Spring 2008

A transcriptome comparison of Caenorhabditis elegans cultured in laboratory and soil-like environments

Richard A. Roy
University of New Hampshire, Durham

A TRANSCRIPTOME COMPARISON OF *CAENORHABDITIS ELEGANS* CULTURED IN LABORATORY AND SOIL-LIKE ENVIRONMENTS

BY

RICHARD A. ROY
BS, Florida Institute of Technology, 1983
MS, Colorado State University, 1985

THESIS

Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of

Master of Science
in
Genetics

May, 2008

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

This thesis has been examined and approved.

Thesis Director, W. Kelley Thomas, Associate
Professor of Biochemistry and Molecular Biology, and
Genetics

R. Daniel Bergeron, Professor of Computer Science

John J. Collins, Associate Professor of Biochemistry
and Molecular Biology, and Genetics

Thomas M. Davis

Thomas M. Davis, Professor of Plant Biology, and
Genetics

15 May 2008

Date

DEDICATION

I dedicate this work to the memory of my friend, the late Professor Charles E. Warren— a scientist and humanitarian who had the courage to strive for excellence in all that he did.

“One machine can do the work of fifty ordinary men. No machine can do the work of one extraordinary man.” — Elbert Hubbard

“The length of your education is less important than its breadth, and the length of your life is less important than its depth.” — Marilyn vos Savant

ACKNOWLEDGEMENTS

I'd like to thank W. Kelley Thomas, my thesis advisor and committee chair, for suggesting this experiment, allowing me to perform it in his lab, and for his other efforts on my behalf.

I wish to express my heartfelt thanks to my other committee members: Dan Bergeron, John Collins, and Tom Davis, for their work on my committee and their other contributions to my education. Each has helped me grow as a scientist.

Likewise, the work of Tom Kocher and the late Charles Warren, my former committee members, is deeply appreciated.

Professor Dee R. Denver of Oregon State University generously offered use of his microarrays and core lab processing services for this experiment. He also made suggestions on the data analysis. I can't thank him enough.

Gratitude is extended to the UNH Genetics Program for teaching assistantships that provided financial support for three semesters of this research and started me on my path as an educator.

I've enjoyed the support of my family and many other loved ones during this degree program. Knowing that they care about me deeply even if they don't particularly care about my research always brings a smile. I especially thank my wife Deborah. The changes over the past twenty years have been amazing. But so are the things that have remained the same! I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A TRANSCRIPTOME COMPARISON OF *CAENORHABDITIS ELEGANS* CULTURED IN LABORATORY AND SOIL-LIKE ENVIRONMENTS

by

Richard A. Roy

University of New Hampshire, May, 2008

Caenorhabditis elegans has been the subject of numerous microarray experiments designed to help understand gene expression and function. Many such experiments have assessed the animal's transcriptional response to simple perturbations of the traditional laboratory environment.

Here, the transcriptomes of worms cultured in two environments, lab and soil-like, were compared using whole-genome tiling microarrays. The results differed significantly between environments with a greater abundance of differentially expressed genes of ambiguous or unknown function in the soil-like environment. Furthermore, the functional categories of genes expressed only in the soil-like environment differed significantly from their lab-only counterparts.

Numerous intergenic regions showed expression. The majority were environment specific but most that were mutually expressed were structurally similar to protein-coding genes. They may well be un-annotated exons or genes. The environment specific regions were significantly shorter, overall, than coding sequences, and may represent polypeptides or non-coding RNA with regulatory or other functions.

INTRODUCTION

Caenorhabditis elegans as a Model Organism

Originally named *Rhabditis elegans* and described by Maupas in 1900[1], the nematode *Caenorhabditis elegans* became the subject of intense scientific interest a little over four decades ago. It has since been used extensively as a model organism in the lab and its genome has been fully sequenced and annotated in great detail[2]. Many thousands of experiments have been performed on it and it is commonly considered the most thoroughly studied and well-understood animal.

Despite its apparent simplicity, *C. elegans* is a sophisticated organism with a repertoire of protein coding genes that is almost as extensive as, and quite similar to, our own—in fact, over 80% of protein coding genes in the worm have homologs in the human genome[3]. The functions of some of these genes are well known and, in many cases, entire pathways have been thoroughly studied and are understood in detail. On the other hand, the functions of many worm genes are a complete mystery and even the signals that elicit expression of these genes are unknown.

Determining gene function and the control of gene expression are two areas of intense research in the life sciences[4-8]. While the purpose of some genes is to regulate the expression of others, the regulatory genes must themselves be regulated by something. In at least some cases, that regulation is caused by external environmental cues. By creating an environment that is different from that in the lab and more like *C. elegans*' natural environment, we may hope to discover new patterns of regulation and the functions of additional genes.

Unfortunately, despite our exhaustive knowledge of many aspects of *C. elegans* biology, very little is known about the worm's natural habitat and lifestyle[9-12]. In fact, the sequenced strain was originally isolated from mushroom compost, not pristine soil. Subsequent efforts to isolate other strains have yielded worms from many places around the world but always in soils that have been subjected to human manipulation. This fact along with other evidence has led some researchers to question whether *C. elegans* might be a strictly human commensal[9]. Even if it is, there is much to learn about the environment we create when we manipulate soil. Learning how *C. elegans* has adapted to that environment is arguably even more interesting. If the worm can live comfortably in a more natural setting however, that might be the most interesting of all.

The purpose of this experiment was to compare the transcriptomes of worms cultured in two environments: a traditional laboratory environment and a soil-like environment containing elements of a "natural" soil that the worm could conceivably encounter outside of the lab. The latter environment might provide challenges and/or metabolic or other opportunities that lab worms have not encountered in thousands of generations. The worm's response to these challenges and opportunities could be mundane, interesting[13-17], or even completely unexpected[18]. It is assumed that some significant subset of such responses will be reflected in the transcriptome[16,19,20].

The concept of culturing *C. elegans* in a soil-like environment is hardly new. In 1977, Anderson and Coleman[17] observed phenotypic differences between nematodes cultured in a lab environment and those cultured in a sand-like medium of glass microbeads. More recently, researchers at Kansas State University have taken an interest in the culturing of nematodes in soil-like environments but have yet to publish results from any experiment even moderately analogous to this one. Some of their work was focused on the culturing of *C. elegans* with a diet of *Serratia marcescens*, a prokaryote common to soil environments. In 2002, another group published their findings on that very topic[21].

While that work is interesting, the purpose was to study the effect of one (worm vs. microbe) interaction. That interaction presumably can and does occur in the wild but the approach and experimental philosophy were very different from those used here[22]. Obviously, a straightforward "species versus species" encounter in the absence of other organisms and complicating factors does not occur in actual soil. Although studying the worm's reactions to such an encounter may well reveal interesting patterns of gene expression, there may also be genes that are only expressed in a more complex environment or in response to specific factors that can only come to exist in such an environment.

C. elegans Microarrays

C. elegans has been used extensively for the study of gene expression and many *C. elegans* microarray experiments have been conducted over the years[5,20,23,24]. Some have focused on studying the change in the transcriptome during the developmental process[5]. Others have focused on the differences in transcriptomes between normal and mutant worms[25]. Still others have studied the effect on the transcriptome when one changes a single aspect of the worm's environment (e.g., oxygen concentration, various toxins, radiation)[24].

It is important to note that the microarrays act as filters in these experiments: they define and limit the results that can be found. Thus, the results of a microarray experiment can never be completely divorced from the array itself. Early microarrays featured modest numbers of putative protein-coding gene sequences. They only measured levels of poly-A transcripts from the target regions (or those with sequence similarity to them)[26]. Subsequent arrays featured larger numbers of spots and better controls. Some arrays were designed to detect the expression of individual exons. Now whole-chromosome[27] or whole-genome tiling arrays allow the detection of transcription from any non-repetitive region[28].

The Affymetrix GeneChip® *C. elegans* Tiling 1.0R Array consists of over 3 million pairs of 25-mer probes corresponding to one strand of virtually the entire *C. elegans* genome. While some probe pairs correspond to more than one location on the genome, none correspond to large numbers of locations. In particular, highly repetitive regions such as telomeres and microsatellites are not represented on the array. On the other hand, larger features such as transposons commonly are at least to some extent.

Each probe pair consists of one "perfect match" and one "mismatch" (PM and MM, respectively) probe; the latter is identical to the former with the exception of one position that is occupied by a different nucleotide. The probe pairs cover or "tile" the genome very thoroughly—probes in successive positions along a chromosome commonly abut one another perfectly. In most of the rest of the cases, they overlap by one or two bases or they have gaps that are one or two bases long. Use of this microarray allows the elucidation of virtually the entire worm transcriptome regardless of the genomic locations transcribed or whether transcription was "expected" at any particular location[28]. Thus, it provides a minimally biased view of transcription.

Although the Affymetrix *C. elegans* Tiling Array is an invaluable tool for genomic research, an understanding of its limitations is essential to interpreting and understanding the information it yields. One important limitation is the fact that when the RNA is processed and hybridized to the array in the usual manner, information regarding which strand was transcribed is lost. Thus, the findings are "strand-agnostic". This is not considered a fatal flaw because most of the transcribed regions coincide with the genomic positions (on one strand or the other) of annotated genes. Transcription in the area of a known exon is simply assumed evidence of transcription of that exon on whichever strand it is found. This is not an unreasonable assumption, of course, and it applies to the great majority of the regions where transcription was found. On the other hand, the strand of transcribed regions that do not coincide with any annotated exon is ambiguous.

Another limitation of the Affymetrix *C. elegans* Tiling Array is positional ambiguity regarding the beginnings and ends of transcribed regions. Because the array consists of probe pairs that are 25 bases long and tile the genome at a spacing of about 25 bases on center, there is little or no overlap between most successive probe pairs. A probe pair corresponding to the end of a particular transcript will usually be only partly covered by that transcript while a neighboring probe pair will be completely spanned by it. At the completely spanned probe pair, the transcript will hybridize well to the PM probe but less well to the corresponding MM probe. The signal intensity level of the PM probe will be significantly higher than that of the MM probe so expression at that pair will be reported as “present”. This may or may not be the case at the probe pair for the end of the transcript. If the transcript does not reach the portion of the MM probe that contains the mismatched base, it will fail to show a significantly different signal intensity level and expression at that location will be reported as “absent”. The overall result is that the length of most transcripts will be under-reported by approximately the width of one probe with roughly half of this error appearing as a starting location being reported later than it should be and an ending position being reported earlier than it should be.

A final important limitation of the Affymetrix *C. elegans* Tiling Array results from sequences that occur at multiple locations within the *C. elegans* genome. When expression is detected at a probe pair that corresponds to multiple locations in the genome, that expression is indicated for all the locations in the genome where that sequence is found. This may or may not be appropriate. Transcriptionally active duplications can certainly exist in genomes. On the other hand, recent duplications can maintain sequence identity for a substantial period even if one copy is no longer actively expressed. An unfortunate consequence of all this is that sequences of just 25 bases (or possibly even a little shorter) that occur in more than one place in the genome may give false-positive indications of expression with arbitrarily high confidence levels and/or levels of expression in places where they are not expressed whatsoever.

The common element in all microarray experiments is a list of signal intensities representing levels of expression. Unfortunately, there are many steps involved in converting a set of RNA samples into a transcriptome comparison and large uncertainties are the norm[23,26,29]. To account for this, differential expression in a microarray experiment is commonly expressed in terms of fold differences with a two-fold or three-fold change being considered the minimum significant difference in expression between two environments or treatments. Even so, these values are based on assumptions that must be made without proof of validity. Thus, they represent an aspect of microarray experiments that is easily attacked and impossible to fully defend.

One assumption that is required in order to make claims about fold differences in expression levels is that the underlying distributions of those levels are the same or very similar. Another major assumption is that the median level of all the genes expressed in one environment is equal to the median level in the other. Without these assumptions (or in any case where they are violated), meaningful fold differences in expression level cannot be calculated.

In this experiment, the inherent uncertainty in expression levels was accounted for in a more statistically conservative way that does not rely on unsupported assumptions about underlying expression distribution patterns. The method used here still allows relevant inferences to be drawn from the data, however, and can be used in more typical microarray experiments, as well.

Objective and Hypotheses

The purpose of this experiment is not to determine the effect of any particular environmental change on the *C. elegans* transcriptome. Instead, it is to get some idea of what that transcriptome could look like in a more complex natural setting. Would it be essentially the same as what is observed in the traditional lab environment? Could it be very different? Might the expression of rarely seen or even novel genes be revealed?

Among genes expressed in both environments, might their levels of expression be re-prioritized from one environment to the other? Can this even be determined in this study?

More formally, the questions being asked in this experiment include:

1. Are the transcriptomes of *C. elegans* different in a soil-like environment than in the lab?
2. Are any genes of unknown function expressed in a soil-like environment?
3. Of genes expressed in both soil-like and lab environments, are there any expressed at a higher priority in one environment than the other?

The corresponding hypotheses are:

1. H_0: The transcriptomes are the same.

   H_1: Some genes are expressed in an environment-specific manner.
2. H_0: There are no genes of unknown function expressed in the soil environment.

   H_1: Some genes of unknown function are expressed in the soil environment.
3. H_0: No genes are prioritized differently in one environment than the other.

   H_1: Some genes are prioritized differently in one environment than the other.

CHAPTER I

MATERIALS & METHODS

For both the lab and soil-like cultures, populations of synchronized worms were created following the protocol of Khan and McFadden[30].

Lab Culture

Synchronized worms were cultured at room temperature on Nematode Growth Medium (NGM) agar with *E. coli* strain OP50 as a food source. In the most typical lab environment, *C. elegans* are cultured in clear plastic disposable plates. In this experiment, however, opaque reusable "Instrument/pipette sterilizing pan" (Nalgene catalog number 6910-0618) trays were used instead. Between uses, they were cleaned and autoclaved.

Approximately 72 hours after the synchronized larvae were put into the trays, they were gravid and some had started to lay eggs. They were washed from the tray into 50 ml conical tubes with M9 solution. Then they were centrifuged at 300 x g for 20 seconds, the supernatant of M9 and *E. coli* was aspirated, and the worms were rinsed with more M9. This process was repeated three times to remove excess *E. coli*. After the third rinsing, the pellet of worms was immediately used to isolate RNA as described below.

Collection of Soil Microbiota

A site (GPS Coordinates: Zone 19T, 0342281 4777622; North Latitude: 43 degrees, 8 minutes, 5.68 seconds, West Longitude: 70 degrees, 56 minutes, 21.22 seconds) near the edge of the UNH campus was selected for the collection of soil microbiota. The site had not been mowed or otherwise maintained and it was completely covered by a thick mat of decaying grasses, weeds, and other vegetation that had grown and died back annually over a period of years. Several small hardwood trees grew in the immediate area; their leaves were also a component of the decaying vegetation. A pine forest bordered the area at a distance of roughly 30 meters and the edge of a wetland was roughly ten meters distant in another direction.

Approximately two liters of sterile S basal solution were prepared and taken to the site with digging tools, two buckets, and a set of sieves on June 18, 2005. The tools, buckets, and sieves had been washed and were surface-sterilized with 95% ethanol just prior to leaving the lab. They air-dried on the way to the collection site.

From a small area, the topmost few inches of soil and some of the overlying vegetable matter were collected in a bucket. The volume was split into two approximately equal portions, one of which was immediately put in a bag. It was subsequently sealed and stored at –80° C. Its mass was found to be 1098 grams. The sterile S basal solution was used to rinse the other volume of soil through the coarsest sieve with the resulting liquid collected in a bucket to be passed through successively finer sieves. The first sieve removed large objects including root balls, rocks, most of the earthworms and insects, and most of the vegetation. Finer sieves removed smaller objects, sand, and eventually, all soil particles except for some of the very finest silts and clays.

The resulting liquid was brought into the lab and filtered through sterile Nitex® mesh with a pore size of 5 microns. As this was a very slow process, it was performed overnight at 4° C in a temperature-controlled room.

A total of about 800 ml of liquid was collected the next morning. It was re-mixed and dispensed equally into 50 ml conical tubes. The tubes were centrifuged at 5500 x g for 30 minutes at 4° C. A substantial pellet formed in each tube. About 2.5 ml glycerol was added to each tube and the pellet re-suspended as well as possible without risking contamination before being frozen at –80° C. Although the tubes were prepared as uniformly as possible, significant variation undoubtedly existed.

Soil-like Culture

Three replications of the experiment were run simultaneously. For each replicate, 75 ml of dry glass micro-beads were spread into an instrument sterilization tray. Fifty ml of NGM solution (prepared like NGM agar without agarose) were added to the tray along with a tube of soil microbiota that had been thawed and re-suspended with the use of a vortex. The synchronized worm larvae were added last.

The worms were allowed to grow for about 78 hours. At that time, many were gravid adults and some had started laying eggs but others were less fully developed, appearing to still be young adult, L4, L3, or even L2 larvae.

To collect the worms from the soil-like environment, the sterilization pan was flooded with large volumes (approximately 600 ml) of S basal solution then agitated as vigorously as possible without spilling the contents. As the glass micro-beads settled to the bottom of the pan, all the remaining liquid and suspended matter were poured into a very large (3L) beaker. This was repeated three times. This process recovered most of the worms but certainly not all of them.

The contents of the 3L beaker were then poured gently through an ASTM #400 stainless steel sieve. The mesh retained the worms while the liquid passed through and was collected for sterilization and disposal. The worms were then rinsed from the sieve into a 50 ml conical tube using a small amount of M9. The 50 ml tube was then centrifuged at 300 x g for 20 seconds and the excess M9 aspirated. The pellet of worms was immediately used to isolate RNA as described below.

RNA Isolation

A P1000 pipette with a large orifice tip was used to transfer the majority of the pellet of worms (and minimal additional fluid) to a high-strength 15 ml conical tube. Depending on the size of the pellet, 7 – 10 ml of Trizol® was added to the tube which was then shaken briefly and immediately put it in a liquid nitrogen bath. Once the contents of the tube were completely frozen, the tube was placed in a water bath at 65° C until it was completely thawed. It was then shaken and refrozen in liquid nitrogen. Twelve freeze/thaw cycles were performed.

After the final thaw, each tube was shaken by hand or with a vortex for 30 seconds and then put on ice for 30 seconds. This process was repeated seven more times and then the tubes were allowed to stand at room temperature for 5 minutes.

Next, 1 – 2ml of chloroform (i.e., 2 ml per ml of packed worms in the original pellet) was added to each tube in a fume hood and the tubes were shaken for 15 seconds by hand then allowed to stand 2 – 3 minutes at room temperature. The tubes were then centrifuged at 5500 x g for 35 minutes at 4° C. The phase containing the total RNA was then transferred to another tube and the RNA precipitated overnight with isopropyl alcohol at –20° C. The next day, the solution was centrifuged at 5500 x g for 25 minutes, then the isopropyl alcohol was poured off leaving a pellet of total RNA that was then washed with 5 ml of 75% ethanol. After centrifuging at 5500 x g for another 8 minutes at 4° C, the ethanol was poured off and the pellet air-dried for 15 minutes before being re-suspended in 500 μl water treated with diethylpyrocarbonate (DEPC) and transferred to an RNAse free 1.5 ml polypropylene Eppendorf tube.

The total RNA was stored at –80° C until it was shipped on dry ice to Oregon State University where the microarray hybridizations were performed.

Data Analysis

Determination of Expressed Regions

The microarray data from the three replicates in the lab environment and three replicates in the soil environment were processed using the Affymetrix Tiling Array Software (TAS) version 1.1.02 and BPMAP file version Ce25b_MR_v02-2_ce4. The analysis parameters were as follows: Bandwidth 70, Max Gap 70, Min Run 45, Threshold 20 (p-value of 0.01). For each probe on the array, TAS determines whether transcription occurred at the corresponding genomic location by considering the signal levels of the PM and MM probes for that location as well as those for neighboring locations.

The bandwidth parameter tells TAS what size neighborhood to use in its calculations. A bandwidth of 30 would include all probes whose midpoints are within 30 bases of the midpoint of the probe being evaluated. Since the great majority of probes are centered ~25 bases apart, the neighborhood defined by a bandwidth of 30 would include the probe being evaluated plus (in most cases) one probe on each side of it. A bandwidth of 15, on the other hand, would rarely include any neighboring probes. Using a large bandwidth reduces the algorithm's sensitivity to noise but can also cause it to overlook the legitimate expression of very short exons. Unfortunately, with a median length of ~150 bases and a modal length even shorter than that[2], the *C. elegans* genome contains numerous relatively short exons. The bandwidth of 70 used in this work defined a neighborhood consisting of the probe in question plus (in most cases) two probes corresponding to adjacent genomic locations in each direction (for a total of five probes). Using a significantly lower bandwidth would have caused TAS to determine expression based on too few probes and/or an inconsistent number of them and would have made the results noisier and less robust.

Not surprisingly, a region is considered to be expressed if probes corresponding to that region are themselves considered to be expressed. The probes "tile" the genome with an average spacing of about 25 bases and TAS has to assume that expression continues across and between adjacent expressed probes. On the other hand, if the adjacent probes are too far apart or if the signal from a probe does not reach significance, the assumption that a single, continuous transcript spans the entire region is less plausible. The Max Gap parameter is the largest space that TAS will tolerate in an expressed region. A space greater than Max Gap will cause the expressed region to be split into two separate regions. Because the gap is measured between the centers of successive probes, a gap of 70 usually corresponds to just two probes, each showing no (or inadequate) expression. Likewise, based on a typical spacing of 25 bases, a gap of 50 would encompass one or two probes with approximately equal frequency. In *C. elegans*, a Max Gap of 70 sometimes causes an entire intron to be "bridged" (a somewhat undesirable situation) but this is an acceptable tradeoff overall because it often compensates for one or two probes that, due to excessive noise, show uncertain probability levels in a region where transcription most likely occurred.

The Minimum Run parameter represents the shortest sequence that can be considered to be expressed. While *C. elegans* has many short exons, using too small a value for Min Run results in excessive noise in the resulting data. A Min Run of 45 bases will virtually always span two successive probes that must each show expression for the region to be considered expressed.

A threshold value of 20 was chosen to limit the amount of noise in the analyzed data and to establish robust statistical support for the experimental findings. The threshold is ten times the log-transformed p-value such that a threshold of 20 corresponds to a p-value of 0.01 (or simply, one percent). This p-value equals the probability that expression was indicated as "present" when in fact it was not but seemed to be due to chance alone.

The parameters used in the TAS analysis were chosen based on the physical design of the Affymetrix *C. elegans* Tiling 1.0R Array and the known properties (i.e., intron and exon size distributions) of the *C. elegans* genome itself. They were not chosen with the intention of showing or hiding the expression of any gene or region. This unbiased approach supports statistical rigor and is consistent with the philosophy of the tiling array, which does not presume the expression of any particular parts of the genome.

Determination of Expression Levels

Using the parameters indicated above, TAS determined which regions (represented by groups of successive probes) showed expression at a statistically significant probability level. The TAS software then created a file containing a list of these regions and an additional file of the signal levels of the probes in those regions. I wrote a script named "Bedder Data" (see Appendix C) to read these two files and create a new file containing the list of regions along with an estimate of the level of expression corresponding to each region.

To account for the possibility that endmost probes might have lower values that could adversely affect the calculation (as later data analysis suggested might be the case), the Bedder Data script disregards the values of the two probes on each end and uses a pseudomedian algorithm (as does TAS itself) on the remaining ones to calculate an overall expression value. The pseudomedian is the most widely used algorithm for tiling array analysis[29].

The pseudomedian algorithm can be computationally expensive because it scales somewhat faster than the square of the number of probes[29]. For long regions (45 or more probes), the time needed to calculate the pseudomedian was prohibitive. On the other hand, large groups of consecutive probes usually have distinct median clusters. The purpose of the pseudomedian calculation is to minimize the effects of noise but in larger data sets, noise tends to cancel itself. For these, the Bedder Data script sorts the values and disregards the highest and lowest eighth. The arithmetic mean of the remaining values is considered the expression level in that region.

Determination of Expressed Genes

The list of expressed regions was compared to a complete list of *C. elegans* exons downloaded from WormBase (build WS 180, 30 July, 2007). This was done using Microsoft Excel with one spreadsheet file per environment for each chromosome. A procedural computer language such as Perl could have performed the comparisons and might have done so more efficiently but the spreadsheet format facilitated algorithm development and visual inspection of intermediate calculations and unanticipated results.

In the WormBase data, each exon is identified by its unique WormBase ID, its starting and ending locations on the chromosome, its strand, transcript type (e.g., coding, rRNA, miRNA, et cetera), prediction status (confirmed, partially confirmed, or predicted), and the starting and ending locations of the gene to which it belongs. Multiple, overlapping versions of some exons are listed in WormBase and they are represented by separate records. Additionally, some exons are associated with more than one gene and they are represented by one record per gene per exon.

Because data from the Affymetrix *C. elegans* Tiling Array are strand-agnostic, the strand that transcripts correspond to can only be inferred. In most cases, expressed regions are expected to coincide with annotated exons so the default assumption is that the transcription is of that exon and therefore corresponds to the same strand as the exon. Expressed regions that overlap an annotated gene but do not coincide with any annotated exon of that gene can be dealt with similarly. In fact, the transcript may represent an un-annotated exon. Either way, the same default assumption can be applied even though the evidence may be considered less compelling.

That default assumption is nullified in the next step of the analysis if the gene(s) that include(s) the region in question have insufficient overall coverage. For example, a gene with a length of 600 bases was not considered to be "expressed" if the transcribed region(s) covered less than 90 bases (15%) of its length. In such cases, the signal indicating expression in that region was assumed to have been a false positive or to have reflected some other effect entirely. Indeed, expression of an antisense transcript for part of a known gene would be indistinguishable from expression of the gene itself in a strand-agnostic setting. In such a case, the signal could indicate the exact opposite meaning as the default assumption—the antisense transcript could actually indicate repression, not expression, of the gene. To address this, genes with an overall coverage of less than 15% were disregarded in this study.

Determination of Differential Expression

As noted above, expression levels were calculated for all expressed regions that had a length of at least 250 bases completely spanning at least nine probes. The calculated expression levels were then assigned to any genes overlapping those regions. Genes spanning more than one region with a calculable expression level were assigned a value equal to the weighted average of its constituents' levels.

Prior transcriptome comparisons have shown that most genes expressed in one environment are also expressed in the other. Typically, the expression levels of such mutually expressed genes are compared to see if there is a significant difference. In this study, any mutually expressed genes whose expression levels could be calculated in both environments were to be compared (see Results). Any mutually expressed genes whose expression levels could not be calculated could not be compared, of course. Genes expressed in just one environment must be regarded as being differentially expressed whether a level of expression can be calculated or not, of course.

In this work, a significantly different rank ordering of the expression levels of a gene between the two environments is referred to as a reprioritization of that gene's expression. Furthermore, no attempt is made to guess any presumed fold-difference in expression level between the two environments. In this way, differential gene expression can be addressed without relying on untenable arguments whose purpose is to make a more specific (but arguably unsupportable) statement about relative levels of expression.

To determine which genes had been reprioritized in the lab environment compared to the soil-like environment, a master list was made of all the genes found to be expressed in the experiment. Then genes that were only expressed in one environment were eliminated from the master list. Next, genes whose level of expression could not be determined were eliminated from the list. Each of the remaining genes was expressed in both environments and had a calculated level of expression in each. The genes were then ranked by their expression level in each environment. Although genes' ranks in the two environments invariably differed, the differences could not necessarily be considered significant due to uncertainty in the levels of expression used to calculate the rankings.

To determine significant differences in rankings (i.e., reprioritization), a new maximum and minimum value were calculated for each gene in each environment. These new extreme values were the original value times 1.5 for the maximum and times 2/3 for the minimum. The maximum and minimum expression values were then used to find maximum and minimum rankings for each gene in each environment. Finally, the range of rankings in one environment was compared to the range in the other for each gene. If they did not overlap, that gene was determined to have been reprioritized.

COG analysis

COGs are Clusters of Orthologous Groups of proteins from prokaryotic and unicellular eukaryotic organisms[31]. There are four COG groups, namely: Cellular Processes and Signaling, Information Storage and Processing, Metabolism, and Poorly Characterized. Each group is divided into categories with each category represented by a single-letter code. There are 25 COG codes currently defined; the letter X is unused (See Appendix B). All the genes in a COG category are assigned the same code. Genes without clear orthologs in multiple distantly related species are not assigned a COG code. KOGs are eukaryote-specific COGs and follow the same conventions[32]. KOGs can be used together with COGs as in WormBase[33].

The COG code (if any) of each gene in this experiment was downloaded from WormBase. Sets of genes were then defined based on whether the genes were expressed in both environments mutually or whether they were expressed in the lab or soil-like environment only. The mutually expressed genes were then subdivided into three sets: those prioritized in the lab environment, those prioritized in the soil-like environment, and those whose prioritization could not be determined or was not significantly different.

A Chi Squared Goodness of Fit (also known as Kolmogorov-Smirnov) test was performed for each COG code to determine whether the fraction of genes with that code in any of the defined subsets was different from the overall fraction. The test was also performed for those genes that did not have a COG code. The five sets of genes represented four degrees of freedom for these tests.

Intergenic Expression

If transcribed regions do not correspond to any annotated gene, they are commonly referred to as "intergenic expression". Intergenic expression has been found in yeast, *C. elegans, Drosophila,* rice, mice, and humans[27,28,34-36]. Given that, and the way the data analysis parameters were chosen, and the fact that the overall results were consistent with previous research, the intergenic expression found in this experiment is presumably legitimate and biologically relevant.

As with the other regions found to be expressed in this experiment, there is no way to know with certainty which strand(s) were transcribed to produce the intergenic expression observed. Informed guesses were made regarding the other expressed regions because they coincided with known or putative genomic features but that is not possible with intergenic expression.

Visualization of transcribed regions including intergenic expression was performed using the Integrated Genome Browser (IGB) program from Affymetrix. IGB can be downloaded free of charge from:

http://www.affymetrix.com/support/developer/tools/download_igb.affx

CHAPTER II

RESULTS

Expressed Regions and Genes

At a p-value of 0.01 (99% confidence level), 42,187 regions were found to be expressed in the lab environment. The expression of 44,812 regions was found in the soil-like environment. In each environment, most of the expressed regions corresponded to the locations of annotated genes. In the lab environment, 3,938 regions (9.33% of the 42,187 expressed) were strictly intergenic while in the soil-like environment, there were 5,140 intergenic regions showing expression (see Table 1).

Environment	Number of Expressed Regions	Number of Genes	Covered Genes	Env-specific genes	Number of Intergenic Regions	Percent of Expressed Regions
Lab	42,187	11,620	8,365	1,221	3,938	9.33
Soil-like	44,812	12,194	8,868	1,724	5,140	11.47

Table 1. Summary of Expressed Regions, Genes, and Intergenic Expression
"Covered Genes" is the number of genes that showed expression on at least 15% of their annotated length. "Env-specific genes" is the number of covered genes expressed in one environment but not the other.

The expressed regions overlapped more than 11,000 annotated genes in each environment but over a quarter of those genes had very low coverage. That is, only a small fraction (less than 15%) of the length of the gene showed expression (e.g., see the "degenerin" gene in Figure 2). Genes with low coverage were eliminated from further consideration or analysis.

Of the 8,365 "covered" genes in the lab environment, 1,221 were specific to that environment. The other 7,144 were mutually expressed in both the lab and soil-like environments. Along with the 1,724 covered genes expressed exclusively in the soil-like environment, a total of 10,089 annotated genes were detected in this experiment (see Figure 1). That number represents almost half of the annotated *C. elegans* genome[33].

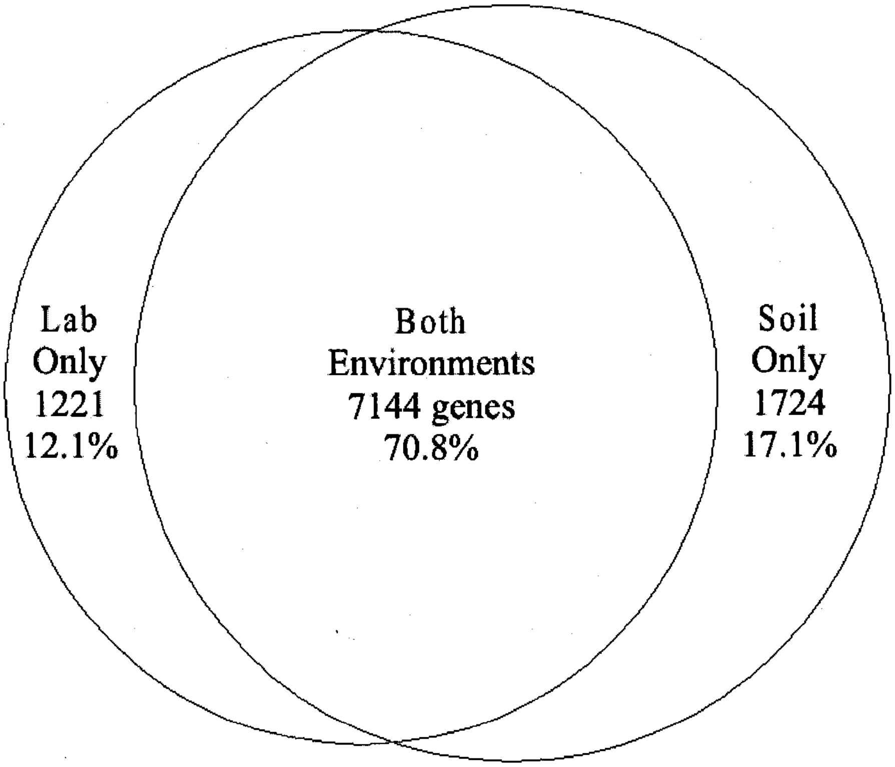

Figure 1. Proportional Venn Diagram of Genes Expressed by Environment
Percentages are based on the total (i.e., 10,089 genes).

Of the 10,089 genes found, 94.3% encode proteins. The balance consists of pseudogenes and various non-coding RNAs. An analysis of the expressed genes on chromosome I failed to show any environment-specific bias in the protein coding or RNA genes.

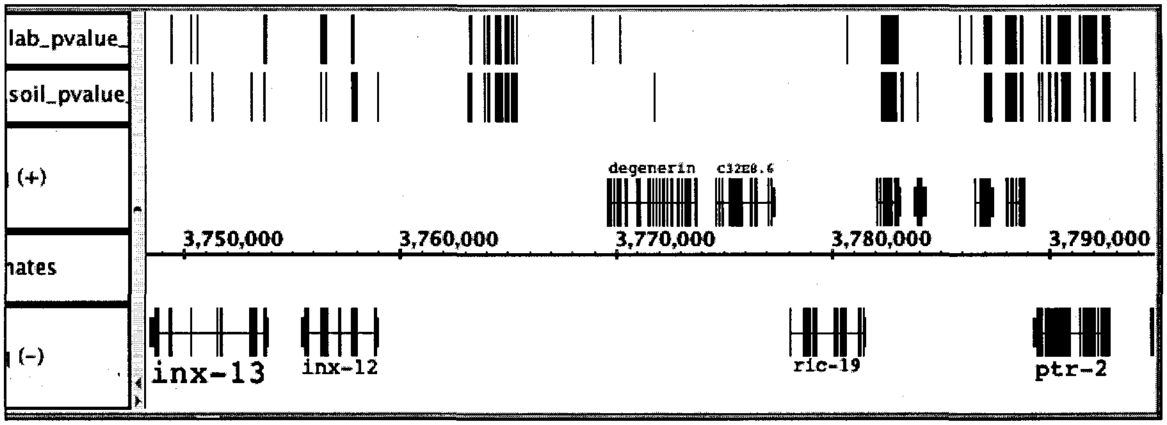

Figure 2. Annotated and Expressed Tracks in the Integrated Genome Browser
This image shows five horizontal "tracks" in IGB. The track labeled "lab_pvalue" depicts transcribed regions in the lab environment as vertical lines or bars. The soil_pvalue track is expression in the soil-like environment. The (+) and (-) tracks depict annotated genes (on the + and – strands, respectively) as vertical lines or bars connected by a horizontal line. The "Coordinates" track shows a horizontal line representing the chromosome marked with genomic locations.

Figure 2 shows a small section of chromosome I as displayed in IGB version 5 (see Materials and Methods). Note the intergenic expression in both environments between 3,760,000 and 3,770,000. Although the cluster appears "gene-like", it is unknown whether that expression does, in fact, represent a protein coding gene. A BLASTN search finds several nearly identical sequences on each *C. elegans* chromosome. A BLASTX search for protein sequences corresponding to the nucleotide sequence finds a few good matches to parts of hypothetical or predicted proteins in *C. elegans*, several for *Nematostella vectensis* (starlet sea anemone), and one each for *Aedes aegypti* (a mosquito) and *Plasmodium falciparum* (the malaria parasite). Nothing else matches with an E-value of 0.1 or better. A careful inspection of the sequence shows that it has several occurrences of the 25 base sequence "aaatattttactctctggcttcacc" and variants thereof. Repetitive sequences are considered by some to be irrelevant "junk DNA".

Of the 10,089 annotated genes expressed in this experiment, 635 (6.3%) are listed as "Predicted" in WormBase[33] — that is, the genome sequences that represent those genes appear to encode proteins but their expression has never been detected by EST or cDNA. The other 9454 genes have been at least partially confirmed. Although the overall frequency of Predicted genes expressed in this experiment was 6.3%, the frequency was higher among genes expressed in just one environment and lower among genes mutually expressed in both environments. Specifically, just 4.2% of the 7144 mutually expressed genes were considered Predicted while 9.7% of the 1,221 lab-only genes, and 12.5% of the 1,724 strictly soil-like genes are Predicted.

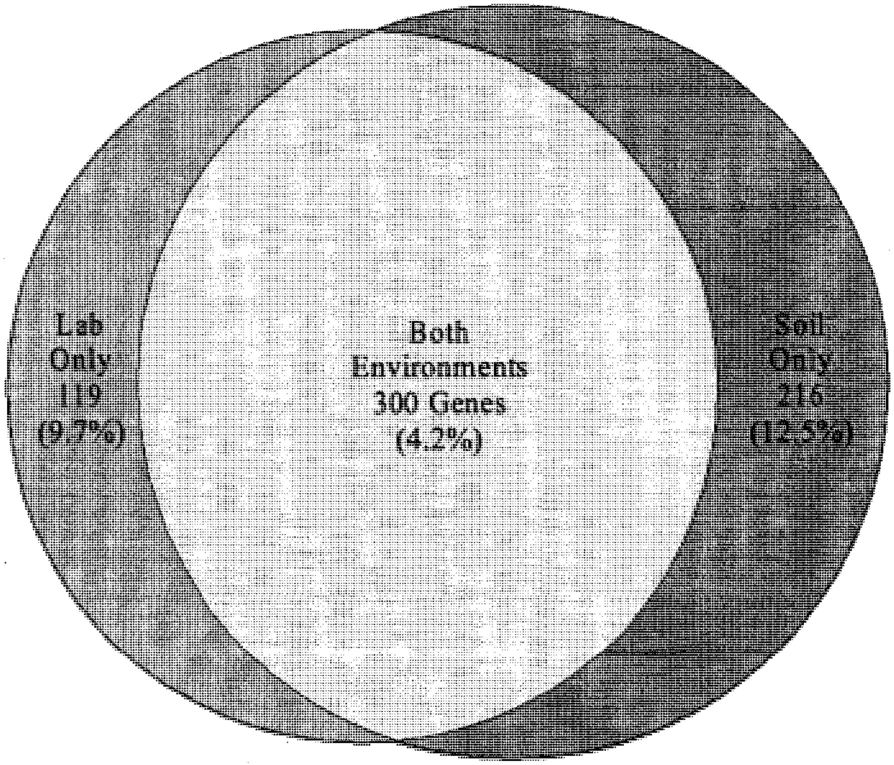

Figure 3. Numbers and Frequencies of the 635 Predicted Genes by Environment
The percentage shown in any section of the diagram is the number of Predicted genes in that section (shown) divided by the total number of genes in that section (see Figure 2).

Expression Levels

Although TAS had normalized the data and merged all three replicates from each environment, the signal levels of probes within any expressed region varied dramatically—very commonly by at least an order of magnitude (see Table 2). In expressed regions represented by a small number of probes, there was no way to calculate a well-supported estimate of the expression level. On the other hand, regions with a length of at least 250 bases fully encompassing at least nine probes tended to have a cluster of median values from which one might calculate a plausible expression value for the region. To the extent that such a cluster existed however, the endmost probes were commonly found to have lower values than the cluster.

The Bedder Data program described in Materials and Methods and listed in Appendix C was effective in using a pseudomedian algorithm to calculate an expression value for expressed regions of sufficient length. For very long expressed regions, it used a less computationally intensive algorithm to determine an expression value.

Table 2 contains a list of several consecutive lengthy expressed regions from an arbitrarily chosen section of the X chromosome. These data were extracted from the log file created by the Bedder Data program. The first region listed starts at position 11,646,792 on the chromosome and the last starts at 11,913,782. For each region, the table shows the number of probes along with the highest and lowest values eliminated from each end of the distribution and the mean of the remaining values. Numerous shorter expressed regions were interspersed in this section of this chromosome but even among these transcripts that span numerous probes, the noise level easily dominates the presumed signal. Note that even after eliminating the extreme one fourth of the values, the remaining ones vary by a factor of two to four. An inspection of Bedder Data log file entries from across the *C. elegans* genome revealed similarly large signal value variations in all cases.

Region Number	1	2	3	4	5	6	7
Number of Probes	56	58	216	49	72	83	145
Lowest signal level	34	58	92	1477	103	53	63
Highest "Low" signal eliminated	199	179	187	2055	169	142	126
Lowest "High" signal eliminated	692	818	711	5113	422	483	440
Highest signal level	1426	961	1154	7767	491	651	658
Calculated Expression Value	409	472	385	3124	281	322	256

Table 2. Ranges of Signal Values for several lengthy Expressed Regions.

Differential Expression

Of the 7144 genes mutually expressed in both the lab and soil-like environments, signal levels could not be calculated for almost half (48.3%). There was undoubtedly differential expression among those genes but it was impossible to quantify in this experiment. Of the mutually expressed genes for which an expression level could be calculated, over 95% had expression levels that described comparable rankings in both environments. These genes could not be described as differentially expressed.

The remaining 175 genes consisted of 74 that were prioritized in the lab environment and 101 that were prioritized in the soil-like environment (see Appendix A). That is, even once a reasonable amount of uncertainty was considered, the gene's expression rank was significantly higher in one environment than in the other. While many microarray experiments find that genes are up- or down-regulated in one environment versus the other, that claim cannot be made here because there is no guarantee that the underlying expression level distributions are the same between the two environments. Instead, it was determined that some genes were expressed at a higher priority in one environment than in the other.

COG Analysis

Of the 10,089 annotated genes whose expression was detected in this experiment, 6,414 have a COG code. Table 3 summarizes these data in columns representing five groups of genes: those expressed in the Lab Only, Lab Prioritized genes, genes mutually expressed in both environments without discernible prioritization, Soil-like Prioritized genes, and Soil-like Only genes. For each group, there are 26 rows: one for each of the 25 COG codes plus one row for genes without a COG code. The values in the table are the numbers of genes expressed in that environment that fall into that COG category.

A Chi Squared Goodness of Fit (also known as Kolmogorov-Smirnov) test with a P-value of 0.05 was performed to determine whether the fraction of genes with any given category is different in any of the defined groups than it is overall. According to the results of that test, fifteen of the values are significantly different than expected. They are framed in the table. Values in the table that are bold are significantly higher than expected; italicized values are significantly lower. Most of the significantly different values are framed with a double line but a few are several times the threshold value and have been framed with a triple line.

Lab Only	Lab Prioritized	Equal Expression	Soil Prioritized	Soil-like Only	COG Group	COG Description
13	0	141	3	12	Cellular Processes and Signaling	Cell cycle control, division, chrom. partitioning
1	0	27	0	**11**		Cell wall/membrane/envelope biogenesis
1	1	35	3	0		Cell motility
57	5	424	7	*43*		Posttranslational mod, protein turnover, chaperones
93	4	551	8	**185**		Signal transduction mechanisms
19	1	205	1	23		Intracellular trafficking, secretion, ves. transport
16	1	74	1	14		Defense mechanisms
21	1	157	6	**43**		Extracellular structures
1	2	26	0	0		Nuclear structure
9	1	127	2	11		Cytoskeleton
15	3	241	1	*12*	Information Storage and Processing	RNA processing and modification
5	**10**	125	0	*2*		Chromatin structure and dynamics
19	1	296	0	*17*		Translation, ribosomal structure and biogenesis
43	4	283	4	*23*		Transcription
28	2	123	2	11		Replication, recombination and repair
16	3	189	0	24	Metabolism	Energy production and conversion
12	0	115	3	17		Amino acid transport and metabolism
10	0	51	0	9		Nucleotide transport and metabolism
24	0	146	1	32		Carbohydrate transport and metabolism
7	1	31	0	4		Coenzyme transport and metabolism
35	2	181	3	33		Lipid transport and metabolism
12	0	84	5	**26**		Inorganic ion transport and metabolism
16	1	40	1	15		Secondary biosynthesis, transport & catabolism
122	9	613	12	119	Poorly Characterized	General function prediction only
92	4	521	6	106		Function unknown
534	18	2159	32	**932**	None	These genes have no COG Code

Table 3. COG Code Analysis of Expressed Genes
Numerical values represent the numbers of genes in that COG category (row) for that expression category (column). The term "Equal Expression" includes genes whose expression levels were comparable in the Lab and Soil-like environments as well as those genes mutually expressed with ambiguous expression levels. "Soil Prioritized" indicates genes expressed with a higher priority in the Soil-like environment.

Most of the results that differ significantly from the expected values are among the genes expressed in the Soil-like Only environment. In the "Cellular Processes and Signaling" COG Group, three categories are over-represented and one is under-represented in the soil-like environment. None of the values in any of the other environments is significantly higher or lower than expected for this COG Group. The "Signal Transduction Mechanisms" category is dramatically over-represented. Gene types in this category include membrane pores and receptors, kinases, and calmodulin genes. Also over-represented are extracellular structures and membrane biogenesis genes such as collagens (presumably for building or maintaining cuticles) and choline kinase. For some reason, protein turnover genes are somewhat under-represented in this group relative to the others.

In the "Information Storage and Processing" COG Group, environment-specific genes are commonly under-represented. Obviously, genes involved in DNA and RNA management tend to be essential and heavily expressed in all environments. It would be somewhat surprising if there were many environment-specific genes of this type. In the lab prioritized group, however, there is a set of ten histone proteins that are expressed at an elevated priority. One possible explanation for this results from the fact that the worms in the lab environment were better synchronized than those in the soil-like environment. If the adult worms or their eggs express a stage-specific set of histones, it would be very possible to see more in the lab environment.

In the "Metabolism" COG Group, one category is over-represented in the lab environment and one in the soil-like environment. These could result from physical differences between the two environments—unlike the lab environment, the soil-like environment had organic and inorganic elements extracted from soil. These differences could be important and could explain differences in gene expression.

The lab environment showed a relative excess of Poorly Characterized proteins. The number was only slightly higher than that in the soil-like environment but it was still statistically significant. The best way to understand this finding may be to compare it to the genes without any COG Group. In that category, both environment-specific groups of expressed genes have large overabundances that are statistically very significant. In other words, genes that are unusual or poorly understood are likely to only be expressed in an environment-specific manner and conversely, genes that are expressed in an environment-specific manner are less likely to be well understood.

Intergenic Expression

Possibly the most surprising findings of this experiment are the expression from intergenic regions. The majority of expressed intergenic regions are specific to one environment or the other but the characteristics of the regions expressed in both environments are strikingly different from those expressed in just one. Table 4 summarizes the relevant statistics.

Environment-specific intergenic transcripts are shorter, overall, than those expressed in both environments. The mode and the median of the former are only two thirds those of the latter and the mean values are only half as great. Although there are 2.5 times as many environment-specific intergenic regions, there are just a couple that are longer than 500 bases while there are 120 of the mutually expressed transcripts that exceed 500 bases. Although long transcripts expressed in one environment would be inherently more likely to overlap long transcripts expressed in the other, these transcripts comprise just a fraction of a percent of the entire genome so it is unlikely to be a coincidence that they occupy the same locations.

	Mutually Expressed		Expressed in just one environment	
	Found in Lab	In Soil-like	Lab Only	Soil-like Only
Count	1282	1240	2656	3900
Shortest	46 (total of 3)	46 (2)	46 (19)	46 (21)
Modal Length	79 (31)	79 (29)	54 (169)	54 (235)
Median Length	107	117.5	64	68
Mean Length	165	181	77	81
5th Longest	1713	1841	299	368
4th Longest	1843	1902	306	373
3rd Longest	1952	2057	366	464
2nd Longest	1981	2273	445	498
Longest	2057	2923	718	595
# > 500	57	63	1	1

Table 4. Summary of Intergenic Expression by Environment
The two columns of numbers on the right represent expressed intergenic regions found in one environment that did not overlap any expressed region in the other environment. The leftmost two columns of numbers represent regions found in one environment that overlap one or more expressed regions in the other. The top row of values is the number of expressed regions in the category. The second row of values lists the shortest region found and the number of regions of that length. All lengths are in bases. The Modal Length is the mode (peak) of the distribution. The value in parentheses is the number of expressed intergenic regions of that length. The bottom row is the number of regions longer than 500 bases in that category.

Figure 4 shows a histogram of the distributions of expressed intergenic regions by environment. The labels on the horizontal axis indicate the sizes of regions in the bin. The first bin includes all transcripts 70 bases or less in length. The next bin is from 71 to 95 bases in length and each successive bin increases the lengths by 25 bases. The lab-only and soil-like only distributions are virtually identical. Likewise, the histograms of expressed regions from one environment that overlap those from the other are quite similar. The interesting aspect of the histograms is that there is no obvious reason why any of the four curves should be any different from the others. The fact that the Soil Overlapping and Lab Overlapping distributions are strikingly different from the environment-specific distributions is simply unknown. If it is not the result of some kind of systematic error in the experiment or analysis, it will presumably be a biologically important finding.

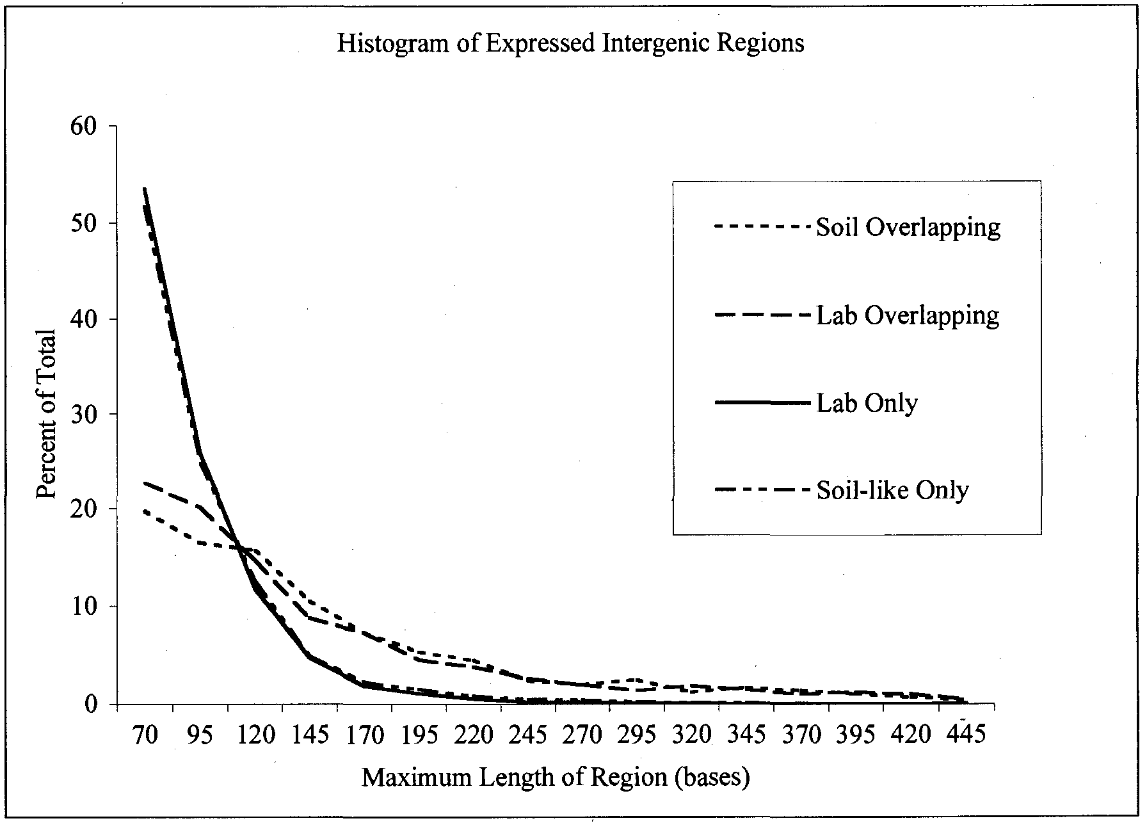

Figure 4. Histogram of Intergenic Length Distributions by Environment

CHAPTER III

DISCUSSION

The purpose of this experiment was to compare the transcriptomes of *C. elegans* cultured in a soil-like environment to those from a traditional lab environment. There was evidence to suggest that worms might express different genes when cultured in a soil-like environment[15-17,21,37-40] and exploring this possibility was an underlying goal of the research.

After developing protocols for the creation of a soil-like environment, then culturing and harvesting the worms, and isolating their RNA, the microarrays were hybridized and the data analysis performed. The Affymetrix *C. elegans* Tiling Array that was used represented a tremendous opportunity but also provided significant technical challenges. Ultimately, the challenges were overcome and important, statistically significant findings resulted.

Using the Affymetrix Tiling Array Software (TAS) with appropriate settings revealed the expression of tens of thousands of genomic regions. This was consistent with the only published paper that used this particular tiling array[28]. The complete list of annotated exons was downloaded from WormBase[33] and used to match the expressed regions to their corresponding genes. A significant number of the expressed regions did not correlate with any annotated sequence but this was consistent with findings from a variety of recent publications[28,34-36,41].

The Integrated Genome Browser from Affymetrix was used to assess the expressed regions and their relationships to the annotated genome. Determining patterns of expression is an important goal that is best approached using a variety of tools. Some visually odd patterns of expression appeared in some sections of the genome and preliminary statistical analyses suggested that some of the data might be incorrect. Affymetrix subsequently issued an updated version of a file used by TAS and the data were re-analyzed. The result was a lower but still significant incidence of intergenic expression and better correlations between the expressed regions and annotated sequences. Analyses based on the earlier version of the TAS file[28] would presumably need to be updated as well to correct any erroneous findings.

During that process, it was also reiterated that some probe pairs on the tiling array correspond to several genomic locations and that this can cause evidence of expression with high confidence to be reported for a particular location even if it resulted from expression at a completely different location.

As part of the data analysis, expression levels were calculated, where possible, for the transcribed regions. The value of the expression data was severely limited by the amount of noise in the signal. This is generally considered typical for microarrays[23,26,29]. Nonetheless, one can hope that with further study, some of the sources of noise (e.g., GC content of the probe sets) that affect the array used in this experiment could be enumerated and better understood or possibly even compensated for to some extent. If that were possible, better estimates of expression levels would allow much better analyses of prioritization of gene expression.

In many microarray experiments, two treatments or samples being analyzed are identical except for one particular experimental element. In that kind of situation, it is plausible that the distributions of expression values and the medians thereof will be essentially the same. This provides a basis for calculating a fold-difference in expression levels. Plausibility is not proof, of course, but the argument is commonly used.

In two very different environments such as those used in this experiment, it is more difficult to build a case that the distributions can be assumed the same. Certainly, the soil-like environment used in this experiment is very different from the lab environment. Furthermore, the lab environment is designed to be completely homogeneous while the soil environment was designed with the opposite goal in mind. In the soil environment, it was hoped that additional biochemical pathways would be activated–whether the levels of expression of the other pathways were scaled down or not. These factors combine to make the commonly used assumptions less likely to be valid in this experiment than they are in most microarray experiments.

During the analysis of microarray data, there are several steps where averaging and normalization occur. Although the raw signal intensity values from the microarray are less-than-perfect proxies for expression levels to begin with, subsequent mathematical processing can change those values very significantly and in non-linear ways. In fact, the most significant statistic that is likely to be preserved throughout the data analysis process may be the rank ordering of the expression levels. Accordingly, findings of differential expression based on the rank ordering (referred to here as differential prioritization) can be made with greater confidence than those based on a presumed relative level of expression.

To the extent that gene expression levels could be determined, an analysis of the expression level ranking showed that some genes were expressed at a higher priority in one environment versus the other. The 175 genes found to be differentially prioritized are listed in Appendix A. In many more cases, differential expression meant expression in just one environment. In this experiment, there were 2945 genes expressed in one environment but not the other.

Among genes that were found to be differentially expressed, there were statistically significant differences in two traits that were considered: the genes' status as Predicted versus Confirmed, and the genes' COG[31,32] codes.

Of the 10,089 genes whose expression was detected in this experiment, there were many (635 or over 6%) whose status is listed in WormBase as "predicted". This means that in experiments to date, no transcripts from those sequences have been found in EST/cDNA libraries[42]. Genes expressed in just one environment were significantly more likely to be listed as Predicted than genes detected in the experiment overall. Additionally, genes expressed only in the soil-like environment were more likely to be considered Predicted than genes expressed only in the lab environment.

Finding that such a gene has been detected in an experiment is interesting, but the situation is complex. In fact, the expression of many of these genes has reportedly been detected in other experiments[43-46] or they may have a recognizable knockout or RNAi phenotype. There are also experiments in which green fluorescent protein (GFP) is expressed at the same time and location as the gene in question. In other words, the "predicted" genes that were detected in this experiment may be relatively mundane even if they are less well understood than others.

The COG codes of differentially expressed genes were also quite different. Most of the COG code differences appeared only in the soil-like environment but some were in the lab environment and one was even in the lab-prioritized group. Differences were found in the metabolic genes expressed from one environment to the other but an even more intriguing group of Cellular Processing and Signaling genes were being expressed in the soil-like environment. Further investigation of that result is certainly warranted. The most striking COG code differences among differentially expressed genes were in those genes that did not have COG codes. This result was not surprising, however.

In cases where there are clear differences in gene expression between the two environments studied in this experiment, these differences cannot *necessarily* be attributed to particular properties of those environments. This experiment began with a large group of animals synchronized at the L1 stage of development. When the animals were collected for analysis, those from the lab environment were still well synchronized but those from the soil-like environment had developed somewhat slower and at different rates. Virtually all of those in the lab environment were adults bearing eggs. The typical animal from the soil-like environment was also an adult bearing eggs but there were significant numbers of young adults without eggs as well as animals that were still in their fourth, third, or even second larval stage of development. Thus, some of the gene expression observed in the soil-like environment could be the stage-specific expression of the less well-developed animals and have nothing whatsoever to do with the soil-like environment *per se*.

Another pitfall to avoid is that of making overly broad assumptions about the implications of the numbers found in this study. It is easy to notice that the number of genes that were significantly expressed in the soil-like environment is larger than the number from the lab environment. It may be tempting to assume that the much greater biological and physical complexity of the soil-like environment was the reason for this. While that may well be a legitimate reason for a greater variety of genes expressed, whether it actually happened cannot be determined from the data. As noted above, stage-specific expression of developmental genes presumably contributed to a larger variety of RNA transcripts in the soil-like environment independently of its complexity.

On the other hand, there is no assurance that there actually *were* a larger number of genes expressed in the soil-like environment. It is possible that there were numerous genes expressed in the lab environment that simply were not detected at statistically significant levels.

The processing of the microarrays for this experiment was done at different times and under slightly different conditions. While most of this was unintentional and unavoidable, some adjustments were made to the processing of the lab environment samples to address issues of noise. The main such issue was that the samples from the lab environment had significant levels of prokaryotic (specifically *E. coli* strain OP50) RNA. Adjustments were made in the processing to help maximize the useful data collected from the microarrays. These adjustments generally resulted in improvements but in one case, a microarray had to be discarded as unreadable and later replaced with one from a new batch. The variability introduced by these adjustments made it somewhat less likely that any particular transcript from the lab environment would be recognized as being statistically significant.

Nonetheless, there were numerous differences in the numbers and types of genes and in the intergenic locations where transcription occurred. Furthermore, there is good statistical support that these findings are "real" and meaningful.

Possibly the most interesting findings from this experiment were in the "intergenic" regions. Statistically, two types of intergenic expressed regions were found. One is specific to one environment or the other and is typically quite short. The other type of intergenic expression is common to both environments and is longer on average. These are more similar in length to the exons of protein coding genes and may be found near annotated genes. The shared transcripts sometimes appear in small groups as was seen in Figure 2. In some cases, these transcripts could be previously unrecognized 5' exons of nearby protein coding genes. Some others might be complete, novel genes.

There are other possibilities as well. Shared, clustered intergenic expression could represent a group of small, non-coding RNA molecules processed from a single transcript—a sort of RNA operon, for example. The search for new types of RNA molecules is very active right now[47]. Another possibility is that these clusters encode non-protein oligopeptide sequences such as those used for quorum sensing in prokaryotes. Oligopeptides have also been found to have vital roles in developmental decisions in Drosophila[48]. On the other hand, clusters of intergenic expression may represent just one part of a larger structure that spans multiple genes.

The fact that there is a correlation between the length and environment-specificity of intergenic transcripts is currently inexplicable.

Given the fact that environment-specific expression was found, the hypothesis that the transcriptomes of worms cultured in the two environments are the same must be rejected. The hypothesis that no genes of unknown function are expressed in the soil environment cannot be rejected. Genes that are not well understood were found but none that can be shown to be completely novel. We also found evidence of differential prioritization in this experiment. Thus, we can reject the hypothesis that there would be none.

The conclusion one might reasonably draw from this experiment is that to search for the expression of genes that are not normally detected or genes whose role is unknown, one should consider exposing the organism to different environments, especially ones that may be similar to those in which it evolved. Doing so may be an effective way of revealing new information about gene expression and function as was done here.

APPENDICES

APPENDIX A

GENES PRIORITIZED BY ENVIRONMENT

The genes listed in the following two tables were determined to have been expressed with different priorities in the two environments. That is, even after accounting for noise, each gene's ranking based on its expression level was significantly higher in one environment than in the other. The first list is the 74 genes that were expressed at higher priority in the Lab environment (conversely, they could be said to have been deprioritized in the Soil-like environment). The second table is the 101 genes that were prioritized in the Soil-like environment. Note that the most interesting finding is not necessarily the one with the higher priority.

Each gene is listed according to its unique WormBase Gene ID, its COG Code (if any), and its KOG Title (if available) or other relevant information downloaded from WormBase. The lists are sorted by COG Code. The list of COG Codes can be found in Appendix B.

WormBase Gene ID	COG Code	KOG Title or Description
WBGene00007111	A	RNA-binding protein SART3 (RRM superfamily)
WBGene00012059	A	ATP-dependent RNA helicase
WBGene00021073	A	tRNA and rRNA cytosine-C5-methylase (NOL1/NOP2)
WBGene00001903	B	Histone H2B
WBGene00001884	B	Histone H4

WBGene00001895	B	Histone 2A
WBGene00001918	B	Histone H2B
WBGene00001924	B	Histone H4
WBGene00001925	B	Histone 2A
WBGene00001928	B	Histone H2B
WBGene00001878	B	Histone H2B
WBGene00001890	B	Histone 2A
WBGene00001941	B	Histone H4
WBGene00011276	C	5'-AMP-activated protein kinase, gamma subunit
WBGene00014258	C	Acetyl-CoA hydrolase
WBGene00017734	C	Electron transfer flavoprotein, beta subunit
WBGene00016061	H	Flavin-containing amine oxidase
WBGene00010872	I	Lecithin:cholesterol acyltransferase (LCAT)
WBGene00019433	I	Short-chain acyl-CoA dehydrogenase
WBGene00015920	J	RNA polymerase I-associated factor - PAF67
WBGene00000899	K	TGFbeta receptor signaling protein SMAD and related proteins
WBGene00020779	K	RNA polymerase I transcription factor UAF
WBGene00003148	K	Transcription factor MBF1
WBGene00003825	K	Predicted transcriptional regulator
WBGene00000226	L	PI-3 kinase family– mitotic growth, DNA repair, recombination
WBGene00003155	L	DNA replication licensing factor, MCM3 component
WBGene00010641	N	Predicted myosin-I-binding protein
WBGene00010557	O	AAA+-type ATPase
WBGene00019619	O	Aspartyl protease
WBGene00003951	O	20S proteasome, regulatory subunit beta PSMB5/PSMB8/PRE2
WBGene00007352	O	AAA+-type ATPase
WBGene00007605	O	Aspartyl protease
WBGene00000985	Q	Dehydrogenases with different specificities
WBGene00010709	R	Uncharacterized conserved prot. sim. to ATP/GTP-binding prot.
WBGene00009397	R	Lectin C-type domain/CUB domain
WBGene00020423	R	mRNA splicing factor
WBGene00001824	R	Zn-finger
WBGene00015181	R	Permease of the major facilitator superfamily
WBGene00000875	R	GTPase-activating protein
WBGene00001333	R	Radixin, moesin and related proteins of the ERM family
WBGene00009396	R	Lectin C-type domain/CUB domain
WBGene00016981	R	Cell membrane glycoprotein
WBGene00014199	S	Uncharacterized coiled-coil containing protein
WBGene00021748	S	Uncharacterized conserved protein
WBGene00017964	S	Uncharacterized protein
WBGene00019209	S	Uncharacterized conserved protein
WBGene00000289	T	Nerve growth factor receptor TRKA and related tyrosine kinases
WBGene00007545	T	Predicted DoH & Cyt. b-561/ferric reductase transmembrane domains
WBGene00000547	T	Cytosolic Ca2+-dependent cysteine protease (calpain), lg subunit
WBGene00011146	T	Cyclic nucleotide phosphodiesterase
WBGene00013238	U	Translocon-associated complex TRAP, delta subunit
WBGene00012253	V	C-type lectin
WBGene00004398	W	Collagens (type IV and type XIII), and related proteins

WBGene00003790	Y	Nuclear pore complex, Nup98 component
WBGene00003791	Y	Nuclear pore complex, rNup107 component (sc Nup84)
WBGene00021009	Z	Actin filament-binding protein Afadin
WBGene00001882		his-8 encodes an H2B histone; histone gene cluster HIS2.
WBGene00001894		his-20 encodes an H2B histone; predicted nucleosome component
WBGene00003977		pes-2 contains a predicted signal sequence and an F-box
WBGene00009372		Partially_confirmed
WBGene00016422		Unnamed protein, Partially_confirmed
WBGene00017066		Partially_confirmed
WBGene00000301		Caveolin
WBGene00013380		Unnamed protein, Confirmed
WBGene00017541		Unnamed protein, Confirmed
WBGene00002068		ify-1 encodes a rapidly evolving protein ligand of FZY-1
WBGene00003022		Confirmed
WBGene00015913		Confirmed
WBGene00020379		Confirmed
WBGene00021005		W03F11.1 may participate in eggshell synthesis and early dev.
WBGene00021468		Confirmed
WBGene00007481		Coding_pseudogene
WBGene00014801		Coding_pseudogene
WBGene00044939		snoRNA

Table 5. Genes Prioritized in the Lab Environment

WormBase Gene ID	COG Code	KOG Title
WBGene00022664	A	mRNA-binding protein Encore
WBGene00002050	D	Nuclear envelope protein lamin, intermediate filament superfamily
WBGene00002054	D	Nuclear envelope protein lamin, intermediate filament superfamily
WBGene00019595	D	Uncharacterized conserved protein
WBGene00003877	E	H+/oligopeptide symporter
WBGene00020788	E	M13 family peptidase
WBGene00007508	E	Aminoacylase ACY1 and related metalloexopeptidases
WBGene00022647	G	Permease of the major facilitator superfamily
WBGene00022200	I	Acyl-CoA reductase
WBGene00008629	I	Carnitine O-acyltransferase CPTI
WBGene00011321	I	Triacylglycerol lipase
WBGene00003163	K	Upstream transcription factor 2/L-myc-2 protein
WBGene00001955	K	bHLH transcription factor
WBGene00019521	K	Transcription factor Doublesex
WBGene00003107	K	Transcriptional corepressor NAB1
WBGene00004340	L	Replication factor C, subunit RFC4
WBGene00022067	L	Tam3-transposase (Ac family)
WBGene00003442	N	Major sperm protein domain
WBGene00003443	N	Major sperm protein domain
WBGene00022760	N	Major sperm protein domain
WBGene00016134	O	Hydrolytic enzymes of the alpha/beta hydrolase fold

WBGene00019986	O	Cysteine proteinase Cathepsin F
WBGene00000781	O	Cysteine proteinase Cathepsin L
WBGene00003956	O	Prolylcarboxypeptidase (angiotensinase C)
WBGene00011932	O	Serine palmitoyltransferase
WBGene00018398	O	Serine palmitoyltransferase
WBGene00021685	O	E3 ubiquitin protein ligase
WBGene00003735	P	Sodium/hydrogen exchanger protein
WBGene00016892	P	p-Nitrophenyl phosphatase
WBGene00018424	P	p-Nitrophenyl phosphatase
WBGene00015660	P	Na+/K+ ATPase, alpha subunit
WBGene00019604	P	p-Nitrophenyl phosphatase
WBGene00010790	Q	Alcohol dehydrogenase, class V
WBGene00014006	R	Predicted alpha-helical protein
WBGene00010238	R	Predicted small molecule kinase
WBGene00012209	R	HMG box-containing protein
WBGene00015074	R	N-acetyltransferase
WBGene00017968	R	Peroxidase/oxygenase
WBGene00021533	R	Leucine rich repeat
WBGene00044071	R	DHHC-type Zn-finger proteins
WBGene00006478	R	Predicted heme/steroid binding protein
WBGene00011066	R	Zn-finger
WBGene00015580	R	CLIP-associating protein
WBGene00017101	R	Uncharacterized protein with ubiquitin fold
WBGene00017430	R	Zn-finger
WBGene00011561	S	Uncharacterized protein with conserved cysteine
WBGene00013239	S	Uncharacterized conserved protein
WBGene00000397	S	Cadherin repeats
WBGene00000030	S	Uncharacterized protein
WBGene00004098	S	Uncharacterized protein
WBGene00020858	S	Conserved protein Mo25
WBGene00016541	T	Casein kinase (serine/threonine/tyrosine protein kinase)
WBGene00001178	T	EGL-Nine (EGLN) protein
WBGene00001196	T	G protein subunit Galphaq/Galphay, small G protein superfamily
WBGene00001725	T	GROUND domains (extracellular cysteine-containing domain)
WBGene00002048	T	Protein tyrosine phosphatase
WBGene00003567	T	Ca2+/Na+ exchanger NCX1 and related proteins
WBGene00016440	T	RGS-GAIP interacting protein GIPC, contains PDZ domain
WBGene00022653	T	Predicted secreted cysteine rich protein found only in C.elegans
WBGene00012765	U	Protein involved in maintenance of Golgi structure and ER-Golgi transport
WBGene00018150	V	p53-mediated apoptosis protein EI24/PIG8
WBGene00009682	W	Uncharacterized protein, contains major sperm protein (MSP) domain
WBGene00000615	W	Collagens (type IV and type XIII), and related proteins
WBGene00000657	W	Collagens (type IV and type XIII), and related proteins
WBGene00000719	W	Collagens (type IV and type XIII), and related proteins
WBGene00009031	W	Uncharacterized protein, contains major sperm protein (MSP) domain
WBGene00009684	W	Uncharacterized protein, contains major sperm protein (MSP) domain
WBGene00003369	Z	Myosin regulatory light chain, EF-Hand protein superfamily
WBGene00003514	Z	Myosin class II heavy chain

WBGene00001699		grd-10 encodes hedgehog-like protein w/signal seq & Ground domain
WBGene00006052		Unnamed protein
WBGene00006053		Unnamed protein
WBGene00010605		Unnamed protein
WBGene00021398		Unnamed protein
WBGene00022745		Unnamed protein
WBGene00004174		Predicted to contain a glutamine/asparagine (Q/N)-rich ('prion') domain
WBGene00006605		tra-2 encodes transmembr receptor; sex determ pathway in XX animals
WBGene00007308		Confirmed
WBGene00013853		Confirmed
WBGene00016752		Unnamed protein
WBGene00017058		Unnamed protein
WBGene00017789		Confirmed
WBGene00020715		Unnamed protein
WBGene00020840		Unnamed protein
WBGene00022410		Confirmed
WBGene00006050		Unnamed protein
WBGene00011748		Unnamed protein
WBGene00017542		Unnamed protein
WBGene00019435		K06A9.1 is a homolog of human TCOF1; Treacher-Collins syndrome
WBGene00021993		Unnamed protein
WBGene00043743		Unnamed protein
WBGene00007516		Partially_confirmed
WBGene00009884		Partially_confirmed
WBGene00013425		Y66A7A.5 encodes a protein with a THAP or THAP-like domain
WBGene00013558		Unnamed protein
WBGene00014179		Unnamed protein
WBGene00015765		Unnamed protein
WBGene00021078		Unnamed protein
WBGene00021625		Partially_confirmed
WBGene00011681		Unnamed protein or ncRNA
WBGene00045165		ncRNA

Table 6. Genes Prioritized in the Soil-like Environment

APPENDIX B

COG CODES SORTED BY GROUP AND BY CODE

Group	Description	Code
Cellular processes and signaling	Cell cycle control, cell division, chromosome partitioning	D
	Cell wall/membrane/envelope biogenesis	M
	Cell motility	N
	Posttranslational modification, protein turnover, chaperones	O
	Signal transduction mechanisms	T
	Intracellular trafficking, secretion, and vesicular transport	U
	Defense mechanisms	V
	Extracellular structures	W
	Nuclear structure	Y
	Cytoskeleton	Z
Information storage and processing	RNA processing and modification	A
	Chromatin structure and dynamics	B
	Translation, ribosomal structure and biogenesis	J
	Transcription	K
	Replication, recombination and repair	L
Metabolism	Energy production and conversion	C
	Amino acid transport and metabolism	E
	Nucleotide transport and metabolism	F
	Carbohydrate transport and metabolism	G
	Coenzyme transport and metabolism	H
	Lipid transport and metabolism	I
	Inorganic ion transport and metabolism	P
	Secondary metabolites biosynthesis, transport and catabolism	Q
Poorly Characterized	General function prediction only	R
	Function unknown	S

Table 7. COG Codes Sorted by Group

Code	Group	Description
A	Information storage and processing	RNA processing and modification
B	Information storage and processing	Chromatin structure and dynamics
C	Metabolism	Energy production and conversion
D	Cellular processes and signaling	Cell cycle control, cell division, chromosome partitioning
E	Metabolism	Amino acid transport and metabolism
F	Metabolism	Nucleotide transport and metabolism
G	Metabolism	Carbohydrate transport and metabolism
H	Metabolism	Coenzyme transport and metabolism
I	Metabolism	Lipid transport and metabolism
J	Information storage and processing	Translation, ribosomal structure and biogenesis
K	Information storage and processing	Transcription
L	Information storage and processing	Replication, recombination and repair
M	Cellular processes and signaling	Cell wall/membrane/envelope biogenesis
N	Cellular processes and signaling	Cell motility
O	Cellular processes and signaling	Posttranslational modification, protein turnover, chaperones
P	Metabolism	Inorganic ion transport and metabolism
Q	Metabolism	Secondary metabolites biosynthesis, transport and catabolism
R	Poorly Characterized	General function prediction only
S	Poorly Characterized	Function unknown
T	Cellular processes and signaling	Signal transduction mechanisms
U	Cellular processes and signaling	Intracellular trafficking, secretion, and vesicular transport
V	Cellular processes and signaling	Defense mechanisms
W	Cellular processes and signaling	Extracellular structures
Y	Cellular processes and signaling	Nuclear structure
Z	Cellular processes and signaling	Cytoskeleton

Table 8. COG Codes Listed Alphabetically

APPENDIX C

THE “Bedder Data” PROGRAM

What follows is the source code for the AppleScript “Bedder Data” program. The program reads a BED file of expressed regions and a file of signal values for the probes in those regions to create a new file listing the regions along with a pseudomedian signal intensity level for each region of sufficient length (i.e., 250 bases). This output file can then be used for further analysis such as determining whether regions expressed in multiple environments have been reprioritized.

The purpose of the IndicateProgress(msg) routine is to provide persistent, on-screen status messages for anyone monitoring the program’s progress.

The PseudoMedian(valList) routine determines the pseudomedian of the list of values passed in. If the list is overly long (i.e., more than 45 values), the routine instead eliminates the most extreme 25% of the values and calculates the mean of the remaining ones. This is much less computationally expensive but gives a numerically similar result.

FindWordInFile(word, file) simply parses a file (e.g., the file of signal values) for a particular word (e.g., a probe location).

The expressionValue(valFile, startStr, endStr) routine parses the file of signal values to create a list of the values at all of the probes between startStr and endStr. It passes this list to the PseudoMedian routine for a best-estimate signal value calculation for the region between the starting and ending probes.

The doParsing(bedFile, valFile, theCode) routine is the main function for parsing the two input files to create the file of merged output. It calls the routines listed above but also does a lot of the work itself. If an expressed region is shorter than some cutoff (250 bases, in this work), doParsing does not attempt to calculate an expression value and simply leaves that field blank.

The parseFile(bedFile, valFile) routine verifies that bedFile can be opened and reads enough to determine the kind of line endings it uses. Then it calls doParsing(bedFile, valFile, delimeter).

The run routine is usually the first routine executed. It asks the user for the files to parse before calling parseFile. In some cases, the files to parse have already been specified when the program is launched. In that case, the routine open(inFiles) is executed instead of run. It does some sanity checks before calling the parseFile routine.

```
global valFilePos

on IndicateProgress(msg)
    tell application "Tex-Edit Plus"
        insert time
        set the selection to ">> " & msg & (ASCII character 13)
    end tell
end IndicateProgress

on PseudoMedian(valList)
    set pseudoPseudoMed to -1
    set valLen to the number of items of valList

    if valLen < 45 then
        -- Generate all n*(n-1)/2 pairwise averages.
        set tempList to []
        set tempLen to valLen * (valLen - 1) / 2
        repeat with ii from 1 to tempLen -- Create a tempList of the appropriate size
            set tempList to tempList & -1
        end repeat

        -- Fill the list with sorted pairwise averages
        set tempLen to 0
        repeat with ii from 1 to valLen - 1
            repeat with jj from ii + 1 to valLen
                set newVal to ((item ii of valList) + (item jj of valList))
                set putItThere to false
                if tempLen > 0 then
                    set kk to 1
                    repeat while kk ≤ tempLen
                        if item kk of tempList > newVal then -- Find where the item goes in the list.
                            repeat with iii from tempLen to kk by -1
                                -- If it comes before any elements, push them down before putting in
the new item.
                                set item (iii + 1) of tempList to item iii of tempList
                            end repeat
                            set item kk of tempList to newVal
                            set putItThere to true
                            set kk to tempLen + 1 -- Done now so stop looping
                        else
                            set kk to kk + 1
                        end if
                    end repeat
                end if
                set tempLen to tempLen + 1
                if putItThere is false then set item tempLen of tempList to newVal
            end repeat
        end repeat
        set pseudoPseudoMed to 0.5 * (the middle item of tempList) -- the pseudomedian
    else -- There are 45+ values. Make a pseudo-pseudomedian by discarding the highest and
lowest 25% (combined) and averaging the rest. Not the same thing but gives us a plausible result
in a finite amount of time!
        set numExtremes to (valLen / 8) as integer
        set hiVal to []
        set loVal to [] -- Create and initialize the extreme value arrays
        repeat with ii from 1 to numExtremes
            set hiVal to hiVal & 0
            set loVal to loVal & 65536
        end repeat

        set groupSum to 0 -- Calculate the sum and find the extreme values
        repeat with ii from 1 to valLen
            set foo to item ii of valList
            set groupSum to groupSum + foo
            if foo < item 1 of loVal then -- Replace an existing loVal
                set targetLoc to numExtremes
                set jj to 2
                repeat while jj ≤ numExtremes
                    if foo > item jj of loVal then
                        set targetLoc to jj - 1 -- Figure out where the new one fits in the lineup
                        set jj to numExtremes + 1
                    else
                        set jj to jj + 1
                    end if
                end repeat
                if targetLoc > 1 then -- If necessary, slide the others down to make room
                    repeat with jj from 1 to targetLoc - 1
                        set item jj of loVal to item (jj + 1) of loVal
                    end repeat
                end if
                set item targetLoc of loVal to foo -- Put in the new value
            end if
            if foo > item 1 of hiVal then -- Replace an existing hiVal
                set targetLoc to numExtremes
                set jj to 2
                repeat while jj ≤ numExtremes
                    if foo < item jj of hiVal then
                        set targetLoc to jj - 1 -- Figure out where the new one fits in the lineup
                        set jj to numExtremes + 1
                    else
                        set jj to jj + 1
                    end if
                end repeat
                if targetLoc > 1 then -- If necessary, slide the others down to make room
                    repeat with jj from 1 to targetLoc - 1
                        set item jj of hiVal to item (jj + 1) of hiVal
                    end repeat
                end if
                set item targetLoc of hiVal to foo -- Put in the new value
            end if
        end repeat
        set groupMean to groupSum / valLen

        -- Subtract away the most extreme data
        repeat with ii from 1 to numExtremes
            set groupSum to groupSum - (item ii of hiVal) - (item ii of loVal)
        end repeat
        set pseudoPseudoMed to groupSum / (valLen - numExtremes - numExtremes)
        IndicateProgress("Lowest value is " & item numExtremes of loVal & ". Highest low value is
" & item 1 of loVal & ". Lowest high value is " & item 1 of hiVal & ". Highest value is " & item
numExtremes of hiVal & ".")
        IndicateProgress("  Pseudo-pseudomedian is " & pseudoPseudoMed & " Mean is " &
groupMean & "(difference is " & 100 * (groupMean / pseudoPseudoMed - 1) & "%).")
    end if
    -- Now round to 2 sig figs.
    if pseudoPseudoMed < 100 then
        set pseudoPseudoMed to pseudoPseudoMed as integer
    else
        if pseudoPseudoMed < 1000 then
            set pseudoPseudoMed to 10 * ((pseudoPseudoMed / 10) as integer)
        else
            if pseudoPseudoMed < 10000 then
                set pseudoPseudoMed to 100 * ((pseudoPseudoMed / 100) as integer)
            else
                set pseudoPseudoMed to 1000 * ((pseudoPseudoMed / 1000) as integer)
            end if
        end if
    end if
    return pseudoPseudoMed
end PseudoMedian

on FindWordInFile(theWord, theFile)
    set stillGoing to true
    repeat while stillGoing
        set fileBuff to read theFile from valFilePos for 32767
        set foo to the offset of theWord in fileBuff
        if foo is 0 then
            set valFilePos to valFilePos + 32767 - the (length of theWord)
        else
            set stillGoing to false
            set valFilePos to valFilePos + foo - 1
        end if
    end repeat
    --display dialog "Found '" & theWord & "' at byte position " & valFilePos & " into the file."
    IndicateProgress("Found '" & theWord & "' at byte position " & valFilePos & " into the file.")
end FindWordInFile

on expressionValue(valFile, startStr, endStr)
    set exprVal to -1
    set stillGoing to true -- Now find the beginning of the data of interest
    repeat while stillGoing
        set valBuff to read valFile from valFilePos for 16384
        set foo to the offset of startStr in valBuff
        if foo is 0 then
            set valFilePos to valFilePos + 16384 - the (length of startStr)
        else
            set stillGoing to false
            set valFilePos to valFilePos + foo - 1 -- Found the start of the data
        end if
    end repeat
    set valBuff to read valFile from valFilePos for 32767
    set maxOffset to the offset of endStr in valBuff
    --display dialog "Found '" & endStr & "' at offset " & maxOffset & " bytes in valBuff (offset from
" & valFilePos & " bytes in the file). Words include: '" & word 1 of valBuff & "', '" & word 3 of
valBuff & "', '" & word 5 of valBuff & "', '" & word 7 of valBuff & "', '" & word 9 of valBuff & "'."
    if maxOffset is 0 then
        set maxOffset to the number of words in valBuff -- Should never happen
    else
        set maxOffset to maxOffset / 4 -- Some absolute minimum average number of characters
(including white space) per word
    end if
    set wordIndex to 1
    set valList to [] as list
    set listItems to 0
    set stillGoing to true
    repeat while stillGoing -- Append items of interest to the list
        set foo to word wordIndex of valBuff
        if (foo = endStr) or (foo = "#") then
            IndicateProgress("Stopping scan at end word '" & foo & "'. There are " & number of
items of valList & " items in valList.")
            --display dialog "Stopping scan at end word '" & foo & "'. There are " & number of items
of valList & " items in valList."
            set stillGoing to false
        else
            set listItems to listItems + 1
            set valList to valList & ((word (wordIndex + 1) of valBuff) as number)
        end if
        set wordIndex to wordIndex + 2
        if wordIndex > maxOffset then
```

```
            IndicateProgress("Stopping scan because wordIndex (" & wordIndex & ") > maxOffset ("
& maxOffset & "). There are " & number of items of valList & " items in valList. The last word read
was '" & foo & "' but I was looking for '" & endStr & "'.")
            --display dialog "Stopping scan because wordIndex (" & wordIndex & ") > maxOffset ("
& maxOffset & "). There are " & number of items of valList & " items in valList. The last word read
was '" & foo & "' but I was looking for '" & endStr & "'."
            set stillGoing to false
        end if
    end repeat
    if listItems < 9 then
        set exprVal to 0
    else -- Trim the list and find the median of those values.
        --display dialog "List items are: " & (items 3 through (listItems - 2) of valList)
        set exprVal to PseudoMedian((items 3 through (listItems - 1) of valList) as list) --
Immediately trim off the endmost values
    end if
    return exprVal
end expressionValue

on doParsing(bedFile, valFile, theCode)
    set lineEnd to ASCII character 13 -- Carriage return (Macintosh line ending)
    set tabChar to ASCII character 9
    set allDone to 0
    set inFileInfo to info for bedFile
    set defName to the name of inFileInfo
    set thePrompt to ("What output file do you want for input file "" & defName & "" ?") as text
    set outFile to choose file name with prompt thePrompt default name (defName & "der")
    display dialog "What minimum length do you want to consider as 'long' exons?" default answer
"250"
    if the button returned of the result is "OK" then
        set minLength to (the text returned of the result) as number
        IndicateProgress("'Long' exons are at least " & minLength & " bytes.")
        set gottaRun to true
        try
            open for access bedFile
            open for access valFile
            set valFilePos to 0
            open for access outFile with write permission
            set chromosomeWas to ""
            repeat while gottaRun
                set inBuff to read bedFile until lineEnd
                if the length of inBuff > 5 then -- Blank lines are shorter, data lines are longer
                    set chrStr to the first word of inBuff
                    if chrStr = "#" then
                        write inBuff to outFile -- Capture all comments and leave intact
                    else
                        if chrStr ≠ chromosomeWas then
                            FindWordInFile(chrStr, valFile)
                            write ("# Chromosome " & chrStr & lineEnd) to outFile
                            set chromosomeWas to chrStr
                        end if
                        set startStr to the second word of inBuff
                        set endStr to the third word of inBuff
                        if endStr ≥ startStr + minLength then
                            set allDone to allDone + 1
                            --if allDone > 50 then set gottaRun to false -- Can be useful for debugging
purposes.
                            set exprVal to expressionValue(valFile, startStr, (endStr - 1) as text)
                        else
                            set exprVal to 0
                        end if
                        if exprVal = 0 then
                            write (startStr & tabChar & endStr & tabChar & lineEnd) to outFile -- leave
exprVal blank
                            IndicateProgress("Chromosome " & chrStr & ", records " & startStr & "
through " & endStr & ".")
                        else
                            write (startStr & tabChar & endStr & tabChar & exprVal & lineEnd) to
outFile
                        end if
                    end if
                end if
            end repeat
        on error parseErr number parseErrNum
            if parseErrNum ≠ -39 then
                IndicateProgress("Error " & parseErrNum & " when parsing file. " & parseErr)
                display dialog "Error " & parseErrNum & " when parsing file. " & parseErr
            else
                IndicateProgress("End of file.")
            end if
        end try
        close access outFile
        close access valFile
        close access bedFile
    end if
end doParsing

on parseFile(bedFile, valFile)
    set successIsMine to true -- I totally rule
    set inBuff to "test"
    try
        open for access bedFile
        set inBuff to read bedFile for 32000 -- bytes (should work just fine in any event!)
    on error inFileErr number inFileErrNum
        -- userCanceledErr = -128, eofErr = -39
        if inFileErrNum ≠ -39 then
            set successIsMine to false
            display dialog "Error " & inFileErrNum & " when preflighting file. " & inFileErr
        end if
    end try
    close access bedFile -- Close and then ...
    if successIsMine then
        set ii to 0
        set jj to 13 -- Probably the case, otherwise must be a 10
        repeat while ii < 32000
            set ii to ii + 1
            set theCode to the ASCII number of character ii of inBuff
            if (theCode < 15) and (theCode ≠ 9) then
                set jj to theCode
                set tempVar to the ASCII number of character (ii + 1) of inBuff
                if (tempVar < 15) and (tempVar ≠ 9) and (tempVar ≠ theCode) then set jj to tempVar
                set ii to 32000
            end if
        end repeat
        --display dialog "The ascii code of the delimiter is " & theCode
        doParsing(bedFile, valFile, jj)
    end if
end parseFile

on run
    set bedFile to choose file with prompt "What 'BED' file do you want?"
    set valFile to choose file with prompt "What signal value file do you want?"
    parseFile(bedFile, valFile)
    IndicateProgress("Parsing complete!")
    beep 3
    say "Stick a fork in me. I'm done!"
end run

on open (inFiles)
    if the (count of the items in inFiles) ≠ 2 then
        -- If the user dropped a bunch of files on the Parser icon, it was probably an accident or the user was just plain confused.
        display dialog "This program only parses a pair of files at a time."
    else
        tell application "Finder"
            copy the kind of item 1 of inFiles to foo1
            copy the file type of item 1 of inFiles to bar1
            copy the kind of item 2 of inFiles to foo2
            copy the file type of item 2 of inFiles to bar2
        end tell
        if foo1 is "Plain text document" or bar1 is "TEXT" then
            set valFile to item 1 of inFiles
            set bedFile to item 2 of inFiles
        else
            set valFile to item 2 of inFiles
            set bedFile to item 1 of inFiles
        end if
        parseFile(bedFile, valFile)
        beep 3
        say "Stick a fork in me. I'm done!"
    end if
end open
```

REFERENCES

1. Maupas, E. Modes et formes de reproduction des nématodes. Archives de Zoologie Experimentale et Generale 8, 463–624 (1900).
2. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282:5396, 2012-8 (1998).
3. Lai, C.-H., Chou, C.-Y., Ch'ang, L.-Y., Liu, C.-S. & Lin, W.-c. Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. Genome Research 10:5, 703 - 13 (2000).
4. Steinmetz, L.M. & Deutschbauer, A.M. Gene function on a genomic scale. Journal of Chromatography B- Analytical Technologies in the Biomedical and Life Sciences 782:1-2, 151-63 (2002).
5. Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G. & Brown, E.L. Genomic analysis of gene expression in *C. elegans*. Science 290, 809 - 12 (2000).
6. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. Science 302:5643, 249-55 (2003).
7. Walhout, A.J.M. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. Current Biology 12:22, 1952-8 (2002).
8. Reinke, V. Functional exploration of the *C. elegans* genome using DNA microarrays. Nature Genetics 32, 541-6 (2002).
9. Fitch, D.H.A. Evolution: An ecological context for *C. elegans*. Curr Biol 15:17, R655-8 (2005).
10. Feder, M.E. & Mitchell-Olds, T. Evolutionary and ecological functional genomics. Nature Reviews Genetics 4:8, 651-7 (2003).
11. Haber, M. Evolutionary history of *Caenorhabditis elegans* inferred from microsatellites: Evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding. Molecular Biology and Evolution 22, 160-73 (2005).
12. Sivasundar, A. & Hey, J. Sampling from natural populations with rnai reveals high outcrossing and population structure in *Caenorhabditis elegans*. Current Biology 15:17, 1598 - 602 (2005).
13. Donkin, S.G. & Dusenbery, D.B. A soil toxicity test using the nematode *Caenorhabditis elegans* and an effective method of recovery. Archives of Environmental Contamination and Toxicology 25, 145-51 (1993).
14. Boyd, W.A. & Williams, P.L. Availability of metals to the nematode *Caenorhabditis elegans*: Toxicity based on total concentrations in soil and extracted fractions. Environ. Toxicol Chem. 22:5, 1100-6 (2003).
15. Boyd, W.A. & Williams, P.L. Comparison of the sensitivity of three nematode species to copper and their utility in aquatic and soil toxicity tests. Environ. Toxicol Chem. 22:11, 2768-74 (2003).

16. Menzel, R., Stürzenbaum, S., Bärenwaldt, A., Kulas, J. & Steinberg, C.E.W. Humic material induces behavioral and global transcriptional responses in the nematode *Caenorhabditis elegans*. Environ Sci Technol 39:21, 8324-32 (2005).
17. Anderson, R.V. & Coleman, D.C. The use of glass microbeads in ecological experiments with bacteriophagic nematodes. Journal of Nematology 9:4, 319-22 (1977).
18. Prahlad, V., Pilgrim, D. & Goodwin, E.B. Roles for mating and environment in *C. elegans* sex determination. Science 302:7, 1046 - 9 (2003).
19. Jones, S.J.M., Riddle, D.L., Pouzyrev, A.T., Velculescu, V.E., Hillier, L., Eddy, S.R., Stricklin, S.L., Baillie, D.L., Waterston, R. & Marra, M.A. Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. Genome Research 11:8, 1346 - 52 (2001).
20. Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. & Davidson, G.S. A gene expression map for *Caenorhabditis elegans*. Science 293:5537, 2087-92 (2001).
21. Mallo, G.V., Kurz, C.L., Couillault, C., Pujol, N., Granjeaud, S., Kohara, Y. & Ewbank, J.J. Inducible antibacterial defense system in *C. elegans*. Curr Biol 12:14, 1209-14 (2002).
22. Bonner, K.M. University of New Hampshire (2007).
23. Quackenbush, J. Microarrays– guilt by association. in Science Vol. 302 240-1 (2003).
24. Sherlock, G. *et al.* The Stanford Microarray Database. Nucleic Acids Research 29:1, 152-5 (2001).
25. Denver, D., Morris, K., Streelman, J., Kim, S., Lynch, M. & Thomas, W. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nature Genetics 37:5, 544-8 (2005).
26. Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D. & Madore, S.J. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Research 28:22, 4552-7 (2000).
27. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:5725, 1149-54 (2005).
28. He, H. *et al.* Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarry. Genome Research 17, 1471-7 (2007).
29. Royce, T.E., Carriero, N.J. & Gerstein, M.B. An efficient pseudomedian filter for tiling microarrays. BMC Bioinformatics 8:186(2007).
30. Khan, F.R. & McFadden, B.A. A rapid method of synchronizing developmental stages of *Caenorhabditis elegans*. (eds. Brill, E.J. & Leiden) 280-3 (1980).
31. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. Science 278:5338, 631 - 7 (1997).
32. Tatusov, R. *et al.* The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4:41(2003).
33. Chen, N. *et al.* WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. Nucleic Acids Research 33:Database Issue, D383 - D9 (2005).

34. David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C., Bofkin, L., Jones, T., Davis, R. & Steinmetz, L. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci 103:14, 5320-5 (2006).

35. Manak, J., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A. & Gingeras, T. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. Nature Genetics 38:10, 1151-8 (2006).

36. Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J. & Deng, X. Genome-wide transcription analyses in rice using tiling microarrays. Nature Genetics 38:1, 124-9 (2006).

37. Garsin, D.A., Sifri, C.D., Mylonakis, E., Qin, X., Singh, K.V., Murray, B.E., Calderwood, S.B. & Ausubel, F.M. A simple model host for identifying gram-positive virulence factors. Proc Natl Acad Sci U S A 98:19, 10892-7 (2001).

38. Dusenbery, D.B., Anderson, G.L. & Anderson, E.A. Thermal acclimation more extensive for behavioral parameters than for oxygen consumption in the nematode *C. elegans*. Journal of Experimental Zoology 206, 191-8 (1978).

39. Donkin, S.G. & Dusenbery, D.B. Using the *Caenorhabditis elegans* soil toxicity test to identify factors affecting toxicity of four metal ions in intact soil. Water Air and Soil Pollution 78, 359-73 (1994).

40. Ahrén, D., Tholander, M., Fekete, C., Rajashekar, B., Friman, E., Johansson, T. & Tunlid, A. Comparison of gene expression in trap cells and vegetative hyphae of the nematophagous fungus *Monacrosporium haptotylum*. Microbiology 151:Pt 3, 789 - 803 (2005).

41. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science 302:5646, 842-6 (2003).

42. Kohara, Y. [large scale analysis of *C. elegans* cDNA]. Tanpakushitsu Kakusan Koso 41:5, 715-20 (1996).

43. Walhout, A.J.M., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S. & Vidal, M. Gateway recombinational cloning: Application to the cloning of large numbers of open reading frames or orfeomes. Methods in Enzymology 328, 575-92 (2000).

44. Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nature Genetics 27:3, 332-6 (2001).

45. Lu, J., Lal, A., Merriman, B., Nelson, S. & Riggins, G. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. Genomics 84:4, 631-6 (2004).

46. McKay, S.J. *et al.* Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. Cold Spring Harb Symp Quant Biol 68, 159-69 (2003).

47. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:5830, 1484-8 (2007).

48. Thorén, P., Persson, D., Karlsson, M. & Nordén, B. The antennapedia peptide penetratin translocates across lipid bilayers - the first direct observation. FEBS Letters 482:3, 265-8 (2000).