previously described (Von Stetina et al. 2007). The following modifications were used for L4 stage larvae and young adults. Gravid adults were obtained from 20 x 150 mm culture plates (8P media, 8X peptone NGM) and treated with hypochlorite to release embryos. Arrested L1 larvae were isolated after hatching overnight at 20 °C in M9 buffer and transferred to Na22 seeded 8P plates for growth at 20 °C for 22-25hrs and then transferred to 23 °C for an additional 24-26hrs to reach mid-L4 stage larvae as shown by the appearance of a tree-shaped vulva (~80%). To obtain Young Adults (YAs), the arrested L1 larvae were grown on Na^{22}-seeded 8P plates at 20 °C for ~72hrs to reach early YA, as evidenced by a mature (everted) vulva in ~80% of animals. Synchronized L4 and YA animals were resuspended in 3 ml homogenization buffer and passed through a French press four times at 6,000 psi to obtain lysate as opposed to three times for L2 larvae. Mock IPs were performed to obtain reference data sets of non-specifically bound RNA for synchronized populations of L2, L4 and YA animals (Von Stetina et al. 2007). At least 3 independent RNA samples were prepared for each cell type and for each of the reference data sets.

RT-PCR to detect novel RNA

Single-stranded cDNA previously generated for microarray analysis was used as template for PCR-based validation of novel TARs. The −RT L2-intestine sample used the same RNA input for amplification, but reverse transcriptase was omitted and dH_2O was added to maintain constant volume. Primers (Table 4.3) were designed to generate small amplicons of 75-150 bp using Batch-Primer3 (You et al. 2008). PCR conditions are as follows: 4 ng ss cDNA, 500 nM each primer, 1.5 µM MgCl2, 2.5 U GoTaq polymerase (Promega), and 200 nM dNTPs in a 50 µl reaction. The reactions were run in a MJ Research Minicycler with the following program: 94 °C 30 sec, 35 cycles of 94 °C sec, 58 °C 30 sec, 72 °C 30 sec. PCR products were electrophoresed on a 2% agarose gel and stained with ethidium bromide (Sigma). The products were visualized with a Bio-rad Gel Doc.

Quantitative PCR validation of novel TAR differential expression

Quantitative PCR (qPCR) was performed on ss cDNA used for microarray analysis. Primers (Table 4.3) were designed to generate amplicons of 75 to 150 bp using Batch-Primer3 (You et al. 2008). Ssofast Eva green reaction mix was used with a 2-step 98 °C 2 sec, 60 °C 5 sec reaction and melting curve on a CFX96 Real Time Thermal Cycler (Bio-rad). Data were normalized to an internal 26S rRNA control using the Pfaffl method of determining relative expression (Pfaffl 2001).

***De novo* transcript identification using mSTAD**

For *de novo* identification of transcriptionally active regions (TARs) we adopted mSTAD, a previously proposed machine-learning based method

(Laubinger et al. 2008; Zeller et al. 2008). For the analysis of cell type samples, a separate mSTAD model was optimized for each developmental stage by training on corresponding reference hybridization data and annotation information belonging to chromosomal chunks, each of which contained one annotated gene with half the intergenic space surrounding it (see Table 4.1). The fitted models were used for transcript identification from all samples belonging to the same developmental stage (*e.g.*, the mSTAD model trained on EE-ref was used for transcript identification in EE-ref, EE BAG neurons, and EE germline precursors).

Table 4.3 Primers used for reverse-transcriptase PCR of novel TARs and real-time PCR validation of differentially expressed novel TARs

Name	length	Tm	GC%	sequence	Amplicon length
TAR_E-pan_77592_F	20	60.23	55	TTCCTCTGGAACTGGACAGG	104
TAR_E-pan_77592_R	20	59.35	55	CCCTGAGCTTTCCACGTAGT	
TAR_E-pan_77593_F	20	59.66	45	CACCCCAAAAATACCTGGAA	131
TAR_E-pan_77593_R	20	59.95	40	TTGATTGCGATGAAAAGCAG	
TAR_E-bwm_63930_F	20	59.94	45	ATCATCCCAAACGCTTTCAC	123
TAR_E-bwm_63930_R	20	58.88	50	TTTCCACTATGCAGCTGACC	
TAR_E-coel_06773_F	20	59.8	45	AAGAGGGTCCAACCGAATTT	121
TAR_E-coel_06773_R	20	59.96	50	CCGGGACTGTGCAAGATAAT	
TAR_L2-int_19313_F	20	59.65	45	GCCGAGATTGAGGAAAAATG	112
TAR_L2-int_19313_R	21	58.78	48	CCGGTACTTATTCGTTTGCTC	
TAR_E-AVE_07153_F	20	59.8	45	AAGAGGGTCCAACCGAATTT	121
TAR_E-AVE_07153_R	20	59.96	50	CCGGGACTGTGCAAGATAAT	
TAR_E-AVE_52539_F	20	59.14	50	GGCTGGTTCTGAAGTCCAAT	137
TAR_E-AVE_52539_R	20	59.88	45	GTGTTGCAGGTTGGGTTTTT	
TAR_L3-L4-dop_35596_F	21	60.13	42.86	TTGAACCCGAAAAAGTGTCTG	97
TAR_L3-L4-dop_35596_R	21	59.27	47.62	TGGAGTCAAGGATTCTGAAGG	
TAR_L3-L4-hypo_34173_F	21	60.13	42.86	TTGAACCCGAAAAAGTGTCTG	97
TAR_L3-L4-hypo_34173_R	21	59.27	47.62	TGGAGTCAAGGATTCTGAAGG	
TAR_L3-L4-hypo_36011_F	20	60.86	45	CACATTGAGCGGGAAATGAT	130
TAR_L3-L4-hypo_36011_R	21	58.46	47.62	TTCTCTTCGGAGATGTTCCTC	
TAR_YA-CEPsh_52288_F	20	60.31	40	ACGTTCCAATCGGAATTCAA	125
TAR_YA-CEPsh_52288_R	20	59.14	45	AGACCACCAGCATGTTCAAA	
TAR_L2-exc-cell_23646_F	20	60.05	40	TCAAATGTGCCCAATGAGAA	115
TAR_L2-exc-cell_23646_R	20	58.49	45	GACCGATTCATGGAAGTTCA	
TAR_L2-exc-cell_40020_F	20	60.28	45	TTTGTGTGTGGCAAGAGGAA	128
TAR_L2-exc-cell_40020_R	20	60.58	50	TGGTCGTACCCCAAATATCG	
TAR_L2-glr_26400_F	20	60.06	50	AGTGTCAACAGCTGCAATCG	134
TAR_L2-glr_26400_R	20	59.99	60	CCAGTCCTCTGCCTGTCTTC	
TAR_L2-A-class_72252_F	20	59.66	55	GCTTCTGGTCCATCCAAGAC	131
TAR_L2-A-class_72252_R	20	60.48	55	GGCACCAGGATAATCTCACG	

TAR_E-hypo_29647_F	24	58.90	45.83	CACTGGTGTAGAAGAACAAGAG GT	94
TAR_E-hypo_29647_R	20	59.17	45	AGGTCGTGCATTTTTCCTTC	

Array data from developmental stages was analyzed with another set of models, each of which was trained on the same array sample for which it identified transcripts (see Table 4.1). One more model was trained for hermaphrodite gonads and TARs for L4 males and L4 hermaphrodite soma were identified with the L4 mSTAD model (see Table 4.1).

Determining overlap between TARs and known genes to identify novel TARs

TARs were compared to the following features annotated in WS199: Protein-coding genes (and their corresponding exon features), pseudogenes (and pseudogenic exons) and non-protein coding genes. We called TARs "unannotated" if they overlapped less than 20 nt with exons (of coding genes and pseudogenes) or with non-coding genes (Figure 4.3D, Figure 4.4). Moreover, we determined the overlap between TARs and genes of the integrated transcript model to obtain "novel" TARs that neither overlapped with annotated features nor with exons of gold standard gene models by ≥20 nt (Figure 4.3E, F).

Non-redundant (nr) TARs resulted from the union of positions inclusive to TARs obtained in any individual sample. Similarly, nr expressed TARs, nr differentially expressed TARs, nr unannotated TARs and nr novel TARs were obtained as the position-wise union of expressed, differentially expressed, unannotated and novel TARs, respectively (Figure 4.3F). For each position within nr expressed TARs, we counted the number of individual samples in which an expressed TAR was detected. Partitioning expressed nrTARs according to overlap with known transcripts resulted in the histograms shown (Figure 4.8C).

Additional correction for multiple testing of (differentially) expressed genes

Since FDRs for genes expressed above background or differentially expressed between samples were calculated for each individual sample (comparison), we applied an additional stringent Bonferroni-style correction for multiple testing (Bonferroni 1935). We divided individual FDR estimates by the number of samples and sample comparisons, respectively, obtaining an adjusted FDR of 1.3×10^{-4} for expression above background and of 7.4×10^{-4} for differential expression (Table 4.2).

Entropy-based detection of selectively enriched genes

Gene-expression entropy was calculated based on the fold change relative to the corresponding reference sample. Fold-changes < $1.0e^{-5}$ were set to a pseudocount of $1.0e^{-5}$ before they were rescaled to the interval [0, 1] by dividing by the sum of fold changes across tissues and cell types for each gene. Afterwards, expression entropy was calculated as described (Schug et al. 2005). Selectively enriched genes were extracted from the set of enriched genes in a given tissue (FDR ≤ 0.05 and FC ≥ 2.0), if additionally (i) their fold change vs. reference was among the upper 40% of the positive FC range observed for this gene across all tissues and (ii) their entropy was among the lower 40% of the distribution observed for all genes (i.e., H ≤ 3.03).

Fold change histograms for differentially expressed genes.

To generate the histograms of expression differences (Figure 4.7A), we first calculated the fold change between expression in a given cell type to the corresponding reference for all genes for which differential expression was

detected in at least one comparison (FDR 0.05). We next determined the maximal fold change across cell types and depending on its direction tabulated the gene either as upregulated or downregulated (relative to reference).

Revealing developmental and cell-type specific expression patterns with self-organizing maps

Self-organizing maps (SOMs)(Kohonen 1982) were constructed using the Matlab SOM toolbox version 2.0 (Vesanto et al. 2000). As an input for SOM training, we selected the subset of genes detected as differentially expressed in the respective samples, applied $\log_2$ transformation and normalized by subtracting the mean expression across conditions for each individual gene (yielding mean-centered log expression). We chose SOM topologies with a hexagonal neighborhood consisting of 30 x 15 and 60 x 60 units for developmental and cell-type data sets, respectively. SOM training proceeded in 100 and 300 epochs with Gaussian neighborhood radius shrinking linearly from 5 to 1 and from 15 to 3 for developmental and cell type data sets, respectively.

Some regions were identified by *k*-means clustering as implemented in the Matlab SOM toolbox. We varied *k*, the pre-chosen number of clusters, from 1 to 15 and 1 to 20 for developmental and cell type data sets, respectively. To obtain a robust clustering, we only retained cluster information that was consistent in 75% of 50 - 100 replicates each of which resulted from the best out of five independent *k*-means runs with randomly initialized cluster centroids. We selected $k = 8$ and $k = 14$ for the developmental and the cell type data set, respectively, based on biological interpretation and silhouette coefficients, a means of assessing which SOM units lie tightly within clusters or which are in

between clusters (Rousseeuw 1987). In addition, we used silhouette coefficients to select the top half of SOM units close to cluster centroids, which resulted in the clusterings shown (Figure 4.10, Figure 4.11).

Hypergeometric tests

To test for significant overlap between separate lists of genes, we applied the hypergeometric test as implemented by Jim Lund (http://elegans.uky.edu/MA/progs/overlap_stats.html). The number of genes in the whole genome was set at 18,451 for the number of genes represented on the *C. elegans* tiling array.

Microscopy

Isolated embryonic cells were imaged using differential interference contrast (DIC) and epifluorescence optics with a Zeiss Axiovert inverted microscope equipped with an ORCA ER (Hamamatsu) high-resolution, cooled CCD camera. Intact animals were imaged with a Zeiss Axioplan compound microscope equipped with an ORCA ER camera and a Leica TCS SP5 confocal microscope.

Results

Strategies for profiling specific cell types and developmental stages

Cell-specific RNA was obtained from GFP-labeled embryonic cells isolated by FACS and from larval cells by use of the mRNA-tagging method (Figure 4.1A, B). Altogether, we generated tiling array profiles from 25 different tissues, with each sample derived from one of five distinct developmental stages. Corresponding reference data sets were collected from all cells for each of these

developmental periods (Table 4.1, see Methods). We also generated an independent developmental series with total RNA isolated from whole animals at seven different ages (EE, LE, L1, L2, L3, L4, YA) (Table 4.1, Figure 4.1). An additional group of tiling array profiles was obtained from young adult hermaphrodite gonads, L4 hermaphrodite somatic cells, and L4 males. Because our samples were isolated by different methods, which could potentially result in biased representation, we used principal component analysis (PCA) to compare tiling array results (see below) obtained from specific cells and from whole animals. PCA shows that tiling array profiles obtained from whole embryos cluster with data sets generated from specific embryonic cells and that larval and adult profiles are grouped with data sets obtained from specific postembryonic and young adult tissues (Figure 4.2). Correlation analysis comparing cell type data sets with developmental series data sets also confirms that expression estimates derived from cell types generally correlate well with the corresponding developmental stage data set generated from total RNA. Thus, a global analysis of our tiling array results suggests that cell-specific profiling preserves overall patterns of temporally regulated gene expression.

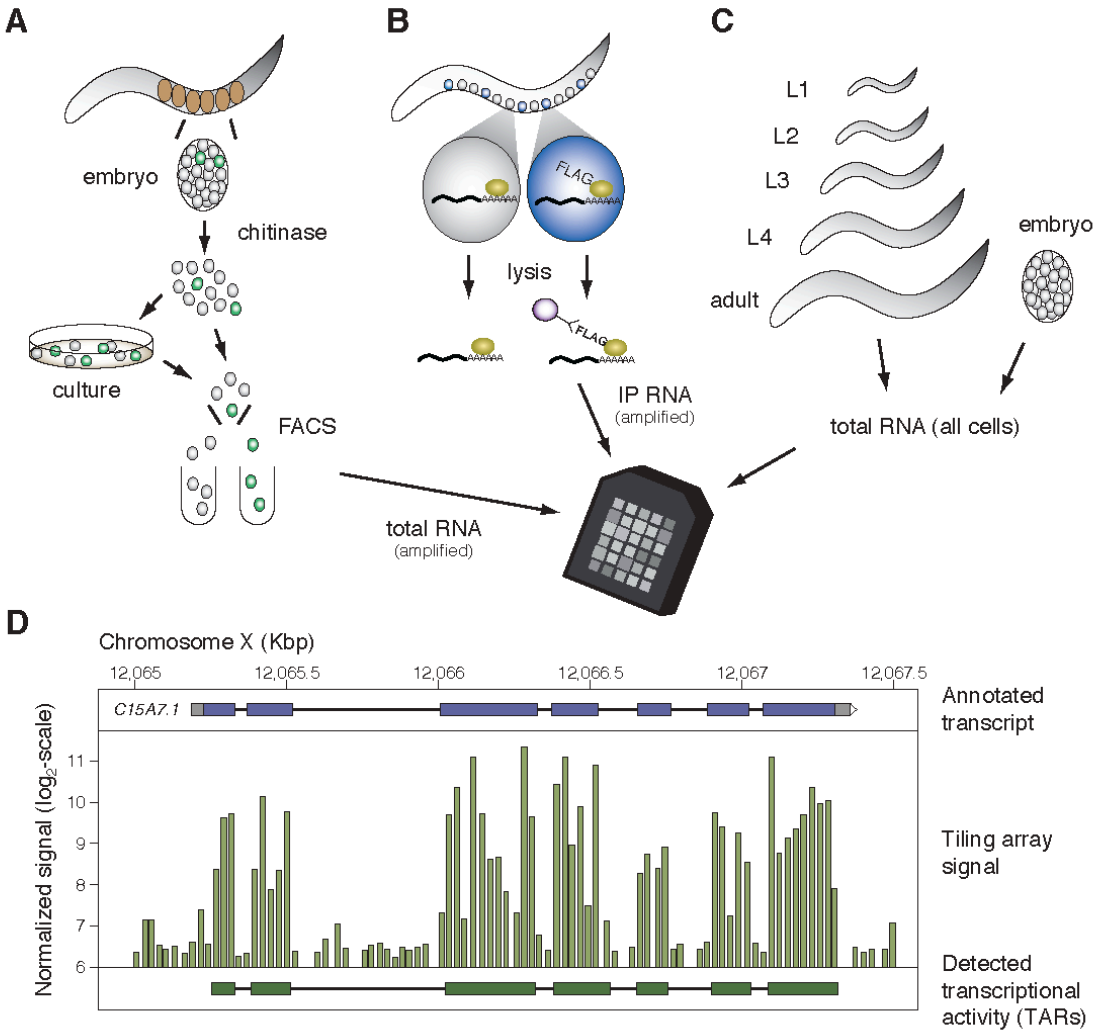


Figure 4.1 Strategies for generating tiling array data sets from specific *C. elegans* cells in embryos and larvae and from whole animals at defined developmental stages.
(A) In the MAPCeL (Micro-array Profiling of *C. elegans* Cells) method, embryos are isolated from gravid adults and blastomeres released by treatment with chitinase. Dissociated embryonic cells are either sorted immediately or cultured for 24 hrs before FACS. Total RNA is amplified for tiling array analysis. (B) The mRNA-tagging strategy was used to isolate RNA from specific larval and adult cells. The epitope-tagged (FLAG) polyA-binding protein (PAB-1) is expressed under the control of cell-specific promoters. The PAB-1:RNA complex is immunoprecipitated and RNA is amplified for tiling array analysis. (C) Total RNA is isolated from synchronized populations of embryonic, larval and adult animals for tiling array analysis. (D) Tiling array data (middle) is shown in a region around the annotated transcript *C15A7.1* (top). Each vertical bar corresponds to the signal of one probe feature. A transcript identified by mSTAD using only the tiling array signal is shown at the bottom.

Table 4.1 Samples used for expression profiling.

Sample Cell types	Stage	Genotype	Description	FACS Cell Purity	DCC #	GEO #	RNA
emb-0hr-ref	embryo 0hr	N2	all viable freshly dissociated (0 hr) embryonic cells	100%	3172	GSE25350	total RNA
emb-BAG	embryo 0hr	nIs242[gcy-33::GFP];lin-15(n765)	embryonic BAG neurons	>82%	2499	GSE23769	total RNA
emb-GLP	embryo 0hr	bnIs1 (pie-1p8::GFP::PGL-1+unc119)	embryonic germ-line precursor cells	>95%	661	GSE23285	total RNA
emb-reference	embryo 24hr	N2	all viable cultured (24 hr) embryonic cells	100%	456	GSE23246	total RNA
emb-AVA	embryo 24hr	OH4326[otEx239(rig-3::GFP) pha-1(e2123)III] VH804[hdIs32(glr-1::DsRed2)	embryonic AVA neurons	>80%	459	GSE23249	total RNA
emb-GABA	embryo 24hr	CZ1200[juIs76(unc-25::GFP) II; lin-15(n765ts)X]	embryonic GABAergic motor neurons	>87%	468	GSE23257	total RNA
emb-bwm-v2	embryo 24hr	ccIs4251 [I;dpy-20(e1282) IV]	embryonic body wall muscle	>97%	470	GSE23260	total RNA
emb-coelomocytes	embryo 24hr	wyIs58 (opt-3::GFP::RAB-3; unc-122::RFP)	embryonic coelomocytes	nd	458	GSE23248	total RNA
emb-dop	embryo 24hr	dat-1::GFP (pRN2003)	embryonic dopaminergic motor neurons	>86%	467	GSE23257	total RNA
emb-intestine	embryo 24hr	wIs84	embryonic intestine	>91%	457	GSE23247	total RNA
emb-panneural	embryo 24hr	evIs111	embryonic neurons	>90%	455	GSE23245	total RNA
emb-A-class	embryo 24hr	wdIs5[unc-4::GFP; dpy-20(e1282)]	embryonic A-class motor neurons	>88%	654	GSE23278	total RNA
emb-hypodermis	embryo 24hr	/+; raIs/[rol-6(SU1006)+pdpy-7::GFP]	embryonic hypodermal cells	>85%	662	GSE23286	total RNA
emb-AVE	embryo 24hr	KM173 (opt-3::GFP[pRF4]); hdIs32 (glr-1::DsRed2)	embryonic AVE neurons	>88%	3173	GSE25351	total RNA
emb-PhM	embryo 24hr	ccIs9753[myo-2::GFP]	embryonic pharyngeal muscle	>91%	2548	GSE23770	total RNA
L2-glr	L2	unc-119 (ed1); [unc-119 (+); glr-1::3XFLAG::PAB-1]	L2 glutamate receptor neurons	na	658	GSE23282	poly A+ / total RNA
L2-A-class	L2	unc-119 (ed1); wdEx257 [unc-119 (+); unc-4::3XFLAG::PAB-1]	L2 A-class motor neurons	na	469	GSE23259	poly A+ / total RNA
L2-GABA_neurons	L2	dpy-5 (e907); wdIs31 [dpy-5 (+); pC04G2.1::3XFLAG::PAB-1]	L2 GABA neurons	na	466	GSE23256	poly A+ / total RNA
L2-bwm	L2	gaIs146 [(myo-3p::FLAG::PAB-1) +	L2 body wall muscle	na	465	GSE23255	poly A+ / total

Table 4.1 Samples used for expression profiling.

Sample	Stage	Genotype	Description	FACS Cell Purity	DCC #	GEO #	RNA
		(sur-5::GFP)]					RNA
L2-excretory_cell	L2	wdIs47 [clh-4::3XFLAG::PAB-1 + rol-6 (su1006)]	L2 excretory cell	na	464	GSE23254	poly A+ / total RNA
L2-intestine	L2	gaIs148 [(ges-1p::FLAG::PAB-1) +(sur-5::GFP)]	L2 intestine	na	463	GSE23253	poly A+ / total RNA
L2-panneural	L2	gaIs153 [(F25B3.3::FLAG::PAB-1) + (sur-5::GFP)]	L2 neurons	na	462	GSE23252	poly A+ / total RNA
L2-coelomocytes	L2	unc-119(ed1); wdEx638 [unc-119(+); unc-122::3XFLAG::PAB-1]	L2 coelomocytes	na	657	GSE23281	poly A+ / total RNA
L2-reference	L2	N2	mock-IP from L2 stage animals	na	461	GSE23251	poly A+ / total RNA
L3-L4-PVD_OLL	L3-L4	unc-119 (ed1); wdEx460 [unc-119 (+); ser-2prom3B::3XFLAG::PAB-1]	L3-L4 PVD and OLL neurons	na	460	GSE23250	poly A+ / total RNA
L3-L4-dop	L3-L4	unc-119 (ed1); wdEx637 [unc-119 (+); dat-1::3XFLAG::PAB-1]	L3-L4 dopaminergic neurons	na	655	GSE23279	poly A+ / total RNA
L3-L4-reference	L3-L4	N2	mock-IP from L3-L4 stage animals	na	659	GSE23283	poly A+ / total RNA
L3-L4-hypodermis	L3-L4	unc-119(ed1); wdEx626[unc-119+; dpy-7::3xFLAG::PAB-1]	L3-L4 hypodermis	na	2454	GSE23287	poly A+ / total RNA
YA-CEPsh	YA	unc-119 (?); nsIs191 [unc-119 (+); hlh-17::3XFLAG::PAB-1]	Young adult CEP sheath cells	na	660	GSE23284	poly A+ / total RNA
YA-ref	YA	N2	Mock-IP from young adult stage animals	na	656	GSE23280	poly A+ / total RNA
Gonad	YA	N2	Dissected gonad from YA hermaphrodite	na	481	GSE23269	total RNA
Whole Animal							
N2EE	early embryo	N2	Early embryos	na	476	GSE23265	total RNA
N2LE	late embryo	N2	Late embryos	na	479	GSE23268	total RNA
L1	L1	N2	L1 animals	na	484	GSE23270	total RNA
L2	L2	N2	L2 animals	na	472	GSE23261	total RNA
L3	L3	N2	L3 animals	na	474	GSE23263	total RNA
L4	L4	N2	L4 animals	na	473	GSE23262	total RNA

Table 4.1 Samples used for expression profiling.

Sample	Stage	Genotype	Description	FACS Cell Purity	DCC #	GEO #	RNA
YA	YA	N2	Young adult animals	na	475	GSE23264	total RNA
soma-only	L4	glp-1(q224)	L4 somatic cells only	na	485	GSE23271	total RNA
male	L4	dpy-28(y1) III; him-8(e1489) IV	L4 males	na	478	GSE23267	total RNA

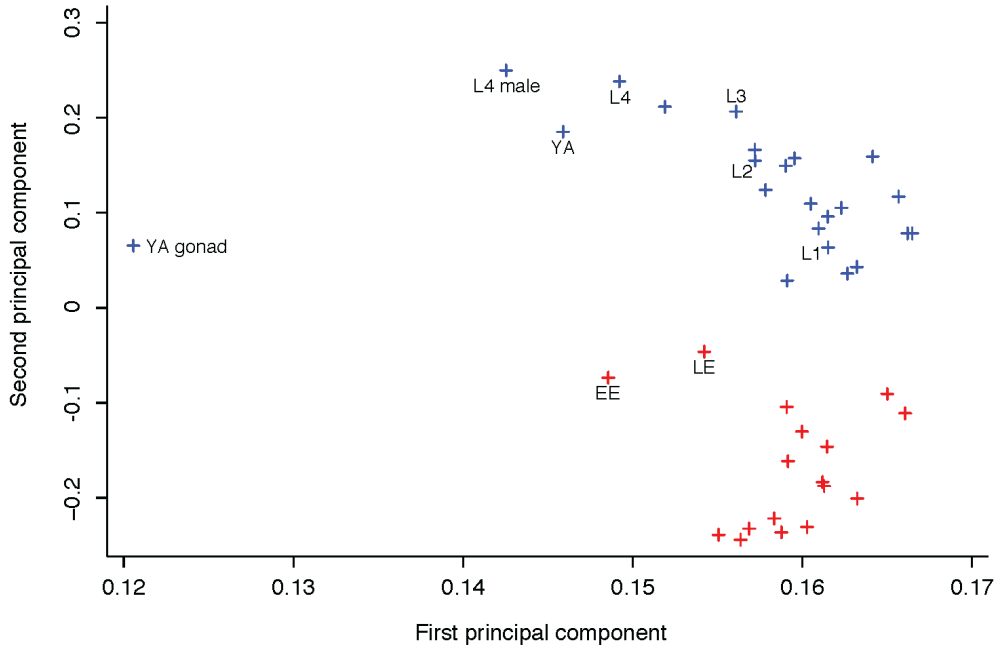


Figure 4.2 Principal component analysis of expression estimates shows agreement in clustering between cell type and developmental stage data. Principal component 1 identifies the striking difference between the gonad and all other profiles. Germ line contribution to other stages (EE, LE, L4, L4 males) separates those profiles from other data sets (*e.g.,* early larval stages L1, L2, L3) along the X-axis. Principal component 2 separates data sets on the Y-axis based on embryonic or postembryonic stages. All embryonic stages are colored red and postembryonic stages are blue.

Expression of annotated genes detected with tiling arrays

To evaluate expression of annotated protein-coding genes, we created probe sets corresponding to the constitutive exons of individual gene models annotated in WormBase and summarized the intensity values for each gene (see Methods). We then identified transcripts that are detectable above background in each sample with a statistical test (see Methods). The union of results derived from all cell types and stages detected a total of 17,452 genes (Table 4.2). Because these results were obtained from multiple independent comparisons, we conservatively adjusted the FDRs to limit the potential accumulation of false positives (see Methods), which resulted in expression detected for 13,149 genes from the union of cell type-specific data sets and for 13,713 genes in at least one of the stage-specific data sets. When both groups of data sets were combined, we detected 14,279 expressed genes (Table 4.2).

Our initial analysis identified an average of 12,228 transcripts in samples derived from a specific cell type or tissue (Table 4.2). To provide a more accurate estimate of genes expressed in each tissue type, we adopted a simple transformation designed to exclude transcripts likely to originate with the minor fraction (3-20%) of unmarked cells isolated by FACS (Table 4.1) or from non-specific RNA generated by the mRNA-tagging protocol (Von Stetina et al. 2007). Transcripts that are highly expressed in a specific tissue might also be detectable at lower levels in profiles derived from other cell types due to this background. Thus, as a conservative measure, we restricted the set of expressed genes for each cell type to transcripts with a higher level of expression measured for a

given cell type than for the corresponding reference (*i.e.*, the "average" cell) at the same stage. This approach effectively excluded, for example, the *myo-3* body muscle-specific transcript (Okkema et al. 1993) from data sets derived from non-muscle cell types while generally retaining "housekeeping" genes such as ribosomal proteins that are likely to be widely expressed in all tissues (Von Stetina et al. 2007). This analysis detected between 4,572 and 7,199 expressed genes in each of the 25 cell types profiled (5,698 genes on average, see Figure 4.5).

Identification of Transcriptionally Active Regions (TARs)

The high probe density (on average every 25 bp) of the tiling array made the non-repetitive portion (~85%) of the genome accessible to *de novo* identification of transcripts in a way that is not biased by potentially incomplete annotations (see Figure 4.1D for an illustration). However, this comprehensive representation of the genome does not allow for optimized probe sequences and consequently results in large variability in hybridization affinity. Our pivotal normalization step thus aimed at reducing probe sequence bias. This correction also improved the signal-to-noise ratio (exon intensity over background) to an even larger extent than observed for another method that additionally exploits reference hybridization to genomic DNA (Figure 4.3A, see Methods for details) (Huber et al. 2006). For the segmentation of hybridization signals into intergenic regions, exons and introns, we used a method called mSTAD (Laubinger et al. 2008; Zeller et al. 2008). Although mSTAD is trained on hybridization signals corresponding to known (mostly protein-coding) genes (see Methods), it

Table 4.2. Gene models detected as expressed above background and with differential expression between cell types and references or between developmental stages.

	Cell types	**Dev. stages**	**Both data sets**
Expressed genes (5% FDR)	17,075	15,822	17,452
			87.7%
Average # expressed genes per data set (5% FDR)	12,228	12,252	12,232
			61.4%
Expressed genes (stringent FDR)	13,149	13,713	14,279
			71.7%
Differentially expressed genes (5% FDR, FC≥2)	10,598	9,552	13,320
			66.9 %
Average # diff. expressed genes per data set (5% FDR, FC≥2)	713	1,321	954
			4.8%
Differentially expressed genes (stringent FDR, FC≥2)	7,983	8,606	11,299
			56.7%

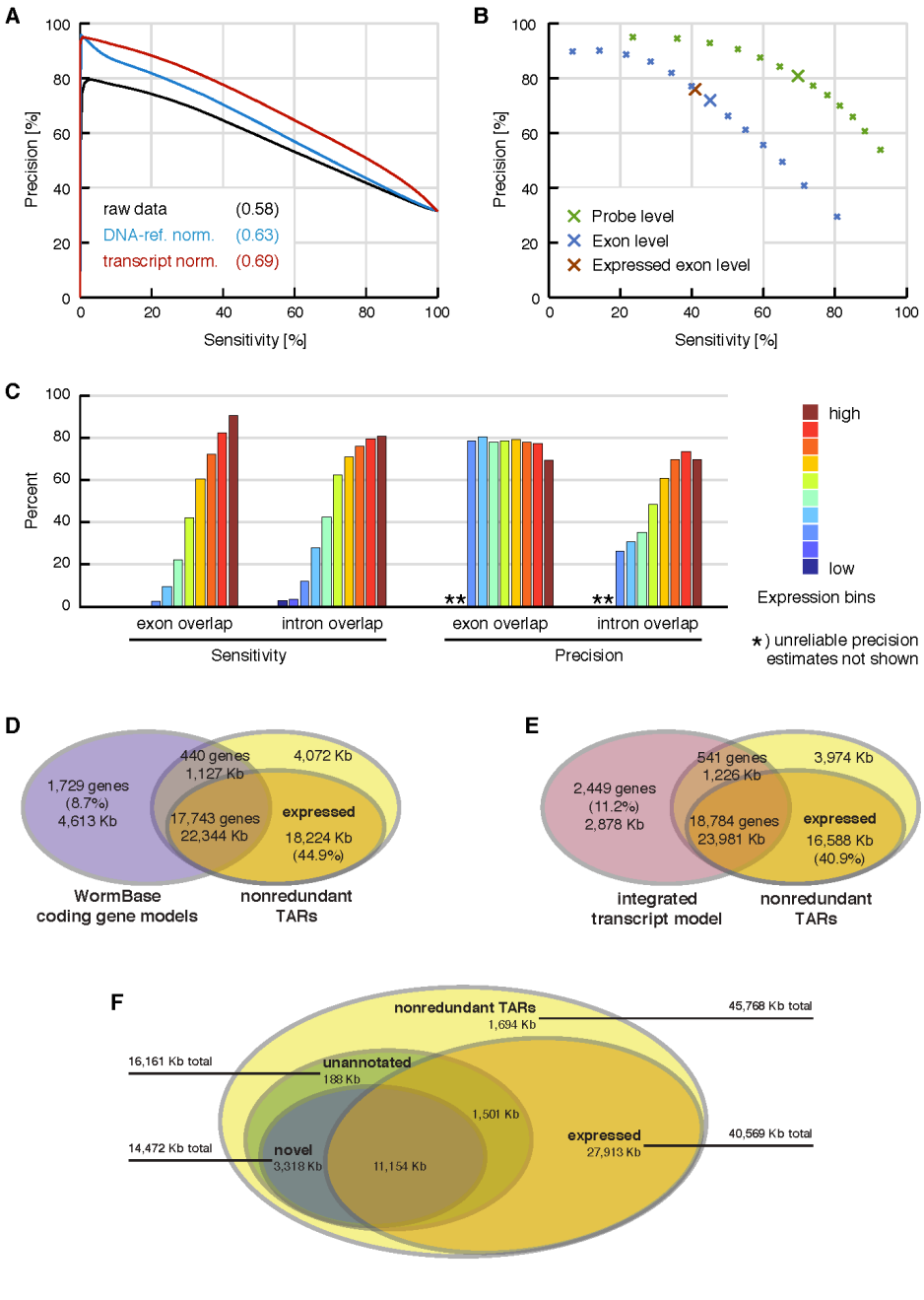

A
Precision [%]
Sensitivity [%]
raw data (0.58)
DNA-ref. norm. (0.63)
transcript norm. (0.69)
B
Precision [%]
Sensitivity [%]
Probe level
Exon level
Expressed exon level
C
Percent
exon overlap
intron overlap
exon overlap
intron overlap
Sensitivity
Precision
high
low
Expression bins
★) unreliable precision estimates not shown
D
1,729 genes (8.7%) 4,613 Kb
440 genes 1,127 Kb
4,072 Kb
17,743 genes 22,344 Kb
expressed 18,224 Kb (44.9%)
WormBase coding gene models
nonredundant TARs
E
2,449 genes (11.2%) 2,878 Kb
541 genes 1,226 Kb
3,974 Kb
18,784 genes 23,981 Kb
expressed 16,588 Kb (40.9%)
integrated transcript model
nonredundant TARs
F
nonredundant TARs 1,694 Kb
45,768 Kb total
16,161 Kb total
unannotated 188 Kb
1,501 Kb
expressed 27,913 Kb
40,569 Kb total
14,472 Kb total
novel 3,318 Kb
11,154 Kb

Figure 4.3. De novo transcript identification with mSTAD and overlap of TARs with annotated and experimentally defined gene models.
(A) Transcript normalization (red) improved exon probe recognition over raw data (black) and compared with normalization using genomic DNA hybridization as reference (blue). Sensitivity and precision were estimated after thresholding the intensity data with increasing cutoffs in a fivefold cross-validation. Sensitivity is defined as the percentage of exon probes with signal above the threshold among all annotated exon probes. Precision is defined as the percentage of annotated exon probes among those with signal above the threshold (see Methods). Values in parentheses indicate area under the curve. Based on data from LE-ref (see Table 4.1). (B) Cross-validation accuracy of mSTAD for probes (green), for exons (blue), and for exons with independently confirmed expression (brown). For exons, sensitivity is defined as the percentage of annotated exons for which all corresponding tiling probes were predicted as exonic by mSTAD. Precision is defined as the percentage of predicted exons for which all probes are annotated as such. Definitions for probes are as in A but with respect to predictions by mSTAD. Exon-level evaluation was repeated with the subset of predicted exons also detected as expressed by a statistical test (see Expressed Exon Level). Enlarged crosses correspond to predictions used for subsequent analysis. Based on data from LE-ref (Table 4.1). (C) Accuracy of exon and intron recognition increased with gene expression. Colored bars correspond to equally sized expression bins. Here exon overlap sensitivity equals the percentage of predicted exons, which overlap by at least 75% of their length with annotated exons. Exon overlap precision equals the percentage of exon predictions overlapping with annotated exons (by 75% or more) among all predicted exons (intron overlap sensitivity and precision are defined analogously with respect to predicted and annotated introns). Based on data from LE-ref (Table 4.1). (D) Overlap between nonredundant TARs (nrTARs), the portion detected as expressed and annotated coding gene models. About 45% of expressed nrTAR bases do not overlap with annotated coding gene models. (E) Overlap between TARs and the modENCODE integrated transcript model (Hillier et al. 2009; Gerstein et al. 2010). About 41% of expressed nrTAR bases do not overlap with the integrated transcript model. (F) Unannotated and novel TARs and their overlap with TARs expressed above array background. Unannotated TARs are defined as TARs without significant overlap (≥20 bp) with exons of annotated coding genes, pseudogenes, and ncRNAs. Novel TARs are defined as the subset of unannotated TARs without significant overlap (≥20 bp) with exons in the integrated transcript model (for details, see main text).

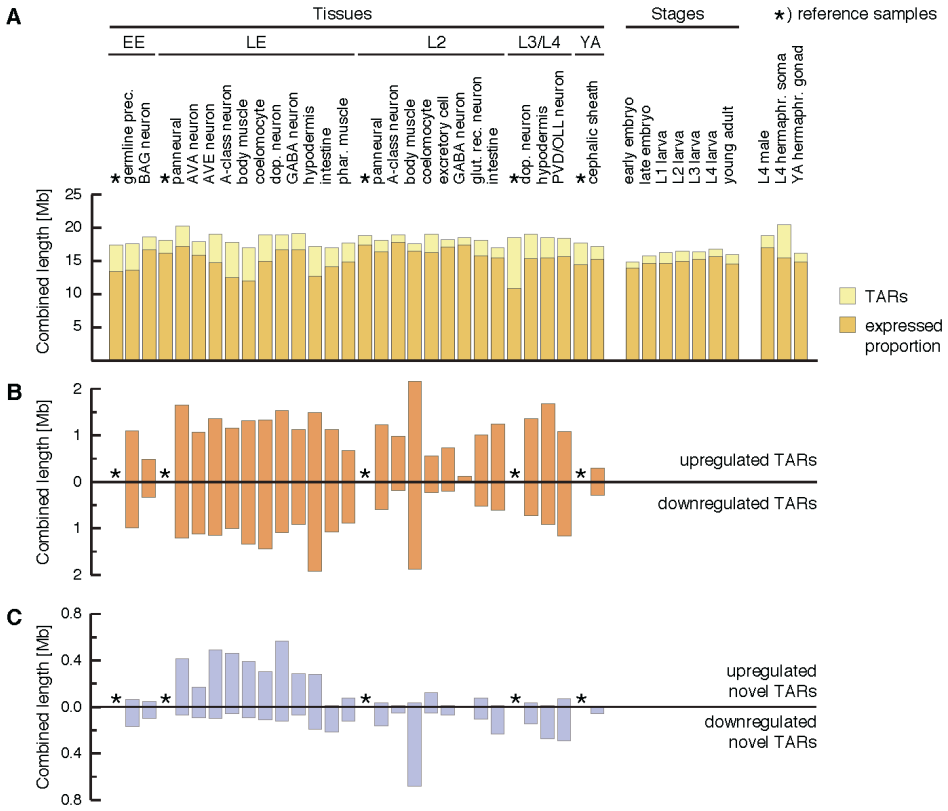


Figure 4.4 TAR predictions
(A) Predicted transcriptionally active regions (TARs) per tissue/cell type for which expression could be confirmed by a statistical test (expressed) (see Methods in main text for details). EE (Early Embryo), LE (Late Embryo), L2 larva, L3/L4 larva, YA (Young Adult). (B) TARs detected as differentially expressed between tissue samples (labels as in A) and reference samples. (C) Novel TARs detected as differentially expressed between tissue samples (labels as in A) and reference samples.

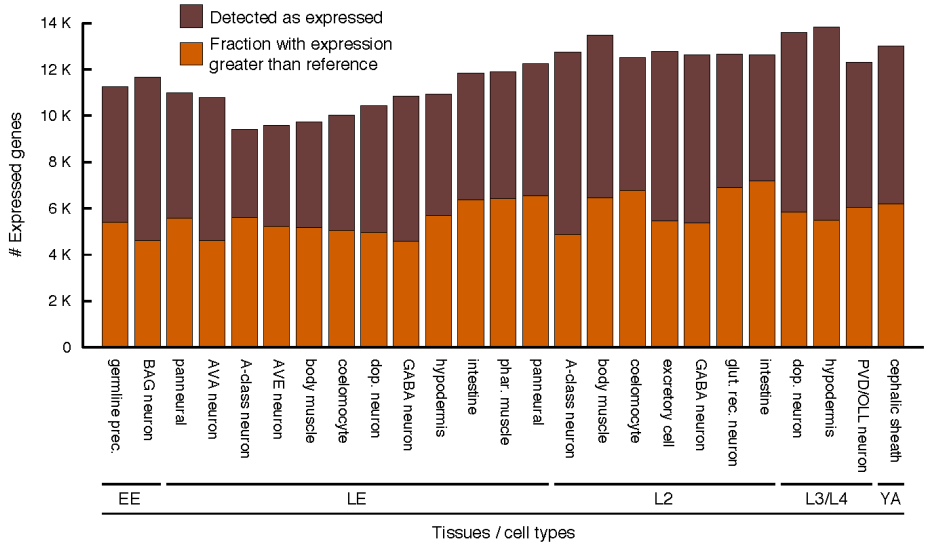


Figure 4.5 Genes expressed in cell type samples
Bar height corresponds to the number of genes detected as expressed above background. The fraction of genes with higher expression in a given cell type than in the corresponding reference sample is indicated (see key).

afterwards predicts transcripts regardless of their annotation status. We first assessed the cross-validation accuracy of these predictions relative to annotated protein-coding gene models (Figure 4.3B). Generally, the sensitivity of these predictions for annotated exons improved with the expression level of the corresponding genes, while high precision (~80%) with respect to overlap with annotated exons was maintained across all expression levels (Figure 4.3C).

Transcriptionally active regions (TARs) map to protein-coding genes and to intergenic domains

From the TAR predictions of individual samples, we constructed non-redundant TARs (nrTARs) containing the union of nucleotides inclusive to a TAR in any of the samples analyzed (30 cell types and reference samples and seven developmental stages, Table 4.1). In total, ~45 Mb were covered by nrTARs, and the subset of expressed nrTARs (*i.e.*, TARs that passed a statistical test for expression above background, see Methods) contained ~40 Mb. We next compared on an individual-nucleotide basis the overlap between known transcripts and nrTARs predicted *de novo* from the tiling array data. In a comparison to protein-coding gene models annotated in WormBase (Rogers et al. 2008), ~84% of nucleotides (*i.e.*, 22,344 kb +/- 1,127 kb) in annotated exons (from >90% of gene models) were also detected in nrTARs (Figure 4.3D). The subset of “expressed” nrTARs covered ~80% of nucleotides in annotated exons from more than 90% of gene models and additionally contained >18 Mb (~45% of expressed nrTARs) outside of exons for annotated protein-coding genes (Figure 4.3D). A similar comparison between nrTARs and the modENCODE integrated

transcript model defined by transcriptome sequencing of polyadenylated RNA (Hillier et al. 2009) and EST evidence from WormBase (Gerstein et al. 2010) detected ~25 Mb of overlapping exons or nrTARs corresponding to ~90% of nucleotides in exons of the integrated transcript model and ~57% of nucleotides in nrTARs (Figure 4.3E). Nearly 41% of nucleotides within the expressed nrTARs were found outside of exons defined by the integrated transcript model (Figure 4.3E). Taken together, most gene models (~90%) were supported by nrTARs, whereas a substantial fraction of nrTARs could not simply be attributed to known transcripts.

Tiling array analysis detects 11 Mb of novel TARs from intergenic regions

We defined “unannotated TARs” as those that did not significantly overlap with exons of any coding gene, ncRNA or pseudogene annotated in WormBase (Rogers et al. 2008) (Figure 4.4). When we additionally required that TARs not overlap with any exons of the integrated transcript model, we obtained “novel TARs” (see Methods). In total, unannotated nrTARs covered ~16 Mb of the genome; ~90% were also novel (Figure 4.3F). Expression above background in any sample could be confirmed by a statistical test (see Methods) for ~11 Mb of novel nrTARs (Figure 4.3F). These results suggest that our extensive profiling of cells and tissues as well as developmental stages revealed a significant fraction of the *C. elegans* transcriptome that went undetected by methods limited by analysis of polyadenylated transcripts or by sampling of fewer conditions (Rogers et al. 2008; Hillier et al. 2009). Our findings parallel results from a previous tiling array study that also detected abundant non-polyA+ transcription from intergenic

regions (He et al. 2007). Although our transcript identification method was originally trained on annotated protein-coding genes, it is based purely on hybridization features (Laubinger et al. 2008; Zeller et al. 2008) and hence is expected to be capable of recognizing non-polyadenylated as well as non-coding transcripts. We verified that TARs identified by mSTAD contain annotated ncRNAs, including snoRNAs, miRNAs and pseudogenes (Figure 4.4). Moreover, on a per-nucleotide basis, almost 60% of the putative long ncRNAs (>2.7 Mb) predicted by (Liu and Deneris 2011) was contained in the set of novel nrTARs described here. However, <20% of nucleotides from the novel nrTARs recognized by our approach were also predicted by (Lu et al. 2011). RT-PCR of a subset of these novel TARs confirmed expression (Figure 4.6A, B).

The majority of *C. elegans* genes are differentially expressed among cell types and developmental stages

We profiled a broad panel of tissues and developmental stages with the idea that this approach could reveal the prevalence of potential gene regulatory mechanisms that might modulate transcript abundance among different cell types or developmental periods. To detect changes in gene expression during development, we performed all pairwise comparisons (total = 21 comparisons) of the seven tiling array data sets obtained from staged embryos (EE, LE), larvae (L1, L2, L3, L4) and young adults (YA) (see Methods). To detect transcripts that are differentially expressed between cell types, we compared each of the 25 cell-specific data sets to its corresponding reference sample (total = 25 comparisons) (Table 4.1). In both cases, these comparisons were designed to detect transcripts that are either significantly depleted or enriched (2-fold, FDR ≤ 0.05)

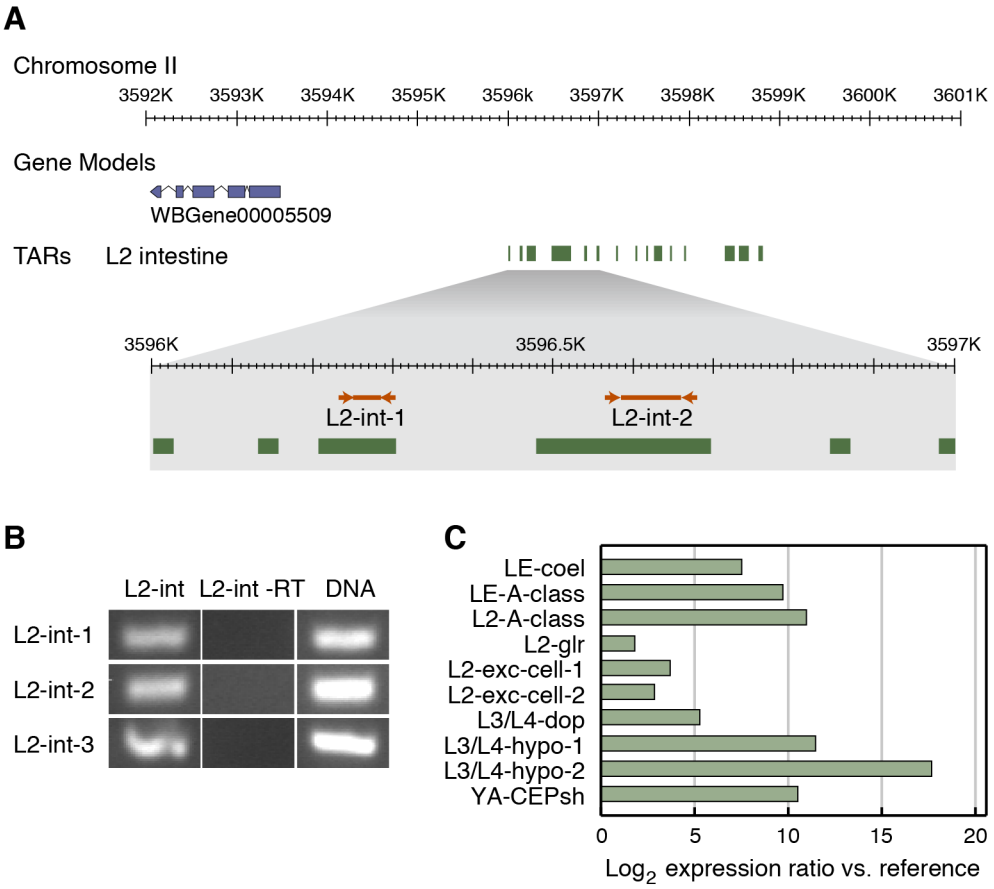


Figure 4.6 mSTAD detects TARs corresponding to protein-coding genes and to novel transcribed regions.
(A) Novel TARs detected in larval L2 intestine. Enlarged region shows location of primers and predicted RT-PCR amplicon from two TARs, L2-int-1 and L2-int-2. (B) RT-PCR detects novel TARs expressed in specific cell types. TARs L2-int-1, 2, 3 are detected in RNA isolated from the larval L2 intestine (L2-int) but are not amplified from RNA in the absence of reverse transcriptase (L2-int-RT). (C) qPCR validates enrichment of novel TARs in specific cell types. Log_2 ratio of enrichment in specific tissue versus corresponding reference samples (Table 4.1).

(Figure 4A). After correcting for multiple testing as above (Methods), this analysis produced conservative estimates of genes differentially expressed between stages (8,606 on average) or between specific cell types and reference samples (7,983) (Table 4.2). On the basis of these results, we conclude that transcripts for a majority of *C. elegans* genes are regulated to achieve different levels of expression during development and between specific types of cells. To validate the enrichment of genes in tissues and cell types as detected here, we compared a select number of our enriched gene lists to similar, independently derived data sets. For example, we identified 318 genes annotated as expressed in the excretory cell by combining GFP expression patterns from WormBase (WS200) and from the Genome BC *C. elegans* Gene Expression Consortium (Hunt-Newbury et al. 2007; Rogers et al. 2008). Our L2 excretory cell enriched gene list (531 genes) contains 61 of these 318 genes, which is significantly higher than expected for a random distribution (6.7X over-represented, $P < 7.1e^{-33}$). Similarly, the list of gene enriched in L2 body-wall muscle generated by our study (1,152 genes) shares 146 genes (2X over-represented, $P < 8.426e^{-17}$) with a comparable list (1,157 genes) obtained from L1 larval body-wall (Roy et al. 2002). A previously produced L2 stage intestine-specific data set (1,925 genes) significantly overlaps with our L2 intestine profile (195 out of 678 genes) (2.8X over-represented, $P < 4.352e^{-42}$) (Pauli et al. 2005). The union of our embryonic and L2 stage intestine enriched data sets contains 1540 transcripts that overlaps with 103 of 153 (8.1 fold over-represented, $P < 2.9e^{-74}$) genes previously identified as intestine-specific in SAGE (serial analysis of gene expression) data

sets derived from embryonic and adult intestine (McGhee et al. 2007; McGhee et al. 2009). A SAGE data set from the young adult gonad (Wang et al. 2009) identified 1,063 genes enriched in the germ line in comparison to the all somatic cells. We generated a tiling array profile also from dissected yound adult gonads and identified 4,363 enriched genes in comparison to the soma. These germ line enriched SAGE and tiling array data sets significantly overlap, sharing 462 genes (1.8X over-representation, $P < 4.016e^{-49}$). A comparison of the previously generated germ line SAGE list to our embryonic Z2/Z3 germ line precursor enriched gene, also shows significant enrichment (3.0X over-representation, $P < 1.051e^{-40}$). These comparisons reinforce the validity of each data set, particularly since the earlier profiles were generated with a variety of methods including GFP reporter imaging and serial analysis of gene expression (SAGE) and also may differ from our samples in developmental age. Lists of cell or tissue-enriched transcripts are available at (http://www.vanderbilt.edu/wormdoc/wormmap).

Specific genes are selectively enriched in certain cell types or tissues

Among the genes that are enriched in a certain tissue, we further sought to distinguish genes that are selectively enriched in the given tissue relative to those with broadly elevated expression in many cell types. The information theoretic concept of Shannon entropy effectively allowed us to define this subset of selectively enriched genes by distinguishing patterns of broad and uniform expression (high entropy) from more restricted ones with a high degree of tissue specificity (low entropy) (Schug et al. 2005). These lists of selectively enriched genes comprised ~20-57% of all genes enriched in the corresponding tissue or

cell type (see Methods). 82% of all genes selectively enriched in any cell type or tissue are specific to only one or two samples. For example, the set of genes selectively enriched in embryonic dopaminergic neurons also shows elevated expression in larval dopaminergic neurons and comprises known dopaminergic genes including the ETS transcription factor, *ast-1,* and its downstream targets the dopamine transporter, *dat-1*, and dopamine biosynthetic enzymes *cat-2* and *cat-4* (Flames and Hobert 2009).

We further defined the set of genes selectively enriched in any of the thirteen neuronal samples, but not enriched in non-neuronal tissue. Strikingly, this combined neuron-selective data set is most strongly enriched for putative 7 transmembrane (7TM) domain G-protein coupled receptor (GPCR)-like proteins (FDR < $5.2e^{-25}$). Our finding is consistent with earlier reports of selective expression of 7TM/GPCR genes in the *C. elegans* nervous system (Troemel et al. 1995; Chen et al. 2005). Cases of 7TM/GPCR genes that are expressed in specific neurons are also evident in our tiling array results. For example, *sra-32* and *sra-36* are uniquely detected in the L2 larval stage A-class neuron data set (Figure 4.8D). Of the 1,512 predicted members of the 7TM/GPCR family, 314 (~21%) were not detected in any RNA-Seq derived data set produced from whole animals (Hillier et al. 2009) for the modENCODE consortium (Gerstein et al. 2010). Among these are 66 family members that are detected in our tiling array assays. Our findings provide an explanation for the relative lack of coverage of the 7TM/GPCR family in the RNA-Seq results and predict that transcripts encoded by other members of this large and diverse gene family could be

detected by expanding our cell-specific profiling strategy to additional neuron types.

Novel TARs are differentially expressed and many are selectively detected in certain cell types

To quantitatively assess expression differences for TARs, in particular novel ones, probes contained within TARs from each cell-specific data set were compared to probes from the same region in the corresponding reference sample (see Table 4.1). This analysis revealed ~5Mb of TARs with significant expression changes between cell types and references or between developmental stages at an FDR ≤ 0.05 and 2-fold expression difference (see Methods). On average, 933 novel TARs are differentially expressed in a particular cell type in comparison to a reference sample of all cells at the corresponding developmental stage (Figure 4.7C). We used quantitative PCR (qPCR) to confirm that ten of these differentially expressed novel TARs indeed show significant enrichment in the specific cell types initially identified in the comparison of tiling array results (Figure 4.6C).

We next explored the extent to which novel TARs are selectively expressed with the goal of cataloging potentially rare transcripts that might be specifically detected in a limited subset of cells or in a discrete developmental period. To investigate the expression patterns of TARs on a per-nucleotide basis, we tabulated the frequency at which a given base was detected as transcribed across cell types and stages. Approximately 15% of bases covered by exons of gene models annotated in WormBase are detected in all 25 cell-types profiled

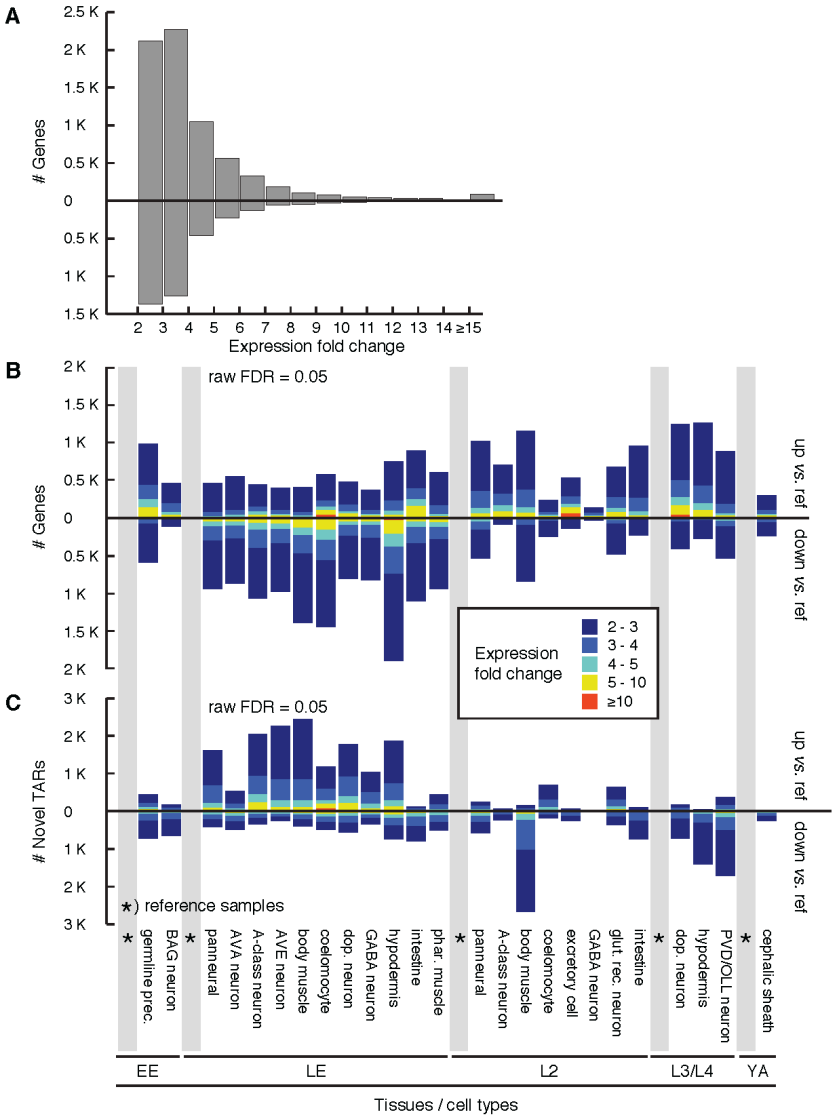


Figure 4.7 Expression fold changes of differentially expressed genes and TARs.
(A) Histogram depicting numbers of gene models binned according to maximal relative expression (fold change) in specific cell types vs. corresponding reference samples derived from all cells (FDR ≤ 0.05) (see supplemental protocol SP24).
(B) Histogram counting gene models differentially expressed between cell types and corresponding reference samples (FDR ≤ 0.05). Expression fold change is color-coded (see key between B and C).
(C) Histogram showing novel TARs that are differentially expressed between cell types and corresponding reference samples (FDR ≤ 0.05). Expression fold change is color-coded (see key between B and C).

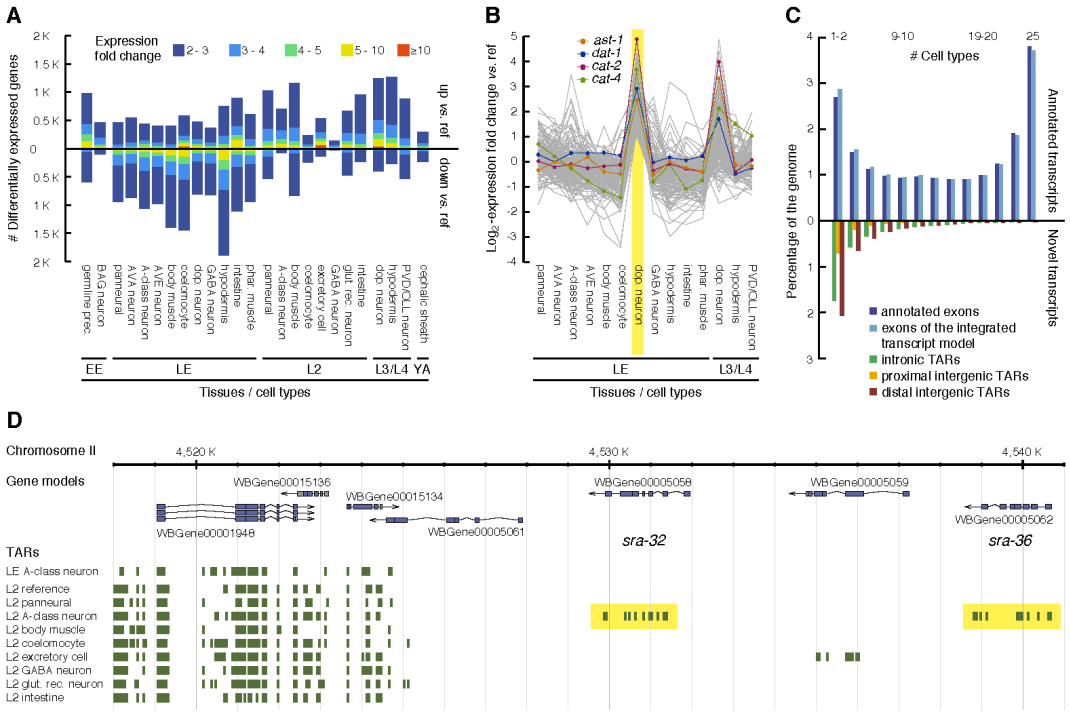


Figure 4.8 Transcripts enriched or depleted in certain cell types.
(A) Genes differentially expressed between a given cell type and the corresponding reference sample (FDR ≤0.05). Bars pointing up and down indicate the number of enriched and depleted genes, respectively, relative to reference. Expression fold change is color-coded (see key). (B) Log_2-expression fold change relative to reference shown as gray lines for genes selectively enriched in LE dopaminergic neurons (highlighted in yellow). Four selectively enriched genes (*ast-1*, *dat-1*, *cat-2*, *cat-4*) with known function in these neurons are plotted in color (see key). (C) Coverage of the genome by expressed transcripts at base-pair resolution. Nucleotides in nonredundant TARs (nrTARs) (for 25 cell-type samples) (Table 4.1) were binned according to the number of samples for which a TAR was detected at the given position. Bars pointing upward correspond to expressed TARs overlapping with exons of annotated coding genes and those defined by the integrated transcript model. Bars pointing downward correspond to nucleotides in expressed novel TARs (for definition, see main text) organized into subgroups according to their location relative to annotated protein coding gene models (see key). Intergenic positions were classified as proximal if within 500 bp of any annotated gene and otherwise as distal. (D) 7TM genes are selectively expressed in a specific neuron. Two members (*sra-32* and *sra-36*) of a tandem array (yellow highlights) of 7TM-encoding genes are selectively enriched in the A-type motor neuron data set derived from L2 larvae.

(Figure 4.8C). A larger fraction of bases derived from gene models (~75%), however, is expressed in at least one, but not all, of the cell types and 11% is detected in no more than two cell types. For stages, we observed that 29% of bases from exons of annotated gene models is detected throughout development *vs.* 8% expressed in no more than one developmental period. Bases corresponding to novel TARs that map to intergenic regions showed a stronger bias for cell or stage-specific expression with >53% detected in either one or two cell types but <1% (~34 kb) detected in all cell types (Figure 4.8C). Bases located >500 nt from a gene model (distal) comprised the majority (~75 %) of novel transcribed intergenic nucleotides uniquely detected in one or two cell types or in a single stage (Figure 4.8C). Given the average intron length of 344 nt for *C. elegans* (Bradnam and Korf 2008), we suggest that these distal bases are more likely to correspond to new transcribed regions as opposed to exons belonging to existing gene models.

Online resources for visualization and data access

To facilitate further study of our tiling array-based expression data, we have made it accessible to the research community through two online visualization tools, both of which are linked from a project website (http://www.vanderbilt.edu/wormdoc/wormmap/). One of the utilities displays expression values across cell types and developmental stages for a user-defined subset of genes (http://jsp.weigelworld.org/wormviz/tileviz.jsp). Additionally, a customized genome browser (http://gbrowse.fml.mpg.de/cgi-bin/gbrowse/ce_WS199), visualizes transcriptionally active regions (TARs) for all

analyzed samples together with gene models and genomic features annotated in WormBase (Rogers et al. 2008). Raw array data have been deposited at GEO (Barrett et al. 2009) (accession numbers GSE23245-GSE23271, GSE23278-GSE23287, GSE23769-GSE23770) and data files are available for download from the project website (http://www.vanderbilt.edu/wormdoc/wormmap/).

Analysis of differentially expressed transcripts reveals cell-specific functions and clusters of co-regulated genes with candidate *cis*-acting motifs

Our quantitative analysis has identified transcripts that are differentially expressed across a broad array of cell types and developmental stages. We expect that these results will provide a useful resource for future studies of cell-specific gene function and for identifying the regulatory elements that define spatial and temporal patterns of gene expression. Below we feature examples of these approaches in order to illustrate potential applications of these data sets.

Genes encoding membrane transporter proteins are highly enriched in the excretory cell

Osmoregulation and excretion are fundamental biological processes that all animals share. In a typical multicellular organism, specialized cell types are assembled into an excretory organ that collects and removes metabolic wastes or functions to maintain ionic balance in changing aqueous environments. In *C. elegans*, these complex physiological tasks are accomplished with a simple excretory system composed of only 4 types of cells; the pore cell, duct cell, gland cells and excretory canal. The largest of these cells, the excretory canal, assumes a unique H-shaped architecture in which elongated tubular processes emanate from the cell soma beneath the posterior bulb of the pharynx, bifurcate

to the right and left sides and then separate again to extend in both anterior and posterior directions along the entire length of the animal (Figure 4.9)(Nelson et al. 1983; Altun 2002-2010). The excretory cell cytoplasm is contained within a cylindrical membrane-bound domain that is penetrated from the interior or basal side by elaborate networks of canals. These “canaliculi” converge on an internal, fluid filled “tunnel” that connects with the duct and pore cells on the ventral side of the head region. Disruption of any one of these cell types, duct, pore or excretory canal, disables osmoregulatory capacity as evidenced by a swollen, lethal phenotype in hypotonic solutions (Nelson et al. 1983). We used the mRNA-tagging strategy to generate a tiling array profile of the excretory cell in L2 larvae, a developmental stage of both active excretory cell growth and essential osmoregulatory function (Table 4.1). This data set identified 531 transcripts that are enriched ($\geq$ 2-fold, FDR $\leq$ 5%) in the excretory cell in comparison to the average L2 larval stage cell (see Methods). GFP reporter genes generated from three genes that are highly enriched in this data set illuminate the elongated anatomy of this unique cell type (Figure 4.9C-E). As would be expected for a cell type with high osmoregulatory activity, molecular function gene ontology (GO) terms for membrane transporter related activities are over-represented in the excretory cell data set (FDR < 0.01, hypergeometric distribution, Figure 4.9B) and thus are indicative of excretory cell specific profile. In addition to detecting genes that code for physiological functions, the enriched profile also includes 17 transcription factors with potential roles in excretory cell differentiation (http://edgedb.umassmed.edu, (Reece-Hoyes et al. 2005)). Indeed, the POU

domain transcription factor, CEH-6, is highly enriched (4-fold) and has been previously shown to control excretory cell morphogenesis and gene expression (Burglin and Ruvkun 2001; Mah et al. 2007; Armstrong and Chamberlin 2010; Mah et al. 2010). All 17 (100%) of the known CEH-6-regulated genes are included in the excretory cell profile (Figure 4.9). Another 79 genes from this list have a perfect match to the CEH-6 binding site octamer, ATTTGCAT, within 1 kb upstream of the translational start site and are thus candidates for additional CEH-6 target genes. Two other members of the homeodomain family in this list, *ceh-26* (3.6-fold) and *ceh-37* (3.7-fold), are known to be expressed in the excretory cell (Lanjuin et al. 2003; Reece-Hoyes et al. 2005) but downstream targets have not been identified. Our finding that multiple transcription factors are expressed in the excretory cell is consistent with the earlier suggestion that excretory cell differentiation likely depends on the gene regulatory roles of multiple transcription factors functioning in parallel pathways (Burglin and Ruvkun 2001; Mah et al. 2007; Armstrong and Chamberlin 2010; Mah et al. 2010). For example, in addition to detecting all of the known *ceh-6* targets, our data set also includes 9/16 (56%) of vacuolar ATPase proton pump subunit genes that are coordinately regulated by the nuclear hormone receptor, *nhr-31* (Hahn-Windgassen and Van Gilst 2009).

Self-Organizing Maps (SOMs) reveal cohorts of co-regulated genes during development and across specific cell types

We used self-organizing maps (SOMs) to seek shared patterns of expression for transcripts derived from coding genes (see Methods). SOMs are a widely applied clustering technique that yields intuitive visualization of high-

dimensional data sets, as *e.g.,* generated with DNA microarrays (Jiang et al. 2001). SOMs are conceptually related to a technique previously proposed to construct a relational map of *C. elegans* gene expression (Kim et al. 2001). In the first instance, we fitted a SOM to the developmental stage data set (Figure 4.10A) and identified eight regions that correspond to genes with shared patterns of either enrichment or depletion in specific developmental periods (Figure 4.10B, see Methods). To demonstrate the variety of developmental expression patterns identified by this approach, we plotted the top 50% of best-fitting genes from each cluster (Figure 4.10C-F). Cluster 1 (CS1) contains genes with elevated expression in the embryo (Figure 4.10B, C). Notable examples from this group include the FoxA transcription factor, *pha-4*, the hunchback homolog, *hbl-1*, (Krause et al. 1997) and the helix-loop-helix transcription factors (bHLH), *hlh-2* and *hlh-3*, for which independent studies have detected peak expression in the embryo (Azzaria et al. 1996; Krause et al. 1997; Fay et al. 1999). Cluster 5 (CS5) contains genes with elevated expression in embryonic stages and in the adult (Figure 4.10B, E). Strikingly similar protein and transcript levels have been previously observed for a member of this group, the FLYWCH transcription factor

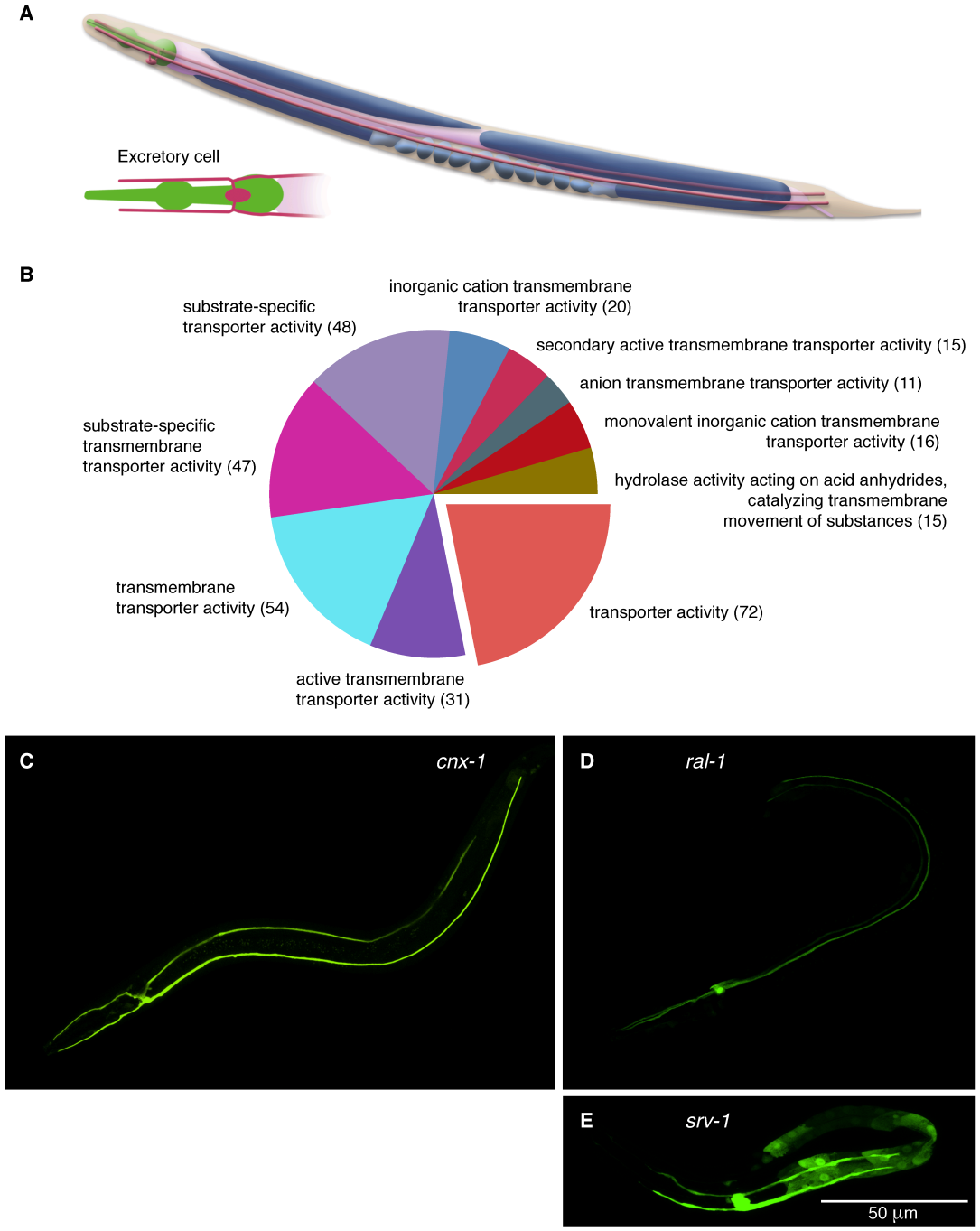


Figure 4.9 The excretory cell expresses many transport-related genes.
(A) The excretory cell body is located ventral to the terminal bulb of the pharynx, and extends canals anteriorly and posteriorly along either side of the body. These canals collect ions and fluid for osmoregulation. (B) Pie chart showing that the top ten GO molecular function categories enriched in the excretory cell profile correspond to transporter proteins (FDR < 0.01). (C) - (E) GFP-reporters selected from excretory cell enriched genes demonstrate robust expression in the excretory canal (C) *cnx-1*, 2.2X, (D) *ral-1*, 3.4X, (E) *srv-1*, 4.5X.

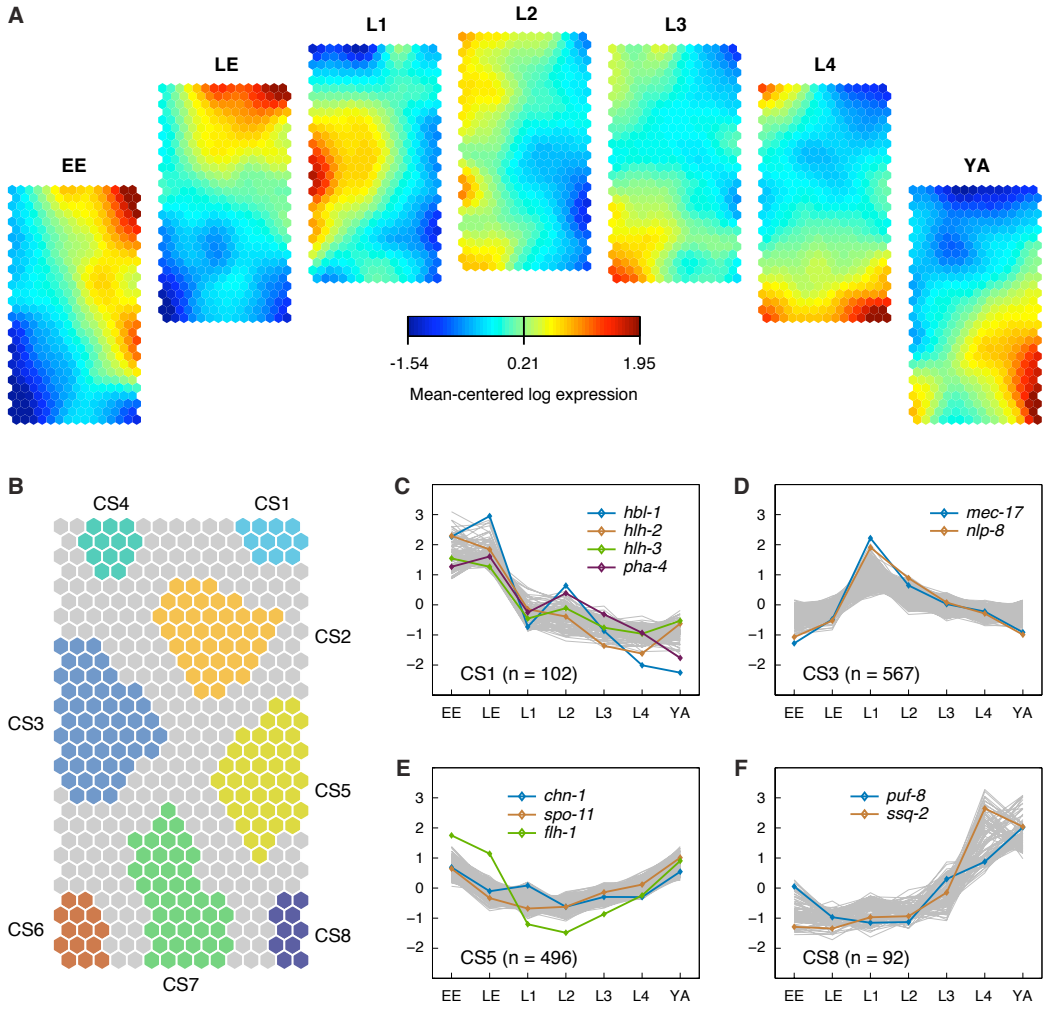


Figure 4.10 Expression patterns during *C. elegans* development.
(A) Component planes of a self-organizing map (SOM) fitted to the developmental stage data set. Each component plane visualizes mean-centered gene expression (log2-scale) in one stage as a color gradient from blue to red indicating low and high expression, respectively (see color bar): EE indicates early embryos; LE, late embryos; L1, larvae stage 1; L2, larvae stage 2; L3, larvae stage 3; L4, larvae stage 4; and YA, young adults. (B) Eight regions (CS1–CS8) of the SOM, which robustly clustered together, are color-coded (see main text for details). (C–F) Mean-centered log2-expression values of genes corresponding to four of the clusters in B are plotted for the 50% of best-fitting genes. Colored lines indicate the expression of a selected subset of genes (see key). *mec-17* and *nlp-8* encode neuron-enriched transcripts; *chn-1* and *spo-11* are highly expressed in the adult hermaphrodite gonad; *puf-8* is highly expressed in embryonic and adult germline; and *ssq-2* encodes a sperm-specific transcript. For other labeled genes, see Results.

flh-1, which blocks expression of specific miRNA genes during embryogenesis (Ogden et al. 2008). We applied a similar SOM clustering procedure to the cell-type specific data sets in order to delineate genes that are co-regulated in different tissues (Figure 4.11A, see Methods). Because these cell types were sampled across a series of developmental stages, we also expected this approach to detect genes with temporally correlated expression. Figure 4.11A depicts the resultant regional clusters superimposed on the SOM. Clusters showing stage-specific expression include C1 (Figure 4.11B), which features genes that are highly expressed in all postembryonic cell types and C7 (Figure 4.11D), which is biased for genes expressed in late embryos and especially in neurons. Cluster C8 is dominated by genes that are highly expressed in neurons, but are depleted or show weak expression in most other cell types (Figure 4.11E). Examples of genes in this group include *ric-4* (snap-25), a synaptic vesicle component that facilitates neurotransmitter release and is known to be exclusively expressed in neurons (Hwang and Lee 2003), and *acy-1* (adenylate cyclase), a key regulator of neuron-dependent behavior (Reynolds et al. 2005). Several clusters detect highly expressed intestinal genes including C1 (postembryonic cell types and larval intestine) (Figure 4.11B) and C11 (embryonic and larval intestine) (Figure 4.11F).

DNA sequence motifs associated with cell-specific and developmentally regulated gene expression

Because each SOM cluster includes genes with similar patterns of expression, we searched for instances in which genes in a specific cluster share common DNA sequence motifs through which *trans*-acting factors might

coordinate their expression. To explore this possibility, we applied the FIRE motif analysis program to the SOM clusters (Elemento et al. 2007). FIRE uses mutual information between the presence or absence of a short nucleotide sequence and the occurrence of a gene in a particular expression cluster to identify over-represented motifs. FIRE produces optimized motifs and links the results to motifs that are available in public databases.

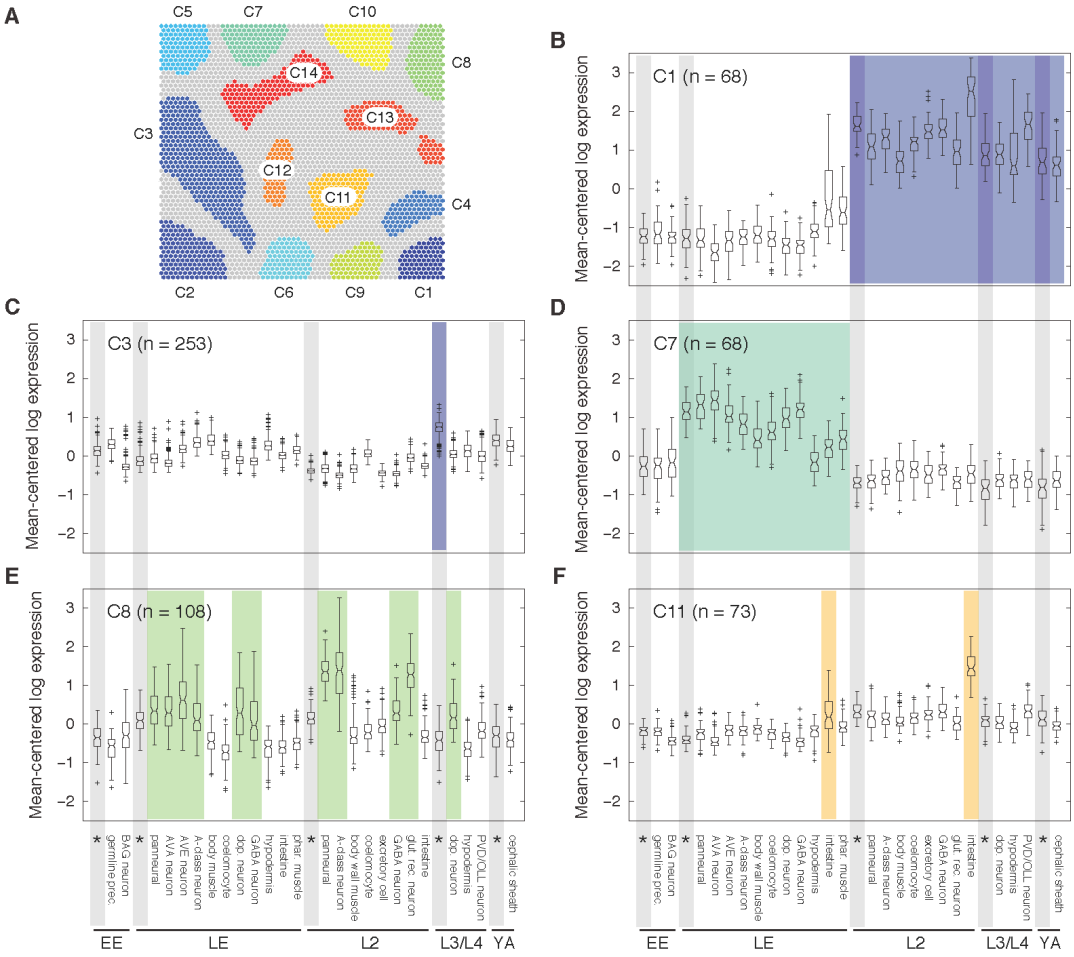


Figure 4.11 SOM clustering of tissue- and cell-type data.
(A) Regions in the SOM for cell-type data as defined by k-means clustering. (B–F) Expression patterns of genes from selected clusters in the SOM. Cell types are indicated at bottom, and reference samples are shaded in gray. Box plots were generated from the mean-centered $\log_2$-expression values of prototypical genes for a given cell type with horizontal lines indicating the median, boxes delineating the interquartile range, and whiskers extending to the most extreme values within 1.5 times the median-quartile range; outliers are depicted as black crosses. Some of the SOM clusters correspond to peaked expression in a subset of cell types and/or developmental stages: (B) higher expression in larval stages and YA compared with embryo with a prominent peak for intestine, (C) elevated expression in L3/L4 reference, (D) high expression in LE neurons, (E) most neurons, and (F) intestine.

FIRE identified 20 upstream promoter motifs and 9 over-represented 3' UTR sequences in genes contained in the SOM clusters derived from developmental stages (Figure 4.10, Figure 4.12A). A canonical E-Box and bHLH-binding site is detected in cluster CS1 which, as noted above, includes transcription factors HLH-2 and HLH-3 (Thellmann et al. 2003). The over-representation of GATA-like transcription factor binding sites in four clusters (CS4, CS5, CS6, CS7) is likely indicative of the broad roles of GATA factors in multiple developmental pathways in *C. elegans* including endodermal and hypodermal cell fate determination and differentiation, germline gene regulation and aging (Koh and Rothman 2001; McGhee et al. 2007; Budovskaya et al. 2008; del Castillo-Olivares et al. 2009). The second highest-ranking motif corresponds to a GC-rich sequence that has been previously identified by computational analysis of germ line expressed genes (Li et al. 2010). This motif is also similar to a putative transcriptional activation site for the E2F homolog, EFL-1, that promotes gene expression in the germ line (Chi and Reinke 2006). Detection of these GATA and E2F sites in cluster CS5 is consistent with our finding that genes which contain these 5' sites and which are enriched in germ line precursor (GLP) cells are also over-represented in this cluster (23 GLP genes with the GATA site are 1.6 fold over-represented, $P < 0.017$; 40 GLP genes with the E2F site are 1.8 fold over-represented, $P < 2.74e^{-04}$). These results validate our approach and suggest that other motifs revealed by this strategy may also correspond to binding sites for transcription factors that regulate developmental gene expression.

For SOM clusters derived from cell-specific profiles (Figure 4.12B), FIRE identified 45 over-represented sequences including 35 upstream motifs and 10 RNA sequences that map to 3' UTR domains (Figure 4.12B). As noted above for the SOM clusters derived from developmental stages, the highest scoring motif matches a GATA transcription factor-binding site. In *C. elegans*, the *elt-2* GATA transcription factor is known to interact with this sequence to drive expression of intestine-specific genes (McGhee et al. 2009). Our results also reflect this role; the GATA motif is overrepresented in cluster C11, which contains transcripts enriched in the embryonic and larval intestine profiles (Figure 4.11F, Figure 4.12B), and in C1 and C4 both of which show peak expression in larval intestine (Figure 4.11B, Figure 4.12B). The accurate identification of the GATA factor-binding site by the FIRE algorithm suggests that other motifs associated with specific SOM clusters may also correspond to specific transcription factor binding sites. An interesting example includes the sequence, TTTCG[AC]AA[CT] (Figure 4.12B), that is over-represented in genes enriched in embryonic neurons in cluster C7 (Figure 4.11D) and also reciprocally depleted in genes that are under-expressed in embryonic neurons in cluster C5. This motif is bound by the vertebrate C/EBP transcription factor (Grange et al. 1991), which has been shown to function with NeuroD to regulate neural gene expression (Sandelin et al. 2004; Calella et al. 2007). It will be interesting to determine whether C/EBP and NeuroD homologs exercise similar functions in *C. elegans* neural develoment.

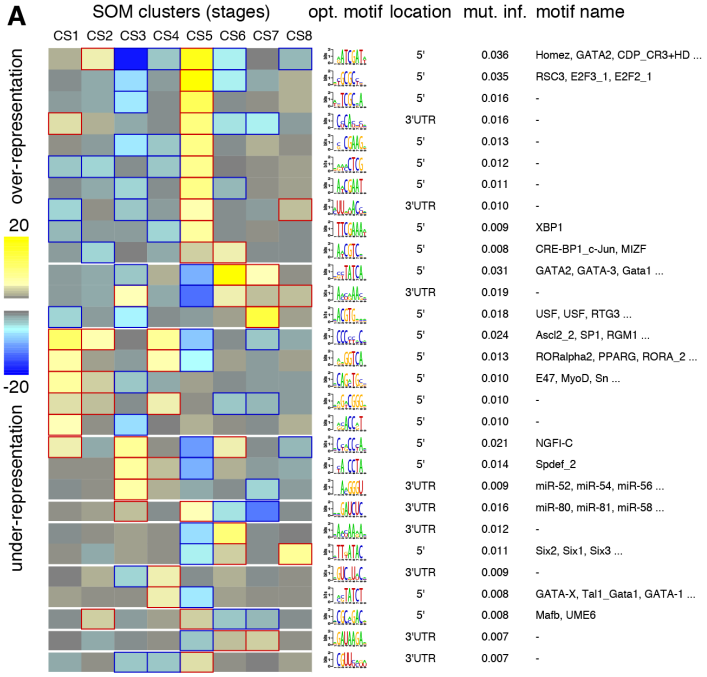
FIRE also identified 3' UTR binding sites for two distinct groups of miRNA genes belonging to the *mir-58* and *mir-51* families (Figure 4.12B). Members of the *mir-58* family (*mir-58*,*-80*,*-81*,*-82*) are abundantly expressed throughout development (Lim et al. 2003; Kato et al. 2009), but assays with promoter::GFP reporter genes have detected cell-specific patterns of expression (Martinez et al. 2008). For instance, *mir-58* is expressed in a broad array of cell types including the excretory canal, intestine, pharynx, and hypodermis, but is excluded from the nervous system (Isik et al. 2010). Emerging evidence indicates that transcript de-stabilization is the principle mechanism whereby miRNAs down-regulate gene expression (Bagga et al. 2005; Guo et al. 2010). Thus, the absence of *mir-58* expression in the nervous system predicts that neuronal transcripts carrying the *mir-58* recognition sequence should escape *mir-58*-induced degradation. And, in fact, our result showing that the *mir-58* sequence is over-represented in neuron-enriched transcripts (cluster C8, Figure 4.11E, Figure 4.12B) is consistent with this model. The motif for the *mir-51* family (*mir-51,-52,-53,-54,-55,-56*) is also over-represented in C8 (Figure 4.11E, Figure 4.12B**)** and in SOM clusters C9 and C10 that are dominated by transcripts enriched in hypodermal cells and neurons. This pattern suggests that *mir-51* family genes may have limited roles in regulating transcript levels in neurons and in the hypodermis. Conversely, the observation that the *mir-58* and *mir-51* motifs are significantly under-represented in C2 and C3 is suggestive of strong regulation by these miRNAs in the tissues that contribute to this cluster. In considering this question, we noted that C2 and C3 include an expression peak for the L3/L4 reference sample (Figure 4.11C).

Because germline tissue is rapidly proliferating at this stage (Kimble 1981), we compared the genes in clusters C2 and C3 to separate tiling array profiles obtained from the adult hermaphrodite gonad, L4 males and all somatic cells at L4 stage. These comparisons show a significant overlap, showing that C2 and C3 genes are largely expressed in the germline and specifically enriched for sperm expression. Thus, we speculate that members of the *mir-58* and *mir-51* gene families may have significant roles in modulating transcript levels in the germ line. The *mir-58* and *mir-51* motifs were previously identified by FIRE analysis of an independent group of whole animal microarray data sets from *C. elegans*.

Discussion

We have used whole genome tiling arrays to profile RNA isolated from specific cells and developmental stages of *C. elegans*. Our strategy of sampling a variety of different cell types and developmental periods was designed to capture potentially rare or transiently expressed transcripts as well as to provide a detailed spatial and temporal map of gene expression.

To monitor expression of individual protein-coding genes, we derived intensity values from aggregated probe sequences corresponding to each annotated gene model. Our combined set of tiling array data from 25 different cell types and seven developmental stages (Figure 4.1) detected ~90% of known protein coding genes (Table 4.2, Figure 4.3).


A
SOM clusters (stages)
opt. motif
location
mut. inf.
motif name
CS1 CS2 CS3 CS4 CS5 CS6 CS7 CS8
over-representation
20
-20
under-representation
5' 0.036 Homez, GATA2, CDP_CR3+HD ...
5' 0.035 RSC3, E2F3_1, E2F2_1
5' 0.016 -
3'UTR 0.016 -
5' 0.013 -
5' 0.012 -
5' 0.011 -
3'UTR 0.010 -
5' 0.009 XBP1
5' 0.008 CRE-BP1_c-Jun, MIZF
5' 0.031 GATA2, GATA-3, Gata1 ...
3'UTR 0.019 -
5' 0.018 USF, USF, RTG3 ...
5' 0.024 Ascl2_2, SP1, RGM1 ...
5' 0.013 RORalpha2, PPARG, RORA_2 ...
5' 0.010 E47, MyoD, Sn ...
5' 0.010 -
5' 0.010 -
5' 0.021 NGFI-C
5' 0.014 Spdef_2
3'UTR 0.009 miR-52, miR-54, miR-56 ...
3'UTR 0.016 miR-80, miR-81, miR-58 ...
3'UTR 0.012 -
5' 0.011 Six2, Six1, Six3 ...
3'UTR 0.009 -
5' 0.008 GATA-X, Tal1_Gata1, GATA-1 ...
5' 0.008 Mafb, UME6
3'UTR 0.007 -
3'UTR 0.007 -

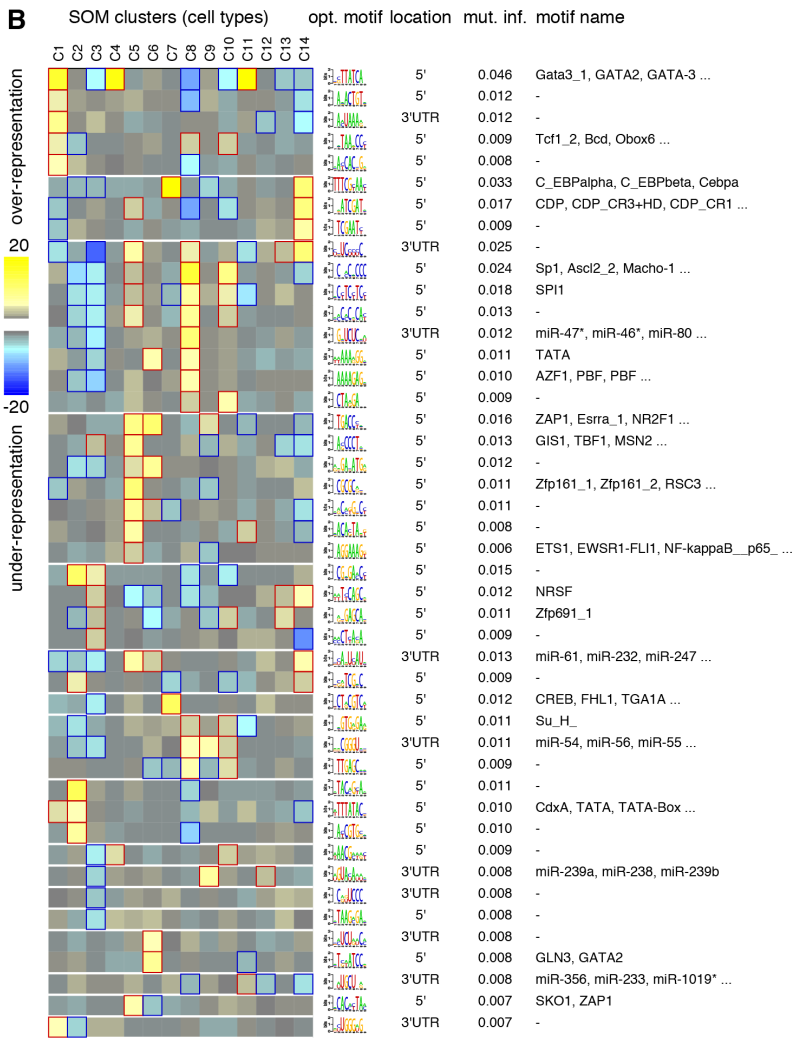

B
SOM clusters (cell types)
opt. motif
location
mut. inf.
motif name
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
over-representation
20
-20
under-representation
5' 0.046 Gata3_1, GATA2, GATA-3 ...
5' 0.012 -
3'UTR 0.012 -
5' 0.009 Tcf1_2, Bcd, Obox6 ...
5' 0.008 -
5' 0.033 C_EBPalpha, C_EBPbeta, Cebpa
5' 0.017 CDP, CDP_CR3+HD, CDP_CR1 ...
5' 0.009 -
3'UTR 0.025 -
5' 0.024 Sp1, Ascl2_2, Macho-1 ...
5' 0.018 SPI1
5' 0.013 -
3'UTR 0.012 miR-47*, miR-46*, miR-80 ...
5' 0.011 TATA
5' 0.010 AZF1, PBF, PBF ...
5' 0.009 -
5' 0.016 ZAP1, Esrra_1, NR2F1 ...
5' 0.013 GIS1, TBF1, MSN2 ...
5' 0.012 -
5' 0.011 Zfp161_1, Zfp161_2, RSC3 ...
5' 0.011 -
5' 0.008 -
5' 0.006 ETS1, EWSR1-FLI1, NF-kappaB__p65_ ...
5' 0.015 -
5' 0.012 NRSF
5' 0.011 Zfp691_1
5' 0.009 -
3'UTR 0.013 miR-61, miR-232, miR-247 ...
5' 0.009 -
5' 0.012 CREB, FHL1, TGA1A ...
5' 0.011 Su_H_
3'UTR 0.011 miR-54, miR-56, miR-55 ...
5' 0.009 -
5' 0.011 -
5' 0.010 CdxA, TATA, TATA-Box ...
5' 0.010 -
5' 0.009 -
3'UTR 0.008 miR-239a, miR-238, miR-239b
3'UTR 0.008 -
5' 0.008 -
3'UTR 0.008 -
5' 0.008 GLN3, GATA2
3'UTR 0.008 miR-356, miR-233, miR-1019* ...
5' 0.007 SKO1, ZAP1
3'UTR 0.007 -

Figure 4.12 Regulatory elements discovered in stage and cell-type expression clusters.
FIRE analysis identifies motifs over- and underrepresented in developmental profile clusters (A) and cell-type profile clusters (B). A heat map indicates whether each motif is overrepresented (yellow) or underrepresented (blue) in each cluster. Motifs are arranged in rows and clusters in columns. Significant overrepresentation is indicated by red box outlines and underrepresentation is indicated by blue outlines ($P \leq 0.05$, Bonferroni-corrected). The optimized motif logo, location of the motif (5′ upstream promoter or 3′ UTR), mutual information with the genes in the cluster, and matching transcription factors and miRNAs listed in public data bases for indicated motifs are shown alongside the heat map. (Elemento et al. 2007).

In addition to detecting expressed genes, our analysis also revealed that ~75% of all detected genes show at least 2-fold, statistically significant differences in transcript levels between cell types or developmental stages (Table 4.2). To document this trend of widespread differential expression among cell types and throughout development, we tabulated the frequency of transcription of a given nucleotide across tissues and stages. This analysis revealed that whereas 15% of exonic sequence is detected in all of the cell types that we sampled, a larger fraction (60%) shows more limited transcription with ~11% in no more than one or two cell types (Figure 4.8C). Our results also indicate that coding sequence is dynamically expressed during development with ~8% of bases from exons uniquely detected in only one embryonic, larval or adult stage. As we extensively sampled the *C. elegans* nervous system, we investigated the subset of genes selectively expressed in neuronal tissue. Among these genes, we noted striking enrichment of members of the 7TM-GPCR family, which is known for highly specific expression in the nervous system (Chen et al. 2005). The restricted expression of 7TM-GPCR genes potentially explains why many members of this family still lack experimental support (Hillier et al. 2009; Schweikert et al. 2009). Our results, however, suggest that profiles of more cell types should confirm expression of additional annotated gene models or genes newly predicted from the genome sequence. Overall, our finding of widespread differential gene expression underscores the conclusion that most *C. elegans* genes are extensively regulated and points to the key role of differential gene expression in the determination of cell fates and developmental progression. In

practice, our data on genes that are selectively enriched in a particular cell type or developmental period should be especially useful for identifying genes with cell- or stage-specific functions (Zhang et al. 2002; Colosimo et al. 2004; Blacque et al. 2005; Cinar et al. 2005; Touroutine et al. 2005; Von Stetina et al. 2007; McGhee et al. 2009; Chatzigeorgiou et al. 2010; Smith et al. 2010; Hallem et al. 2011).

In addition to using our tiling array results to identify genes expressed in specific cell types or developmental periods, we also sought evidence for more complex patterns in which cohorts of genes might be similarly regulated across tissues or among different development stages. For this purpose, we used the unbiased strategy of self-organizing maps (SOMs) to cluster co-expressed genes (Figure 4.10, Figure 4.11). This approach revealed, for example, a striking cluster with consistently elevated transcript levels in both embryonic and larval neurons that is largely comprised of genes with established neuron-specific functions (Figure 4.11E). Other clusters could reflect genes with common functions in a wide array of cell types during a particular developmental period (Figure 4.11B, D). Thus, our approach has confirmed known groups of co-regulated genes as well as suggested novel clusters that could point to previously unstudied biological roles for batteries of co-expressed genes. In addition to providing a direct read-out of cell-specific gene expression, our microarray data should also substantially enhance the accuracy of SVM-based strategies that rely on gold standard training sets for *ab initio* identification of cell-specific expression from whole animal microarray data (Chikina et al. 2009). Motif analysis of the SOM

results derived from our data sets identified highly over-represented flanking sequences in genes belonging to specific clusters (Figure 4.12). Each case could be indicative of a regulatory mechanism involving a shared *trans*-acting factor. For example, a consensus binding site for a GATA factor with a broad role in regulating intestine-specific genes in *C. elegans* (McGhee et al. 2009) was specifically over-represented in SOM clusters defined by transcripts with high expression levels in tiling array data sets derived from intestinal cells (Figure 4.12B). Over-represented motifs in the 3' UTR regions include recognition sites for two large and highly expressed groups of closely related miRNAs, the *mir-58* and *mir-51* families (Figure 4.12B). Our analysis of these results points to potential roles for both *mir-58* and *mir-51* in regulating transcript abundance in the germ line, a suggestion consistent with the recent observation that the Drosophila ortholog of the *mir-58* family, bantam, is required for germline stem cell fate (Yang et al. 2009).

Our tiling array results confirm expression of the vast majority (~90%) of *C. elegans* protein coding genes recently identified by RNA-Seq analysis (Hillier et al. 2009). Additionally, our machine-learning algorithm also identified a substantial number of TARs arising from intergenic regions (Figure 4.3D-F). A conservative treatment of these data that uses a statistical test for expression above background, leads to the estimate that ~11 Mb of intergenic sequence, or ~10% of the *C. elegans* genome, encodes novel RNAs that have not been previously annotated in WormBase or detected by RNA-Seq. One explanation for this difference is that we assayed total RNA from embryonic cells and

developmental stages and that the poly-A+ pull-downs that we used for sampling postembryonic cell types (Figure 4.1B) also include a significant non-polyA+ fraction (Von Stetina et al. 2007). In contrast, recent RNA-Seq results for *C. elegans* were limited to purified poly-A+ RNA (Hillier et al. 2009). Because the known families of short non-coding RNAs (ncRNAs) were manually excluded from our list of intergenic RNAs, we propose that these transcripts define potentially new types of non-coding RNA. An independent analysis of *C. elegans* transcriptomics data that includes the tiling array results used in this work, also reports a substantial number (~4.6 Mb) of putative non-coding RNAs from intergenic regions with a large overlap (>2.5 Mb) to our ncRNA predictions (Lu et al. 2011). Our analysis indicates that transcription of these novel TARs shows an even stronger bias for cell-specific expression than transcripts derived from protein coding genes (Figure 4.8C). In this respect, our findings are similar to an earlier report that a majority of unannotated human transcripts are expressed in only one of the eleven different cell lines sampled (Consortium et al. 2007). Although the extent of intergenic transcription from the mammalian genome is controversial (van Bakel et al. 2010), mounting evidence points to multifaceted roles for long intergenic ncRNAs (lincRNAs) including transcriptional control, imprinting, dosage compensation and maintenance and remodeling of chromatin structure (Rinn et al. 2007; Hirota et al. 2008; Wilusz et al. 2009; Tsai et al. 2010). Nevertheless, in every case, definitive tests are required to establish specific functions for candidate regulatory ncRNAs. The tissue-specific patterns of ncRNA expression (Figure 4.8C) that we have revealed for *C. elegans* should

provide a valuable guide to the likely focus of mutant phenotypes that perturb expression of specific ncRNAs (Mercer et al. 2008). We note for example, that the recent discovery of an *in vivo* role for the lincRNA, Evf2, in neuronal differentiation hinged on prior knowledge of Evf2 expression in a specific brain region (Bond et al. 2009).

Although the tiling array results reported here should provide a useful resource for defining the roles of specific genes in cell fate and development, RNA-Seq data derived from these cell specific RNA samples would offer a more accurate representation of gene structure and substantially greater dynamic range for measuring differential gene expression. With the recent development of effective methods for excluding ribosomal RNA from sequencing templates (Armour et al. 2009; Albrecht et al. 2010), it should now be feasible to use RNA-Seq for a direct test of the non-coding RNA transcripts predicted by our tiling array results (see Chapter V). The fact that cell-specific tiling arrays detected predicted coding genes that were not touched by RNA-Seq analysis of *C. elegans* transcripts derived from the whole animal, also suggests that deep sequencing of RNA isolated from individual cell types could reveal additional protein-coding genes.

CHAPTER V

APPLICATIONS OF MASSIVELY PARALLEL SEQUENCING FOR TRANSCRIPTOME PROFILING AND WHOLE-GENOME SEQUENCING IN *C. ELEGANS*

Introduction

The nematode, *Caenorhabditis elegans*, was the first multicellular organism to have its genome fully sequenced (Consortium 1998). The completed genome sequence has served as the basis for efforts to define protein-coding genes, non-coding genes, and other structural elements that are involved in any number of cellular and organismal processes. Sequencing the genome relied on the traditional Sanger method that is low-throughput and has a high cost per read. In the past decade, the method of sequencing by synthesis has provided the basis for new technology that can produce a tremendous number of sequence reads in parallel (Margulies et al. 2005; Shendure et al. 2005; Bentley et al. 2008). There are many applications of this technology, but two of the most popular applications are whole genome (re)sequencing (WGS) and RNA sequencing (RNA-Seq).

WGS enables researchers to isolate genomic DNA from any organism and produce high-quality, high-coverage sequence data at a relatively low cost. Previously, genetic studies relied on complicated, laborious, and time-consuming mapping strategies to identify sequence variants that are produced via

mutagenesis during a genetic screen. With the introduction of high-throughput WGS, researchers can now sequence an individual's genome to identify unique variants, or perform a genetic screen and sequence the genome of individuals with a phenotype of interest (Bentley et al. 2008; Hillier et al. 2008; Sarin et al. 2008; Shen et al. 2008; Doitsidou et al. 2010).

Profiling the entire complement of RNA transcripts produced within whole animals, tissues, and specific cell-types provides a resource for researchers to identify when and where gene products are expressed and make predictions for further experimental testing. Microarrays have been the most widely used method for assaying global gene expression. Due to limited resolution as a consequence of probe length and density, moderate dynamic range, and potential cross-hybridization artifacts, microarrays are being increasingly supplanted by high-throughput sequencing of cDNA libraries as the technology of choice for global expression studies. Currently, methods exist for isolating mRNA, miRNAs, and ncRNAs and for generating cDNA libraries for sequencing on a variety of platforms. It is also possible to directly sequence the 3' end of single RNA molecules, although availability of the sequencing machine required for this highly sensitive approach is limited (Ozsolak et al. 2009; Ozsolak et al. 2010).

A primary concern for RNA sequencing projects is the ribosomal RNA (rRNA) content of the template. While microarrays largely do not have this problem, since RNAs (or cDNAs) can bind to their target probes and produce a fluorescent signal independently of other probes, the reads produced using RNA-

Seq directly correspond to the content of the cDNA library. So if the cDNA library consists of 90% ribosomal RNA, then approximately 90% of reads will map to rRNA transcripts. It is therefore desirable to remove rRNAs from total RNA samples before initiating high throughput sequencing in order to achieve efficient and cost effective sequencing output for lower abundance transcripts derived from other genes. Although rRNAs can be effectively excluded by constructing sequencing libraries from poly-A transcripts, this approach also removes other novel noncoding RNAs, which can have important biological functions.

In this chapter, I will describe my efforts to analyze WGS data for the identification of sequence variants produced by genetic screens performed in the laboratory. The primary focus of this chapter will be on my efforts to deplete rRNA from total RNA preparations of *C. elegans* RNA and apply massively parallel RNA-Seq to identify transcripts, quantify levels of transcription, and perform differential expression analysis.

Materials and Methods

Nematode strains

Animals were grown as described (Brenner 1974). The strains NC1975 [*unc-4*(*e120*); *blr-2*(*wd77*); *blr-1*(*wd76*)], NC1961 [*blr-9*(*wd88*); *unc-4*(*e120*); *wd87*] and NC2135 [*ceh-12*(*gk391*); *unc-4*(e2320); wd87; wd95] were grown and genomic DNA was isolated using a Qiagen genomic DNA isolation kit.

WGS quality control, read mapping, and variant identification

Genomic DNA libraries were prepared using standard Illumina protocols by the Genome Technology Core at Vanderbilt. Each DNA library was

sequenced using one flow cell lane on an Illumina Genome Analyzer IIx. Sequence read quality was analyzed using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). The reads were required to have quality scores >30 and normal GC content (~35-38% GC). Reads were mapped to the *C. elegans* reference genome using MaqGene (Bigelow et al. 2009). Mapped reads were allowed to have up to 3 mismatches and variants were required to have a depth of coverage >4. For each strain, the flat file describing all variants identified was uploaded to a MySQL database. For identifying duplicate variants across strains, the chromosome, base position, and called base had to be identical. For visualization of genome coverage, reads were mapped to the genome using BWA (Li and Durbin 2009). To visualize mapped data, SAM files were converted to BigWig format and uploaded to the UCSC Genome Browser (Kent et al. 2002; Kent et al. 2010; Rhead et al. 2010) or BAM files were viewed in the Integrated Genome Viewer (Robinson et al. 2011).

Ribosomal RNA depletion methods

Ribominus: Total RNA from whole animals was isolated using Trizol and processed with a Ribominus eukaryote kit (Invitrogen) as described by the manufacturer except as noted. One microgram of total RNA was bound to the rRNA probes and passed over the streptavidin-coated beads one time (and not twice as recommended in the protocol for larger amounts of starting total RNA).

RNAse H digestion: First strand cDNA was generated using Superscript II & III (Invitrogen) with the following primers designed for 18S and 26S transcripts:

rrn-1-3' GGTTCACCTACAGCTACCTTGTTAC, 18S-R1 ATCTCGTTATTGCTGCGGTT, *rrn-3.1*-3' AAGGATAGTCTCAACAGATCGCAG, rrn3_R1: CAACTAAGCGACCAGTCACCAA, *rrn-3.1*-R3 TTTCGCCCCTATACCCAAGTC. RNAse H enzyme was obtained from Promega. After RNAse H treatment, DNA was degraded with DNAse I and RNA purified using the DNA-free RNA kit (Zymo Research).

Terminator Exonuclease (Terminator): Whole animal total RNA was treated with Terminator exonuclease (Epicentre) for 30 minutes at 37 °C in a thermal cycler. Epicentre has since altered the protocol to recommend an incubation time and temperature to 60 minutes at 30 °C.

RNA quality analysis

RNA concentrations were measured using UV/Vis spectroscopy (NanoDrop, Thermo Fisher). RNA quality was assayed using the Bioanalyzer 2100 (Agilent).

Quantitative real-time PCR

qPCR was performed as described in Chapter IV. Primers are shown in Table 5.1.

RNA-Seq library construction and sequencing

RNA isolated from the AVA neuron and other cells (see Chapters II & III) was amplified using a developmental version of the NuGEN Ovation RNA-Seq kit. Double-stranded cDNA was generated using a modified Exon module (NuGEN) designed to produce short fragments of ds cDNA (~200 bp) to avoid sonication during library preparation. The final ds cDNA was end-polished, A-

tailed, and adapters were ligated using standard Illumina protocols by the Genome Technology Core at Vanderbilt or by the laboratory of Robert Waterston at the University of Washington. Libraries for the AVA, pan neural, embryonic reference, whole animal, Ribominus, RNAse H, and Terminator samples were sequenced using 1-3 lanes of an Illumina Genome Analyzer II(x). The AVA library was also sequenced using 1 lane on an Illumina HiSeq 2000.

RNA-Seq quality control and read mapping

Sequence read quality was confirmed by FastQC as above. Reads were mapped to the *C. elegans* reference genome using BWA (Li and Durbin 2009) with default settings along with <2 mismatches and a unique alignment. The SAM output files were converted to BAM files using Samtools (Li and Durbin 2009). To visualize mapped data, BAM files were converted to BigWig files and uploaded to the UCSC Genome Browser (Kent et al. 2002; Kent et al. 2010; Rhead et al. 2010).

RNA-Seq transcript quantification, differential expression, and splice site analysis

The SAM files generated from mapping reads to the genome were used as input for Cufflinks and Cuffdiff. The annotation used for assigning mapped reads to gene models was obtained from Ensembl as a GTF file for assembly version WS190 (Hubbard et al. 2002), which matches the reference genome used and allows for visualization in the UCSC Genome browser. Differential expression was tested using Cuffdiff with fold-change ≥2 and false discovery rate (FDR) ≤ 5%. To analyze splice junctions, MapSplice was used with default settings except for the following: minimum intron length ≥1, maximum intron

length ≤ 20,000, and segment length = 25. Splice junctions were output as BED files and upload to the UCSC Genome browser for visualization.

Table 5.1 qPCR primers

rrn-1-5'	GATTGATTCTGTCAGCGCGATATGC
rrn-1_R1	TTGCGTTGGGGTATAGTTG
rrn-1_F1	GTGTCTGCCCTTTCAACTAGAT
rrn-1_R2	TAAGTTTCGCGCCTGCTG
18S-F1	AAGGAGAGGGCAAGTCTGGT
18S-R1	AACCGCAGCAATAACGAGAT
18S-F2	TTCTTCCATGTCCGGGATAG
rrn-1-3'	GGTTCACCTACAGCTACCTTGTTAC
rrn-3.1-5'	AGTCGTGATTACCCGCTGAAC
rrn-3_R4	GGCTCTTCCCGTTTCACT
rrn-3_F1	TTGTGATCGTTGCCGGGT
rrn-3_R2	AACTAAACGCTAGCCGCC
26S-F1	ATTGGTTCAGCCAGAGATGG
26S-R1	CGTTCAAAGAGCACGAGACA
26S-F2	TGTCGGGAGGCATCTCTATC
26S-R2	CGTCGCAGAATTCACTACGA
UNC54-F3	TACCGATCAACTCGGAGAGG
UNC54-R3	CCAAAGCGTGTTGGAGTTCT
UNC54-F4	GCCAACTTGAACCTCCAGAA
UNC54-R4	TCGCATCTTTGAGAGGGAGT
GPD3-F1	GAATCAACGGATTCGGAAGA
GPD3-R1	GACGGCAACAACATTGACAC
GPD3-F2	CAGCTTCCCTCGATGACATT
GPD3-R2	TCCTCGGTGTAAGCGAGAAT

Results

Whole genome re-sequencing of *C. elegans* mutant strains

Another graduate student in the laboratory, Judsen Schneider, performed a genetic screen to identify mutations that suppress the Unc-4 backward movement defect. Many independent mutations were isolated, characterized, and mapped to chromosomes or chromosomal regions (J. Schneider, R. Skelton, and D. Miller, manuscript in preparation). To identify the putative phenotype-causing variation, six strains were selected for whole genome sequencing (see Methods). Strains with mutations located on separate chromosomes were crossed together to create 3 strains with 2 alleles per strain. All 3 strains contained a known mutation in *unc-4* and 2 of these strains also contained a known deletion in *ceh-12* (see Figure 5.1). Genomic DNA was isolated and sequenced on an Illumina Genome Analyzer IIx. Each strain was sequenced using 86 base reads on 1 lane of the flow cell generating approximately 28 million reads and a total of 2.4 gigabases of data per strain. Reads were mapped to the WormBase reference genome (WS190) using the Maq short read alignment program as implemented in MaqGene for variant calling and BWA for coverage (Li et al. 2008; Bigelow et al. 2009; Li and Durbin 2009). See Table 5.2 for sequencing and mapping summary statistics. Roughly 75% of reads mapped to the *C. elegans* genome resulting in an average depth of coverage of 20X. The mutagen used in the screen, ethylmethanesulfonate (EMS), largely induces single base changes, so analysis of variants was focused on single-nucleotide variations (SNVs). Initially, SNV analysis was performed using MaqGene, which

required setup of a MySQL database and Apache webserver running on Linux. Once installed, MaqGene provides a simple web-based interface for initiating the analysis pipeline consisting of read mapping, SNV calling, and determining whether SNVs located in protein-coding genes are likely to be deleterious. Results are generated in tab-delimited text files allowing easy filtering with a spreadsheet program (e.g. Microsoft Excel).

As an initial validation of the sequencing data, I examined whether the known mutations contained in each strain was correctly detected. All three strains contained the *unc-4(e120)* allele, which is a mutation in the splice acceptor of intron 5. The *wd76_wd77* and *wd87_wd95* strains also contained the *ceh-12(gk391)* allele, which is a deletion in the first exon of the gene. Both mutations were easily detected in their respective strains (Figure 5.1). MaqGene correctly called the *e120* splice site mutation as a C->T transition in the splice acceptor (Figure 1A) and the *gk391* deletion as an uncovered region affecting the *ceh-12* gene (Figure 1B). These results provide validation for the quality of the sequence data and utility of the MaqGene software.

A summary of variation information for each strain from MaqGene is presented in Table 5.3. A surprisingly large number of intergenic variants were detected in each strain. It is not clear how many represent real SNPs vs sequencing errors. The most attractive variants (premature stop codons, splicing variants, and non starts) are much more rare and easily filtered by manual inspection of read alignments in a genome browser. It is known that there are a large number of genomic variations in strains cultivated in individual labs

Table 5.2 Sequencing and mapping statistics

Strain	wd76_wd77	wd87_wd88	wd87_wd95
Total reads (million)	30.6	30	32.2
Mapped (million reads/Gb)	27.1/2.3	28.3/2.4	24/2.1
Avg. depth-coverage	23X	24X	21X

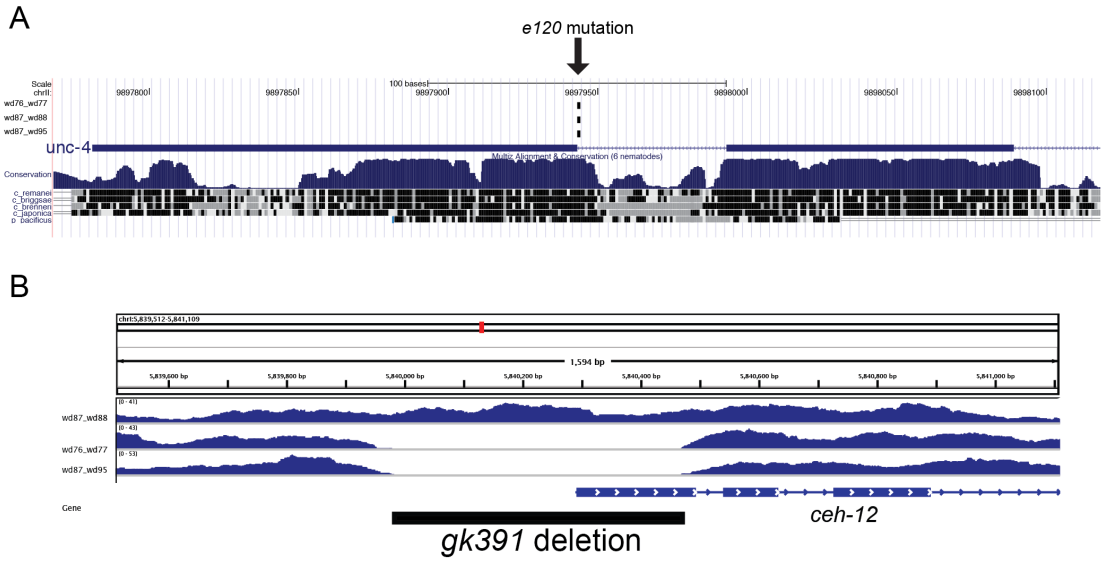


Figure 5.1. Known mutations in *unc-4* and *ceh-12* are accurately detected by WGS.
(A) The *unc-4(e120)* lesion is a C->T point mutation in the splice acceptor of intron 5. (B) The *ceh-12(gk391)* lesion is a 481 bp deletion beginning upstream of the ATG start site through most of the first exon. The *wd87_wd88* strain that did not contain the *ceh-12(gk391)* allele shows normal coverage across the same region.

Table 5.3. Summary of MaqGene variations.

class	wd76_wd77	wd87_wd88	wd87_wd95
5’ UTR	1,307	1,337	1,334
Frame shift	37	56	48
In-frame	11	9	12
Mis-sense	6,328	6,640	6,656
ncRNA	895	924	898
Intergenic	128,403	137,270	133,866
Non start	9	10	10
Premature stop	696	698	698
readthrough	23	29	23
silent	7,479	7,638	7,683
SNP	1,057	1,095	1,098
Splice acceptor	49	53	49
Splice donor	51	50	54
3’ UTR	3,616	3,731	3,690
Uncovered regions	663	1,589	387

compared to the reference *C. elegans* genome (Bentley et al. 2008; Hillier et al. 2008; Sarin et al. 2008; Shen et al. 2008; Doitsidou et al. 2010). Prior to mutagenesis, the parental strain already contains many variations that are not causative for the phenotype obtained through the genetic screen. Therefore, comparing variant information between sequenced strains is necessary to filter out variants that likely existed in the parental strain. To compare the detected variants, I created a MySQL database containing all SNV data generated by MaqGene. Then, each SNV was compared to all other SNVs from each of the three strains sequenced using the genomic location and called base (see Methods). Since I also had genetic mapping data available for each strain (Schneider, Skelton, Miller), SNVs were filtered to the genomic region corresponding to the known genetic interval expanding the interval to the next megabase coordinate (Table 5.4). These filtering steps drastically reduced the number of candidate SNVs for each allele.

The most obvious first candidate SNV for the *wd76* allele is a mutation that creates a premature stop codon early in the first exon of F57G8.7 (Table 5.5, Figure 5.2B). The F57G8.7 gene encodes a protein that is not conserved outside of nematodes and is completely uncharacterized (WormBase). While the lack of conservation and known function does not provide any useful information, F57G8.7 may have a unique function in the nervous system. Another candidate mutation is an uncovered region affecting the *srh-136* gene (Table 5.5, Figure 5.2A). *srh-136* encodes a putative seven transmembrane (7-TM) G-protein coupled receptor (GPCR, WormBase). 7-TM GPCRs are not well conserved, but

Table 5.4 SNVs filtered for uniqueness to that allele and genetic interval.

class	*wd76*	*wd77*	*wd87*	*wd88*	*wd95*
5’ UTR	0	3	0	0	1
In-frame	0	0	0	0	0
Mis-sense	14	1	10	11	16
ncRNA	0	0	0	0	0
Non-genic	29	7	112	4	25
Premature stop	1	0	0	1	0
readthrough	0	0	0	0	0
silent	4	2	3	3	13
SNP	3	1	2	0	3
3’ UTR	0	1	1	1	0
uncovered	12	14	70	125	23

do show functional conservation in some cases (Troemel et al. 1995). *srh-136* is an attractive candidate gene since G-protein signaling is involved in many cellular functions. The best candidate mutation for the *wd77* allele is a missense mutation in the *srsx-14* gene that encodes a 7-TM GPCR (Table 5.5). The mutation is a G->A conversion, which changes a methionine to isoleucine. At this point, it is not obvious how this amino acid change would affect *srsx-14* function or localization, but the possibility of two 7-TM GPCRs found in the same screen is intriguing.

There are two possible candidates for *wd87*. *ymel-1* has mutation in the intron 2 splice donor that would likely generate a truncated protein product. *ymel-1* is a conserved mitochondrial-localized protease, which can suppress alpha-synuclein inclusions in adult animals (van Ham et al. 2008). The other candidate is, *hpr-9*, a homolog of the Rad9 9-1-1 complex subunit. The mutation is a GGA->AGA[Gly->Arg] conversion and it is not clear if this amino acid change would alter function or localization.

The *wd88* allele has two possible candidates. *ccb-2* encodes a beta subunit of the L-type voltage-gated calcium channel (WormBase). The *ccb-2* mutation is a GAT->AAT[Asp->Asn] conversion (Table 5.5). The possibility of a modified calcium channel that suppresses a backward movement phenotype is intriguing. The other candidate is Y71F9AL.17, which encodes an alpha subunit of the coatomer complex (COPI, WormBase). The mutation results in a CTT->TTT[Leu->Phe] change (Table 5.5). Although protein trafficking would be an interesting process to study in the motor circuit, it is not clear how retrograde

Table 5.5 Best candidate genes in mapped regions.

	gene	class	
wd76	F57G8.7	premature stop	
	srh-136	deletion	
wd77	*srsx-14*	missense	ATG->ATA[Met->Ile]
wd87	*ymel-1*	splice donor	
	hpr-9	missense	GGA->AGA[Gly->Arg]
wd88	*ccb-2*	missense	GAT->AAT[Asp->Asn]
	Y71F9AL.17	missense	CTT->TTT[Leu->Phe]
wd95	F35F10.1	missense	ACG->AGG[Thr->Arg]

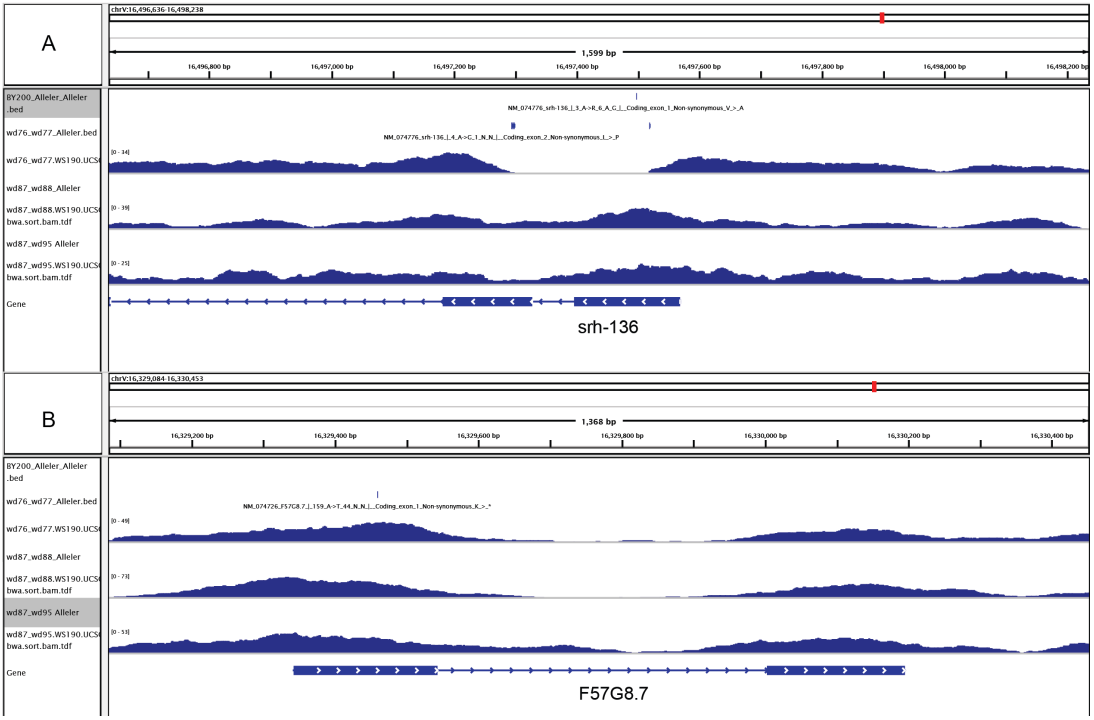


Figure 5.2 *wd76* candidate genes.
(A) *srh-136* shows a deletion from the first to the second exons. *srh-136* encodes a 7-TM GPCR. (B) F57G8.7 has a premature stop codon and encodes an unconserved, uncharacterized protein.

transport from the *trans*-golgi to the *cis*-golgi or ER would play a role in neural specificity.

There is one major candidate for the *wd95* allele. The F35F10.1 gene encodes a nematode-specific predicted protein N-glycanase (WormBase), which has a missense mutation resulting in a threonine to arginine conversion (Table 5.5). Protein N-glycanases are predicted to be involved proteasome-dependent removal of misfolded glycosylated proteins (Suzuki et al. 2000). Since many transmembrane proteins are glycosylated, aberrant localization of membrane-bound proteins could be involved in the suppression of the Unc-4 backward movement phenotype.

rRNA-depletion and single-cell RNA quantification by RNA-Seq

A major goal of the modENCODE project was to identify all RNA transcripts produced from the *C. elegans* genome (Celniker et al. 2009). Over the course of this project, the application of deep sequencing for transcriptome profiling has matured and has become a routine assay of gene expression. As our role in the consortium was to identify transcripts expressed in single cell-types, we standardized on tiling microarrays since the quantities of RNA isolated from rare populations of cells was very low and at the beginning of this project there was no method available to sequence minute quantities of mRNA or non-poly(A) RNA (see Chapter III). As mentioned previously, a major problem with sequencing total RNA is the rRNA content of the sample. If a total RNA sample consists of 90% rRNA transcripts, then approximately 90% of RNA-Seq reads will map to rRNA. To apply RNA-Seq to our minute quantities of RNA, we either

needed to selectively amplify non-rRNA transcripts, or selectively deplete rRNA from our samples. The first option was most readily available due to a long-standing collaboration with NuGEN Technologies (San Carlos, CA). To generate enough cDNA for sequencing, we worked with NuGEN to test developmental versions of their amplification strategy designed for RNA-Seq. NuGEN uses custom primers in their amplification procedure that inhibit amplification of rRNA transcripts. Our previous experience with this amplification strategy detected mRNA transcripts as well as novel transcripts (see Chapter III), suggesting we should be able to apply the same strategy for RNA-Seq and detect all non-rRNA transcripts with greater resolution and sensitivity. To test this strategy, we amplified total RNA from three samples: embryonic AVA neurons, all embryonic neurons, and an all-embryonic cell reference sample. The AVA neuron sample was sequenced using three lanes of a flow cell on an Illumina Genome Analyzer II. The pan neural and reference samples were sequenced using two lanes each. This generated over 11 million reads for the AVA sample, 8 million reads for the pan neural sample, and 9 million reads for the reference sample (Table 5.7). By mapping the reads to the genome, I compared the number of reads mapping to ribosomal DNA regions and all other genomic regions. The reference sample contained 76% rRNA reads, while the AVA and pan neural samples contained 84% rRNA reads, suggesting the NuGEN amplification strategy robustly amplified rRNA transcripts from *C. elegans* RNA samples. This result is unique compared to human and mouse samples that showed very low amounts of rRNA reads (Shawn Levy and NuGEN, personal communication) (Head et al. 2011).

Thus, the NuGen amplification strategy does not efficiently exclude rRNA sequences during amplification of *C. elegans* RNA although this protocol is apparently highly biased against mammalian rRNA amplification. It is not clear whether the NuGEN amplification would amplify rRNA of another invertebrate species, such as *Drosophila*. Because of these results, I decided to test alternative methods for depleting rRNA from our C. elegans total RNA samples.

An option for rRNA depletion became available when Invitrogen released the Ribominus kit. This strategy uses biotin-labeled locked-nucleic acid (LNA) probes complementary to eukaryotic rRNA sequences. When the probes are hybridized to the sample rRNA, the biotin-labeled probe:rRNA transcript complex is bound to streptavidin-coated beads and removed from the remainder of the RNA sample. The standard protocol for this method suggested using > 2 µg of total RNA as input and performing two passes over the avidin-beads. As a first step in scaling the procedure down to lower amounts of input RNA, I used 1 µg of whole animal total RNA and performed one pass over the avidin-beads. The residual RNA (5 ng) was amplified by the NuGEN method (WT-Ovation-Pico). Each of the two RNA samples were sequenced with one lane of the Illumina GAII. 7 million reads were obtained for each sample but the rRNA content of the Ribominus treated sample was not significantly depleted and therefore did not result in a enhanced coverage of mRNA transcripts (Table 5.7).

Since there was no other commercially-available method to remove rRNA transcripts from total RNA samples, I pursued a strategy originally described in the Affymetrix Expression Handbook and applied by Rosenow, et al. 2001 to

bacterial RNA (Affymetrix, Santa Clara, CA. ; Rosenow et al. 2001). RNAse H is a ribonuclease that specifically degrades the RNA component of an RNA:DNA heteroduplex. I designed primers to three regions of the 26S rRNA transcript and two regions of the 18S transcript. The rRNA primers were used for single-strand cDNA synthesis to generate cDNA specifically for rRNA transcripts.This approach should result in the formation of a RNA:DNA duplex strictly for rRNA transcripts but not for other RNA species. This sample was initially treated with RNAse H (Promega) to degrade the rRNA and then with DNAse I to remove the residual rRNA cDNA strands. I tested rRNA depletion by qPCR, which showed a strong reduction for the 5' end of the 18S transcript and moderate reduction in the 3' end of the 26S transcript (Table 5.6). This RNA sample was amplified with the NuGen WT-Ovation-Pico method and sequenced using one lane on an Illumina GAII. Over 6 million reads were generated, but surprisingly, the sample consisted of 99% rRNA reads. One explanation for this result could be non-specific degradation of mRNA by the RNAse H enzyme enriching the abundant rRNA in the sample. The RNA-Seq result contradicts the qPCR result, but the regions targeted for PCR are small and may not provide an accurate assessment for the transcript as a whole. It was not clear that through optimization of this protocol that sufficient rRNA depletion would occur and generate the desired results in a timely manner, so I focused on another method that became commercially available.

A new enzyme called Terminator exonuclease was released by Epicentre that specifically degrades RNA with a 5'-monophosphate. The 5' end of

ribosomal RNA contains a 5'-monophosphate, which would allow it to be degraded. RNAs with a 5'-cap, 5'-hydroxyl, or 5'-triphosphate are not degraded by Terminator exonuclease, thereby allowing mRNAs to remain intact. To test this method, I treated 500 ng of whole animal total RNA with the Terminator exonuclease and used another 500 ng sample of total RNA omitting the enzyme as a mock treatment negative control. The enzyme treatment resulted in a 50% reduction in total RNA quantity as determined by UV/Vis spectroscopy. Prior to submitting for deep sequencing, I used qPCR to measure the relative transcript levels of the 18S and 26S rRNAs and two mRNAs, a muscle myosin *unc-54* and the GAPDH transcript *gpd-1* (Figure 5.3). The rRNA transcripts show a robust decrease of up to 80-fold compared to the mock-treated RNA. The *unc-54* and *gpd-1* mRNA transcripts also showed a corresponding increase of around 5-fold. These results are indicative of a strong depletion of rRNA from the total RNA sample with a relative increase in mRNA levels. Next, we sequenced the amplified cDNA from the Terminator-treated sample using a single lane on an Illumina GAIIx. Over 28 million reads were generated and the rRNA reads were limited to 22% of all mapped reads. With almost 80% of reads mapping to non-rDNA regions of the genome, gene models showed robust coverage and novel transcripts outside of annotated gene models were detected (Figure 5.4). To test this approach on a normal low amount input RNA, I used 1 ng of an embryonic A-class motor neuron total RNA sample. I used the Terminator treatment as described and sequenced the amplified cDNA on one lane of the GAIIx. Approximately 28 million reads were obtained and just over 20% mapped to

rRNA. Oddly, the other 80% of reads did not map to the *C. elegans* genome, but to the *Bacillus anthracis* genome. It is not clear where this template originated. Due to the scant amount of mRNA contained in a 1 ng sample of total RNA, the mRNA transcripts could have been degraded by non-specific Terminator exonuclease activity and/or lost in subsequent RNA purifications. To test whether Terminator exonuclease can degrade rRNA from samples less than 500 ng, I treated 500, 250, 100, 50, and 25 ng quantities of total RNA from a whole animal sample. Using the resulting RNA quantity after enzyme treatment as an indicator, only the 500 ng and 250 ng showed a robust decrease in material (data not shown). This result suggested that additional optimization of the exonuclease treatment protocol would be required to scale down the amount of input RNA used to levels that are necessary for rare cell-specific samples.

Recently, Illumina released a new sequencing machine called the HiSeq 2000. A combination of hardware and sequencing chemistry enhancements allows the machine to generate 80-100 million reads per lane of a flow cell. Based on previous results, we predicted that a sample with 85% rRNA content should yield ~12-15 million reads from one lane on a HiSeq 2000 and thus provide robust coverage of non-rRNA transcripts in that sample. To test this approach, we used one lane on a HiSeq 2000 to sequence the same cDNA library generated from total AVA neuron RNA used previously. Over 80 million reads were obtained with 66% mapping to the genome. The percentage of rRNA reads remained the same at 85%. This resulted in 7.6 million reads mapping to non-rDNA regions of the genome (Table 5.7).

Qualitative inspection of these results in the genome browser showed that genes known to be highly expressed in AVA neurons (e.g., *rig-3*, Figure 5.5A) show robust coverage whereas genes normally restricted to other tissues showed little to no coverage (Figure 5.5b). RNA-Seq results obtained from a poly(A)+ RNA whole embryo sample generated by another group in our modENCODE project is shown (embryo poly(A)+). To quantify gene expression, I used Cufflinks to normalize the AVA HiSeq data with the whole embryo poly(A)+ data (Trapnell et al. 2010; Roberts et al. 2011) (see Methods). To measure transcript abundance, Cufflinks uses the Fragments Per Kilobase of exon per Million fragments mapped (FPKM), which is analogous to RPKM (Mortazavi et al. 2008). The whole embryo poly(A)+ sample detected 14,325 protein-coding genes with FPKM > 0 and the AVA HiSeq sample detected 10,654 protein-coding genes with FPKM > 0 suggesting that sequencing total RNA is able to detect gene expression with sensitivity approaching purified mRNA procedures. To identify what genes are differentially expressed between AVA and the whole embryo reference, I used the Cuffdiff program, which tests for differences in the FPKM log ratio between two samples (Trapnell et al. 2010). Despite only providing one replicate per data set, Cuffdiff identified 3,292 genes as differentially expressed between the AVA and whole embryo reference sample. There are 1,117 genes enriched in AVA vs. the whole embryo with fold change ≥ 2 and 2175 genes relatively depleted with fold change ≤ -2. The AVA enriched genes include many known AVA expressed genes including *flp-18*, *glr-1*, *glr-4*, *glr-5*, *unc-42*, and *rig-3*. Obtaining a whole-embryo reference sample using the

same method as that used to generate the AVA library and producing at least two replicates for each data set, should provide more reliable results. Since we previously quantified gene expression levels using tiling microarrays for AVA (see Chapters II & III), I correlated the AVA RNA-Seq FPKM values vs. the AVA microarray expression values as shown in Figure 5.6. The Spearman's *rho* correlation between the data sets is 0.87 demonstrating very good agreement between the techniques.

To utilize the additional resolution provided by RNA-Seq, I analyzed splice junctions using the program MapSplice (Wang et al. 2010). MapSplice segments sequence reads and maps the segments to the genome. The segments are matched to the original intact read, coverage of the detected splice is computed, and sequence features are analyzed to determine whether the junction represents a canonical or non-canonical splice junction. This method has been shown to be more accurate and sensitive than another splice junction analysis program, Tophat (Wang et al. 2010). In total, the MapSplice algorithm detects 47,879 splice junctions in the AVA total RNA-Seq data and 99,952 junctions in the whole embryo mRNA-Seq data. The embryonic mRNA-Seq data contains 40 million mapped reads (5X the AVA RNA-Seq data), which provides greater depth of coverage across more transcripts than the AVA RNA-Seq data set and allows MapSplice to detect more splice junctions. Optimization of parameters used by the MapSplice algorithm should detect more bona fide splice junctions. Although these data provide support for constitutive splice junctions as well as junctions generated through alternative splicing, the junctions are not assigned to a

transcript isoform, so additional analysis steps are required to provide stronger evidence for expression of particular transcript isoforms.

Discussion

Massively parallel sequencing provides a platform for analysis of the genome at single-base resolution. This technology advances two areas of research in our laboratory: The identification of mutant alleles isolated from genetic screens by whole genome re-sequencing and high-resolution transcriptome analysis of specific cell-types. While the benefit of this technology is readily apparent, evaluating data analysis approaches for whole genome re-sequencing and developing sample preparation methods and data analyses for RNA-Seq has proven challenging.

Methods for mutant allele identification

Genetic screens have provided biologists with the ability to identify mutant variants that alter specific biological processes. While classical methods (*e.g.*, EMS mutagenesis) are still used for generating random mutation, methods for mapping the mutant alleles have evolved greatly. From the previous century to today, researchers rely on genetic markers to perform 2-factor and 3-factor mapping. By crossing the strain carrying an identified allele with another strain carrying one or two markers on a chromosome, genetic intervals where the mutation is located can be defined by analyzing recombination rates between the allele and the known markers. After narrowing the genetic interval to a manageable range, phenotypic non-complementation and/or genetic rescue with genomic fragments can identify the gene where the mutation is located. By

sequencing through the region using traditional Sanger sequencing, the lesion is identified. This process is reliable, but can be labor intensive and slow depending on the phenotype of the isolated mutant. The advent of whole genome re-sequencing can dramatically shorten this process particularly for organisms with compact genomes such as *C. elegans* and *Drosophila* (Sarin et al. 2008; Blumenstiel et al. 2009; Doitsidou et al. 2010).

In this chapter, I have described the implementation of data analysis methods for WGS of *C. elegans*. These methods depend on the generation of high-quality sequencing reads to prevent systematic error from influencing variant calling and sufficient depth of coverage on the genome to sample as much of the genome as possible. The MaqGene pipeline automates a large portion of the analysis from mapping to variant calling and, depending on the size of the data set and target genome, the analysis can be completed in a few hours (Bigelow et al. 2009). However, additional analysis is necessary to remove variants that existed in the parental strain. This filtering step reduces the number of candidate variants from thousands to <100, which can then be easily scanned for likely deleterious mutations (*e.g.,* premature stop codons). While the focus of this chapter has been on data analysis of WGS, there are several interesting aspects of the candidate genes identified and how they can be involved in expression of the Unc-4 backward movement phenotype. The *unc-4* mutation results in a strong backward movement defect due to de-repression of target genes that alter connectivity in the motor circuit. Thus, loss of function mutations in these downstream genes should at least partially restore backward locomotion

to an *unc-4* mutant. Another student in the Miller lab, Rachel Skelton, has shown that one such gene, *goa-1*/Gαo) strongly suppresses Unc-4 movement. A candidate for each of the *wd76* and *wd77* alleles encodes a putative 7-TM GPCR. When GPCRs are activated, they function as guanine exchange factors for the activation of G-proteins. Based on these results, we can formulate the hypothesis that *wd76* and *wd77* are alleles of the GPCRs, *srh-136* and *srsx-14*, respectively, and function to activate the Gαo, *goa-1*. Therefore, loss of function mutations in either the GPCR genes or in *goa-1* should result in Unc-4 suppression. Although additional experiments are required to test these predictions, the development of WGS methods for detecting genetic variants should provide a rapid and efficient strategy for the identification of authentic *unc-4* pathway genes.

Development and application of RNA-Seq methods for cell-specific transcriptome profiling

The Miller laboratory has had a long-standing interest in utilizing global gene expression profiling for identifying candidate genes involved in specific cellular functions. The relative ease of creating transgenic animals and the implementation of advanced molecular genetic tools has provided the opportunity to study gene expression in specific cell-types from embryonic stages through adulthood (Fox et al. 2005; Fox et al. 2007; Stetina et al. 2007; Watson et al. 2008; Smith et al. 2010; Spencer et al. 2011). These studies have relied on the Affymetrix microarray platform for measuring expression of annotated protein-coding genes (Fox et al. 2005; Fox et al. 2007; Von Stetina et al. 2007; Watson et al. 2008; Smith et al. 2010) as well as for detecting transcription from non-

coding sequence regions (Spencer et al. 2011). One of the major criticisms of microarrays is the potential for artifactual signals due to non-specific cross-hybridization (Wu et al. 2005; Zhang et al. 2005). Additionally, the resolution of microarrays is limited to probe density and at present, stands at 25 nt for the Affymetrix *C. elegans* tiling array. The application of RNA-Seq for transcriptome profiling largely removes these limitations. To date, most RNA-Seq studies have focused on the mRNA fraction of the transcriptome by purifying poly(A)+ RNA (Mortazavi et al. 2008; Hillier et al. 2009; Gerstein et al. 2010; Graveley et al. 2011). These studies obtain RNA samples from cell-lines or whole animals that provide relatively large quantities (> 1 µg) of total RNA for purification of the poly(A)+ fraction. Because much smaller quantities of total RNA can be feasibly isolated from specific C. elegans cell types (< 10 ng), it was necessary to develop alternative strategies for obtaining useful RNA-Seq results from these samples. In this chapter, I have described the empirical testing of several methods for depleting rRNA from total RNA samples. These techniques still routinely require quantities of input RNA that preclude the use of RNA from rare cell-types and currently only the Terminator exonuclease method has succeeded in efficiently reducing rRNA levels from *C. elegans* whole animal RNA (Figure 5.4, Table 5.7). Thus, additional effort will be required to deplete rRNA from individual cell-type samples while leaving all other RNAs intact.

The release of the Illumina HiSeq 2000 has partially obviated the need for rRNA-depletion strategies through brute-force sequencing of a complete total

Table 5.6 Real-time PCR analysis of RNAse H depletion of rRNA.

	- RNAse H	+ RNAse H	
amplicon	Ct	Ct	Fold change
gpd-3 5' end	28.3	27	+2.5
gpd-3 3' end	28.4	27.5	+1.9
unc-54 5' end	28.6	27.7	+1.9
unc-54 3' end	26.4	26.1	+1.2
18S 5' end	10.9	16.8	-60
18S 3' end	9.5	13.2	-13
26S 5' end	13	13.7	-1.6
26S 3' end	12.2	17.1	-30

Table 5.7 Summary of sequencing reads for rRNA depletion experiments and rRNA content.

Sample	# of reads	% rRNA
AVA neuron GAIIx	11.5M	84
pan neural	8.1M	84
reference	9.1M	76
Total RNA	7.1M	97
1-pass Ribominus	7.2M	85
RNAse H	6.6M	99
Terminator exonuclease	28M	22
AVA neuron HiSeq 2000	80M	85

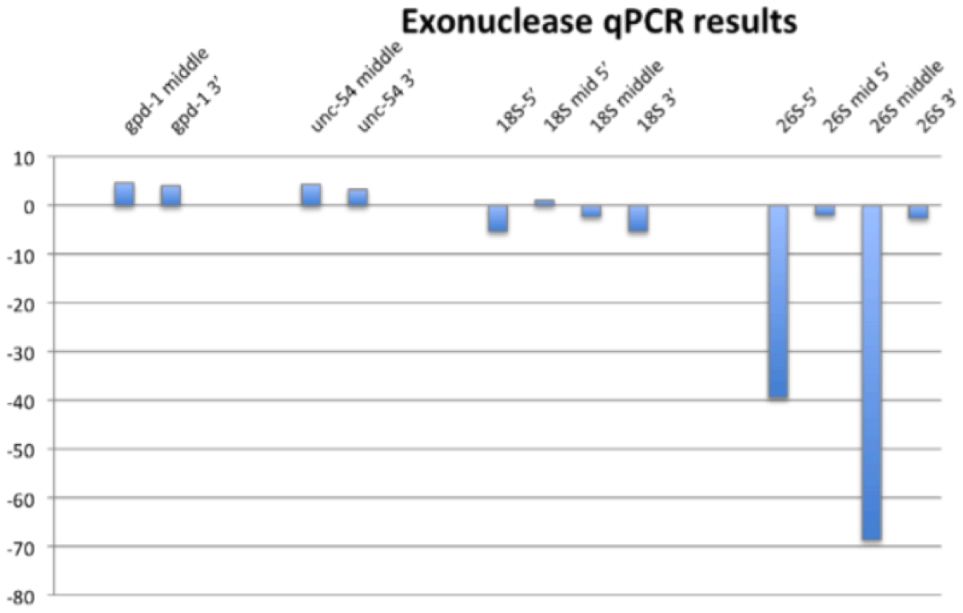


Figure 5.3. Quantitative PCR analysis of transcript levels in Terminator exonuclease-treated vs. mock-treated total RNA.

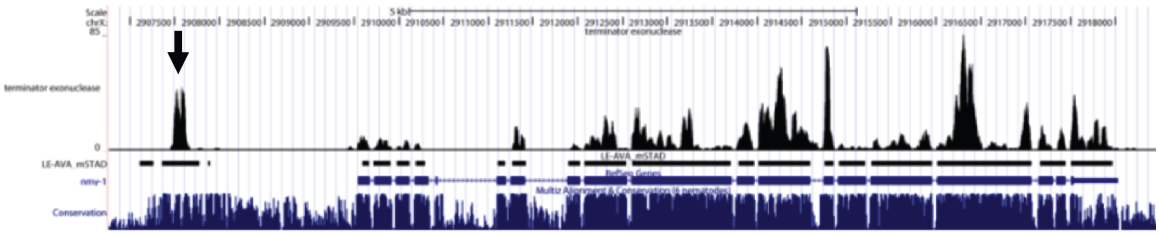


Figure 5.4. RNA-Seq analysis of rRNA-depleted whole animal total RNA. The non-muscle myosin gene *nmy-1* shows robust coverage and a novel transcript is detected in the 5' upstream region, which corresponds to a novel TAR detected from tiling array analysis of the embryonic AVA neuron (arrow).

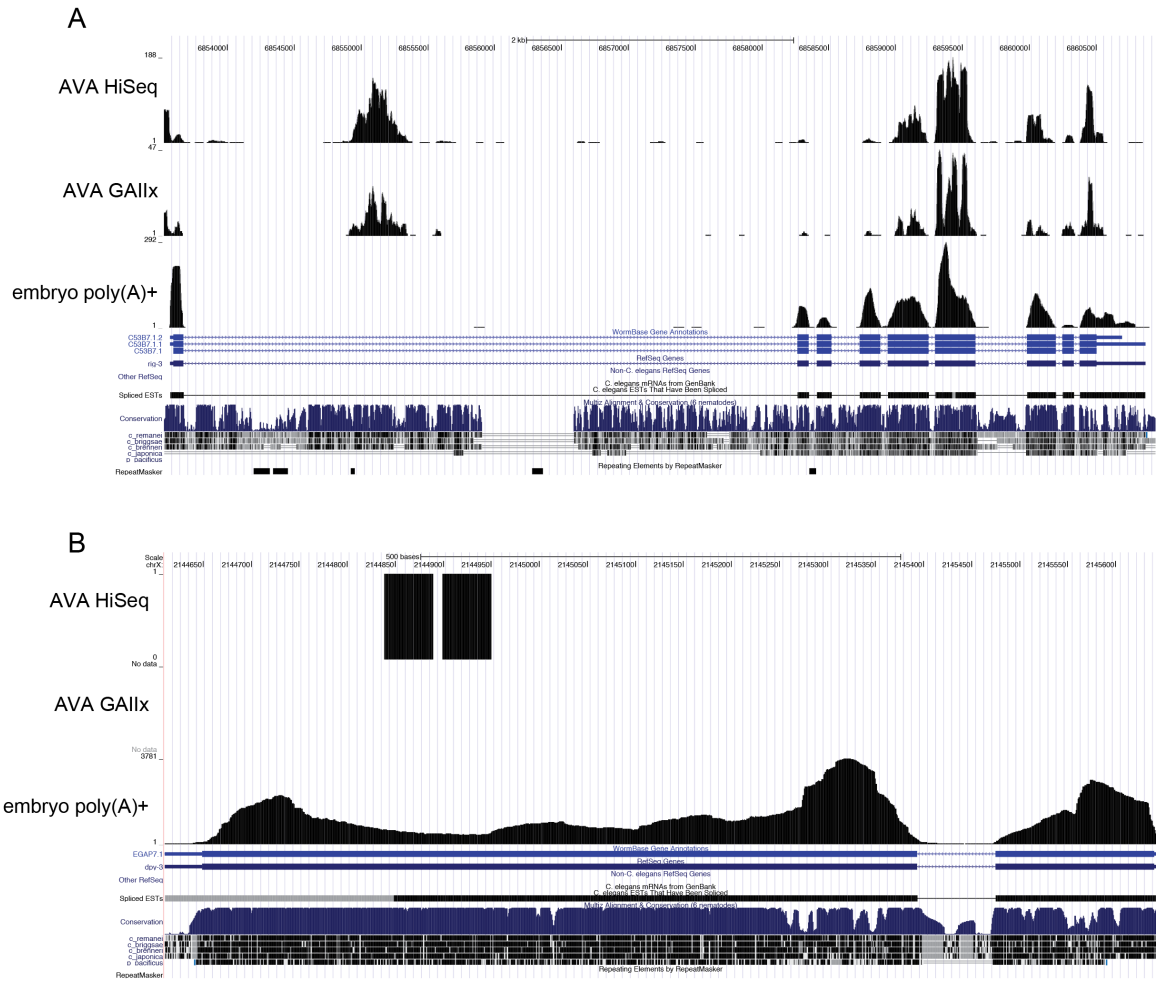


Figure 5.5 Comparison of gene model coverage for an AVA-enriched gene (A) *rig-3* and a hypodermal enriched gene (B) *dpy-3*. A novel transcript is detected in the first intron of *rig-3* that was initially found in the AVA tiling array data. Note that the *dpy-3* transcript is relatively more abundant in the total embryo poly(A)+ RNA-Seq data than in the AVA results that includes only two reads.

RNA sample. By sequencing a cDNA library generated from the AVA command interneuron that comprises 0.35% of all cells in the embryo (2 neurons/550 cells), I have provided evidence that RNA-Seq can accurately detect transcripts from a rare cell-type. Although a single method for expression quantification and differential expression analysis was presented (Cufflinks/CuffDiff), several alternatives exist that must be compared to determine the optimal statistical approach depending on the experiment being performed (Anders and Huber 2010; Bohnert and Rätsch 2010; Gao et al. 2010; Wang et al. 2010). Approaches used for normalization of RNA-Seq data are also maturing. In this work, I used upper-quartile normalization on the AVA and embryonic RNA-Seq data. This method removes all transcripts with zero read counts, and the read counts for transcripts expressed in the highest quartile (75^{th} percentile) are used to scale transcript expression values of all transcripts in each data set being normalized together. This allows comparison between data sets generated independently and with little to no bias (Bullard et al. 2010).

A major advantage of RNA-Seq analysis is the ability to determine the actual transcript isoform being expressed along with splice site usage. A specific isoform may be expressed and function in one cell-type, while an alternative isoform is expressed and functions in another cell-type. By performing RNA-Seq analysis on each cell-type, isoform expression can be determined. Then, additional experiments can be designed to test for function and determine the regulatory elements necessary for control of cell-specific expression and alternative splicing. As a first step towards this goal, I analyzed splice junctions in

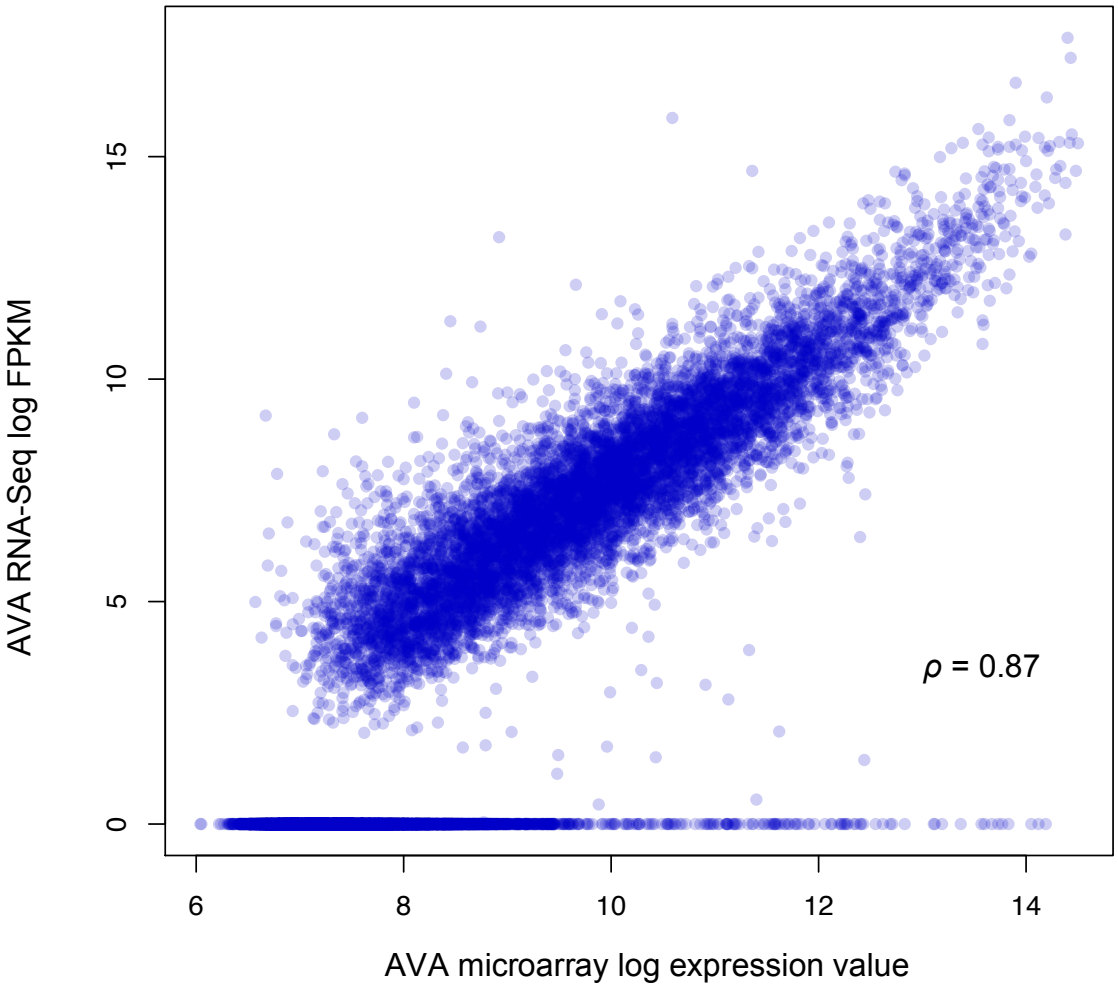


Figure 5.6 Correlation of AVA RNA-Seq vs. microarray gene expression values. Many genes are not detected as expressed in the RNA-Seq data resulting in zero values that have a corresponding expression value > 6 in the microarray data set.

the AVA neuron data and whole embryo data. Many junctions are detected in each data set and typically correspond to known splice junctions (Figure 5.7). Additional analyses are necessary to determine whether the detected splice junctions validate known and predicted splice junctions and the transcript isoform associated with the junction. Also, the junctions detected in the AVA data can be compared to the whole embryo data to identify junctions that are uniquely detected in the AVA neuron. As additional RNA-Seq data for specific cell-types is generated, it will be interesting to transcript-level expression differences and alternative splicing patterns to provide further evidence for gene products that function in each cell-type.

Massively parallel sequencing has provided a tremendous opportunity to advance several aspects of biomedical research. Through whole-genome sequencing, unique genomic variations can be identified in isogenic strains of *C. elegans* as well as individual humans. These variations can be correlated with phenotypes and tested experimentally to delineate the molecular determinants of the phenotype. RNA-Seq allows researchers to sample the dynamics of gene expression, from the production of an unspliced transcript to an mRNA being actively translated, with unprecedented resolution. While sample preparation methods have matured, the subsequent data analysis continues to develop. As occurred with microarray technology, new algorithms will be developed that take advantage of the available data and will likely extract additional useful information.

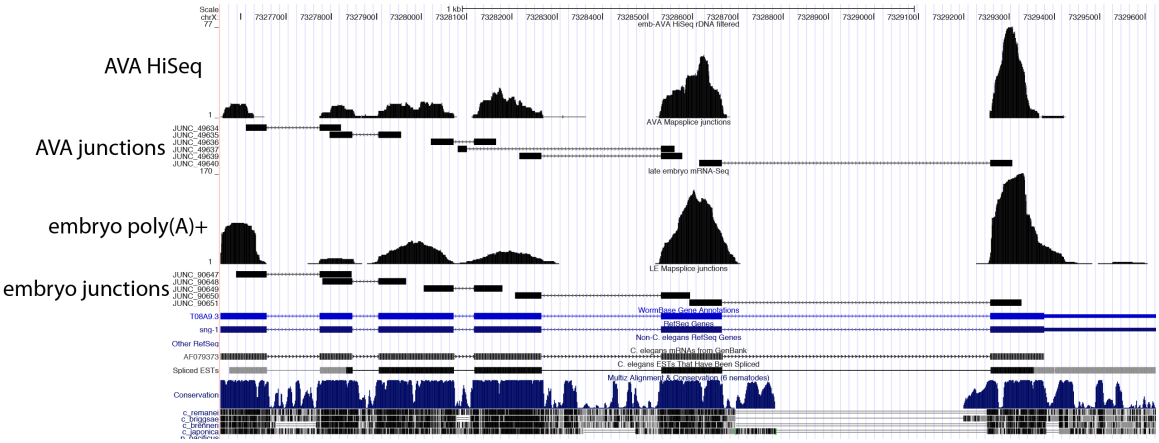


Figure 5.7 Splice junction analysis of RNA-Seq accurately identifies known splice sites.
The example shows the gene model of *sng-1*, the Synaptogyrin ortholog. All known splice junctions are detected in the AVA and whole embryo RNA-Seq data.

CHAPTER VI

GENERAL DISCUSSION AND FUTURE DIRECTIONS

The goal of this dissertation project was to define and analyze gene expression profiles of single cell-types and to identify molecules that control synaptic specificity in the *C. elegans* motor circuit. My work presented here describes an enhancement of the MAPCeL method to facilitate isolation of single cells that cannot otherwise be marked by the use of a single promoter. This advance allowed the gene expression profiling of two interneurons, AVA and AVE that drive backward locomotion in *C. elegans*. These expression profiles revealed strong expression of a transcript encoding the RIG-3 adhesion molecule which I then demonstrated by genetic analysis is required for the creation of synapses between AVA and the A-class motor neurons in the backward movement circuit. To expand our knowledge of animal development and cellular function, the application of cell-specific expression profiling and advanced data analyses to over 30 cell-types provides a detailed view of transcriptional activity across the anatomy of *C. elegans*. The introduction of new technology allows finer grained analysis of cellular and molecular processes. The application of second generation sequencing methods enables gene expression profiling at single-nucleotide resolution. This discussion will highlight how applying advanced methods enhances our understanding of cellular function and identity.

Profiling the AVA and AVE command interneurons

The Miller laboratory previously implemented a method for cell-specific gene expression profiling of *C. elegans* embryonic cells that relied on expressing a fluorescent reporter in the cell of interest (Fox et al. 2005; Fox et al. 2007). Because the command AVA command interneuron lacked a cell-specific promoter, I optimized the use of a unique combination of two fluorescent reporters to mark each one for isolation by FACS (Chapter II). To apply this approach to another neuron, I worked with Rebecca McWhirter to profile the AVE neuron. Both expression profiles show enrichment for neuronal transcripts and many of the known AVA and AVE expressed genes. Since AVA and AVE function as interneurons, they are expected to receive a synaptic input from a variety of sensory neurons and other interneurons. Indeed, AVA and AVE express a variety of neurotransmitter receptors including: glutamate, acetylcholine, GABA, dopamine, serotonin, and neuropeptide receptors. The evidence for how AVA and AVE communicate with postsynaptic neurons is less clear with the exception of peptide signaling. Both AVA and AVE express a number of neuropeptide and FMRFamide-like peptides that could activate peptide receptors on the A-class motor neurons. Neuropeptides have been shown to have effects on pharyngeal pumping in *C. elegans* and control feeding behavior in *Aplysia* (Sweedler et al. 2002; Papaioannou et al. 2005). Typically, neuropeptide receptors function as GPCRs to modulate the function of neurotransmitter receptors, thus it is possible that peptidergic signaling could regulate backward locomotion (Liu et al. 2003).

A major motivation for profiling AVA and AVE was to identify molecules that could be involved in synaptic specificity. Previous efforts in the lab have focused on pathways functioning in the A-class motor neurons. To form synaptic connections between the A-type command interneurons (AVA, AVD, AVE) and the A-class motor neurons, the interneurons may secrete a signal or present an adhesion molecule which when bound by the corresponding receptor expressed in motor neurons would initiate synaptic formation. AVA and AVE express several adhesion molecules, which could play a role in axon guidance or synaptic connectivity. Both neurons express the *zig-8*, *lad-2*, and *rig-3* IgCAMs. The *zig-8* gene encodes a secreted protein with two Ig domains (Benard et al. 2009). A *zig-8* mutant was tested for axon guidance defects in several classes of neurons, but did not have a significant effect. The command interneurons were not tested, so it would be interesting to ask if *zig-8* is involved in AVA/AVE axon guidance or synaptic connectivity. *lad-2* encodes the single L1CAM homolog in *C. elegans* and is required for normal axon guidance in the SMD, PLN, and SDQ neurons (Wang et al. 2008). Although AVA and AVE were not tested, *lad-2* could have a similar role in the command interneurons. GFP reporter results initially suggested that *rig-3* is expressed in AVA (WormBase). The expression profiles generated in this work confirmed expression of *rig-3* in AVA and also showed expression in AVE. The existing *rig-3* GFP-reporter, however did not express in AVE which suggests that the promoter element used in the rig-3::reporter lacks regulatory elements necessary for AVE expression.

RIG-3 is required for synapse formation in the motor circuit

Roger Sperry proposed the chemoaffinity hypothesis to explain how neurons correctly identify their postsynaptic partners and form neural circuits (Sperry 1963). The hypothesis proposes that neurons express specific combinations of molecules on their surface that act as identifiers that effectively distinguish each neuron from every other neuron in the same region. The neurons that form synaptic connections become attached by affinity of the unique molecules expressed on their surface. Later, cell adhesion molecules were proposed to perform the role of the identifier molecules. Further experimental analysis has shown that some adhesion molecules, such as the sidekicks, SynCAM, and SYG-1/SYG-2, do function as synaptic specificity determinants (Biederer et al. 2002; Yamagata et al. 2002; Shen and Bargmann 2003; Shen et al. 2004). Given this evidence, RIG-3 could act as a unique determinant of connectivity between A-type command interneurons (AVA, AVE) and the A-class motor neurons (Chapter III). I showed that AVA synapses to A-Class motor neurons are disrupted in *rig-3* mutants. Although AVA axonal morphology appears normal in the *rig-3* mutant the placement of A-class neuron processes in the ventral cord has not been examined directly and therefore could potentially account for the disruption of AVA to A-class synaptogenesis. Other neuron-specific synapses in the ventral nerve cord should also be tested to distinguish between the possibilities that RIG-3 is selectively required for synapses in the backward motor circuit *vs* having a more general role in synaptogenesis. In any case, the finding of a novel cell adhesion molecule that is required for

synaptogenesis is exciting and promises to reveal a potentially new mechanism for the creation of synapses in a specific neural circuit.

Global analysis of gene expression across the anatomy of *C. elegans*

Gene expression profiling has proven to be an invaluable method for building, refining, and testing experimental hypotheses. As shown in Chapter IV, we have defined gene expression patterns for all genes across many tissues and cell-types (Spencer et al. 2011). This information provided two surprising findings: about 75% of genes are differentially expressed and over 10 Mb of intergenic sequence in the genome is actively transcribed. The number of differentially expressed genes reveals the complexity of gene regulation and the need for such regulatory control for multicellular organism development. As more single cell expression profiles are obtained, the number of differentially expressed genes can only increase. For example, the BAG neurons are a left-right pair of sensory neurons that respond to CO_2 (Hallem and Sternberg 2008). We collaborated with the Sternberg lab to profile the BAG neurons to identify the receptor necessary for CO_2 sensation. One of the most highly enriched transcripts in the BAG neuron profile encodes a receptor-type guanylate cyclase, GCY-9. The Sternberg lab used this clue to show that *gcy-9* is required for CO_2 sensation and likely corresponds to the CO_2 receptor (Hallem et al. 2011). This gene is uniquely enriched in the BAG neurons and demonstrates how specifically genes can be expressed (see Figure 6.1).

Figure 6.1 The BAG neurons uniquely express the receptor-type guanylate cyclase gene, *gcy-9*.
The red line indicates normalized gene expression values for all microarray data sets analyzed in Chapter IV. Sample names are on the left. For most data sets, the expression value for *gcy-9* is around the mean value of 1, except for the BAG neurons, where the expression value is 8.6.

rRNA-depletion strategies and RNA-Seq analysis of the AVA command interneuron

To further refine our gene expression profiling efforts, I evaluated several rRNA-depletion strategies for RNA-Seq (Chapter V). These methods proved difficult to scale to low amounts of input RNA and with the exception of Terminator exonuclease failed to efficiently deplete *C. elegans* rRNA from total RNA samples. With additional effort, it is possible the Terminator exonuclease reaction could be scaled down to use modest amounts of cell-specific RNA. Additional methods for rRNA depletion have been released by Illumina (DSN) and Epicentre (Ribo-Zero). The Illumina DSN method performs cDNA library normalization using a duplex-specific nuclease. Double-stranded nucleic acids will anneal in a concentration-dependent manner. The most abundant molecules will anneal first followed by less abundant molecules. Since rRNA will represent over 90% of the cDNA library, the rRNA ds cDNA is allowed to anneal, then the sample is treated with a duplex-specific nuclease to degrade the annealed cDNA. This method was tested by Valerie Reinke and Bob Waterston, our collaborators in the modENCODE project but only moderate depletion of rRNA was achieved. The Epicentre Ribo-Zero method adopts a similar strategy to the Invitrogen Ribominus approach. Ribo-Zero uses probes for rRNA attached to microbeads to extract rRNA from a solution of total RNA. At this point, it is not clear if this method would perform better than the Ribominus method, which in my hands was not very efficient for depleting *C. elegans* rRNAs.

By sequencing the AVA cDNA library using the Illumina HiSeq 2000, I showed that generating a high number of sequence reads (80 million) partially

negates the need to deplete rRNA. However, since most of the reads (~85%) generated by this approach are discarded, an effective method of rRNA depletion would certainly reduce costs by allowing the use of sample-specific barcodes to for multiplex sequencing. In addition, the number of non-rRNA reads generated (~8M) is likely not sufficient to accurately measure the transcriptome of a single cell. We have noted uneven coverage across gene models, but this may be due to bias by the NuGEN WT-Pico amplification. As seen in the examples for the whole embryo samples in Chapter V, the unamplified sample shows much more even coverage across gene models. Only 2 ng of ds cDNA library is necessary for RNA-Seq, so it may be possible to directly synthesize ds cDNA from our cell-specific RNA samples and therefore avoid this potential artifact of RNA amplification.

Despite the need for increased RNA-Seq coverage, the tiling array results are highly correlated with the AVA RNA-Seq data (Figure 5.6) which suggests that RNA-Seq is likely as sensitive as microarrays. Sequencing to a greater depth should increase the sensitivity for rare transcripts. For certain experiments, it is not necessary to detect ncRNAs (noncoding RNAs). For these circumstances purifying or selectively amplifying poly(A) RNAs would be sufficient. A method for amplifying mRNAs from single cells has been developed for the ABI SOLiD platform (Tang et al. 2010). We are collaborating with Kris Gunsalus and Paul Scheid at NYU to evaluate this method for our cell-specific samples. This protocol uses oligo(dT) primers and limited PCR amplification to produce a cDNA library. A similar method developed by Clontech and Illumina uses ligation-

dependent PCR to amplify mRNA transcripts (http://www.clontech.com). Since both methods initially use oligo(dT) primers, it will likely be important to compare them to determine the most accurate and quantitative procedure.

Future Directions

The identification of RIG-3 as a positive regulator of AVA to A-class synapse formation suggests the AVA/AVE expression profiles contain other molecules necessary for synapse formation and neuronal function. The expression profiles could be screened by RNAi for genes with defects in synapse formation. The A-class motor neuron expression profiles (Fox et al. 2005; Von Stetina et al. 2007) also identify candidate adhesion molecules that could interact with RIG-3 or the other adhesion molecules expressed in AVA and AVE. As mentioned in Chapter III, embryonic A-class neurons are enriched for *rig-3*. Although *rig-3* is not enriched in the larval A-class neuron profile two other IgCAM genes, *syg-1* and *rig-6* are highly expressed (Von Stetina et al. 2007). *syg-1* is the IgCAM expressed in the HSN motor neuron that innervates vulval muscle. *syg-1* interacts with *syg-2* expressed in surrounding epithelial cells to define the location of the HSN neuromuscular junction (Shen and Bargmann 2003; Shen et al. 2004). *rig-6* is the homolog of contactin and is required for septate junction formation in *Drosophila* and cerebellar development and neurite formation in mouse (Berglund et al. 1999; Falk et al. 2002; Hu et al. 2003; Faivre-Sarrailh et al. 2004; Katidou et al. 2008). Since *rig-3* is not expressed in larval A-class motor neurons, it could be possible that RIG-3 is required in AVA and DA motor neurons for initiation of synaptic connectivity, but when VA motor neurons

are born in the second larval stage (L2), RIG-3 (or LAD-2) in AVA could interact with SYG-1 or RIG-6 in VAs to initiate AVA to VA synaptic connectivity.

The AVA and AVE command interneuron expression profiles will provide a foundation for future studies in the *C. elegans* motor circuit. To further our understanding of how the motor circuit forms, the remaining command interneurons could be profiled with a variety of methods. For example, I have generated a strain that uniquely marks the PVC forward movement command interneuron (see Figure 6.2). PVC innervates the B-class motor neurons to drive forward locomotion (White et al. 1986). Comparing differences between the A-type and B-type command interneuron expression could reveal mutually exclusive molecules that control synaptic specificity. The AVD and AVB command interneurons are more difficult to mark using available reporters. Testing the combinations suggested in Chapter II could provide a useful strain for these purposes. Additionally, now it should be possible to isolate these neurons using FACS based on cell recently developing protocol for dissociating cells from *C. elegans* larvae (Zhang et al. 2011). Previously, the tough cuticle of larval stage animals prevented isolation of intact cells. With a combination of detergent and the protease Pronase, the cuticle is gently degraded and viable larval cells are released. We have begun to test this approach with larval muscle cells and the serotonergic neuron NSM (Rebecca McWhirter, Clay Spencer, David Miller). Our results have confirmed that body muscle cells can be easily isolated as previously reported. Although the NSM neuron is embedded in the muscular pharynx suggesting it might be difficult to extract, GFP-positive NSM neurons are

easily identifiable in our cultures. Due to the ease of this procedure, it will likely supplant mRNA-tagging for larval cell profiling in most experiments.

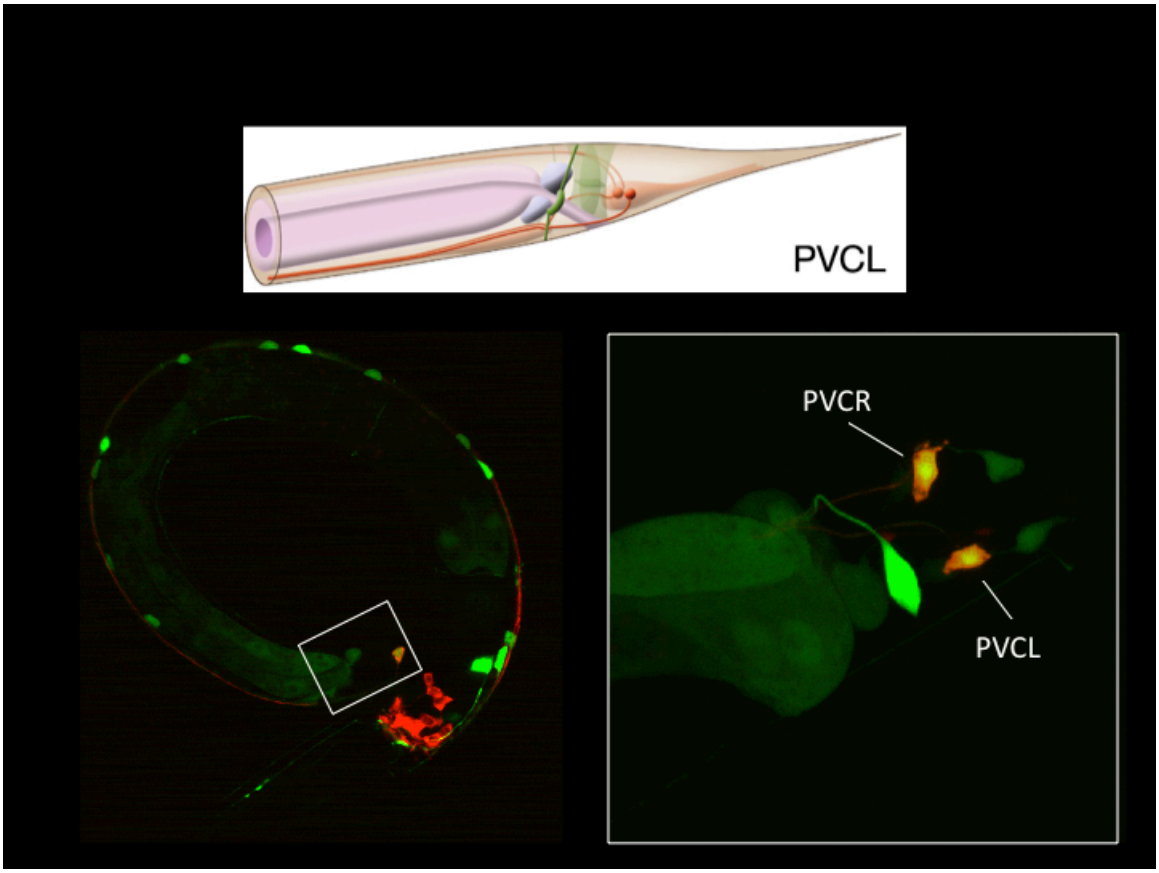


Figure 6.2 P*flp-11*::GFP;P*glr-1*::DsRed2 uniquely mark the PVC command interneuron.
(top) PVC cell bodies are located in the tail and send processes in the ventral nerve cord to the nerve ring in the head. (bottom left) Whole animal view P*flp-11*::GFP;P*glr-1*::DsRed2 expression pattern in a larval animal. (bottom right) Magnified view of PVC cell bodies in the tail showing co-expression of GFP and DsRed2.

REFERENCES

Abbas L. 2003. Synapse formation: let's stick together. *Curr Biol* **13**(1): R25-27.

Abouelhoda M. 2004. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms* **2**: 53-86.

Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T. 2010. Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic acids research* **38**: 868-877.

Altun ZF, Herndon, L.A., Crocker, C., Lints, R. and Hall, D.H. (ed.s). 2002-2010. WormAtlas.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**: R106.

Antoine JC, Ternynck T, Rodrigot M, Avrameas S. 1978. Lymphoid cell fractionation on magnetic polyacrylamide--agarose beads. *Immunochemistry* **15**(7): 443-452.

Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM et al. 2009. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**: 647-649.

Armstrong KR, Chamberlin HM. 2010. Coordinate regulation of gene expression in the C. elegans excretory cell by the POU domain protein CEH-6. *Mol Genet Genomics* **283**(1): 73-87.

Azzaria M, Goszczynski B, Chung MA, Kalb JM, McGhee JD. 1996. A fork head/HNF-3 homolog expressed in the pharynx and intestine of the Caenorhabditis elegans embryo. *Developmental Biology* **178**(2): 289-303.

Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli AE. 2005. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**(4): 553-563.

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC genomics* **7**: 246.

Bamji SX, Shimazu K, Kimes N, Huelsken J, Birchmeier W, Lu B, Reichardt LF. 2003. Role of beta-catenin in synaptic vesicle localization and presynaptic assembly. *Neuron* **40**(4): 719-731.

Baran R, Aronoff R, Garriga G. 1999. The C. elegans homeodomain gene unc-42 regulates chemosensory and glutamate receptor expression. *Development* **126**(10): 2241-2251.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2): 215-233.

Benard C, Tjoe N, Boulin T, Recio J, Hobert O. 2009. The Small, Secreted Immunoglobulin Protein ZIG-3 Maintains Axon Position in Caenorhabditis elegans. *Genetics* **183**(3): 917-927.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**: 289-300.
Benson DL, Tanaka H. 1998. N-cadherin redistribution during synaptogenesis in hippocampal neurons. *J Neurosci* **18**(17): 6892-6904.
Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.
Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics* **37**(7): 766-770.
Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC. 2007. Mammalian mirtron genes. *Mol Cell* **28**(2): 328-336.
Berglund EO, Murai KK, Fredette B, Sekerkova G, Marturano B, Weber L, Mugnaini E, Ranscht B. 1999. Ataxia and abnormal cerebellar microorganization in mice with ablated contactin gene expression. *Neuron* **24**(3): 739-750.
Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705): 2242-2246.
Biederer T, Sara Y, Mozhayeva M, Atasoy D, Liu X, Kavalali ET, Südhof TC. 2002. SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science* **297**(5586): 1525-1531.
Bigelow H, Doitsidou M, Sarin S, Hobert O. 2009. MAQGene: software to facilitate C. elegans mutant genome sequence analysis. *Nature methods* **6**: 549.
Blacque OE, Perens EA, Boroevich KA, Inglis PN, Li C, Warner A, Khattra J, Holt RA, Ou G, Mah AK et al. 2005. Functional genomics of the cilium, a sensory organelle. *Curr Biol* **15**(10): 935-941.
Blencowe BJ. 2006. Alternative splicing: new insights from global analyses. *Cell* **126**(1): 37-47.
Blumenstiel JP, Noll AC, Griffiths JA, Perera AG, Walton KN, Gilliland WD, Hawley RS, Staehling-Hampton K. 2009. Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics* **182**: 25-32.
Bohnert R, Rätsch G. 2010. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic acids research* **38 Suppl**: W348-351.
Bolstad BM, Irizarry Ra, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* **19**: 185-193.
Bond AM, Vangompel MJ, Sametsky EA, Clark MF, Savage JC, Disterhoft JF, Kohtz JD. 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**(8): 1020-1027.

Bonferroni CE. 1935. *Il calcolo delle assicurazioni su gruppi di teste.*, Rome, Italy.
Bonner WA, Hulett HR, Sweet RG, Herzenberg LA. 1972. Fluorescence activated cell sorting. *Rev Sci Instrum* **43**(3): 404-409.
Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ. 2010. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* **11**: 282.
Bradnam KR, Korf I. 2008. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3**(8): e3093.
Breitling R, Armengaud P, Amtmann A, Herzyk P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**: 83-92.
Brenner S. 1974. The genetics of Caenorhabditis elegans. *Genetics* **77**(1): 71-94.
Brockie PJ, Madsen DM, Zheng Y, Mellem J, Maricq AV. 2001a. Differential expression of glutamate receptor subunits in the nervous system of Caenorhabditis elegans and their regulation by the homeodomain protein UNC-42. *J Neurosci* **21**(5): 1510-1522.
Brockie PJ, Maricq AV. 2003. Ionotropic glutamate receptors in Caenorhabditis elegans. *Neurosignals* **12**(3): 108-125.
Brockie PJ, Mellem JE, Hills T, Madsen DM, Maricq AV. 2001b. The C. elegans glutamate receptor subunit NMR-1 is required for slow NMDA-activated currents that regulate reversal frequency during locomotion. *Neuron* **31**(4): 617-630.
Budovskaya YV, Wu K, Southworth LK, Jiang M, Tedesco P, Johnson TE, Kim SK. 2008. An elt-3/elt-5/elt-6 GATA Transcription Circuit Guides Aging in C. elegans. *Cell* **Vol 134**: 291-303.
Bukalo O, Fentrop N, Lee AY, Salmen B, Law JW, Wotjak CT, Schweizer M, Dityatev A, Schachner M. 2004. Conditional ablation of the neural cell adhesion molecule reduces precision of spatial learning, long-term potentiation, and depression in the CA1 subfield of mouse hippocampus. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **24**(7): 1565-1577.
Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11**: 94.
Burglin TR, Ruvkun G. 2001. Regulation of ectodermal and excretory function by the C. elegans POU homeobox gene ceh-6. *Development* **128**(5): 779-790.
Caday CG, Klagsbrun M, Fanning PJ, Mirzabegian A, Finklestein SP. 1990. Fibroblast growth factor (FGF) levels in the developing rat brain. *Brain research Developmental brain research* **52**(1-2): 241-246.
Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA et al. 2008. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *The Journal of*

neuroscience : the official journal of the Society for Neuroscience **28**(1): 264-278.
Cai X, Hagedorn CH, Cullen BR. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna* **10**(12): 1957-1966.
Calella AM, Nerlov C, Lopez RG, Sciarretta C, von Bohlen und Halbach O, Bereshchenko O, Minichiello L. 2007. Neurotrophin/Trk receptor signaling mediates C/EBPalpha, -beta and NeuroD recruitment to immediate-early gene promoters in neuronal cells and requires C/EBPs to induce immediate-early gene transcription. *Neural development* **2**: 4.
Caretti E, Devarajan K, Coudry R, Ross E, Clapper ML, Cooper HS, Bellacosa A. 2007. Comparison of RNA amplification methods and chip platforms for microarray analysis of samples processed by laser capture microdissection. *J Cell Biochem*.
Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM et al. 2009. Unlocking the secrets of the genome. *Nature* **459**: 927-930.
Chalasani SH, Chronis N, Tsunozaki M, Gray JM, Ramot D, Goodman MB, Bargmann CI. 2007. Dissecting a circuit for olfactory behaviour in Caenorhabditis elegans. *Nature* **450**(7166): 63-70.
Chalasani SH, Kato S, Albrecht DR, Nakagawa T, Abbott LF, Bargmann CI. 2010. Neuropeptide feedback modifies odor-evoked dynamics in Caenorhabditis elegans olfactory neurons. *Nature Neuroscience* **13**(5): 615-621.
Chalfie M, Sulston JE, White JG, Southgate E, Thomson JN, Brenner S. 1985. The neural circuit for touch sensitivity in Caenorhabditis elegans. *J Neurosci* **5**(4): 956-964.
Chase DL, Pepper JS, Koelle MR. 2004. Mechanism of extrasynaptic dopamine signaling in Caenorhabditis elegans. *Nat Neurosci* **7**(10): 1096-1103.
Chatzigeorgiou M, Yoo S, Watson JD, Lee W-H, Spencer WC, Kindt KS, Hwang SW, Miller III DM, Treinin M, Driscoll M et al. 2010. Specific roles for DEG/ENaC and TRP channels in touch and thermosensation in C. elegans nociceptors. *NATURE NEUROSCIENCE* **13**: 861-U106.
Chen BL, Hall DH, Chklovskii DB. 2006. Wiring optimization can relate neuronal structure and function. *Proc Natl Acad Sci U S A* **103**(12): 4723-4728.
Chen N, Pai S, Zhao Z, Mah A, Newbury R, Johnsen RC, Altun Z, Moerman DG, Baillie DL, Stein LD. 2005. Identification of a nematode chemosensory gene family. *Proc Natl Acad Sci U S A* **102**(1): 146-151.
Cherry TJ, Trimarchi JM, Stadler MB, Cepko CL. 2009. Development and diversification of retinal amacrine interneurons at single cell resolution. *Proc Natl Acad Sci USA* **106**: 9495-9500.
Chi W, Reinke V. 2006. Promotion of oogenesis and embryogenesis in the C. elegans gonad by EFL-1/DPL-1 (E2F) does not require LIN-35 (pRB). *Development* **133**(16): 3147-3157.

Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG. 2009. Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput Biol* **5**(6): e1000417.
Christensen M, Estevez A, Yin X, Fox R, Morrison R, McDonnell M, Gleason C, Miller DM, 3rd, Strange K. 2002. A primary culture system for functional analysis of C. elegans neurons and muscle cells. *Neuron* **33**(4): 503-514.
Christensen M, Strange K. 2001. Developmental regulation of a novel outwardly rectifying mechanosensitive anion channel in Caenorhabditis elegans. *J Biol Chem* **276**(48): 45024-45030.
Chung CY, Seo H, Sonntag KC, Brooks A, Lin L, Isacson O. 2005. Cell type-specific gene expression of midbrain dopaminergic neurons reveals molecules involved in their vulnerability and protection. *Hum Mol Genet* **14**(13): 1709-1725.
Cinar H, Keles S, Jin Y. 2005. Expression profiling of GABAergic motor neurons in Caenorhabditis elegans. *Curr Biol* **15**(4): 340-346.
Colosimo ME, Brown A, Mukhopadhyay S, Gabel C, Lanjuin AE, Samuel AD, Sengupta P. 2004. Identification of thermosensory and olfactory neuron-specific genes via expression profiling of single neuron types. *Curr Biol* **14**(24): 2245-2251.
Consortium E Birney E Stamatoyannopoulos JA Dutta A Guigó R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.
Consortium TCeS. 1998. Genome Sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**: 2012-2018.
Cremer H, Chazal G, Goridis C, Represa A. 1997. NCAM is essential for axonal growth and fasciculation in the hippocampus. *Molecular and Cellular Neurosciences* **8**(5): 323-335.
CRICK FH, BARNETT L, BRENNER S, WATTS-TOBIN RJ. 1961. General nature of the genetic code for proteins. *Nature* **192**: 1227-1232.
David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**(14): 5320-5325.
Davis RE, Stretton AO. 2001. Structure-activity relationships of 18 endogenous neuropeptides on the motor nervous system of the nematode Ascaris suum. *Peptides* **22**(1): 7-23.
del Castillo-Olivares A, Kulkarni M, Smith HE. 2009. Regulation of sperm gene expression by the GATA factor ELT-1. *Developmental Biology* **333**(2): 397-408.
Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C et al. 2008. Annotating genomes with massive-scale RNA sequencing. *Genome biology* **9**: R175.
Ding M, Chao D, Wang G, Shen K. 2007. Spatial regulation of an E3 ubiquitin ligase directs selective synapse elimination. *Science* **317**(5840): 947-951.

Dityatev A, Dityateva G, Schachner M. 2000. Synaptic strength as a function of post- versus presynaptic expression of the neural cell adhesion molecule NCAM. *Neuron* **26**(1): 207-217.

Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O. 2010. C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PloS one* **5**: e15435.

Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML et al. 2008. Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell* **135**(4): 749-762.

Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A et al. 2007. Genome-scale analysis of in vivo spatiotemporal promoter activity in Caenorhabditis elegans. *Nat Biotechnol* **25**(6): 663-668.

Eckhardt M, Bukalo O, Chazal G, Wang L, Goridis C, Schachner M, Gerardy-Schahn R, Cremer H, Dityatev A. 2000. Mice deficient in the polysialyltransferase ST8SiaIV/PST-1 allow discrimination of the roles of neural cell adhesion molecule protein and polysialic acid in neural development and synaptic plasticity. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **20**(14): 5234-5244.

Elemento O, Slonim N, Tavazoie S. 2007. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**(2): 337-350.

Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA. 1996. Laser capture microdissection. *Science* **274**(5289): 998-1001.

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome research* **17**: 69-73.

Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ, Marra MA, Holt RA, Moerman DG, Hobert O. 2007. The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes & Development* **21**(13): 1653-1674.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**(3): 186-194.

Faivre-Sarrailh C, Banerjee S, Li J, Hortsch M, Laval M, Bhat MA. 2004. Drosophila contactin, a homolog of vertebrate contactin, is required for septate junction organization and paracellular barrier function. *Development* **131**(20): 4931-4942.

Falk J, Bonnon C, Girault JA, Faivre-Sarrailh C. 2002. F3/contactin, a neuronal cell adhesion molecule implicated in axogenesis and myelination. *Biology of the Cell / Under the Auspices of the European Cell Biology Organization* **94**(6): 327-334.

Farkas RH, Qian J, Goldberg JL, Quigley HA, Zack DJ. 2004. Gene expression profiling of purified rat retinal ganglion cells. *Investigative ophthalmology & visual science* **45**(8): 2503-2513.

Fay DS, Stanley HM, Han M, Wood WB. 1999. A Caenorhabditis elegans homologue of hunchback is required for late stages of development but not early embryonic patterning. *Developmental Biology* **205**(2): 240-253.
Feinberg EH, Vanhoven MK, Bendesky A, Wang G, Fetter RD, Shen K, Bargmann CI. 2008. GFP Reconstitution Across Synaptic Partners (GRASP) defines cell contacts and synapses in living nervous systems. *Neuron* **57**(3): 353-363.
Feng Z, Li W, Ward A, Piggott BJ, Larkspur ER, Sternberg PW, Xu XZ. 2006. A C. elegans model of nicotine-dependent behavior: regulation by TRP-family channels. *Cell* **127**(3): 621-633.
Flames N, Hobert O. 2009. Gene regulatory logic of dopamine neuron differentiation. *Nature* **458**(7240): 885-889.
Fox RM, Von Stetina SE, Barlow SJ, Shaffer C, Olszewski KL, Moore JH, Dupuy D, Vidal M, Miller DM. 2005. A gene expression fingerprint of C. elegans embryonic motor neurons. *BMC Genomics* **6**(1): 42.
Fox RM, Watson JD, Von Stetina SE, McDermott J, Brodigan TM, Fukushige T, Krause M, Miller DM. 2007. The embryonic muscle transcriptome of Caenorhabditis elegans. *Genome Biol* **8**(9): R188.
Frohlich J, Konig H. 2000. New techniques for isolation of single prokaryotic cells. *FEMS microbiology reviews* **24**(5): 567-572.
Fu AY, Spence C, Scherer A, Arnold FH, Quake SR. 1999. A microfabricated fluorescence-activated cell sorter. *Nature Biotechnology* **17**(11): 1109-1111.
Galliot B, de Vargas C, Miller D. 1999. Evolution of homeobox genes: Q50 Paired-like genes founded the Paired class. *Development genes and evolution* **209**(3): 186-197.
Gao D, Kim J, Kim H, Phang TL, Selby H, Tan AC, Tong T. 2010. A survey of statistical software for analysing RNA-seq data. *Human genomics* **5**: 56-60.
Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**: 469-477.
Gautier L, Cope L, Bolstad BM, Irizarry Ra. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)* **20**: 307-315.
Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T et al. 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology* **26**: 317-325.
Gerrow K, El-Husseini A. 2006. Cell adhesion molecules at the synapse. *Front Biosci* **11**: 2400-2419.
Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**(6012): 1775-1787.

Gotenstein JR, Swale RE, Fukuda T, Wu Z, Giurumescu CA, Goncharov A, Jin Y, Chisholm AD. 2010. The C. elegans peroxidasin PXN-2 is essential for embryonic morphogenesis and inhibits adult axon regeneration. *Development* **137**(21): 3603-3613.

Gowda M, Jantasuriyarat C, Dean RA, Wang G-L. 2004. Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant physiology* **134**: 890-897.

Grange T, Roux J, Rigaud G, Pictet R. 1991. Cell-type specific activity of two glucocorticoid responsive units of rat tyrosine aminotransferase gene is associated with multiple binding sites for C/EBP and a novel liver-specific nuclear factor. *Nucleic acids research* **19**(1): 131-139.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**(7339): 473-479.

Gregory RI, Chendrimada TP, Shiekhattar R. 2006. MicroRNA biogenesis: isolation and characterization of the microprocessor complex. *Methods in molecular biology* **342**: 33-47.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**(7308): 835-840.

Hahn-Windgassen A, Van Gilst MR. 2009. The Caenorhabditis elegans HNF4alpha Homolog, NHR-31, mediates excretory tube growth and function through coordinate regulation of the vacuolar ATPase. *PLoS Genet* **5**(7): e1000553.

Hallem EA, Spencer WC, McWhirter RD, Zeller G, Henz SR, Ratsch G, Miller DM, 3rd, Horvitz HR, Sternberg PW, Ringstad N. 2011. Receptor-type guanylate cyclase is required for carbon dioxide sensation by Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* **108**(1): 254-259.

Hallem EA, Sternberg PW. 2008. Acute carbon dioxide avoidance in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* **105**(23): 8038-8043.

Hart AC, Sims S, Kaplan JM. 1995. Synaptic code for sensory modalities revealed by C. elegans GLR-1 glutamate receptor. *Nature* **378**(6552): 82-85.

Hatta K, Takeichi M. 1986. Expression of N-cadherin adhesion molecules associated with early morphogenetic events in chick development. *Nature* **320**(6061): 447-449.

Hattori D, Demir E, Kim HW, Viragh E, Zipursky SL, Dickson BJ. 2007. Dscam diversity is essential for neuronal wiring and self-recognition. *Nature* **449**(7159): 223-227.

He H, Wang J, Liu T, Liu XS, Li T, Wang Y, Qian Z, Zheng H, Zhu X, Wu T et al. 2007. Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Res* **17**(10): 1471-1477.

Head SR, Komori HK, Hart GT, Shimashita J, Schaffer L, Salomon DR, Ordoukhanian PT. 2011. Method for improved Illumina sequencing library

preparation using NuGEN Ovation RNA-Seq System. *BioTechniques* **50**(3): 177-180.
Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, Suarez-Farinas M, Schwarz C, Stephan DA, Surmeier DJ et al. 2008. A translational profiling approach for the molecular characterization of CNS cell types. *Cell* **135**(4): 738-748.
Hill E, Broadbent ID, Chothia C, Pettitt J. 2001. Cadherin superfamily proteins in Caenorhabditis elegans and Drosophila melanogaster. *J Mol Biol* **305**(5): 1011-1024.
Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in C. elegans: so many genes, such a little worm. *Genome Res* **15**(12): 1651-1660.
Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* **5**(2): 183-188.
Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. 2009. Massively parallel sequencing of the poly-adenylated transcriptome of C. elegans. *Genome Res*.
Hills T, Brockie PJ, Maricq AV. 2004. Dopamine and glutamate control area-restricted search behavior in Caenorhabditis elegans. *J Neurosci* **24**(5): 1217-1225.
Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, Ohta K. 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**(7218): 130-134.
Hochreiter S, Clevert DA, Obermayer K. 2006. A new summarization method for Affymetrix probe level data. *Bioinformatics* **22**(8): 943-949.
Hu QD, Ang BT, Karsak M, Hu WP, Cui XY, Duka T, Takeda Y, Chia W, Sankar N, Ng YK et al. 2003. F3/contactin acts as a functional ligand for Notch during oligodendrocyte maturation. *Cell* **115**(2): 163-175.
Hu X, Bessette PH, Qian J, Meinhart CD, Daugherty PS, Soh HT. 2005. Marker-specific sorting of rare cells using dielectrophoresis. *Proceedings of the National Academy of Sciences of the United States of America* **102**(44): 15757-15761.
Huang P, Pleasance ED, Maydan JS, Hunt-Newbury R, O'neil NJ, Mah A, Baillie DL, Marra MA, Moerman DG, Jones SJ. 2007. Identification and analysis of internal promoters in Caenorhabditis elegans operons. *Genome Res*.
Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T et al. 2002. The Ensembl genome database project. *Nucleic acids research* **30**: 38-41.
Huber W, Toedling J, Steinmetz LM. 2006. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*.
Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, Fang L, Halfnight E, Lee D, Lin J, Lorch A et al. 2007. High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. *PLoS Biol* **5**(9): e237.

Hutter H. 2004. Five-colour in vivo imaging of neurons in Caenorhabditis elegans. *Journal of microscopy* **215**(Pt 2): 213-218.
Hutter H, Vogel BE, Plenefisch JD, Norris CR, Proenca RB, Spieth J, Guo C, Mastwal S, Zhu X, Scheel J et al. 2000. Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**(5455): 989-994.
Hwang SB, Lee J. 2003. Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode Caenorhabditis elegans. *J Mol Biol* **333**(2): 237-247.
Il sC.
Inaki M, Yoshikawa S, Thomas JB, Aburatani H, Nose A. 2007. Wnt4 Is a Local Repulsive Cue that Determines Synaptic Target Specificity. *Curr Biol* **17**(18): 1574-1579.
Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**(4): e15.
Ishii S, Tago K, Senoo K. 2010. Single-cell analysis and isolation for microbiology and biotechnology: methods and applications. *Appl Microbiol Biotechnol* **86**(5): 1281-1292.
Isik M, Korswagen HC, Berezikov E. 2010. Expression patterns of intronic microRNAs in Caenorhabditis elegans. *Silence* **1**(1): 5.
Ivanov D, Dvoriantchikova G, Nathanson L, McKinnon SJ, Shestopalov VI. 2006. Microarray analysis of gene expression in adult retinal ganglion cells. *FEBS Letters* **580**(1): 331-335.
Ji X. 2008. The mechanism of RNase III action: how dicer dices. *Current topics in microbiology and immunology* **320**: 99-116.
Jiang M, Ryu J, Kiraly M, Duke K, Reinke V, Kim SK. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* **98**(1): 218-223.
Jiao Y, Moon SJ, Montell C. 2007. A Drosophila gustatory receptor required for the responses to sucrose, glucose, and maltose identified by mRNA tagging. *Proc Natl Acad Sci U S A* **104**(35): 14110-14115.
Jontes JD, Phillips GR. 2006. Selective stabilization and synaptic specificity: a new cell-biological model. *Trends Neurosci* **29**(4): 186-191.
Karin M. 1990. Too many transcription factors: positive and negative interactions. *The New Biologist* **2**(2): 126-131.
Katidou M, Vidaki M, Strigini M, Karagogeos D. 2008. The immunoglobulin superfamily of neuronal cell adhesion molecules: lessons from animal models and correlation with human disease. *Biotechnol J* **3**(12): 1564-1580.
Kato M, de Lencastre A, Pincus Z, Slack FJ. 2009. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol* **10**(5): R54.
Keene JD, Komisarow JM, Friedersdorf MB. 2006. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of

ribonucleoprotein complexes from cell extracts. *Nature protocols* **1**(1): 302-307.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler aD. 2002. The Human Genome Browser at UCSC. *Genome Research* **12**: 996-1006.
Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed data sets. *Bioinformatics (Oxford, England)* **26**: 2204-2207.
Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. 2001. A gene expression map for Caenorhabditis elegans. *Science* **293**(5537): 2087-2092.
Kimble J. 1981. Alterations in cell lineage following laser ablation of cells in the somatic gonad of Caenorhabditis elegans. *Dev Biol* **87**: 286-300.
Kiss T. 2001. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO journal* **20**: 3617-3622.
Kleene R, Schachner M. 2004. Glycans and neural cell interactions. *Nature reviews Neuroscience* **5**(3): 195-208.
Koh K, Rothman JH. 2001. ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in C. elegans. *Development* **128**(15): 2867-2880.
Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**: 59-69.
Krause M, Park M, Zhang JM, Yuan J, Harfe B, Xu SQ, Greenwald I, Cole M, Paterson B, Fire A. 1997. A C. elegans E/Daughterless bHLH protein marks neuronal but not striated muscle development. *Development* **124**(11): 2179-2189.
Kube DM, Savci-Heijink CD, Lamblin AF, Kosari F, Vasmatzis G, Cheville JC, Connelly DP, Klee GG. 2007. Optimization of laser capture microdissection and RNA amplification for gene expression profiling of prostate cancer. *BMC Mol Biol* **8**: 25.
Kubiak TM, Larsen MJ, Bowman JW, Geary TG, Lowery DE. 2008. FMRFamide-like peptides encoded on the flp-18 precursor gene activate two isoforms of the orphan Caenorhabditis elegans G-protein-coupled receptor Y58G8A.4 heterologously expressed in mammalian cells. *Biopolymers* **90**(3): 339-348.
Kunitomo H, Uesugi H, Kohara Y, Iino Y. 2005. Identification of ciliated sensory neuron-expressed genes in Caenorhabditis elegans using targeted pull-down of poly(A) tails. *Genome Biol* **6**(2): R17.
Lanjuin A, VanHoven MK, Bargmann CI, Thompson JK, Sengupta P. 2003. Otx/otd homeobox genes specify distinct sensory neuron identities in C. elegans. *Dev Cell* **5**(4): 621-633.
Latchman DS. 1997. Transcription factors: an overview. *The International Journal of Biochemistry & Cell Biology* **29**(12): 1305-1312.
Laubinger S, Zeller G, Henz SR, Sachsenberg T, Widmer CK, Naouar N, Vuylsteke M, Schölkopf B, Rätsch G, Weigel D. 2008. At-TAX: a whole

genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome biology* **9**: R112.
Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**(5543): 862-864.
Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annual review of genetics* **34**: 77-137.
Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal* **23**(20): 4051-4060.
Leethanakul C, Patel V, Gillespie J, Pallente M, Ensley JF, Koontongkaew S, Liotta LA, Emmert-Buck M, Gutkind JS. 2000. Distinct pattern of expression of differentiation and growth-related genes in squamous cell carcinomas of the head and neck revealed by the use of laser capture microdissection and cDNA arrays. *Oncogene* **19**(28): 3220-3224.
Li C, Kim K, Nelson LS. 1999a. FMRFamide-related neuropeptide gene family in Caenorhabditis elegans. *Brain Res* **848**(1-2): 26-34.
Li C, Nelson LS, Kim K, Nathoo A, Hart AC. 1999b. Neuropeptide gene families in the nematode Caenorhabditis elegans. *Ann N Y Acad Sci* **897**: 239-252.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**: 1851-1858.
Li L, Mao J, Sun L, Liu W, Wu D. 2002. Second cysteine-rich domain of Dickkopf-2 activates canonical Wnt signaling pathway via LRP-6 independently of dishevelled. *J Biol Chem* **277**(8): 5977-5981.
Li X, Panea C, Wiggins CH, Reinke V, Leslie C. 2010. Learning "graph-mer" motifs that predict gene expression trajectories in development. *PLoS Computational Biology* **6**(4): e1000761.
Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**(7027): 769-773.
Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev* **17**(8): 991-1008.
Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. 1999. High density synthetic oligonucleotide arrays. *Nature Genetics* **21**(1 Suppl): 20-24.
Liu C, Deneris ES. 2011. Transcriptional control of serotonin-modulated behavior and physiology. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **36**: 361-362.
Liu X, Long F, Peng H, Aerni SJ, Jiang M, Sanchez-Blanco A, Murray JI, Preston E, Mericle B, Batzoglou S et al. 2009. Analysis of cell fate from single-cell gene expression profiles in C. elegans. *Cell* **139**(3): 623-633.
Liu Z, Geng L, Li R, He X, Zheng JQ, Xie Z. 2003. Frequency modulation of synchronized Ca2+ spikes in cultured hippocampal networks through G-

protein-coupled receptors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **23**(10): 4156-4163.
Lobo MK, Karsten SL, Gray M, Geschwind DH, Yang XW. 2006. FACS-array profiling of striatal projection neuron subtypes in juvenile and adult mouse brains. *Nat Neurosci* **9**(3): 443-452.
Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C et al. 2011. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome research* **21**(2): 276-285.
Lund E, Dahlberg JE. 2006. Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs. *Cold Spring Harbor symposia on quantitative biology* **71**: 59-66.
Maduro M, Pilgrim D. 1995. Identification and cloning of unc-119, a gene expressed in the Caenorhabditis elegans nervous system. *Genetics* **141**(3): 977-988.
Mah AK, Armstrong KR, Chew DS, Chu JS, Tu DK, Johnsen RC, Chen N, Chamberlin HM, Baillie DL. 2007. Transcriptional regulation of AQP-8, a Caenorhabditis elegans aquaporin exclusively expressed in the excretory system, by the POU homeobox transcription factor CEH-6. *J Biol Chem* **282**(38): 28074-28086.
Mah AK, Tu DK, Johnsen RC, Chu JS, Chen N, Baillie DL. 2010. Characterization of the octamer, a cis-regulatory element that modulates excretory cell gene-expression in Caenorhabditis elegans. *BMC Mol Biol* **11**: 19.
Malone JH, Oliver B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology* **9**: 34.
Mane SP, Evans C, Cooper KL, Crasta OR, Folkerts O, Hutchison SK, Harkins TT, Thierry-Mieg D, Thierry-Mieg J, Jensen RV. 2009. Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC genomics* **10**: 264.
Mann F, Peuckert C, Dehner F, Zhou R, Bolz J. 2002. Ephrins regulate the formation of terminal axonal arbors during the development of thalamocortical projections. *Development* **129**(16): 3945-3955.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
Maricq AV, Peckol E, Driscoll M, Bargmann CI. 1995. Mechanosensory signalling in C. elegans mediated by the GLR-1 glutamate receptor. *Nature* **378**(6552): 78-81.
Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18**: 1509-1517.
Martinez NJ, Ow MC, Reece-Hoyes JS, Barrasa MI, Ambros VR, Walhout AJ. 2008. Genome-scale spatiotemporal analysis of Caenorhabditis elegans microRNA promoter activity. *Genome Res* **18**(12): 2005-2015.

Masai I, Lele Z, Yamaguchi M, Komori A, Nakata A, Nishiwaki Y, Wada H, Tanaka H, Nojima Y, Hammerschmidt M et al. 2003. N-cadherin mediates retinal lamination, maintenance of forebrain compartments and patterning of retinal neurites. *Development* **130**(11): 2479-2494.
Matranga C, Tomari Y, Shin C, Bartel DP, Zamore PD. 2005. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123**(4): 607-620.
McGhee JD, Fukushige T, Krause MW, Minnema SE, Goszczynski B, Gaudet J, Kohara Y, Bossinger O, Zhao Y, Khattra J et al. 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the C. elegans intestine, from embryo to adult. *Developmental biology* **327**: 551-565.
McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA et al. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev Biol* **302**(2): 627-645.
McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL, Chan S, Dube N, Fang L, Goszczynski B, Ha E et al. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. *Cold Spring Harb Symp Quant Biol* **68**: 159-169.
McManus MT, Petersen CP, Haines BB, Chen J, Sharp PA. 2002. Gene silencing using micro-RNA designed hairpins. *Rna* **8**(6): 842-850.
McMiller TL, Johnson CM. 2005. Molecular characterization of HLH-17, a C. elegans bHLH protein required for normal larval development. *Gene* **356**: 1-10.
Meissner B, Warner A, Wong K, Dube N, Lorch A, McKay SJ, Khattra J, Rogalski T, Somasiri A, Chaudhry I et al. 2009. An integrated strategy to study muscle development and myofilament structure in Caenorhabditis elegans. *PLoS genetics* **5**(6): e1000537.
Mellem JE, Brockie PJ, Zheng Y, Madsen DM, Maricq AV. 2002. Decoding of polymodal sensory stimuli by postsynaptic glutamate receptors in C. elegans. *Neuron* **36**(5): 933-944.
Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**(2): 716-721.
Michel V, Yuan Z, Ramsubir S, Bakovic M. 2006. Choline transport for phospholipid synthesis. *Experimental biology and medicine (Maywood, NJ)* **231**: 490-504.
Miller D, Niemeyer C. 1995. Expression of the unc-4 homeoprotein in Caenorhabditis elegans motor neurons specifies presynaptic input. *Development* **121**(9): 2877-2886.
Miller DM, Niemeyer CJ, Chitkara P. 1993. Dominant unc-37 mutations suppress the movement defect of a homeodomain mutation in unc-4, a neural specificity gene in Caenorhabditis elegans. *Genetics* **135**(3): 741-753.
Miller DM, Shen MM, Shamu CE, Bürglin TR, Ruvkun G, Dubois ML, Ghee M, Wilson L. 1992. C. elegans unc-4 gene encodes a homeodomain protein

that determines the pattern of synaptic input to specific motor neurons. *Nature* **355**(6363): 841-845.
Miller J, Knorr R, Ferrone M, Houdei R, Carron CP, Dustin ML. 1995. Intercellular adhesion molecule-1 dimerization and its consequences for adhesion mediated by lymphocyte function associated-1. *The Journal of experimental medicine* **182**: 1231-1241.
Mitchell PJ, Tjian R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**(4916): 371-378.
Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu A-L, Zhao Y, McDonald H, Zeng T, Hirst M et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research* **18**: 610-621.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**: 621-628.
Moss EG. 2002. MicroRNAs: hidden in the genome. *Current biology : CB* **12**(4): R138-140.
Much JW, Slade DJ, Klampert K, Garriga G, Wightman B. 2000. The fax-1 nuclear hormone receptor regulates axon pathfinding and neurotransmitter expression [In Process Citation]. *Development* **127**(4): 703-712.
Muller BM, Kistner U, Kindler S, Chung WJ, Kuhlendahl S, Fenster SD, Lau LF, Veh RW, Huganir RL, Gundelfinger ED et al. 1996. SAP102, a novel postsynaptic protein that interacts with NMDA receptor complexes in vivo. *Neuron* **17**(2): 255-265.
Murray JI, Bao Z, Boyle TJ, Boeck ME, Mericle BL, Nicholas TJ, Zhao Z, Sandel MJ, Waterston RH. 2008. Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nat Methods* **5**(8): 703-709.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, NY)* **320**: 1344-1349.
Nelson FK, Albert PS, Riddle DL. 1983. Fine structure of the Caenorhabditis elegans secretory-excretory system. *J Ultrastruct Res* **82**(2): 156-171.
Nguyen DN, Liu Y, Litsky ML, Reinke R. 1997. The sidekick gene, a member of the immunoglobulin superfamily, is required for pattern formation in the Drosophila eye. *Development* **124**(17): 3303-3312.
Nyrén P, Pettersson B, Uhlén M. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical biochemistry* **208**: 171-175.
Ogden SK, Fei DL, Schilling NS, Ahmed YF, Hwa J, Robbins DJ. 2008. G protein Galphai functions immediately downstream of Smoothened in Hedgehog signalling. *Nature* **456**(7224): 967-970.
Okaty BW, Sugino K, Nelson SB. 2011. A Quantitative Comparison of Cell-Type-Specific Microarray Gene Expression Profiling Methods in the Mouse Brain. *PLoS ONE* **6**: e16493.

Okkema PG, Harrison SW, Plunger V, Aryana A, Fire A. 1993. Sequence Requirements for myosin gene expression and regulation in Caenorhabditis elegans. *Genetics* **135**: 385-404.
Orphanides G, Lagrange T, Reinberg D. 1996. The general transcription factors of RNA polymerase II. *Genes & Development* **10**(21): 2657-2683.
Oshlack A, Robinson MD, Young MD. 2010. From RNA-seq reads to differential expression results. *Genome Biology* **11**: 220.
Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, Milos PM. 2010. Digital transcriptome profiling from attomole-level RNA samples. *Genome research* **20**(4): 519-525.
Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. *Nature* **461**: 814-818.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**(12): 1413-1415.
Pan Y, Ouyang Z, Wong WH, Baker JC. 2011. A new FACS approach isolates hESC derived endoderm using transcription factors. *PLoS ONE* **6**(3): e17536.
Papaioannou S, Marsden D, Franks CJ, Walker RJ, Holden-Dye L. 2005. Role of a FMRFamide-like family of neuropeptides in the pharyngeal nervous system of Caenorhabditis elegans. *J Neurobiol* **65**(3): 304-319.
Pauli F, Liu Y, Kim YA, Chen PJ, Kim SK. 2005. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. *Development*.
Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**(9): e45.
Pflugrad A, Meir JY, Barnes TM, Miller DM. 1997. The Groucho-like transcription factor UNC-37 functions with the neural specificity gene unc-4 to govern motor neuron identity in C. elegans. *Development* **124**(9): 1699-1709.
Pysh JJ, Wiley RG. 1974. Synaptic vesicle depletion and recovery in cat sympathetic ganglia electrically stimulated in vivo. Evidence for transmitter secretion by exocytosis. *The Journal of Cell Biology* **60**(2): 365-374.
Ramakrishnan N, Tadepalli S, Watson LT, Helm RF, Antoniotti M, Mishra B. 2010. Reverse engineering dynamic temporal models of biological processes and their relationships. *Proceedings of the National Academy of Sciences of the United States of America* **107**(28): 12511-12516.
Rana TM. 2007. Illuminating the silence: understanding the structure and function of small RNAs. *Nature reviews Molecular cell biology* **8**(1): 23-36.
Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout AJM. 2005. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome biology* **6**: R110.
Reynolds NK, Schade MA, Miller KG. 2005. Convergent, RIC-8-Dependent G{alpha} Signaling Pathways in the Caenorhabditis elegans Synaptic Signaling Network. *Genetics* **169**(2): 651-670.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ et al. 2010. The UCSC Genome Browser database: update 2010. *Nucleic acids research* **38**: D613-619.
Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7): 1311-1323.
Rio DC. 1993. Splicing of pre-mRNA: mechanism, regulation and role in development. *Current Opinion in Genetics & Development* **3**(4): 574-584.
Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* **12**: R22.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**: 24-26.
Rogers A, Antoshechkin I, Bieri T, Blasiar D, Bastiani C, Canaran P, Chan J, Chen WJ, Davis P, Fernandes J et al. 2008. WormBase 2007. *Nucleic acids research* **36**: D612-617.
Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**: 84-89.
Rosenow C, Saxena R, Durst M, Gingeras T. 2001. Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic Acids Res* **29**: e112.
Rossner MJ, Hirrlinger J, Wichert SP, Boehm C, Newrzella D, Hiemisch H, Eisenhardt G, Stuenkel C, von Ahsen O, Nave KA. 2006. Global transcriptome analysis of genetically identified neurons in the adult cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **26**(39): 9956-9966.
Roush SF, Slack FJ. 2006. Micromanagement: a role for microRNAs in mRNA stability. *ACS Chem Biol* **1**(3): 132-134.
Rousseeuw P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**: 53-65.
Roy PJ, Stuart JM, Lund J, Kim SK. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418**(6901): 975-979.
Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**(6): 1193-1207.
Ruby JG, Jan CH, Bartel DP. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**(7149): 83-86.
Rutishauser U, Watanabe M, Silver J, Troy FA, Vimr ER. 1985. Specific alteration of NCAM-mediated cell adhesion by an endoneuraminidase. *The Journal of Cell Biology* **101**(5 Pt 1): 1842-1849.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. 2002. Using the transcriptome to annotate the genome. *Nature biotechnology* **20**: 508-512.

Salinas PC. 2005. Signaling at the vertebrate synapse: new roles for embryonic morphogens? *J Neurobiol* **64**(4): 435-445.

Sandelin M, Zabihi S, Liu L, Wicher G, Kozlova EN. 2004. Metastasis-associated S100A4 (Mts1) protein is expressed in subpopulations of sensory and autonomic neurons and in Schwann cells of the adult rat. *The Journal of comparative neurology* **473**(2): 233-243.

Sarafi-Reinach TR, Melkman T, Hobert O, Sengupta P. 2001. The lin-11 LIM homeobox gene specifies olfactory and chemosensory neuron fates in C. elegans. *Development* **128**(17): 3269-3281.

Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. 2008. Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat Methods* **5**(10): 865-867.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, NY)* **270**: 467-470.

Schmitz C, Kinge P, Hutter H. 2007. Axon guidance genes identified in a large-scale RNAi screen using the RNAi-hypersensitive Caenorhabditis elegans strain nre-1(hd20) lin-15b(hd126). *Proc Natl Acad Sci USA* **104**(3): 834-839.

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ, Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6**(4): R33.

Schwarz V, Pan J, Voltmer-Irsch S, Hutter H. 2009. IgCAMs redundantly control axon navigation in Caenorhabditis elegans. *Neural development* **4**: 13.

Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, Philips P, De Bona F, Hartmann L, Bohlen A et al. 2009. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res* **19**(11): 2133-2143.

Seggerson K, Tang L, Moss EG. 2002. Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. *Developmental Biology* **243**(2): 215-225.

Shen K, Bargmann CI. 2003. The immunoglobulin superfamily protein SYG-1 determines the location of specific synapses in C. elegans. *Cell* **112**(5): 619-630.

Shen K, Fetter RD, Bargmann CI. 2004. Synaptic specificity is generated by the synaptic guidepost protein SYG-2 and its receptor, SYG-1. *Cell* **116**(6): 869-881.

Shen Y, Sarin S, Liu Y, Hobert O, Pe'er I. 2008. Comparing platforms for C. elegans mutant identification using high-throughput whole-genome sequencing. *PloS one* **3**: e4012.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex

polony sequencing of an evolved bacterial genome. *Science (New York, NY)* **309**: 1728-1732.
Simmer F, Tijsterman M, Parrish S, Koushika SP, Nonet ML, Fire A, Ahringer J, Plasterk RH. 2002. Loss of the putative RNA-directed RNA polymerase RRF-3 makes C. elegans hypersensitive to RNAi. *Curr Biol* **12**(15): 1317-1319.
Smith CJ, Watson JD, Spencer WC, O'Brien T, Cha B, Albeg A, Treinin M, Miller DM. 2010. Time-lapse imaging and cell-specific expression profiling reveal dynamic branching and molecular determinants of a multi-dendritic nociceptor in C. elegans. *Developmental biology* **345**: 18-33.
Smith RD, Malley JD, Schechter AN. 2000. Quantitative analysis of globin gene induction in single human erythroleukemic cells. *Nucleic acids research* **28**(24): 4998-5004.
Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.
Smyth GK. 2005. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, (ed. R Gentleman, V Carey, S Dudoit, R Irizarry, W Huber), pp. 397-420. Springer, New York.
Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J et al. 2011. A spatial and temporal map of C. elegans gene expression. *Genome research* **21**(2): 325-341.
Sperry RW. 1963. CHEMOAFFINITY IN THE ORDERLY GROWTH OF NERVE FIBER PATTERNS AND CONNECTIONS. *Proc Natl Acad Sci USA* **50**: 703-710.
Staal FJ, van der Burg M, Wessels LF, Barendregt BH, Baert MR, van den Burg CM, van Huffel C, Langerak AW, van der Velden VH, Reinders MJ et al. 2003. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* **17**(7): 1324-1332.
Stetina SEV, Watson JD, Fox RM, Olszewski KL, Spencer WC, Roy PJ, Miller DM. 2007. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. *Genome Biol* **8**: R135.
Sulston JE, Horvitz HR. 1977. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol* **56**(1): 110-156.
Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**(1): 64-119.
Suzuki T, Park H, Hollingsworth NM, Sternglanz R, Lennarz WJ. 2000. PNG1, a yeast gene encoding a highly conserved peptide:N-glycanase. *The Journal of cell biology* **149**: 1039-1052.

Sweedler JV, Li L, Rubakhin SS, Alexeeva V, Dembrow NC, Dowling O, Jing J, Weiss KR, Vilim FS. 2002. Identification and characterization of the feeding circuit-activating peptides, a novel neuropeptide family of aplysia. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **22**(17): 7797-7808.

Takeichi M, Abe K. 2005. Synaptic contact dynamics controlled by cadherin and catenins. *Trends Cell Biol* **15**(4): 216-221.

Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, Lao K, Surani MA. 2010. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature protocols* **5**: 516-535.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**: 377.

Teichmann SA, Chothia C. 2000. Immunoglobulin superfamily proteins in Caenorhabditis elegans. *J Mol Biol* **296**(5): 1367-1383.

Thellmann M, Hatzold J, Conradt B. 2003. The Snail-like CES-1 protein of C. elegans can block the expression of the BH3-only cell-death activator gene egl-1 by antagonizing the function of bHLH proteins. *Development* **130**(17): 4057-4071.

Tian L, Nyman H, Kilgannon P, Yoshihara Y, Mori K, Andersson LC, Kaukinen S, Rauvala H, Gallatin WM, Gahmberg CG. 2000. Intercellular adhesion molecule-5 induces dendritic outgrowth by homophilic adhesion. *The Journal of cell biology* **150**: 243-252.

Tietjen I, Rihel J, Dulac CG. 2005. Single-cell transcriptional profiles and spatial patterning of the mammalian olfactory epithelium. *Int J Dev Biol* **49**(2-3): 201-207.

Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. 2003. Single-cell transcriptional analysis of neuronal progenitors. *Neuron* **38**(2): 161-175.

Torres TT, Metta M, Ottenwälder B, Schlötterer C. 2008. Gene expression profiling by massively parallel sequencing. *Genome research* **18**: 172-177.

Tosney KW, Watanabe M, Landmesser L, Rutishauser U. 1986. The distribution of NCAM in the chick hindlimb during axon outgrowth and synaptogenesis. *Developmental Biology* **114**(2): 437-452.

Touroutine D, Fox RM, Von Stetina SE, Burdina A, Miller DM, Richmond JE. 2005. acr-16 encodes an essential subunit of the levamisole-resistant nicotinic receptor at the Caenorhabditis elegans neuromuscular junction. *J Biol Chem* **280**(29): 27013-27021.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25**: 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511-515.

Troemel ER, Chou JH, Dwyer ND, Colbert HA, Bargmann CI. 1995. Divergent seven transmembrane receptors are candidate chemosensory receptors in C. elegans. *Cell* **83**(2): 207-218.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**(5992): 689-693.
Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**(5785): 320-324.
van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**(5): e1000371.
van Ham TJ, Thijssen KL, Breitling R, Hofstra RMW, Plasterk RHA, Nollen EAA. 2008. C. elegans model identifies genetic modifiers of alpha-synuclein inclusion formation during aging. *PLoS genetics* **4**: e1000027.
Varadan V, Miller DM, Anastassiou D. 2006. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* **22**(14): e497-506.
Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science (New York, NY)* **270**: 484-487.
Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. 2000. SOM Toolbox for Matlab 5.
Vogel C, Teichmann SA, Chothia C. 2003. The immunoglobulin superfamily in Drosophila melanogaster and Caenorhabditis elegans and the evolution of complexity. *Development* **130**(25): 6317-6328.
Von Stetina SE, Watson JD, Fox RM, Olszewski KL, Spencer WC, Roy PJ, Miller DM. 2007. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. *Genome Biol* **8**(7): R135.
Waites CL, Craig AM, Garner CC. 2005. Mechanisms of vertebrate synaptogenesis. *Annu Rev Neurosci* **28**: 251-274.
Walker CS, Brockie PJ, Madsen DM, Francis MM, Zheng Y, Koduri S, Mellem JE, Strutz-Seebohm N, Maricq AV. 2006a. Reconstitution of invertebrate glutamate receptor function depends on stargazin-like proteins. *Proc Natl Acad Sci USA* **103**(28): 10781-10786.
Walker CS, Francis MM, Brockie PJ, Madsen DM, Zheng Y, Maricq AV. 2006b. Conserved SOL-1 proteins regulate ionotropic glutamate receptor desensitization. *Proc Natl Acad Sci USA* **103**(28): 10787-10792.
Walker JM, Rapley R, Richards MP. 2005. *Medical Biomethods Handbook*. Humana Press, Totowa, NJ.
Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM et al. 2010. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* **38**: e178.
Wang X, Zhang W, Cheever T, Schwarz V, Opperman K, Hutter H, Koepp D, Chen L. 2008. The C. elegans L1CAM homologue LAD-2 functions as a coreceptor in MAB-20/Sema2 mediated axon guidance. *J Cell Biol* **180**(1): 233-246.
Wang X, Zhao Y, Wong K, Ehlers P, Kohara Y, Jones SJ, Marra MA, Holt RA, Moerman DG, Hansen D. 2009. Identification of genes expressed in the

hermaphrodite germ line of C. elegans using SAGE. *BMC Genomics* **10**: 213.
Washbourne P, Dityatev A, Scheiffele P, Biederer T, Weiner JA, Christopherson KS, El-Husseini A. 2004. Cell adhesion molecules in synapse formation. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **24**: 9244-9249.
Watson JD, Wang S, Von Stetina SE, Spencer WC, Levy S, Dexheimer PJ, Kurn N, Heath JD, Miller DM. 2008. Complementary RNA amplification methods enhance microarray identification of transcripts expressed in the C. elegans nervous system. *BMC Genomics* **9**: 84.
Wettenhall JM, Smyth GK. 2004. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**(18): 3705-3706.
White J, Southgate E, Thomson J. 1992. Mutations in the Caenorhabditis elegans unc-4 gene alter the synaptic input to ventral cord motor neurons. *Nature* **355**(6363): 838-841.
White JG, Southgate E, Thomson JN, Brenner S. 1976. The structure of the ventral nerve cord of Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* **275**(938): 327-348.
White JG, Southgate E, Thomson JN, Brenner S. 1986. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London* **B314**: 1-340.
Wightman B, Baran R, Garriga G. 1997. Genes that guide growth cones along the C. elegans ventral nerve cord. *Development* **124**(13): 2571-2580.
Wightman B, Ebert B, Carmean N, Weber K, Clever S. 2005. The C. elegans nuclear receptor gene fax-1 and homeobox gene unc-42 coordinate interneuron identity by regulating the expression of glutamate receptor subunits and other neuron-specific genes. *Dev Biol* **287**(1): 74-85.
Wilusz JE, Sunwoo H, Spector DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23**(13): 1494-1504.
Wormbase. 2006. http://www.wormbase.org. **WS155**.
Wu C, Carta R, Zhang L. 2005. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic acids research* **33**: e84.
Wurtman RJ. 1992. Choline metabolism as a basis for the selective vulnerability of cholinergic neurons. *Trends in neurosciences* **15**: 117-122.
Yamagata M, Sanes JR. 2008. Dscam and Sidekick proteins direct lamina-specific synaptic connections in vertebrate retina. *Nature* **451**(7177): 465-469.
Yamagata M, Sanes JR, Weiner JA. 2003. Synaptic adhesion molecules. *Curr Opin Cell Biol* **15**(5): 621-632.
Yamagata M, Weiner JA, Sanes JR. 2002. Sidekicks: synaptic adhesion molecules that promote lamina-specific connectivity in the retina. *Cell* **110**(5): 649-660.
Yang Y, Xu S, Xia L, Wang J, Wen S, Jin P, Chen D. 2009. The bantam microRNA is associated with drosophila fragile X mental retardation protein and regulates the fate of germline stem cells. *PLoS genetics* **5**(4): e1000444.

Yang Z, Edenberg HJ, Davis RL. 2005. Isolation of mRNA from specific tissues of Drosophila by mRNA tagging. *Nucleic Acids Res* **33**(17): e148.
Yen CL, Mar MH, Meeker RB, Fernandes A, Zeisel SH. 2001. Choline deficiency induces apoptosis in primary cultures of fetal neurons. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **15**: 1704-1710.
You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD. 2008. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**: 253.
Zeller G, Henz SR, Laubinger S, Weigel D, Rätsch G. 2008. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* **538**: 527-538.
Zhang J, Finney RP, Clifford RJ, Derr LK, Buetow KH. 2005. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85**: 297-308.
Zhang S, Banerjee D, Kuhn JR. 2011. Isolation and Culture of Larval Cells from C. elegans. *PLoS ONE* **6**: e19505.
Zhang S, Ma C, Chalfie M. 2004. Combinatorial marking of cells and organelles with reconstituted fluorescent proteins. *Cell* **119**(1): 137-144.
Zhang Y, Ma C, Delohery T, Nasipak B, Foat BC, Bounoutas A, Bussemaker HJ, Kim SK, Chalfie M. 2002. Identification of genes expressed in C. elegans touch receptor neurons. *Nature* **418**(6895): 331-335.
Zheng S, Chen L. 2009. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic acids research* **37**: e75.
Zheng Y, Brockie PJ, Mellem JE, Madsen DM, Maricq AV. 1999. Neuronal control of locomotion in C. elegans is modified by a dominant mutation in the GLR-1 ionotropic glutamate receptor. *Neuron* **24**(2): 347-361.
Zheng Y, Mellem JE, Brockie PJ, Madsen DM, Maricq AV. 2004. SOL-1 is a CUB-domain protein required for GLR-1 glutamate receptor function in C. elegans. *Nature* **427**(6973): 451-457.
Zhou X, Ruan J, Wang G, Zhang W. 2007. Characterization and identification of microRNA core promoters in four model species. *PLoS Computational Biology* **3**(3): e37.
Zipursky SL, Wojtowicz WM, Hattori D. 2006. Got diversity? Wiring the fly brain with Dscam. *Trends Biochem Sci* **31**(10): 581-588.