Gene Expression Profiling of Cells, Tissues, and Developmental Stages of the Nematode *C. elegans*

S.J. McKay,* R. Johnsen,† J. Khattra,* J. Asano,* D.L. Baillie,† S. Chan,* N. Dube,¶ L. Fang,† B. Goszczynski,‡ E. Ha,† E. Halfnight,¶ R. Hollebakken,† P. Huang,* K. Hung,† V. Jensen,† S.J.M. Jones,* H. Kai,¶ D. Li,† A. Mah,† M. Marra,* J. McGhee,‡ R. Newbury,¶ A. Pouzyrev,§ D.L. Riddle,§ E. Sonnhammer,** H. Tian, ‡ D. Tu,† J.R. Tyson,† G. Vatcher,* A. Warner,¶ K. Wong,* Z. Zhao,† and D.G. Moerman¶

**Genome Sciences Centre, BC Cancer Agency, Vancouver, B.C., Canada, V6T 1Z4; †Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, B.C., Canada, V5A 1S6; ‡Department of Biochemistry and Molecular Biology, University of Calgary, Calgary, Alberta, Canada T2N 4N1; ¶Department of Zoology, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4; §Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211-7400; **Karolinska Institute, Stockholm, Sweden*

Completion of the DNA sequences of the human genome and that of the nematode *Caenorhabditis elegans* allows the large-scale identification and analysis of orthologs of human genes in an organism amenable to detailed genetic and molecular analyses. We are determining gene expression profiles in specific cells, tissues, and developmental stages in *C. elegans*. Our ultimate goal is not only to describe detailed gene expression profiles, but also to gain a greater understanding of the organization of gene regulatory networks and to determine how they control cell function during development and differentiation.

The use of *C. elegans* as a platform to investigate the details of gene regulatory networks has several major advantages. Two key advantages are that it is the simplest multicellular organism for which there is a complete sequence (*C. elegans* Sequencing Consortium 1998), and it is the only multicellular organism for which there is a completely documented cell lineage (Sulston and Horvitz 1977; Sulston et al. 1983). *C. elegans* is amenable to both forward and reverse genetics (for review, see Riddle et al. 1997). A 2-week life span and generation time of just 3 days for *C. elegans* allows experimental procedures to be much shorter, more flexible, and more cost-effective compared to the use of mouse or zebrafish models for genomic analyses. Finally, the small size, transparency, and limited cell number of the worm make it possible to observe many complex cellular and developmental processes that cannot easily be observed in more complex organisms. Morphogenesis of organs and tissues can be observed at the level of a single cell (White et al. 1986). As events have shown, investigating the details of *C. elegans* biology can lead to fundamental observations about human health and biology (Sulston 1976; Hedgecock et al. 1983; Ellis and Horvitz 1986).

We are using complementary approaches to examine gene expression in *C. elegans*. We are constructing transgenic animals containing promoter green fluorescent protein (GFP) fusions of nematode orthologs of human genes. These transgenic animals are examined to determine the time and tissue expression pattern of the promoter::GFP constructs. Concurrently, we are undertaking serial analysis of gene expression (SAGE) on all developmental stages of intact animals and on selected purified cells. Tissues and selected cells are isolated using a fluorescence activated cell sorter (FACS) to sort promoter::GFP marked cell populations. To date we have purified to near homogeneity cell populations for embryonic muscle, gut, and a subset of neurons. The SAGE and promoter::GFP expression data are publicly available at http://elegans.bcgsc.bc.ca.

PROMOTER::GFP FUSIONS AS INDICATORS OF SPECIFIC TISSUE AND TEMPORAL GENE EXPRESSION

Our ultimate goal is to examine the in vivo spatial and temporal expression profiles of as many genes in the *C. elegans* genome as possible. Presently, the most effective methods for determining expression patterns in the worm are either antibodies or reporter fusion constructs. We have opted to use the more cost-effective promoter::GFP fusion technique. GFP reporter constructs are exquisitely sensitive and can detect expression at the resolution of a single cell (Chalfie et al. 1994; Chalfie 1995). The *C. elegans* community is fortunate to have an excellent GFP insertion vector kit available (developed by Dr. Andrew Fire, Carnegie Institution, http://www.ciwemb.edu/pages/firelab.html). We have been preceded in our approach by others, in particular, the laboratory of Ian Hope (Hope 1991; Lynch et al. 1995), where 350 expressing reporter gene fusions have been constructed (http://bgypc086.leeds.ac.uk/). Although the use of GFP fusions as expression reporters is not novel, the scale of our project is unprecedented.

To make a viable high-throughput approach for GFP fusion constructs, we needed a method that was both fast and efficient. Over the past year, we have demonstrated that fusion-PCR, also known as "stitching," enables construction of GFP fusions on a genome-wide scale. This PCR stitching technique has been used successfully by at least two groups (Cassata, Kagoshima et al. 1998; Hobert 2002) and we demonstrate here that it is scalable.

Choosing Candidate *C. elegans* Genes for Promoter GFP Analysis

Our study focuses on nematode homologs of human genes. A comparison of the two predicted proteomes with INPARANOID (Remm et al. 2001) identified 4367 *C. elegans* proteins with probable human orthologs (http://inparanoid.cgb.ki.se). This list of genes provides an excellent opportunity to use the worm to infer biological information for genes potentially relevant to human biology and health care. Of particular interest are predicted worm/human homologs for which there are no data concerning function; more than half of the worm orthologs have no functional annotation associated with them. These are particularly important gene targets, as they may form a new set of "Rosetta stone" proteins.

Most of the genome annotations used in the selection of our list of target genes were obtained from WormBase (www.wormbase.org; Stein et al. 2001; Harris et al. 2003). The list was filtered to remove rRNA genes and genes with SL2 trans-splice acceptor sites, which are associated with operons (Blumenthal 1995; Blumenthal et al. 2002). Also removed were genes with characterized mRNAs, an indication that the gene was already well studied. Preference was given to genes with EST-confirmed 5′ ends and those identified as embryonically expressed in Intronerator (Kent and Zahler 2000). We did not remove genes for which other researchers have constructed reporter fusions, because such genes act as a control set for our work. Indeed, thus far, at least four examples of expression patterns we have observed with our promoter::GFP constructs are identical to those observed by other investigators using either antibodies or functional GFP fusions.

The PCR stitching technique uses a two-step approach. First, the promoterless GFP gene and the putative *C. elegans* promoter region are PCR-amplified separately. In a second-round PCR, a complementary region engineered into the 3′ primer of the promoter amplicon and the 5′ primer of the GFP amplicon allows them to prime each other to form a chimeric amplicon containing a complete expression cassette. The PCR experiments were designed to capture putative promoter regions by amplifying about 3 kb of genomic DNA sequence immediately upstream of the predicted ATG initiator site. When an upstream gene was within 3 kb, the size of the amplicon was adjusted downward. Early PCR experiments were designed semimanually with the aid of primer3 (Rozen and Skaletsky 2000). To facilitate scale-up, we took advantage of the excellent *C. elegans* genome informatics resources to automate the PCR experimental design process. We used Perl and AcePerl (Stein and Thierry-Mieg 1998) to extract *C. elegans* genomic DNA sequence and annotations from wormbase to tie them together with the primer design and validation programs primer3 and e-PCR (Schuler 1997). To provide flexible, real-time design of PCR-based GFP fusion experiments, an interactive Web version of the program is available (http://elegans.bcgsc.bc.ca; S. McKay et al., in prep,).

Constructing Transgenic Animals with Heritable Promoter::GFP Constructs

Transgenic worms were generated by a modification of the method described by Mello et al. (1991). Promoter::GFP constructs and *dpy-5*(+) plasmid (pCeh-361) (kindly provided by C. Thacker and A. Rose) were used to construct transgenic strains. Transformants were identified by rescue of the Dpy-5 mutant phenotype. In *C. elegans*, transgene constructs usually form large extrachromosomal arrays. Due to the holokinetic nature of *C. elegans* chromosomes, these arrays can be partitioned during mitosis as though they were small chromosomes. However, extrachromosomal arrays must be large to be heritable (Stinchcomb et al. 1985; Clark et al. 1990; Mello and Fire 1995; Mello et al. 1991). Heritability of the GFP transgene construct is of considerable importance here, as somatic mosaicism or loss of the construct during gametogenesis could confound inferred gene expression patterns.

To determine whether our GFP transgenes form sufficiently large concatemeric arrays in vivo, we used quantitative PCR to estimate the copy number of the promoter::GFP constructs and plasmids in 20 different transgenic strains. We estimate that there are about 5–10 copies of promoter::GFP and 100–600 copies of the *dpy-5* plasmid in the heritable arrays. Although linear GFP DNA appears to be incorporated into arrays an order of magnitude less efficiently than circular plasmids, the sensitivity of the GFP assay does not require high copy numbers. To date, we have generated transgenic lines representing more than 1000 genes.

The ultimate in stable inheritance is ensured by chromosomal integration of the transgene, a process that can be induced by creating double-stranded breaks in chromosomes with ionizing radiation. Although the necessary handling and strain cleanup steps make this process less amenable to scale-up, we are constructing chromosomal integrant strains for a subset of the GFP constructs using low-dose X-ray irradiation (1500R). To date, such strains have been constructed for 80 genes. All of the strains generated from this study will be made available through the *Caenorhabditis* Genetics Center (biosci.umn.edu/CGC/CGChomepage.htm).

Expression Analysis of Promoter::GFP Constructs

As transformants carrying GFP fusions become available, they are subjected to a detailed in vivo analysis. As a first pass, we determine the developmental stage, tissues, and where possible, the individual cells where GFP expression is observed (Table 1). To date, we have observed GFP expression for 450 (56%) of 802 different transgenic lines. Possible reasons why no GFP expression was observed in the remaining lines include (1) germ-line silencing (Kelly et al. 1997; for review, see Seydoux and Schedl 2001), (2) absence of promoter::GFP in the heritable arrays, (3) conditional gene expression, or (4) failure to capture the entire transcription control element. The PCR experiments were designed to amplify as much

Table 1. Temporal and Tissue-specific Expression of Promoter::GFP Fusions

Tissue	Larval exclusive	Adult exclusive	Both larval and adult stages
Pharynx	2	11	59
Intestinal	23	3	66
Vulval	0	33	1
Spermatheca	0	6	1
Body wall muscle	3	6	40
Hypodermis	3	1	17
Seam cells	0	0	2
Anal sphincter and depressor muscle	0	9	12
Excretory cell	0	3	7
Nerve ring	5	0	36
Ventral nerve cord	6	1	23
Dorsal nerve cord	0	1	3
Head neurons	6	2	45
Tail neurons	5	5	48
Body neurons	2	2	9

of the potential promoter region as possible (up to 3 kb). Although it is rare in *C. elegans*, there are cases where important transcription control elements lie outside this 3-kb range and therefore preclude expression of the GFP construct.

Preliminary classification of GFP expression is done using a low-power GFP dissecting microscope. More detailed follow-up is done using a standard or confocal microscope equipped with epifluorescence and Nomarski optics. Ultimately, detailed expression patterns and gene activation in embryos are captured with live, two-channel four-dimensional microscopy. The fourth dimension is time; Z-stacks of developing embryos are recorded using Nomarski microscopy every 30–45 seconds. Interspersed with the normal Z-stacks we record GFP fluorescence in specific cells, which are then mapped and identified relative to the Nomarksi images. Software that supports this recording and analysis has been developed (Schnabel et al. 1997; also see Fire 1994; Thomas and White 1998; Bürglin 2000), and we are using programs derived from the study by Schnabel et al. (1997).

A survey of temporal GFP expression patterns is shown in Table 1, and some illustrative examples are displayed in Figure 1. We have detected GFP at all developmental stages and have identified expressed GFP in all major tissues except the germinal gonad. We did not expect to observe germ-line expression with any of our extrachromosomal array constructs, because germ-line silencing affects genes in extrachromosomal arrays (for review, see Seydoux and Schedl 2001). So far, we have not observed germ-line expression in any of the integrated lines derived from extrachromosomal arrays. A majority of the promoters we have examined thus far drive GFP expression in the intestine (92) and the nervous system (70), many exclusively in one of these tissues (Table 1). The large number of genes expressed in the intestine, the functional equivalent of the human stomach, intestine, and liver, agrees with our findings using SAGE on adult dissected intestine (see below and Table 2). Besides the intestine and nervous system, other major tissues including muscle and hypodermis are well represented in our data set. Subsets of cells and tissues within these broad categories are also delineated; we have observed GFP expression specific to the nerve ring, sensory neurons, ventral nerve cord, pharynx, seam cells, the excretory canal, the spermatheca, and anal sphincter muscles (Table 1). Our single biggest challenge in determining cell identity concerns the 302 cells that comprise the nematode nervous system (White et al. 1986). Neural expression patterns display a myriad of combinatorial possibilities, a fraction of which are represented in Table 1.

Table 2. SAGE Libraries

Stage	Tissue	Tags		Genes
		total	unique	
Embryo 14-bp tags	whole	133,825	25,885	8,187
Embryo 21-bp tags	whole	220,032	44,992	8,929
L1 larvae starved	whole	116,363	19,494	6,429
L1 larvae normal	whole	109,994	17,532	6,705
L2 larvae	whole	130,209	24,658	7,264
L3 larvae	whole	127,924	24,039	7,667
L4 larvae	whole	141,878	25,701	8,046
Young adult	whole	119,222	23,128	6,302
Adult (glp-4)	dissected gut	138,346	14,386	4,892
Adult (glp-4)	whole	117,529	19,140	6,974
Embryo (myo-3::GFP)	FACS sorted muscle	58,147	16,967	4,850
6-day adult (fer-15)	whole	110,306	19,861	6,758
1-day adult (fer-15;daf-2)	whole	101,939	16,960	5,159
6-day adult (fer-15;daf-2)	whole	100,737	14,004	4,687
10-day adult (fer-15;daf-2)	whole	116,336	19,183	5,594
Mixed stage[a]	whole	175,995	37,894	9,222
Dauer larvae[a]	whole	65,828	18,136	5,373
Meta library (14-bp tags)		1,806,431	130,112	14,661
Meta library (21-bp tags)		278,179	53,738	10,896

The total and unique tag numbers are for the raw tag collection prior to filtering. Libraries were filtered to remove tags with low sequence quality (below phred20) and tags originating from duplicate ditags (possible PCR artifacts). The number of genes refers to genes whose expression was detected by the presence of one or more tags mapped unambiguously to a single mRNA.

[a]From Jones et al. (2001). Only tags with known sequence quality were considered.

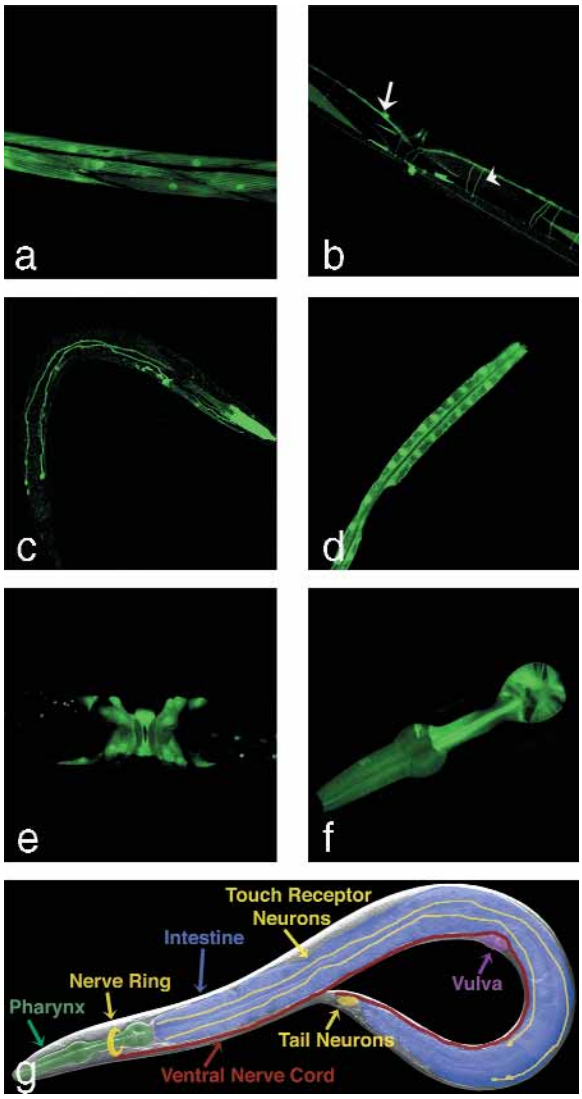

Figure 1. A gallery of promoter::GFP expression patterns for *C. elegans*. (*a*) Body wall muscle—gene B0228.4. (*b*) Ventral cord neurons and commissures—gene Y102A11A.2. Arrow indicates ventral cord and arrowhead points to a commissure. (*c*) Touch cells and pharynx—gene F32F2.1. (*d*) Intestine—gene Y102A11A.2. (*e*) vulval cells—gene Y47G6A.7. (*f*) Pharynx—gene C09G4.1. (*g*) Diagram illustrating location of some of the cells shown in panels *a–f*.

Because neuronal cells have been assigned to 118 classes (White et al. 1986), it is perhaps not surprising that there are many different neuronal gene expression patterns.

The GFP expression strains developed as part of this project are intended to become useful reagents for the biomedical research community. If enough promoters from different genes can be analyzed, we hope to be able to deduce logical rules regulating gene expression. Certainly this goal will be achievable if we combine this GFP expression set with the SAGE studies described below. A first step toward understanding the coordinate regulation of genes whose promoters drive similar expression patterns is to both computationally and biologically dissect the function of the promoter region. To facilitate this, a new comparative tool has emerged in the form of DNA sequence alignments between *C. elegans* and *C. briggsae* genomes (L. Stein et al. 2001). These two species are sufficiently diverged (80 to 100 million years) that noncoding sequences have diverged, but coding and other functional sequences remain conserved. This property can be exploited to refine existing gene models and create new ones, as well as to help identify potentially important *cis*-regulatory elements upstream of conserved genes. The latter could prove invaluable for dissecting the function of promoter regions of genes that we select for further study.

SERIAL ANALYSIS OF GENE EXPRESSION: TEMPORAL AND TISSUE-SPECIFIC EXPRESSION PROFILING

Several studies using either DNA microarray analysis or serial analysis of gene expression (SAGE) have been done to examine the expression of *C. elegans* genes in the whole organism (see, e.g., Hill et al. 2000; Reinke et al. 2000; Jones et al. 2001: Kim et al. 2001). SAGE is complementary to microarray analysis and, at present, is the most sensitive and specific method for obtaining qualitative and quantitative information on expressed RNAs (Velculescu et al. 1995). Using this approach, we can establish the portion of the genome that is transcribed and contributes to the protein profile at various time points during growth and development. Within the RNA profile, we can also identify many genes that do not encode proteins, but produce only RNA products. Finally, we can gain useful insight into alternatively spliced mRNA isoforms, their changes over time, and relative abundance. Ours is the first study to use SAGE to examine all the developmental stages of *C. elegans*. Thus far, we have constructed 17 libraries spanning all developmental stages from embryo to adult and also representing tissue, cell-type, and mutation-specific populations (Table 2). Taken together, these libraries include ~1.8 million observed tags.

Tag to Gene Mapping: Building the "Conceptual" Transcriptome

For a SAGE tag to be associated with a specific gene, it is first necessary to build a conceptual transcriptome representing the processed transcripts of all known genes and predicted gene models. Although tags corresponding to the mitochondrial and noncoding transcriptomes are also represented in our *C. elegans* SAGE libraries, most tags correspond to the predicted nuclear transcriptome. The process we used to build the conceptual transcriptome is illustrated in Figure 2. An examination of WormBase (www.wormbase.org; release WS110) , the public repository of information on the biology and genome of *C. elegans*, reveals that nearly 40% of the 22,156 *C. elegans* gene models have no EST evidence to confirm gene structure or expression. About 44% of WormBase genes have sufficient EST coverage to determine the extent of 3′UTRs (untranslated regions) in the processed transcripts.

Because the SAGE technique captures transcripts by their poly-A tail, and the tags are usually anchored at the

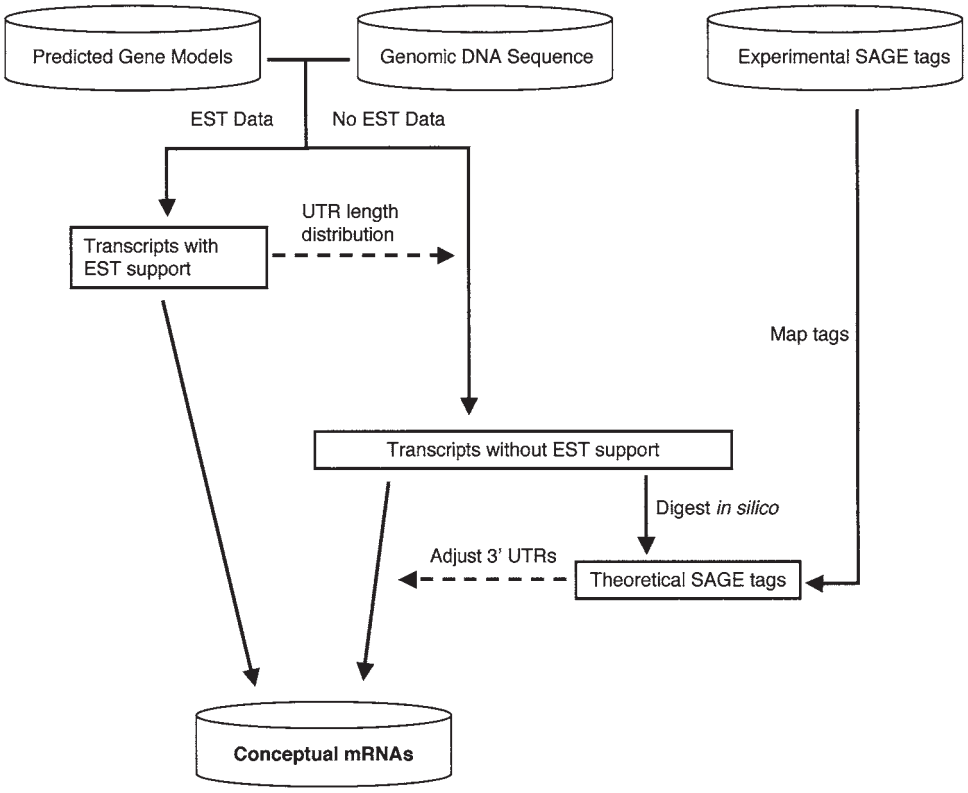

Figure 2. Building the conceptual transcriptome. Conceptual transcripts were assembled with known UTRs for genes with EST coverage and predicted UTRs for other genes based on the distribution of known UTR lengths. Introns were excised from the coding DNA and UTRs. Predicted 3′UTRs were adjusted according to potential polyadenylation signals, and both 3′ and 5′UTRs were truncated where required to avoid overlapping other genes. In some cases, overestimated 3′UTR lengths were detected by abundant experimentally observed SAGE tags occurring at the penultimate *Nla*III site (position 2). These predicted UTRs were truncated accordingly.

3′-most *Nla*III site, mRNAs with a cut site in their 3′UTR would be missed if coding sequences alone were used to map tags. For the 12,272 gene models lacking confirmed 3′UTRs, the untranslated regions of processed transcripts were predicted using a method modified from that of Pleasance et al. (2003). UTR lengths were estimated based on size distributions that cover 95% of known UTRs. About 5,550 of the predicted 3′UTRs include a *Nla*III site. Because the highest frequency SAGE tag for a transcript occurs at the first tag position, we used pooled SAGE data from more than a million SAGE tags to further refine the 3′UTR predictions for 1,449 gene models.

To determine how many transcripts we can identify, a meta-library of ~1.8 million tags was constructed by pooling all of the SAGE libraries (excluding longSAGE) in Table 2. A "specific" tag is defined as a tag that uniquely matches to a single gene or that can be resolved to a single gene by taking the lowest position match. To minimize the potential impact of sequencing errors, only tags with a cumulative phred score of 20 (Ewing and Green 1998) were considered. A score of Phred20 corresponds to a 99% probability that a base is called correctly. In this case, the score represents the average sequence quality of the entire tag sequence. A total of 26,682 specific tags corresponding to mRNAs for nuclear genes were observed. The total number of genes whose expression was detected by a SAGE tag for at least one transcript was 14,661. A distinct advantage of the SAGE technique is its ability to discriminate between alternative splice variants. Indeed, 7,073 (49%) of the detected genes are represented by two or more tags. A subset of just 1,126 (8%) of these genes have previously observed alternative splice variants documented in WormBase. Even among these previously well-studied genes, over 800 have multiple tags, potentially representing previously unobserved splice variants.

A Comparison of Short (14 bp) Versus Long (21 bp) SAGE Tags

Until very recently, all SAGE libraries were constructed using the tagging enzyme *Bsm*FI, which generates a 14-bp tag. Theoretically, a 14-bp tag is sufficient to unambiguously identify any gene in the *C. elegans* genome. In practice, not all tags map unambiguously to a single location. Two factors contribute to this ambiguity. First, there are multigene families stemming from ancestral sequence duplications; these related genes can share similar 3′ ends. Second, there appears to be some sequence compositional bias in 3′UTRs that tend to be AT-

rich. Based on a theoretical analysis of the *C. elegans* transcriptome, Pleasance et al. (2003) observed that, of all *C. elegans* genes that have an *Nla*III site in their conceptual transcript, about 12% would not be unambiguously identified by 14-bp tags. With an additional three nucleotides, they predicted that this number could be reduced to about 6% but, beyond 17 bp, there was no substantial reduction in ambiguity. Although 17-bp SAGE tags are not currently available, the need for longer tags has recently been addressed by the new longSAGE technique, which uses the enzyme *Mme*I to generate 21-bp tags (Saha et al. 2002). Now that there is a means for generating 21-bp tags, why would one not always use it? It comes down to cost, the bulk of which is in sequencing. It is possible to obtain greater sample depth of sequencing with 14-bp tags than for the same amount of sequencing with 21-bp tags. There is a trade-off between sampling deep enough to detect low-abundance transcripts and sequencing longer tags to reduce ambiguity.

To empirically determine the benefits of longer tags, we examined the same embryonic mRNA sample with both normal SAGE and longSAGE (Table 2). The two approaches identified 6,118 common genes (Fig. 3A) but also identified a substantial number of specific tags unique to one library. Since the majority of the nonoverlapping transcripts were of low abundance (Fig. 3B), it appears that such transcripts were detected stochastically at the sampling depth used. LongSAGE emerged as the method of choice, however, as it met the theoretical expectation of a twofold reduction in the ambiguities observed in assigning 14-bp tags to genes. Using longSAGE, it was possible to infer 2,896 specific genomic sites for tags shared by both libraries, but for which no unambiguous single sight could be assigned using a 14-bp tag. On the basis of the differences in tag ambiguity between the two protocols, we conclude that it is necessary to utilize longSAGE if resolving power is of utmost importance.

A Comparison of DNA Microarrays and SAGE

There are close to ten published studies using DNA microarray analysis to profile *C. elegans* gene expression (for review, see Reinke 2002; also see Stuart et al. 2003), but so far only two SAGE studies (Jones et al. 2001; Holt and Riddle 2003). It is therefore important to compare the two approaches, as they both have advantages and pitfalls. The Affymetrix GeneChip™ array for *C. elegans* was designed to represent 22,500 *C. elegans* transcripts or EST clusters. Sequence information for probe design came from the December 05, 2000 Sanger Center ACeDB database release and GenBank release 121, and was re-annotated by Affymetrix. We remapped the Affymetrix probe sets to our current conceptual transcriptome to allow direct comparison of transcript profiles with SAGE. Because of changing gene models and genomic DNA annotations, not all transcripts predicted in 2000 can be compared directly to the 2003 version. However, ~ 90% of the Affymetrix probe sets can poten-

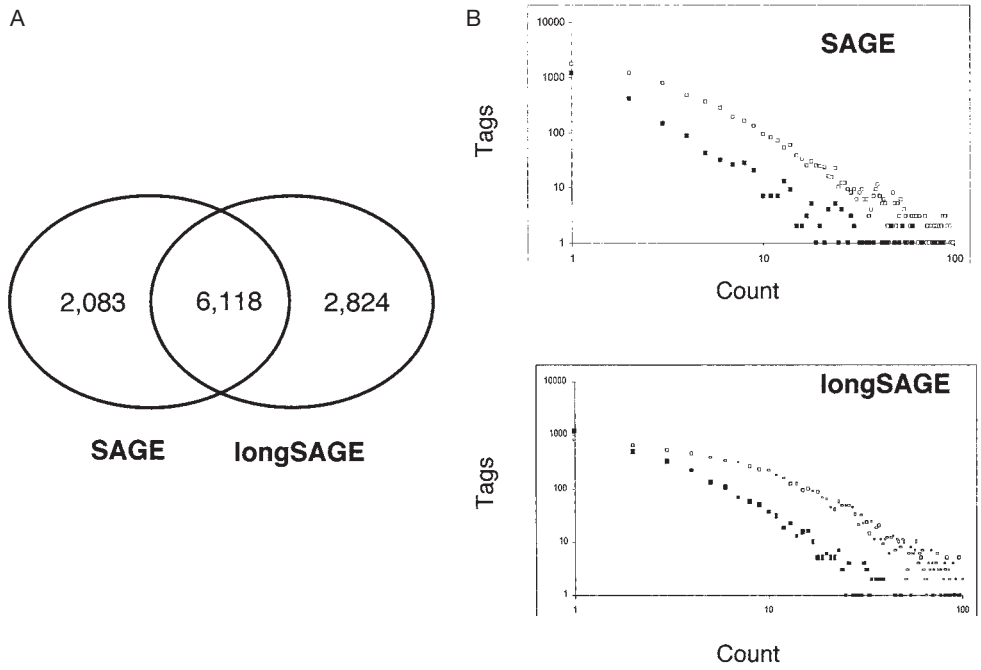

Figure 3. SAGE vs. longSAGE comparison. SAGE and longSAGE tags were filtered to remove duplicate ditags, linker tags, or tags with low-quality sequence. Only tags mapping unambiguously to the positive strand of a single transcript, or that could be resolved to a single sequence by taking the lowest position match (specific tags), were considered. Both libraries were constructed from the same mRNA sample, extracted from a synchronized embryonic population (Table 2). (*A*) A Venn diagram comparison of genes identified by SAGE and longSAGE. The region of overlap indicates genes for which specific tags were observed in both libraries. (*B*) Log X log plots of tag count distributions of shared tags (with the same 14-nt 5′ end) and unique tags. Due to the logarithmic scale, tag counts of 0 and 1 are not distinguishable. Note that the tags unique to the SAGE or longSAGE libraries (*filled squares*) are much less abundant.

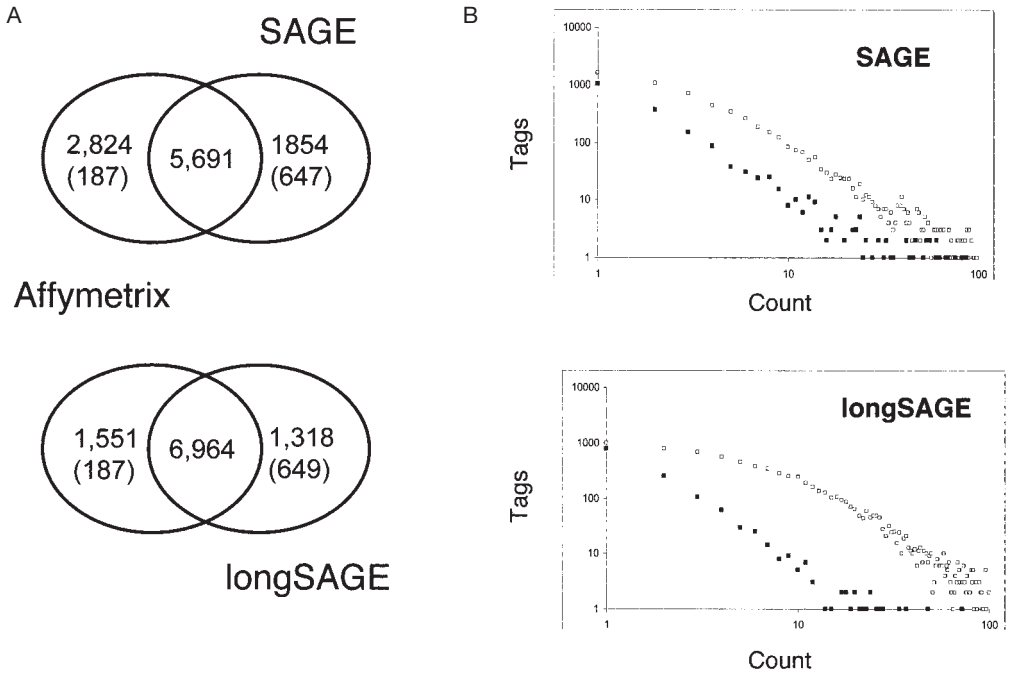

Figure 4. SAGE vs. AffyMetrix GeneChip™. Each 25mer probe in an Affymetrix probe set (obtained from www.affymetrix.com) was compared to all transcripts in the conceptual transcriptome. There are 20,291 probe sets that map specifically to 17,147 conceptual mRNAs. The remaining 2,257 probe sets mapped to multiple transcripts, or did not map to any. Only specific SAGE or longSAGE tags were considered. Mitochondrial transcripts, which are absent from the Affymetrix chip, were not considered. Oligonucleotide sequences from each probe set on the Affymetrix chip that was called as "present" in three replicate Affymetrix GeneChip experiments were mapped to transcripts in our virtual transcriptome. (*A*) Comparison of expressed genes detected by SAGE and longSAGE vs. the Affymetrix chip. For the SAGE methods, the number in parentheses represents transcripts not detectable by the chip. For Affymetrix, the number in parentheses represents transcripts without *Nla*III sites (undetected by SAGE). (*B*) Log x log plots of SAGE tag count distributions of shared and unique tags. Transcripts unique to the SAGE methods (*filled squares*) are of lower abundance. Only transcripts for which a direct SAGE/Affymetrix comparison was possible were considered.

tially be identified by SAGE, and 17,123 map unambiguously to single conceptual transcripts. To compare the chip and the two SAGE methods empirically, we hybridized the same mRNA samples we used for the whole-embryo SAGE and longSAGE libraries to Affymetrix chips. The pool of transcripts detected by both methods is limited: (1) Transcripts lacking an *Nla*III site or for which no specific tags were observed are excluded from SAGE and (2) probe sets that do not unambiguously detect a single transcript (or set of alternatively spliced transcripts) from our conceptual transcriptome are excluded from Affymetrix. Even with these caveats, a comparison of transcription profiles of embryonic libraries derived from SAGE and DNA chip analyses reveals a great deal of concordance (Fig. 4A), with more than half of all transcripts detected present in both data sets. As with the comparison between SAGE and longSAGE, most of the disagreement between methods is due to low-abundance transcripts (Figs. 3B and 4B). What is clear from the SAGE/Affymetrix comparison is that, because of the improved specificity in tag-to-gene mapping, longSAGE improves correspondence to the Affymetrix chip data (Fig. 4A). Given appropriate filtering to avoid sequencing errors and other artifacts, a rare, positive observation of a specific tag indicates that a transcript is likely present, albeit at low abundance. However, failing to observe a specific SAGE tag does not unequivocally demonstrate the absence of a transcript, as very rare transcripts would not be consistently detectable at normal sampling depths. The fact that rare transcripts can be observed at all is an important advantage of SAGE over microarrays because the signal from such low-abundance transcripts would be difficult to distinguish from background noise. An equally important advantage is that SAGE does not require a priori understanding of the transcriptome in order to detect transcripts. Among the sets of transcripts found only by SAGE are those transcription units or alternative splice variants that are not currently represented on the Affymetrix chip because they are novel. Finally, an important advantage that microarray analysis has over SAGE is that microarrays are less costly. The next generation of chips for transcription profiling stands to be greatly improved by the addition of novel transcripts identified by SAGE. Bearing in mind that not all genes are currently suitable for direct comparison between SAGE and Affymetrix analyses, the intersect between chip and longSAGE (Fig. 4) sets a conservative estimate of at least 7,000 as the minimal number of different transcripts expressed during embryogenesis, a full third of *C. elegans* predicted genes.

Embryonic Muscle and Adult Intestine: Examples of Tissue-specific Expression Profiling

For analysis of cellular function during development and growth, we are performing SAGE analysis on purified or enriched samples from specific cell populations and tissues. We use two different protocols depending on whether we are purifying tissue from embryos or from adult animals. To examine developing embryos, we used enzymatic digestion and mechanical shearing to free individual cells (Christensen et al. 2002). If cells of interest are labeled with a GFP marker, they can be isolated using a fluorescence activated cell sorter (FACS). Depending on when the GFP tag is expressed, labeled cells can be isolated directly from a fragmented embryo, or the cells can be plated and allowed to differentiate further prior to sorting. There are GFP tags available for every major tissue during development (hypodermis, nervous system, intestine, and muscle), and for subpopulations of those tissues. The isolation of developing gut cells after sorting is shown in Figure 5. We now have the means to isolate and purify analyzable quantities of specific cell populations from *C. elegans* embryos.

Our first embryonic tissue-specific SAGE library was for embryonic muscle using *myo-3*::GFP as a tissue-specific marker (Okkema et al. 1993). The *myo-3* myosin heavy-chain gene is expressed during late embryogenesis in nascent body wall muscle cells. It is first detected as the cells migrate from a lateral position to muscle quadrants located on the dorsal and ventral sides of the embryo (Epstein et al. 1993). We were able to fragment embryos and obtain individual muscle cells via FACS in sufficient quantities to extract mRNA and construct a longSAGE library. This library allowed us to detect 4,850 different genes (Table 2). Among this set of transcripts were many of the genes one expects to find, including body wall myosins, actins, and several components associated with sarcomere assembly. Although it is not surprising to detect mRNA for major structural proteins because they are expected to be relatively abundant, the sensitivity of tissue enrichment and SAGE is demonstrated by detection of the relatively rare mRNA for *hlh-1*, the nematode homolog of myoD (Krause et al. 1990). We currently have tissue-specific embryonic libraries under construction for all the major germ layers including embryonic gut, the developing nervous system, and the hypodermis.

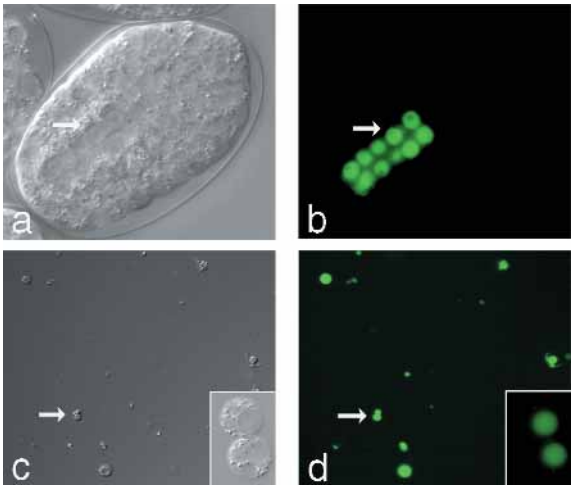

Figure 5. FACS of promoter::GFP marked embryonic intestine cells. (*a*) Late-stage living embryo viewed using Nomarski optics. Arrow points to double row of intestinal cells. This is a dorsal view, and anterior is to the top right corner. (*b*) Same embryo viewed using fluorescence microscopy. The promoter for the Elt-2 transcription factor is fused to GFP and acts as a marker for developing intestinal cells. (*c,d*) Disaggregated embryo cells enriched for GFP expression after FAC sorting. *c* shows cells viewed using Nomarski optics and *d* shows the same field of cells viewed using fluorescence optics. Arrow points to identical cells in both views.

Although isolation of most tissues or organs from adult worms has not been possible, hand dissection of a few adult intestines has been used to study vitellogenin synthesis (Kimble and Sharrock 1983). We used the temperature-sensitive *glp-4* mutant, *bn2*, which lacks a gonad when raised at 25ºC (Beanan and Strome 1992), thereby removing one of the major internal organs of the worm and making gut dissection much easier. We constructed a SAGE library from 1,863 dissected adult intestines. As a control, a library was also made from whole *glp-4(bn2)* worms grown under identical conditions. Both show the expected distribution of transcripts, with a few transcripts present at very high levels (1,000–2,000 tags per library) and many transcripts present at 1 tag per library (Table 2) (S. McKay et al., unpubl.). The quality of the dissection is judged to be good based on the low-to-undetectable level of tags in the gut library corresponding to transcripts that are known to be expressed outside the gut (e.g., cuticular collagens, major sperm proteins, muscle proteins). A preliminary estimate of the number of different transcripts detected in the adult intestine is about 4,900 (Table 2).

There are two ways in which to view SAGE profiles from specific tissues. The first view is that it provides an enduring archive, an inventory of genes that are needed to make embryonic muscle or an adult worm gut. The second view derives from our interest in gene regulation, where there is significant value in knowing whether a particular gene is expressed only in a particular tissue. As an example of this approach, we have used the gut SAGE data. By comparing the number of tags to the gut-specific vitellogenin genes in the gut library and in the intact *glp-4(bn2)* library, we estimate that 1,000–2,000 of the genes expressed in the adult gut are gut-specific. Even at this early stage, several conclusions can be drawn. Perhaps not surprisingly, many of the genes expressed at the highest level only in the gut are digestive enzymes, in particular aspartic proteases. The *asp-1* gene encodes such a protease and has previously been demonstrated to be expressed strongly and specifically in the intestine (Tcherepanova et al. 2000); the current results certainly confirm this at >2,000 tags in the gut library. Other members of the same protease family are expressed at comparable or even higher levels. We have a special interest in gut transcription factors, and here the SAGE list is proving invaluable. As expected, transcripts for transcription factors are present at reasonably low levels (a few dozen

or fewer tags per library). Tags corresponding to the gut-specific GATA-type zinc-finger factor *elt-2* (Hawkins and McGhee 1995; Fukushige et al. 1998) are present at the highest level of any recognizable transcription factor in the gut library. The library provides an intriguing list of a dozen or more transcription factors that, by the level of transcript enrichment, are judged to be gut-specific, and yet nothing is yet known about them.

Perhaps the most intriguing finding in either tissue-specific library is the presence of experimentally unverified gene models derived from computational analysis for which there is no functional annotation. These genes promise much new territory to explore. As many of these predicted genes have human homologs, they are of particular relevance to the themes of this symposium.

Exploiting the SAGE Data Sets: Developmental Profiling and Gene Discovery

Throughout the life cycle of any organism there are dynamic changes in the expression profile of the genome. Tracking and displaying these changes in a way that leads to further understanding of individual gene function and overall gene regulation is one of the most significant challenges of 21st-century biology. We are exploring how best to mine the SAGE data for information pertaining to gene regulation and gene pathways and also to explore how best to present the data for exploitation by others.

The following are three examples of how one might track and display a large gene family through development (Fig. 6). In the first example, we examined the large zinc-finger gene family. WormBase identifies 785 potential zinc-finger-encoding transcripts. From our studies of all developmental stages, we identified 1,299 specific SAGE tags corresponding to 625 genes. Their expression profile is illustrated in Figure 6A. We have done a similar study of cuticle collagens (SAGE identifies 167 of 206 potential collagen genes annotated in WormBase, Fig. 6B) and kinases (SAGE identifies 652 of 734 potential kinase-encoding genes annotated in WormBase, Fig. 6B). In each of these large families, we were able to track at least two-thirds of the genes. Although this is already a significant achievement, we should be able to detect even more members of these large families if we use enriched tissues.

Tracking known genes is an important use of expression profiling data, but SAGE also enables gene discovery. In the embryonic SAGE and longSAGE libraries, a total of 1,070 14-bp tags and 2,730 21-bp tags map to unique locations in the genomic DNA but do not map to any known nuclear, mitochondrial, or rRNA transcript. The larger number of unambiguous long tags demonstrates the resolving power of longSAGE, which resolved the ambiguity of one-third of the ambiguous 14-bp tags in this class. The SAGE protocol involves DNase treatment of an RNA sample, so tags that map to genomic DNA but not to predicted transcripts can be used to infer novel transcribed sequences, or undocumented alternative splice variants, or UTRs of known genes. For example, Jones et al. (2001) identified two novel transcribed sequences possessing telomeric repeat-like sequences and no obvious open reading frame that are present at high abundance in dauer larvae but not in other life stages. Among the 14-bp "genomic" tags, 445 map to introns of known genes; this could be explained in large part by previously unknown exons or, more rarely, by small genes nested entirely within the intron. The remaining 625 genomic tags map to regions for which there are no annotated transcribed sequences, and likely represent novel transcription units or, if they are near known genes, alter-

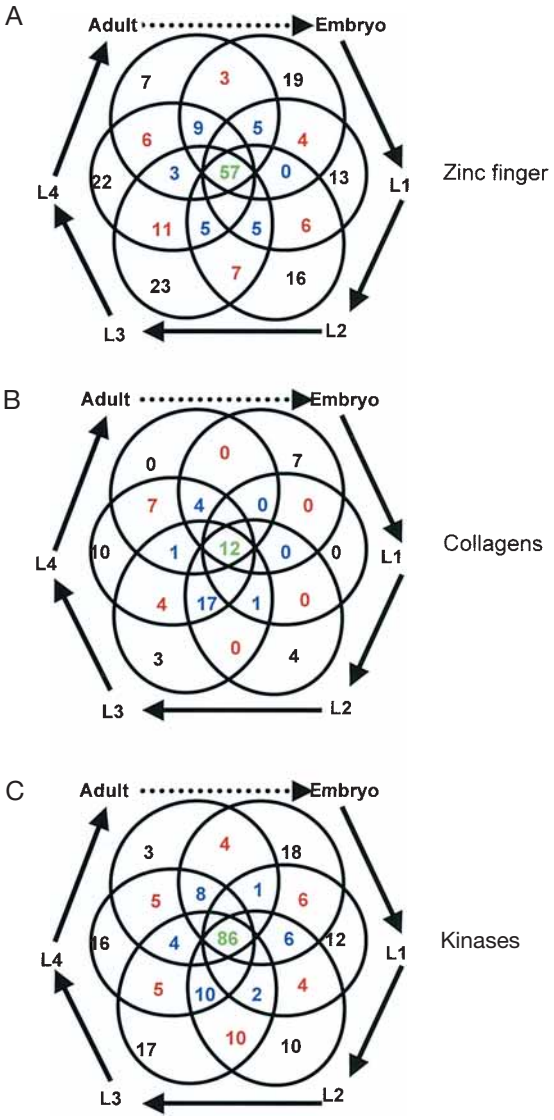

Figure 6. Venn diagrams showing transcription profiles for six stages of *C. elegans* development. Only specific SAGE tags were considered. Transcripts are counted by virtue of presence or absence rather than relative abundance. Putative alternative splice variants identified by different SAGE tags for the same gene are counted separately. The criterion for presence of a transcript is the observation of at least one tag of high-quality sequence (phred40). Regions of overlap indicate the number of transcripts common to all affected stages. Numbers in black, red, blue, and green correspond to the count of transcripts observed in one, two, three, or all developmental stages. (*A*) Zinc-finger genes. (*B*) Collagen genes. (*C*) Kinase genes.

native polyadenylation signals or UTRs not fully represented by ESTs. The improved resolution of longSAGE allowed us to identify 1,257 tags that map within introns of known genes and 1,473 that occur in intergenic regions. Many of the novel tags overlap with regions of sequence similarity in orthologous regions of the genome of the related nematode *C. briggsae,* suggesting conserved regions may have functional significance. Comparative genomics and directed RT-PCR experiments will be required to characterize the novel transcribed sequences whose presence we infer with genomic SAGE tags.

CONCLUSIONS

1. The *C. elegans* and human genomes are estimated to have at least 4,300 orthologous gene pairs.
2. The function of many of these genes is unknown in either organism, but the simplicity of nematode anatomy coupled with powerful genetic tools should contribute to the understanding of their function.
3. There are temporal and tissue-specific promoters, but few or no single-cell promoters. Individual cell identity within a tissue would then appear to result from combinatorial overlaps.
4. SAGE technology confirms the expression of at least 14,600 genes in the nematode. This number will increase as the technology is refined.
5. SAGE reveals multiple different tags for half of the genes in *C. elegans*, suggesting that alternative splicing of genes is common in this organism.
6. Many of the SAGE tags map to unannotated regions of the nematode genome and thus may identify new genes.
7. The studies outlined here using promoter::GFP constructs and SAGE will lead to the establishment of a gene expression database that can be interrogated to understand temporal and spatial patterning during development.
8. We have established the minimum number of genes required for nematode embryogenesis, 7,000, and the number of genes required to determine and maintain the function of a specific tissue, about 5,000.

ACKNOWLEDGMENTS

We thank David Miller III and his colleagues at Vanderbilt for providing us with their protocol for fragmenting embryos. Funding for this project was provided by Genome British Columbia and Genome Canada to D.L.B., S.J.M.J., M.M., and D.G.M. J.T. is the recipient of an International Prize Travelling Research Fellowship from the Wellcome Trust. S.J. and M.M. are Michael Smith Foundation Health Research Scholars. Additional funding was provided by National Institutes of Health operating grants AG-12689 and GM-60151 to D.L.R. and a Canadian Institute of Health Research grant to J.M.

REFERENCES

Beanan M.J. and Strome S. 1992. Characterization of a germ-line proliferation mutation in *C. elegans. Development* **116:** 755.

Blumenthal T. 1995. Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* **11:** 132.

Blumenthal T., Evans D., Link C.D., Guffanti A., Lawson D., Thierry-Mieg J., Thierry-Mieg D., Chiu W.L., Duke K., Kiraly M., and Kim S.K. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417:** 851.

Bürglin T.R. 2000. A two-channel four-dimensional image recording and viewing system with automatic drift correction. *J. Microsc.* **200:** 75.

Cassata G., Kagoshima H., Pretot R.F., Aspock G., Niklaus G., and Bürglin T.R. 1998. Rapid expression screening of *Caenorhabditis elegans* homeobox open reading frames using a two-step polymerase chain reaction promoter-gfp reporter construction technique. *Gene* **212:** 127.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. *Science* **282:** 2012.

Chalfie M. 1995. Green fluorescent protein. *Photochem. Photobiol.* **62:** 651.

Chalfie M., Tu Y., Euskirchen G., Ward W.W., and Prasher D.C. 1994. Green fluorescent protein as a marker for gene expression. *Science* **263:** 802.

Christensen M., Estevez A., Yin X., Fox R., Morrison R., McDonnell M., Gleason C., Miller D.M., III, and Strange K. 2002. A primary culture system for functional analysis of *C. elegans* neurons and muscle cells. *Neuron* **33:** 503.

Clark D.V., Johnsen R.C., McKim K.S., and Baillie D.L. 1990. Analysis of lethal mutations induced in a mutator strain that activates transposable elements in *Caehorhabditis elegans. Genome* **33:** 109.

Ellis H.M. and Horvitz H.R. 1986. Genetic control of programmed cell death in the nematode *C. elegans. Cell* **44:** 817.

Epstein H.F., Casey D.L., and Ortiz I. 1993. Myosin and paramyosin of *Caenorhabditis elegans* embryos assemble into nascent structures distinct from thick filaments and multifilament assemblages. *J. Cell Biol*. **122:** 845.

Ewing B. and Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186.

Fire A. 1994. A four dimensional digital archiving system for cell lineage tracing and retrospective embryology. *Comput. Appl. Biosci.* **10:** 443.

Fukushige T., Hawkins M.G., and McGhee J.D. 1998. The GATA-factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. *Dev. Biol*. **198:** 286.

Harris T.W., Lee R., Schwarz E., Bradnam K., Lawson D., Chen W., Blasier D., Kenny E., Cunningham F., Kishore R., Chan J., Muller H.M., Petcherski A., Thorisson G., Day A., Bieri T., Rogers A., Chen C.K., Spieth J., Sternberg P., Durbin R., and Stein L.D. 2003. WormBase: A cross-species database for comparative genomics. *Nucleic Acids Res.* **31:** 133.

Hawkins M.G. and McGhee J.D. 1995. elt-2, a second Gata factor from the nematode *Caenorhabditis elegans*. *J. Biol. Chem.* **270:** 14666.

Hedgecock E.M., Sulston J.E., and Thomson J.N. 1983. Mutations affecting programmed cell deaths in the nematode *Caenorhabditis elegans. Science* **220:** 1277.

Hill A.A., Hunter C.P., Tsung B.T., Tucker-Kellogg G., and Brown E.L. 2000. Genomic analysis of gene expression in *C. elegans. Science* **290:** 809.

Hobert O. 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32:** 728.

Holt S.J. and Riddle D.L. 2003. SAGE surveys *C. elegans* carbohydrate metabolism: Evidence for an anaerobic shift in the long-lived dauer larva. *Mech. Ageing Dev.* **124:** 779.

Hope I.A. 1991. "Promoter trapping" in *Caenorhabditis elegans*. *Development* **113:** 399.

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., and Marra M.A. 2001. Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. *Genome Res.* **11:** 1346.

Kelly W.G., Xu S., Montgomery M.K., and Fire A. 1997. Dis-

tinct requirements for somatic and germline expression of a generally expressed *Caenorhabditis elegans* gene. *Genetics* **146:** 227.

Kent W.J. and Zahler A.M. 2000. The intronerator: Exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res.* **28:** 91.

Kim S.K., Lund J., Kiraly M., Duke K., Jiang M., Stuart J.M., Eizinger A., Wylie B.N., and Davidson G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087.

Kimble J. and Sharrock W.J. 1983. Tissue-specific synthesis of yolk proteins in *Caenorhabditis elegans*. *Dev. Biol.* **96:** 189.

Krause M., Fire A., Harrison S.W., Priess J., and Weintraub H. 1990. CeMyoD accumulation defines the body wall muscle cell fate during *C. elegans* embryogenesis. *Cell* **63:** 907.

Lynch A.S., Briggs D., and Hope I.A. 1995. Developmental expression pattern screen for genes predicted in the *C. elegans* genome sequencing project. *Nat. Genet.* **11:** 309.

Mello C. and Fire A. 1995. DNA transformation. *Methods Cell Biol*. **48:** 451.

Mello C.C., Kramer J.M., Stincomb D., and Ambros V. 1991. Efficient gene transfer in *C. elegans:* Extrachromosomal maintenance and integration of transforming sequences. *EMBO J*. **10:** 3959.

Okkema P.G., Harrison S.W., Plunger V., Aryana A., and Fire A. 1993. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans. Genetics* **135:** 385.

Pleasance E.D., Marra M.A., and Jones S.J. 2003. Assessment of SAGE in transcript identification. *Genome Res.* **13:** 1203.

Reinke V. 2002. Functional exploration of the *C. elegans* genome using DNA microarrays. *Nat. Genet.* **32:** 541.

Reinke V., Smith H.E., Nance J., Wang J., Van Doren C., Begley R., Jones S.J., Davis E.B., Scherer S., Ward S., and Kim S.K. 2000. A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6:** 605.

Remm M., Storm C.E., and Sonnhammer E.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314:** 1041.

Riddle D.L., Blumenthal T., Meyer B.J., and Preiss J.R. 1997. *C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Rozen S. and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132:** 365.

Saha S., Sparks A.B., Rago C., Akmaev V., Wang C.J., Vogelstein B., Kinzler K.W., and Velculescu V.E. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20:** 508.

Schnabel R., Hutter H., Moerman D.G., and Schnabel H. 1997. Assessing normal embryogenesis in *Caenorhabditis elegans* using a 4D microscope: Variability of development and regional specification. *Dev. Biol.* **184:** 234.

Schuler G.D. 1997. Sequence mapping by electronic PCR. *Genome Res.* **7:** 541.

Seydoux G. and Schedl T. 2001. The germline in *C. elegans:* Origins, proliferation, and silencing. *Int. Rev. Cytol*. **203:** 139.

Stein L.D. and Thierry-Mieg J. 1998. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* **8:** 1308.

Stein L., Sternberg P., Durbin R., Thierry-Mieg J., and Spieth J. 2001. WormBase: Network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* **29:** 82.

Stein L.D., Bao Z., Blasiar D., Blumenthal T., Brent M.R., Chen N., Chinwalla A., Clarke L., Clee C., Coghlan A., Coulson A., D'Eustachio P., Fitch D.H., Fulton L.A., Fulton R.E., Griffiths-Jones S., Harris T.W., Hillier L.W., Kamath R., Kuwabara P.E., Mardis E.R., Marra M.A., Miner T.L., Minx P., and Mullikin J.C., et al. 2003. The genome sequence of *Caenorhabditis briggsae:* A platform for comparative genomics. *PLoS Biol.* **1:** E45.

Stinchcomb D.T., Shaw J.E., Carr S.H., and Hirsh D. 1985. Extrachromosomal DNA transformation of *Caenorhabditis elegans. Mol. Cell. Biol*. **5:** 3484.

Stuart J.M., Segal E., Koller D., and Kim S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302:** 249.

Sulston J.E. 1976. Post-embryonic development in the ventral cord of *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **275:** 287.

Sulston J.E. and Horvitz H.R. 1977. Postembryonic cell lineages of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **82:** 41.

Sulston J.E., Schierenberg E., White J.G., and Thomson J.N. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100:** 64.

Tcherepanova I., Bhattacharyya L., Rubin C.S., and Freedman J.H. 2000. Aspartic proteases from the nematode *Caenorhabditis elegans*. Structural organization and developmental and cell-specific expression of asp-1. *J. Biol. Chem.* **275:** 26359.

Thomas C.F. and White J.G. 1998. Four dimensional imaging: The exploration of space and time. *Trends Biotechnol.* **16:** 175.

Velculescu V.E., Zhang L., Vogelstein B., and Kinzler K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484.

White J.G., Southgate E., Thomson J.N., and Brenner S. 1986. The structure of the nervous system of the nematode *Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314:** 1.