A Spatial and Temporal Map of *C. elegans* Gene Expression

William Clayton Spencer, Georg Zeller, Joseph D. Watson, et al.

Genome Res. published online December 22, 2010
Access the most recent version at doi:10.1101/gr.114595.110

A Spatial and Temporal Map of *C. elegans* Gene Expression

W. Clay Spencer[1†], Georg Zeller[2,3,4†], Joseph D. Watson[1,8], Stefan R. Henz[3], Kathie L. Watkins[1], Rebecca D. McWhirter[1], Sarah Petersen[1], Vipin T. Sreedharan[2], Christian Widmer[2], Jeanyoung Jo[5], Valerie Reinke[5], Lisa Petrella[6], Susan Strome[6], Stephen E. Von Stetina[1,7], Menachem Katz[9], Shai Shaham[9], Gunnar Rätsch[2], David M. Miller III[1*]

[1]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN 37232

[2]Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany,

[3]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[4]Present address: European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[5]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520

[6]Department of MCD Biology, University of California Santa Cruz, Santa Cruz, CA 95064

[7]Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

[8]Present address: Department of Biochemistry & Biophysics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599

[9]Laboratory of Developmental Genetics, The Rockefeller University, New York, NY 10065

†These authors contributed equally to this work.

*Corresponding author:

David M. Miller, III

465 21st Ave S

3120 MRBIII

Nashville, TN 37232

david.miller@vanderbilt.edu

(615) 343-3447

Running Title: CELL-SPECIFIC GENE EXPRESSION IN *C. ELEGANS*

Keywords: *C. elegans*, genomics, gene expression, transcription, microarray

ABSTRACT

The *C. elegans* genome has been completely sequenced and the developmental anatomy of this model organism is described at single-cell resolution. Here we utilize strategies that exploit this precisely defined architecture to link gene expression to cell type. We obtained RNAs from specific cells and from each developmental stage using tissue-specific promoters to mark cells for isolation by FACS or for mRNA extraction by the mRNA-tagging method. We then generated gene expression profiles of >30 different cells and developmental stages using tiling arrays. Machine-learning-based analysis detected transcripts corresponding to established gene models and revealed novel transcriptionally active regions (TARs) in non-coding domains that comprise at least 10% of the total *C. elegans* genome. Our results show that ~75% of transcripts with detectable expression are differentially expressed among developmental stages and across cell types. Examination of known tissue- and cell-specific transcripts validates these data sets and suggests that newly identified TARs may exercise cell-specific functions. Additionally, we used self-organizing maps to define groups of co-regulated transcripts and applied regulatory element analysis to identify known transcription factor- and miRNA-binding sites, as well as novel motifs that likely function to control subsets of these genes. By using cell-specific, whole-genome profiling strategies, we have detected a large number of novel transcripts and produced high-resolution gene expression maps that provide a basis for establishing the roles of individual genes in cellular differentiation.

Data supporting the manuscript:

Accession numbers are provided for microarray data sets. Supplementary material will be provided with the manuscript and hosted on an FTP server.

INTRODUCTION

The generation of specific cell types depends on spatial and temporal control of gene expression. The nematode *C. elegans* has been widely utilized to address this question because of its simple body plan and fully sequenced genome (Hillier et al. 2005). Although comprised of fewer than 1,000 somatic cells, the tissues of *C. elegans* adults include cell types characteristic of all metazoans such as muscle, nerve, intestine, skin, *etc* (Altun 2002-2010). Moreover, the developmental origin of each of these cells is fully described in a complete map of cell divisions from fertilized zygote to sexually mature adult (Sulston and Horvitz 1977; Sulston et al. 1983). The *C. elegans* genome sequence is also precisely defined, and at ~100 Mb is about 1/30 the size of the human genome (Hillier et al. 2005). However, with 20,168 predicted genes (http://wiki.wormbase.org/index.php/WS200), the *C. elegans* protein-coding genome is only slightly smaller than that of humans (http://www.sanger.ac.uk/PostGenomics/encode/stats.html). Major classes of non-coding RNAs (ncRNAs) such as microRNAs (miRNAs) are also represented in *C. elegans* (Ruby et al. 2006; Kato et al. 2009). Thus, *C. elegans* provides a simple but representative model of development that depends on differential expression of a compact, well-described genome. Although *C. elegans* is completely sequenced, some predicted genes lack direct evidence of transcription and other cryptic protein-coding genes and ncRNAs are likely to have been overlooked by gene prediction software (Hillier et al. 2009; Schweikert et al. 2009). In addition, the cell-specific expression patterns of the majority of *C. elegans* genes are unknown. Thus, the anatomy and development of the animal is defined at the resolution of the single cell but a comparably precise atlas of gene expression is not currently available.

The goal of a comprehensive gene expression map has been achieved in part by analysis of promoter::GFP fusions for a broad array of protein coding genes (Dupuy et al. 2007; Hunt-Newbury et al. 2007; Murray et al. 2008; Liu et al. 2009). This methodology, however, is generally not quantitative and can be misleading if key regulatory elements are omitted from the reporter genes (Hunt-Newbury et al. 2007). We have adopted the alternative strategy of measuring native transcripts from a broad array of specific tissues and cell-types. In addition,

we used whole genome tiling arrays in order to sample the entire non-repetitive genome and therefore achieve an unbiased approach to transcript discovery. In addition to assigning gene expression to identified tissues and stages, our approach of analyzing different cell types and developmental periods also ensures detection of RNAs that may be selectively expressed during discrete temporal intervals or in limited numbers of cells. We accomplished this goal by utilizing recently developed methods for obtaining RNA from specific *C. elegans* cells (Roy et al. 2002; Zhang et al. 2002; Fox et al. 2005). Altogether, we sampled 13 embryonic cell types and 12 larval and adult tissues. We also produced tiling array data sets for whole animal RNA isolated from seven different developmental stages. Additional profiling results were obtained from larval males and from the hermaphrodite gonad and soma. Thus, our datasets significantly enhance a growing body of tissue and stage specific gene expression for *C. elegans* (McKay et al. 2003; Pauli et al. 2005; Von Stetina et al. 2007 #6182; Meissner et al. 2009). Our results indicate that most protein coding genes (~75%) are differentially expressed among the stages and cell types that we sampled. In addition to providing evidence of extensive gene regulation, these results should also greatly aid genetic analysis by suggesting cell types or developmental stages in which highly expressed transcripts are likely to function. For example, our results provide the first comprehensive description of gene expression in *C. elegans* primordial germ cells and led to the discovery that proteins encoded by a subset of these genes are expressed well before their established roles in meiosis and oogenesis. To identify novel transcripts, we utilized a recently developed computational method for recognizing transcribed regions irrespective of their annotation status (Laubinger et al. 2008; Zeller et al. 2008). This approach revealed a large number of previously unannotated transcripts encoded by at least 10% of the *C. elegans* genome. These novel transcripts show striking cell specificity that may be indicative of tissue-specific functions. To facilitate the use of these data for future studies of gene function, we provide online resources for visualizing transcribed regions in a genome browser and for estimating relative gene expression levels across tissue types and developmental stages.

RESULTS

Strategies for profiling specific cell types and developmental stages

Cell-specific RNA was obtained from GFP-labeled embryonic cells isolated by FACS and from larval cells by use of the mRNA-tagging method (**Fig. 1A**, **B**, see also **Supplemental Fig. S1**). Altogether, we generated tiling array profiles from 25 different tissues, with each sample derived from one of five distinct developmental stages. The cell specificity of these data sets is reflected in the strong enrichment of native transcripts for the corresponding gene promoter used to mark each cell type for profiling (**Supplemental Fig. S2**). Corresponding reference data sets were collected from all cells for each of these developmental periods (**Table 1,** see Methods). We also generated an independent developmental series with total RNA isolated from whole animals at seven different ages (EE, LE, L1, L2, L3, L4, YA) (**Table 1**, **Fig. 1**, **Supplemental Table S1**). An additional group of tiling array profiles was obtained from young adult hermaphrodite gonads, L4 hermaphrodite somatic cells, and L4 males. Because our samples were isolated by different methods, which could potentially result in biased representation, we used principal component analysis (PCA) to compare tiling array results (see below) obtained from specific cells and from whole animals. PCA shows that tiling array profiles obtained from whole embryos cluster with data sets generated from specific embryonic cells and that larval and adult profiles are grouped with data sets obtained from specific postembryonic and young adult tissues (**Supplemental Fig. S3**). Correlation analysis comparing cell type data sets with developmental series data sets also confirms that expression estimates derived from cell types generally correlate well with the corresponding developmental stage data set generated from total RNA (**Supplemental Fig. S4**). Thus, a global analysis of our tiling array results suggests that cell-specific profiling preserves overall patterns of temporally regulated gene expression.

Expression of annotated genes detected with tiling arrays

To evaluate expression of annotated protein-coding genes, we created probe sets corresponding to the constitutive exons of individual gene models annotated in WormBase and summarized the intensity values for each gene (see Methods). We then identified transcripts that are detectable above background in each sample with a statistical test (see Methods). The

union of results derived from all cell types and stages detected a total of 17,452 genes (**Table 2**). Because these results were obtained from multiple independent comparisons, we conservatively adjusted the FDRs to limit the potential accumulation of false positives (see Methods), which resulted in expression detected for 13,149 genes from the union of cell type-specific data sets and for 13,713 genes in at least one of the stage-specific data sets. When both groups of data sets were combined, we detected 14,279 expressed genes (**Table 2**).

Our initial analysis identified an average of 12,228 transcripts in samples derived from a specific cell type or tissue (Table 2). To provide a more accurate estimate of genes expressed in each tissue type, we adopted a simple transformation designed to exclude transcripts likely to originate with the minor fraction (3-20%) of unmarked cells isolated by FACS (**Supplemental Table S1**) or from non-specific RNA generated by the mRNA-tagging protocol (Von Stetina et al. 2007). Transcripts that are highly expressed in a specific tissue might also be detectable at lower levels in profiles derived from other cell types due to this background. Thus, as a conservative measure, we restricted the set of expressed genes for each cell type to transcripts with a higher level of expression measured for a given cell type than for the corresponding reference (*i.e.*, the "average" cell) at the same stage. This approach effectively excluded, for example, the *myo-3* body muscle-specific transcript (Okkema et al. 1993) from data sets derived from non-muscle cell types while generally retaining "housekeeping" genes such as ribosomal proteins that are likely to be widely expressed in all tissues (Von Stetina et al. 2007). This analysis detected between 4,572 and 7,199 expressed genes in each of the 25 cell types profiled (5,698 genes on average, see **Supplemental Fig. S5**).

Identification of Transcriptionally Active Regions (TARs)

The high probe density (on average every 25 bp) of the tiling array made the non-repetitive portion (~85%) of the genome accessible to *de novo* identification of transcripts in a way that is not biased by potentially incomplete annotations (see **Fig. 1D** for an illustration). However, this comprehensive representation of the genome does not allow for optimized probe sequences and consequently results in large variability in hybridization affinity. Our pivotal normalization step thus aimed at reducing probe sequence bias. This correction also improved the signal-to-

noise ratio (exon intensity over background) to an even larger extent than observed for another method that additionally exploits reference hybridization to genomic DNA (**Fig. 2A**, see Methods for details) (Huber et al. 2006). For the segmentation of hybridization signals into intergenic regions, exons and introns, we used a method called mSTAD (Laubinger et al. 2008; Zeller et al. 2008)). Although mSTAD is trained on hybridization signals corresponding to known (mostly protein-coding) genes (see Methods), it afterwards predicts transcripts regardless of their annotation status. We first assessed the cross-validation accuracy of these predictions relative to annotated protein-coding gene models (**Fig. 2B**). Generally, the sensitivity of these predictions for annotated exons improved with the expression level of the corresponding genes, while high precision (~80%) with respect to overlap with annotated exons was maintained across all expression levels (**Fig. 2C**). An additional evaluation of transcripts predicted by mSTAD in comparison to another tiling array-based method (Gerstein et al., *in press*) further corroborated the conclusion that mSTAD accurately reconstructs expressed transcripts (**Supplemental Fig. S6**, **Supplemental Fig. S7**).

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

From the TAR predictions of individual samples, we constructed non-redundant TARs (nrTARs) containing the union of nucleotides inclusive to a TAR in any of the samples analyzed (30 cell type and reference samples and seven developmental stages, **Table 1**). In total, ~45 Mb were covered by nrTARs, and the subset of expressed nrTARs (*i.e.*, TARs that passed a statistical test for expression above background, see Methods) contained ~40 Mb (**Supplemental Table S2**). We next compared on an individual-nucleotide basis the overlap between known transcripts and nrTARs predicted *de novo* from the tiling array data. In a comparison to protein-coding gene models annotated in WormBase (Rogers et al. 2008), ~84% of nucleotides (*i.e.*, 22,344 kb +/- 1,127 kb) in annotated exons (from >90% of gene models) were also detected in nrTARs (**Fig. 2D**). The subset of "expressed" nrTARs covered ~80% of nucleotides in annotated exons from more than 90% of gene models (**Supplemental Table S3**) and additionally contained >18 Mb (~45% of expressed nrTARs) outside of exons for annotated protein-coding genes (**Fig. 2D**). A similar comparison between nrTARs and the modENCODE integrated transcript model

defined by transcriptome sequencing of polyadenylated RNA (Hillier et al. 2009) and EST evidence from WormBase (Gerstein et al., *in press*) detected ~25 Mb of overlapping exons or nrTARs corresponding to ~90% of nucleotides in exons of the integrated transcript model and ~57% of nucleotides in nrTARs (**Fig. 2E**). Nearly 41% of nucleotides within the expressed nrTARs were found outside of exons defined by the integrated transcript model (**Fig. 2E**). Taken together, most gene models (~90%) were supported by nrTARs, whereas a substantial fraction of nrTARs could not simply be attributed to known transcripts.

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

We defined "unannotated TARs" as those that did not significantly overlap with exons of any coding gene, ncRNA or pseudogene annotated in WormBase (Rogers et al. 2008) (**Supplemental Fig. S8**). When we additionally required that TARs not overlap with any exons of the integrated transcript model, we obtained "novel TARs" (see Methods). In total, unannotated nrTARs covered ~16 Mb of the genome; ~90% were also novel (**Fig. 2F**). Expression above background in any sample could be confirmed by a statistical test (see Methods) for ~11 Mb of novel nrTARs (**Fig. 2F**, **Supplemental Table S2**, **Supplemental Fig. S9A**). These results suggest that our extensive profiling of cells and tissues as well as developmental stages revealed a significant fraction of the *C. elegans* transcriptome that went undetected by methods limited by analysis of polyadenylated transcripts or by sampling of fewer conditions (Rogers et al. 2008; Hillier et al. 2009). Our findings parallel results from a previous tiling array study that also detected abundant non-polyA+ transcription from intergenic regions (He et al. 2007). Although our transcript identification method was originally trained on annotated protein-coding genes, it is based purely on hybridization features (Laubinger et al. 2008; Zeller et al. 2008) and hence is expected to be capable of recognizing non-polyadenylated as well as non-coding transcripts. We verified that TARs identified by mSTAD contain annotated ncRNAs, including snoRNAs, miRNAs and pseudogenes (**Supplemental Fig. S8**, **Supplemental Table S4**). Moreover, on a per-nucleotide basis, almost 60% of the putative long ncRNAs (>2.7 Mb) predicted by Liu et al. (Gen Res, *in press*) was contained in the set of novel nrTARs described here (**Supplemental Fig. S11**). However, <20% of nucleotides from the novel nrTARs recognized by our approach were also

predicted by Liu et al. (Gen Res, *in press*) (**Supplemental Fig. S11**). RT-PCR of a subset of these novel TARs confirmed expression (**Fig. 3A, B**).

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

We profiled a broad panel of tissues and developmental stages with the idea that this approach could reveal the prevalence of potential gene regulatory mechanisms that might modulate transcript abundance among different cell types or developmental periods. To detect changes in gene expression during development, we performed all pairwise comparisons (total = 21 comparisons) of the seven tiling array data sets obtained from staged embryos (EE, LE), larvae (L1, L2, L3, L4) and young adults (YA) (see Methods). To detect transcripts that are differentially expressed between cell types, we compared each of the 25 cell-specific data sets to its corresponding reference sample (total = 25 comparisons) (**Table 1**). In both cases, these comparisons were designed to detect transcripts that are either significantly depleted or enriched (≥ 2-fold, FDR ≤ 0.05) (**Fig. 4A).** After correcting for multiple testing as above (Methods), this analysis produced conservative estimates of genes differentially expressed between stages (8,606 on average) or between specific cell types and reference samples (7,983) (**Table 2**, **Supplemental Fig. S12**). A combined, corrected threshold of FDR < 0.11% for all 46 comparisons yielded 11,299 differentially expressed genes among all stages and cell types tested (**Table 2**). Although this stringent correction effectively reduced our estimate of the overall number of expressed transcripts from 88% to 72% of all *C. elegans* genes, the ratio of differentially expressed genes among detected genes in both cases is stable at ~75% (**Table 2**, **Supplemental Fig. S13**). On the basis of these results, we conclude that transcripts for a majority of *C. elegans* genes are regulated to achieve different levels of expression during development and between specific types of cells.

To validate the enrichment of genes in tissues and cell types as detected here, we compared a select number of our enriched gene lists to similar, independently derived data sets. Comparison of enriched gene sets from this study for the L2 excretory cell, embryonic and L2 intestine, L2 body wall muscle and YA gonad with comparable tissue-specific profiles generated by other groups produced highly significant overlaps (1.8- to 8.1-fold over-representations, all

hypergeometric p-values < 1e-15, see Supplemental Results SR1). These comparisons reinforce the validity of each data set, particularly since the earlier profiles were generated with a variety of methods including GFP reporter imaging and serial analysis of gene expression (SAGE) (see Supplemental Results SR1) and also may differ from our samples in developmental age. Lists of cell or tissue-enriched transcripts are included as supplemental data (http://www.vanderbilt.edu/wormdoc/wormmap/) and should be useful for identifying genes with key roles in the corresponding cell type (see also Supplemental Results SR2). Examples of this application are provided for the embryonic germ cell precursors (see below) and for the larval excretory cell (Supplemental Result SR3).

Specific genes are selectively enriched in certain cell types or tissues

Among the genes that are enriched in a certain tissue, we further sought to distinguish genes that are selectively enriched in the given tissue relative to those with broadly elevated expression in many cell types. The information theoretic concept of Shannon entropy effectively allowed us to define this subset of selectively enriched genes by distinguishing patterns of broad and uniform expression (high entropy) from more restricted ones with a high degree of tissue specificity (low entropy) (Schug et al. 2005). These lists of selectively enriched genes comprised ~20-57% of all genes enriched in the corresponding tissue or cell type (see Methods). 82% of all genes selectively enriched in any cell type or tissue are specific to only one or two samples. Moreover, the overlap between tissues sampled at two time points is generally larger than between different tissues or cell types (**Supplemental Fig. S14**). For example, the set of genes selectively enriched in embryonic dopaminergic neurons also shows elevated expression in larval dopaminergic neurons and comprises known dopaminergic genes including the ETS transcription factor, *ast-1,* and its downstream targets the dopamine transporter, *dat-1*, and dopamine biosynthetic enzymes *cat-2* and *cat-4*. (**Supplemental Result SR2**)(Flames and Hobert 2009).

We further defined the set of genes selectively enriched in any of the thirteen neuronal samples, but not enriched in non-neuronal tissue. Strikingly, this combined neuron-selective data set is most strongly enriched for putative 7 transmembrane (7TM) domain G-protein coupled receptor (GPCR)-like proteins (FDR < 5.2e-25, **Supplemental Fig. S15**)

Our finding is consistent with earlier reports of selective expression of 7TM/GPCR genes in the *C. elegans* nervous system (Troemel et al. 1995; Chen et al. 2005). Cases of 7TM/GPCR genes that are expressed in specific neurons are also evident in our tiling array results. For example, *sra-32* and *sra-36* are uniquely detected in the L2 larval stage A-class neuron data set (**Fig. 4D**). Of the 1,512 predicted members of the 7TM/GPCR family, 314 (~21%) were not detected in any RNA-seq derived data set produced from whole animals (Hillier et al. 2009) for the modENCODE consortium (Gerstein, et al. *in press*). Among these are 66 family members that are detected in our tiling array assays (Supplemental File #1 7TM genes). Our findings provide an explanation for the relative lack of coverage of the 7TM/GPCR family in the RNA-Seq results and predict that transcripts encoded by other members of this large and diverse gene family could be detected by expanding our cell-specific profiling strategy to additional neuron types (**Supplemental Fig. S16**).

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

To quantitatively assess expression differences for TARs, in particular novel ones, probes contained within TARs from each cell-specific data set were compared to probes from the same region in the corresponding reference sample (see **Table 1**). This analysis revealed ~5Mb of TARs with significant expression changes between cell types and references or between developmental stages at an FDR ≤ 0.05 and 2-fold expression difference (see Methods, **Supplemental Fig. S9B, Supplemental Fig. S10**). On average, 933 novel TARs are differentially expressed in a particular cell type in comparison to a reference sample of all cells at the corresponding developmental stage (**Supplemental Fig. S9C**, **Supplemental Fig. S12C**). We used quantitative PCR (qPCR) to confirm that ten of these differentially expressed novel TARs indeed show significant enrichment in the specific cell types initially identified in the comparison of tiling array results (**Fig. 3C**).

We next explored the extent to which novel TARs are selectively expressed with the goal of cataloging potentially rare transcripts that might be specifically detected in a limited subset of cells or in a discrete developmental period. To investigate the expression patterns of TARs on a

per-nucleotide basis, we tabulated the frequency at which a given base was detected as transcribed across cell types and stages. Approximately 15% of bases covered by exons of gene models annotated in WormBase are detected in all 25 cell-types profiled (**Fig. 4C**). A larger fraction of bases derived from gene models (~75%), however, is expressed in at least one, but not all, of the cell types and 11% is detected in no more than two cell types. For stages, we observed that 29% of bases from exons of annotated gene models is detected throughout development *vs.* 8% expressed in no more than one developmental period (**Supplemental Fig. S17**). Bases corresponding to novel TARs that map to intergenic regions showed a stronger bias for cell or stage-specific expression with >53% detected in either one or two cell types but <1% (~34 kb) detected in all cell types (**Fig. 4C**). Bases located >500 nt from a gene model (distal) comprised the majority (~75 %) of novel transcribed intergenic nucleotides uniquely detected in one or two cell types or in a single stage (**Fig. 4C**, **Supplemental Fig. S17**). Given the average intron length of 344 nt for *C. elegans* (Bradnam and Korf 2008), we suggest that these distal bases are more likely to correspond to new transcribed regions as opposed to exons belonging to existing gene models.

Online resources for visualization and data access

To facilitate further study of our tiling array-based expression data, we have made it accessible to the research community through two online visualization tools, both of which are linked from a project website (http://www.vanderbilt.edu/wormdoc/wormmap/). One of utilities displays expression values across cell types and developmental stages for a user-defined subset of genes (http://fml.mpg.de/raetsch/wormviz/tileviz.jsp, see also **Supplemental Fig. S18**). Additionally, a customized genome browser (http://gbrowse.fml.mpg.de/cgi-bin/gbrowse/ce_WS199), visualizes transcriptionally active regions (TARs) for all analyzed samples together with gene models and genomic features annotated in WormBase (see **Supplemental Fig. S7**) (Rogers et al. 2008). Raw array data have been deposited at GEO (Barrett et al. 2009) (accession numbers GSE23245-GSE23271, GSE23278-GSE23287, GSE23769-GSE23770) and supplemental data files are available for download from the project website (http://www.vanderbilt.edu/wormdoc/wormmap/) and from modMine (http://intermine.modencode.org).

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

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

Cell-specific expression profiling of embryonic primordial germ cells reveals X-chromosome silencing and early activation of meiosis and oogenesis genes.

Two primordial germ cells, Z2 and Z3, are born from the P4 blastomere within two hours of fertilization and ultimately give rise to all germ cells in the adult (Sulston et al. 1983). Z2 and Z3 are mitotically dormant and largely transcriptionally quiescent throughout embryogenesis, even as somatic cells rapidly proliferate and actively transcribe genes. Although some genes are known to be expressed in Z2/Z3 in older embryos (*e.g.*, *nos-2* and *pgl-1*; (Kawasaki et al. 1998; Subramaniam and Seydoux 1999; Kawasaki et al. 2004), the full complement of genes actively transcribed in Z2/Z3 during this period has not been previously defined. Comparison of the Z2/Z3 expression data set to a reference profile obtained from all embryonic cells (EE, **Table 1**) identified 979 genes with significantly enriched (≥ 2-fold, FDR ≤ 5%) expression in Z2/Z3 (Supplemental File Z2/Z3 #2). Because the germline reporter (P*pie-1*::GFP::PGL-1) (Table 1) used for marking Z2/Z3 for FACS shows expression in hypodermal cells of older embryos (unpublished data, see also **Supplemental Fig. S14**), we removed 335 genes from the Z2/Z3 list that were also enriched in an independent profile of embryonic hypodermal cells (**Table 1**, see supplemental protocol SP13). We refer to the resultant list of 644 genes as the "core Z2/Z3" data set (Supplemental File Z2/Z3 #2). This stringent treatment effectively excludes hypodermal genes from the core Z2/Z3 list but also eliminates other transcripts that are known to serve roles in germline development and are in fact expressed in Z2/Z3 (*i.e.*, *deps-1, ima-2, pgl-1, htp-3, cpg-2, daz-1;* **Fig. 5**, **Supplemental Table S5**).

Two observations indicate that the core Z2/Z3 data set is strongly enriched for authentic Z2/Z3 transcripts. First, we detected significant overlap between the core Z2/Z3 data set and previously defined sets of transcripts produced in the adult germline (Reinke et al. 2000; Reinke et al. 2004) and maternally loaded into embryos (Baugh et al. 2003) (397 of 644 Z2/Z3 genes, $p < 8e\text{-}72$, hypergeometric test, see Methods). The core Z2/Z3 data set also significantly overlaps with a previously generated germ line SAGE enriched gene data set (125 of 644 Z2/Z3 genes, $p < 1.62$) (Wang et al. 2009). Second, X-linked genes are as under-represented in the core Z2/Z3 dataset as they are in the adult germ line (Reinke et al. 2004) (**Fig. 5A**, Supplemental File Z2/Z3 #2). This result provides the first evidence that the X chromosomes are under-expressed in primordial germ cells, as has been previously documented in larval and adult germ cells. The specificity of this effect for germ cells is underscored by tiling array results obtained from adult hermaphrodite somatic cells ("soma only", **Fig. 5A**), which instead show over-representation of X-linked transcripts.

Strikingly, many genes expressed in Z2/Z3 have known roles in germ cell proliferation, meiosis and oocyte differentiation (**Supplemental Table S5**, **Supplemental Fig. S19**) events that do not take place in the germ line until larval and adult stages (Kimble 2005). By contrast, genes expressed during spermatogenesis are under-represented in the Z2/Z3 data set (Supplemental File Z2/Z3 #2). This observation suggests that the gene expression program for oocyte differentiation is activated in the primordial germ cells, whereas the program for spermatogenesis remains quiescent, even though spermatogenesis occurs prior to oogenesis in *C. elegans* hermaphrodites (Hubbard 2005; Shakes et al. 2009). Alternatively, many of the RNAs detected in this analysis could reflect the persistence of maternally-provided transcripts in Z2 and Z3. To distinguish between maternal loading and zygotic transcription, we manually examined existing *in situ* hybridization images (NextDB; http://nematode.lab.nig.ac.jp) for genes in the core Z2/Z3 list (**Supplemental Table S6**). We found that 14 of 100 randomly selected genes in the core Z2/Z3 list show *in situ* signal in PGCs in embryos and/or L1s, compared to 3 of 100 randomly selected genes in the genome ($p = 0.0047$, Fisher's exact test). Of the 14 genes from the core Z2/Z3 list, 12 genes show new appearance of transcript signal in

Z2/Z3 after a previously negative stage (*i.e.,* P4 or an early stage of Z2/Z3 lacked signal) (*prg-1, glh-1, ppw-2, lsl-1, pas-5, asb-1, iff-1, ucr-2.3, cpg-1, ife-3, rpl-11.1, hil-4*). Of those 12 genes, 10 show signal in Z2/Z3 in newly hatched first stage larvae (L1s), and 6 show signal in embryos. Of the 3 genes from the control list, all 3 show signal in Z2/Z3 in L1s (*top-2, cra-1, mel-46*), and none show signal in Z2/Z3 in embryos. These observations confirm that at least a subset of transcripts detected in the core Z2/Z3 data set are products of zygotic transcription in the primordial germ cells of embryos.

Transcription in Z2/Z3 of genes with protein products known to act later in development raised the question of whether zygotically-expressed transcripts are translated in Z2/Z3. We addressed this possibility by immunostaining embryos with antibodies against several meiotic proteins. Strikingly, immunostaining for both HTP-3 and REC-8 (Pasierbek et al. 2001; Goodyer et al. 2008) is easily detected in embryonic Z2/Z3. Both proteins appear to be maternally loaded in early embryos, diminish significantly by the P4 stage, and then turn on in Z2/Z3 (**Fig. 5B**). Thus, at least some of the zygotically-expressed transcripts are indeed translated in the primordial germ cells of embryos. This finding confirms the validity of our Z2/Z3 transcription profiling and provides a clear example of how these data can lead to new discoveries about germ cell biology.

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

We used self-organizing maps (SOMs) to seek shared patterns of expression for transcripts derived from coding genes (see Methods). SOMs are a widely applied clustering technique that yields intuitive visualization of high-dimensional data sets, as *e.g.,* generated with DNA microarrays (Jiang et al. 2001). SOMs are conceptually related to a technique previously proposed to construct a relational map of *C. elegans* gene expression (Kim et al. 2001). In the first instance, we fitted a SOM to the developmental stage data set (**Fig. 6A**) and identified eight regions that correspond to genes with shared patterns of either enrichment or depletion in specific developmental periods (**Fig. 6B**, see Methods). To demonstrate the variety of developmental expression patterns identified by this approach, we plotted the top 50% of best-

fitting genes from each cluster (**Fig. 6C-F**, see **Supplemental Fig. S20** for additional clusters). Cluster 1 (CS1) contains genes with elevated expression in the embryo (**Fig. 6B**, **C**). Notable examples from this group include the FoxA transcription factor, *pha-4*, the hunchback homolog, *hbl-1*, (Krause et al. 1997)(Krause et al. 1997)and the helix-loop-helix transcription factors (bHLH), *hlh-2* and *hlh-3*, for which independent studies have detected peak expression in the embryo (Azzaria et al. 1996; Krause et al. 1997; Fay et al. 1999). Cluster 5 (CS5) contains genes with elevated expression in embryonic stages and in the adult (**Fig. 6B**, **E**). Strikingly similar protein and transcript levels have been previously observed for a member of this group, the FLYWCH transcription factor *flh-1*, which blocks expression of specific miRNA genes during embryogenesis (Ow et al. 2008).

We applied a similar SOM clustering procedure to the cell-type specific data sets in order to delineate genes that are co-regulated in different tissues (**Fig. 7A**, **Supplemental Fig. S21**, see Methods). Because these cell types were sampled across a series of developmental stages, we also expected this approach to detect genes with temporally correlated expression. **Fig. 7A** depicts the resultant regional clusters superimposed on the SOM. Clusters showing stage-specific expression include C1 (**Fig. 7B**), which features genes that are highly expressed in all postembryonic cell types and C7 (**Fig. 7D**), which is biased for genes expressed in late embryos and especially in neurons. C8 is dominated by genes that are highly expressed in neurons, but are depleted or show weak expression in most other cell types (**Fig. 7E**). Examples of genes in this group include *ric-4* (snap-25), a synaptic vesicle component that facilitates neurotransmitter release and is known to be exclusively expressed in neurons (Hwang and Lee 2003), and *acy-1* (adenylate cyclase), a key regulator of neuron-dependent behavior (Reynolds et al. 2005). Several clusters detect highly expressed intestinal genes including C1 (postembryonic cell types and larval intestine) (**Fig. 7B**) and C11 (embryonic and larval intestine) (**Fig. 7F**, see **Supplemental Fig. S22** for additional examples).

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

Because each SOM cluster includes genes with similar patterns of expression, we searched for instances in which genes in a specific cluster share common DNA sequence motifs through which *trans*-acting factors might coordinate their expression. To explore this possibility, we applied the FIRE motif analysis program to the SOM clusters (Elemento et al. 2007). FIRE uses mutual information between the presence or absence of a short nucleotide sequence and the occurrence of a gene in a particular expression cluster to identify over-represented motifs. FIRE produces optimized motifs and links the results to motifs that are available in public databases.

FIRE identified 20 upstream promoter motifs and 9 over-represented 3' UTR sequences in genes contained in the SOM clusters derived from developmental stages (**Fig. 6**, **Fig. 8A**, **Supplemental Fig. S23**, **Supplemental Fig. S24**). A canonical EBox and bHLH binding site is detected in cluster CS1 which, as noted above, includes transcription factors HLH-2 and HLH-3 (Thellmann et al. 2003). The over-representation of GATA-like transcription factor binding sites in four clusters (CS4, CS5, CS6, CS7) is likely indicative of the broad roles of GATA factors in multiple developmental pathways in *C. elegans* including endodermal and hypodermal cell fate determination and differentiation, germline gene regulation and aging (Koh and Rothman 2001; McGhee et al. 2007; Budovskaya et al. 2008; del Castillo-Olivares et al. 2009). The second highest-ranking motif corresponds to a GC rich sequence that has been previously identified by computational analysis of germ line expressed genes (Li et al. 2010). This motif is also similar to a putative transcriptional activation site for the E2F homolog, EFL-1, that promotes gene expression in the germ line (Chi and Reinke 2006). Detection of these GATA and E2F sites in cluster CS5 is consistent with our finding that genes which contain these 5' sites and which are enriched in germ line precursor (GLP) cells (Supplemental File Z2/Z3 #2) are also over-represented in this cluster (23 GLP genes with the GATA site are 1.6 fold over-represented, $p < 0.017$; 40 GLP genes with the E2F site are 1.8 fold over-represented, $p < 2.74e-04$). These results validate our approach and suggest that other motifs revealed by this strategy may also correspond to binding sites for transcription factors that regulate developmental gene expression.

For SOM clusters derived from cell-specific profiles (**Fig. 8B**), FIRE identified 45 over-represented sequences including 35 upstream motifs and 10 RNA sequences that map to 3' UTR domains (**Fig. 8B**, **Supplemental Fig. S23**, **Supplemental Fig. S24**). As noted above for the SOM clusters derived from developmental stages, the highest scoring motif matches a GATA transcription factor-binding site. In *C. elegans*, the *elt-2* GATA transcription factor is known to interact with this sequence to drive expression of intestine-specific genes (McGhee et al. 2009). Our results also reflect this role; the GATA motif is overrepresented in cluster C11, which contains transcripts enriched in the embryonic and larval intestine profiles (**Fig. 7F**, **Fig. 8B**), and in C1 and C4 both of which show peak expression in larval intestine (**Fig. 7B**, **Supplemental Fig. S22**, **Fig. 8B**). The accurate identification of the GATA factor-binding site by the FIRE algorithm suggests that other motifs associated with specific SOM clusters may also correspond to specific transcription factor binding sites. An interesting example includes the sequence, TTTCG[AC]AA[CT] (**Fig. 8B**), that is over-represented in genes enriched in embryonic neurons in cluster C7 (**Fig. 7D**) and also reciprocally depleted in genes that are under-expressed in embryonic neurons in cluster C5 (**Supplemental Fig. S22**). This motif is bound by the vertebrate C/EBPα transcription factor (Grange et al. 1991), which has been shown to function with NeuroD to regulate neural gene expression (Sandelin et al. 2004; Calella et al. 2007). It will be interesting to determine whether C/EBP and NeuroD homologs exercise similar functions in *C. elegans* neural develoment.

FIRE also identified 3' UTR binding sites for two distinct groups of miRNA genes belonging to the *mir-58* and *mir-51* families (**Fig. 8B**, **Supplemental Fig. S23**, **Supplemental Fig. S24**). Members of the *mir-58* family (*mir-58*,*-80*,*-81*,*-82*) are abundantly expressed throughout development (Lim et al. 2003; Kato et al. 2009), but assays with promoter::GFP reporter genes have detected cell-specific patterns of expression (Martinez et al. 2008). For instance, *mir-58* is expressed in a broad array of cell types including the excretory canal, intestine, pharynx, and hypodermis, but is excluded from the nervous system (Isik et al. 2010). Emerging evidence indicates that transcript de-stabilization is the principle mechanism whereby miRNAs down-regulate gene expression (Bagga et al. 2005; Guo et al. 2010). Thus, the absence of *mir-58* expression in the

nervous system predicts that neuronal transcripts carrying the *mir-58* recognition sequence should escape *mir-58*-induced degradation. And, in fact, our result showing that the *mir-58* sequence is over-represented in neuron-enriched transcripts (cluster C8, **Fig. 7E**, **Fig. 8B**) is consistent with this model. The motif for the *mir-51* family (*mir-51,-52,-53,-54,-55,-56*) is also over-represented in C8 (**Fig. 7E**, **Fig. 8B)** and in SOM clusters C9 and C10 that are dominated by transcripts enriched in hypodermal cells and neurons (**Supplemental Fig. S22**). This pattern suggests that *mir-51* family genes may have limited roles in regulating transcript levels in neurons and in the hypodermis. Conversely, the observation that the *mir-58* and *mir-51* motifs are significantly under-represented in C2 and C3 is suggestive of strong regulation by these miRNAs in the tissues that contribute to this cluster. In considering this question, we noted that C2 and C3 include an expression peak for the L3/L4 reference sample (**Fig. 7C**). Because germ line tissue is rapidly proliferating at this stage (Kimble 1981), we compared the genes in clusters C2 and C3 to separate tiling array profiles obtained from the adult hermaphrodite gonad, L4 males and all somatic cells at L4 stage (see Supplemental Result SR5). These comparisons show a significant overlap, showing that C2 and C3 genes are largely expressed in the germline and specifically enriched for sperm expression. Thus, we speculate that members of the *mir-58* and *mir-51* gene families may have significant roles in modulating transcript levels in the germ line. The *mir-58* and *mir-51* motifs were previously identified by FIRE analysis of an independent group of whole animal microarray data sets from *C. elegans* (Elemento et al. 2007). Our results have now confirmed that each of these motifs is associated with a group of co-regulated genes and have also provided additional clues pointing to the specific cell types in which *mir-51* and *mir-58* might function.

DISCUSSION

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

To monitor expression of individual protein-coding genes, we derived intensity values from aggregated probe sequences corresponding to each annotated gene model. Our combined set of tiling array data from 25 different cell types and seven developmental stages (**Fig. 1**) detected ~90% of known protein coding genes (**Table 2**, **Fig. 2**, **Supplemental Table S3**).
In addition to detecting expressed genes, our analysis also revealed that ~75% of all detected genes show at least 2-fold, statistically significant differences in transcript levels between cell types or developmental stages (**Table 2, Supplemental Fig. S13**). To document this trend of widespread differential expression among cell types and throughout development, we tabulated the frequency of transcription of a given nucleotide across tissues and stages. This analysis revealed that whereas 15% of exonic sequence is detected in all of the cell types that we sampled, a larger fraction (60%) shows more limited transcription with ~11% in no more than one or two cell types (**Fig. 4C**). Our results also indicate that coding sequence is dynamically expressed during development with ~8% of bases from exons uniquely detected in only one embryonic, larval or adult stage (**Supplemental Fig. S17**). As we extensively sampled the *C. elegans* nervous system, we investigated the subset of genes selectively expressed in neuronal tissue. Among these genes, we noted striking enrichment of members of the 7TM-GPCR family, which is known for highly specific expression in the nervous system (Chen et al. 2005). The restricted expression of 7TM-GPCR genes potentially explains why many members of this family still lack experimental support (Hillier et al. 2009; Schweikert et al. 2009). Our results, however, suggest that profiles of more cell types should confirm expression of additional annotated gene models (see **Supplemental Fig. S16**) or genes newly predicted from the genome sequence. Overall, our finding of widespread differential gene expression underscores the conclusion that most *C. elegans* genes are extensively regulated and points to the key role of differential gene expression in the determination of cell fates and developmental progression. In practice, our data on genes that are selectively enriched in a particular cell type or developmental period should be especially useful for identifying genes with cell- or stage-specific functions (Zhang et al. 2002; Colosimo et al. 2004; Blacque et al. 2005; Cinar et al. 2005; Touroutine et al. 2005; Von Stetina et al. 2007; McGhee et al. 2009; Chatzigeorgiou et al. 2010; Smith et al. 2010; Hallem et al., *in press*). For example, our analysis

of the transcripts enriched in the primordial germ cells (Z2/Z3) revealed two new features of germline development. First, compared to the autosomes, relatively few X-linked genes are expressed in *C. elegans* primordial germ cells (**Fig. 5A**). This embryonic silencing of X-linked genes mirrors a similar effect in larval and adult germ cells (Reinke et al. 2004). Thus, except for a brief period of activation of X-linked genes in oocytes, the X chromosome is under-expressed throughout all stages of germ cell development. Second, Z2/Z3 express genes involved in meiosis and oogenesis, well in advance of when those processes actually occur. Moreover, at least some of those transcripts are translated into protein (**Fig. 5B**). This result raises the interesting question of whether those proteins serve additional non-meiosis and non-oogenesis roles or instead whether primordial germ cells express many more of their repertoire of gene products than they need at that stage. Notably, the primordial germ cells of *Drosophila* and mice also express meiosis proteins, indicating that early expression of this protein class is conserved (Baltus et al. 2006; Mukai et al. 2006; Rogers et al. 2008). More detailed investigation of the individual genes expressed in Z2/Z3 should provide a wealth of new information about the processes occurring in primordial germ cells, including mitotic quiescence and chromatin regulation. As one example, cyclin B is important for mitotic quiescence of germ cells in other systems such as *Drosophila* (Deshpande et al. 1999). Of the several orthologs of cyclin B in the *C. elegans* genome, only *cyb-2.2* is enriched in the Z2/Z3 data set, making it a candidate to mediate mitotic quiescence in Z2/Z3.

In addition to using our tiling array results to identify genes expressed in specific cell types or developmental periods, we also sought evidence for more complex patterns in which cohorts of genes might be similarly regulated across tissues or among different development stages. For this purpose, we used the unbiased strategy of self-organizing maps (SOMs) to cluster co-expressed genes (**Fig. 6**, **Fig. 7**). This approach revealed, for example, a striking cluster with consistently elevated transcript levels in both embryonic and larval neurons that is largely comprised of genes with established neuron-specific functions (**Fig. 7E**). Other clusters could reflect genes with common functions in a wide array of cell types during a particular developmental period (**Fig. 7B**, **D**). Thus, our approach has confirmed known groups of co-

regulated genes as well as suggested novel clusters that could point to previously unstudied biological roles for batteries of co-expressed genes. In addition to providing a direct read-out of cell-specific gene expression, our microarray data should also substantially enhance the accuracy of SVM-based strategies that rely on gold standard training sets for *ab initio* identification of cell-specific expression from whole animal microarray data (Chikina et al. 2009). Motif analysis of the SOM results derived from our data sets identified highly over-represented flanking sequences in genes belonging to specific clusters (**Fig. 8**, **Supplemental Fig. S23**). Each case could be indicative of a regulatory mechanism involving a shared *trans*-acting factor. For example, a consensus binding site for a GATA factor with a broad role in regulating intestine-specific genes in *C. elegans* (McGhee et al. 2009) was specifically over-represented in SOM clusters defined by transcripts with high expression levels in tiling array data sets derived from intestinal cells (**Fig. 8B**). Over-represented motifs in the 3' UTR regions include recognition sites for two large and highly expressed groups of closely related miRNAs, the *mir-58* and *mir-51* families (**Fig. 8B**). Our analysis of these results points to potential roles for both *mir-58* and *mir-51* in regulating transcript abundance in the germ line, a suggestion consistent with the recent observation that the Drosophila ortholog of the *mir-58* family, bantam, is required for germline stem cell fate (Yang et al. 2009).

Our tiling array results confirm expression of the vast majority (~90%) of *C. elegans* protein coding genes recently identified by RNA-Seq analysis (**Supplemental Table S3**) (Hillier et al. 2009). Additionally, our machine-learning algorithm also identified a substantial number of TARs arising from intergenic regions (**Fig. 2D-F**). A conservative treatment of these data that uses a statistical test for expression above background, leads to the estimate that ~11 Mb of intergenic sequence, or ~10% of the *C. elegans* genome, encodes novel RNAs that have not been previously annotated in WormBase or detected by RNA-Seq (**Supplemental Table S2**). One explanation for this difference is that we assayed total RNA from embryonic cells and developmental stages and that the poly-A+ pull-downs that we used for sampling postembryonic cell types (**Fig. 1B**) also include a significant non-polyA+ fraction (Von Stetina et al. 2007). In contrast, recent RNA-Seq results for *C. elegans* were limited to purified poly-A+

RNA (Hillier et al. 2009). Because the known families of short non-coding RNAs (ncRNAs) were manually excluded from our list of intergenic RNAs, we propose that these transcripts define potentially new types of non-coding RNA. An independent analysis of *C. elegans* transcriptomics data that includes the tiling array results used in this work, also reports a substantial number (~4.6 Mb) of putative non-coding RNAs from intergenic regions with a large overlap (>2.5 Mb) to our ncRNA predictions (**Supplemental Fig. S11**) (Liu et al. *in press*). Our analysis indicates that transcription of these novel TARs shows an even stronger bias for cell-specific expression than transcripts derived from protein coding genes (**Fig. 4C**). In this respect, our findings are similar to an earlier report that a majority of unannotated human transcripts are expressed in only one of the eleven different cell lines sampled (Birney et al. 2007). Although the extent of intergenic transcription from the mammalian genome is controversial (van Bakel et al. 2010), mounting evidence points to multifaceted roles for long intergenic ncRNAs (lincRNAs) including transcriptional control, imprinting, dosage compensation and maintenance and remodeling of chromatin structure (Rinn et al. 2007; Hirota et al. 2008; Wilusz et al. 2009; Tsai et al. 2010). Nevertheless, in every case, definitive tests are required to establish specific functions for candidate regulatory ncRNAs. The tissue-specific patterns of ncRNA expression (**Fig. 4C**, **Supplemental Fig. S12C**) that we have revealed for *C. elegans* should provide a valuable guide to the likely focus of mutant phenotypes that perturb expression of specific ncRNAs (Mercer et al. 2008). We note for example, that the recent discovery of an *in vivo* role for the lincRNA, Evf2, in neuronal differentiation hinged on prior knowledge of Evf2 expression in a specific brain region (Bond et al. 2009).

Although the tiling array results reported here should provide a useful resource for defining the roles of specific genes in cell fate and development, RNA-Seq data derived from these cell specific RNA samples would offer a more accurate representation of gene structure and substantially greater dynamic range for measuring differential gene expression. With the recent development of effective methods for excluding ribosomal RNA from sequencing templates (Armour et al. 2009; Albrecht et al. 2010), it should now be feasible to use RNA-Seq for a direct test of the non-coding RNA transcripts predicted by our tiling array results (WC Spencer and DM Miller, unpublished). The fact that cell-specific tiling arrays detected predicted coding genes

that were not touched by RNA-Seq analysis of *C. elegans* transcripts derived from the whole animal, also suggests that deep sequencing of RNA isolated from individual cell types could reveal additional protein-coding genes (see **Supplemental Fig. S16**).

METHODS

Sample production

Nematode culture. *C. elegans* strains were maintained as described (Brenner 1974). We used N2 as the wildtype strain. Other strains used in this study are listed in **Supplemental Table S1**.

Construction of cell-specific 3XFLAG::PAB-1 plasmids. To express 3XFLAG-tagged PAB-1 in specific cell-types, promoters were amplified and cloned into the *pSV41(Pgateway::3XFLAG::PAB-1 + unc-119 minigene)* plasmid using the Gateway cloning system (Invitrogen). Transgenics were obtained by microparticle bombardment or by microinjection (see supplemental protocols SP1 - SP4).

Isolation of cell-specific RNA by the mRNA-tagging method. Cell-specific RNA was isolated from transgenics expressing 3XFLAG-tagged PAB-1 using the mRNA-tagging strategy (Roy et al. 2002) described in (Von Stetina et al. 2007) (see supplemental protocol SP5).

Preparation and primary cell culture of embryonic cells and isolation of fluorescently-labeled embryonic cells by FACS. Embryonic cells were isolated by FACS as previously described (Christensen et al. 2002; Fox et al. 2005; Fox et al. 2007). Cell types were sorted to a fractional purity ranging from 80-97% (**Supplemental Table S1**). RNA was extracted from sorted cells in Trizol LS, treated with DNaseI and purified using the DNA-free RNA kit from Zymo Research (see supplemental protocols SP6 - SP8).

RNA amplification and microarray hybridization. RNA from sorted cells and mRNA-tagging lines was amplified and labeled using the WT-Ovation Pico, WT-Ovation Exon and Encore Biotin

kits from NuGEN Inc. for application to *C. elegans* tiling arrays (Affymetrix). Pearson correlation coefficients between replicates were determined to confirm consistent microarray data quality (see supplemental protocols SP9, SP10).

RT-PCR and quantitative PCR to validate novel and differentially expressed transcripts. Primers were designed to produce short amplicons (75-150 bp) using Batch-Primer3 (You et al. 2008). RT-PCR was performed using the same cDNA produced for microarray analysis for template and GoTaq polymerase (Promega). Quantitative PCR (qPCR) was performed using the same cDNA produced for microarray analysis for template and Sso-Fast Eva green reaction mix on a CFX96 real-time PCR machine (Bio-Rad) (**Fig. 3**, see supplemental protocols SP11, SP12).

Immunochemistry for Z2/Z3 protein expression. Embryos were fixed using methanol/acetone (Strome and Wood 1983). Images were acquired with a Volocity spinning disk confocal system (Perkin-Elmer/Improvision, Norwalk, CT) fitted on a Nikon Eclipse TE2000-E inverted microscope (**Fig. 5B**, see supplemental protocol SP13 for details).

Computational analyses of tiling array data

Array annotation. Tiling array features were mapped to the *C. elegans* genome and WormBase gene annotation (Rogers et al. 2008). Additionally repetitive tiling probes were flagged (see supplemental protocol SP14). Based on annotated protein-coding gene models, tiling probes were annotated into exonic, intronic, intergenic and ambiguous categories.

Normalization and transcript identification. Raw tiling array data were normalized to correct for uneven background (Borevitz et al. 2003; Zeller et al. 2009), between-array variability with quantile normalization (Bolstad et al. 2003) and probe-sequence effects with transcript normalization (Zeller et al. 2008). We evaluated the extent to which normalizing for probe sequence effects improved subsequent transcript recognition in comparison to DNA reference normalization (Huber et al. 2006) on the basis of the above probe annotation (**Fig. 2A**). In this context, we defined sensitivity as the percentage of tiling probes with signal above a cutoff (true positives, TP) among all annotated exon probes and undetected ones (false negatives,

FN): Sn = TP / (TP + FN). Precision was defined as the percentage of annotated exon probes (TP) among those with signal above the cutoff (including true and false positives, FP): Pr = TP / (TP + FP). Varying the threshold parameter across the whole range of measured array intensities resulted in curves showing different trade-offs between precision and sensitivity (**Fig. 2A,** see supplemental protocol SP15).

For *de novo* identification of transcriptionally active regions (TARs) from tiling array data, we employed mSTAD (margin-based segmentation of tiling array data), a machine-learning based method (Laubinger et al. 2008; Zeller et al. 2008). Its internal parameters were trained on hybridization patterns and tiling probe annotations in regions around experimentally confirmed genes. Genome-wide TAR predictions were generated in a two-fold cross-validation scheme. Cross-validation accuracy was assessed with respect to annotated genes confirmed by full-length cDNA sequences (**Fig. 2B**, **C**) as well as to the modENCODE integrated transcript model (Gerstein et al., *in press*) (**Supplemental Fig. S6**, Supplemental Protocols SP16, SP17). Additionally, accuracy was compared to modMine TARs (**Supplemental Fig. S6**, Supplemental Protocol SP17, Supplemental Result SR4).

Identification of new transcripts. "Unannotated" TARs were identified in comparison to coding and non-coding genes and pseudogenes annotated in WormBase as those with < 20 nt overlap to annotated exons. If additionally a given TAR did not overlap by ≥20 nt with exons of the integrated transcript model, we called it a "novel" transcript (**Fig. 2D-F**, see supplemental protocol SP18). Taking the per-nucleotide union of TARs obtained in individual samples, we obtained non-redundant (nr) TARS (analogously for expressed nrTARs, differentially expressed nrTARs, unannotated nrTARs and novel nrTARs) (**Fig. 2F**, see supplemental protocol SP18). For each position within expressed nrTARs, we counted the number of samples in which a TAR was detected to generate histograms of sample specificity (**Fig. 4C**, **Supplemental Fig. S17**, see supplemental protocol SP18).

Detection of expressed transcripts and significant expression differences. For each annotated protein-coding gene and predicted TAR, we constructed a probe set for expression

summarization (see supplemental protocol SP19). Subsequently, transcript expression was estimated using a customized RMA pipeline (Bolstad et al. 2003; Irizarry et al. 2003; Gautier et al. 2004) (see supplemental protocol SP19). A Mann-Whitney U test with an empirical background model and FDR correction for multiple testing was used to detect expressed transcripts (Benjamini and Hochberg 1995). Genes and TARs with an FDR ≤ 0.05 were reported as expressed above background (see **Table 2**, **Supplemental Fig. S9A**, see also supplemental protocol SP20). We detected differentially expressed transcripts using a method based on linear models (Smyth 2004). Genes and TARs were called differentially expressed if the FDR was ≤ 0.05 and the fold change (FC) ≥ 2.0 (**Table 2**, **Fig. 4A**, **Supplemental Fig. S9B**, **C**, **Supplemental Fig. S12**, see supplemental protocol SP21). To more strictly correct for potential false-positives resulting from multiple sample comparisons, we divided individual FDR estimates by the number of samples or sample comparisons, respectively. This resulted in an adjusted FDR of 1.3e-4 for expression above background and of 7.4e-4 for differential expression (**Table 2**, **Supplemental Fig. S12**, supplemental protocol SP22). We called genes "selectively enriched" in a given tissue (see Results) if they met the following requirements: (i) enriched expression in a given tissue (FDR ≤ 0.05 and FC ≥ 2.0), (ii) fold change *vs.* reference among the upper 40% of the positive FC range observed for this gene across all tissues, (iii) fold-change entropy among the lower 40% of the distribution observed for all genes (see supplemental protocol SP23, (Schug et al. 2005).

Self-organizing maps. We adopted self-organizing maps (SOMs) (Kohonen 1982) as a means of discovering, clustering and visualizing gene expression similarity with respect to cell types or developmental stages. One SOM was fitted to mean-normalized log2-transformed gene expression estimates from the developmental stage data set (**Fig. 6**, **Supplemental Fig. S20**) and another one to those from cell type samples (**Fig. 7**, **Supplemental Fig. S21**, see supplemental protocol SP25). Regions in the SOM corresponding to characteristic and coherent expression patterns were afterwards identified by *k*-means clustering of the SOM units (with $k = 8$ and $k = 14$ for the developmental and the cell type data set, respectively, see supplemental protocol SP25). The top half of more coherent SOM units were identified by means of silhouette

coefficients resulting in the clusterings shown (Rousseeuw 1987) (**Fig. 6, Fig. 7**, supplemental protocol SP25). Finally, we visualized prototypical gene expression patterns for each SOM region. Plotted are genes with a best-matching SOM unit within one of these regions and a quality error below the 50th and 20th percentile for developmental and cell-type data sets, respectively (**Fig. 6**, **Fig. 7**, **Supplemental Fig. S20**, **Supplemental Fig. S22**).

Motif discovery. Regulatory elements were identified using the FIRE algorithm (Elemento et al. 2007). Gene clusters produced from SOM analyses were submitted to the Integrated Genomics Exploration Tools (IGET) website (http://iget.princeton.edu) for analysis using FIRE.

GO/Protein domain enrichment analysis. Gene lists were tested for gene ontology (GO) or protein domain enrichment using the enrichment widgets on the modMINE website (http://intermine.modencode.org). Significance of enrichment was determined using hypergeometric tests and p-values were corrected for multiple-testing using FDR (Benjamini and Hochberg 1995).

ACKNOWLEDGEMENTS

We thank E. Hallem, N. Ringstand and P. Sternberg for P*gcy-33*::GFP, B. Grant for P*unc-122*::RFP, O. Hobert for P*rig-3::GFP*, Harald Hutter for P*glr-1::dsRed*, S. Kim for the P*ges-1* mRNA tagging line, P. Roy for the P*myo-3* mRNA tagging line, D. Marlee and A. George for help with building the P*clh-4* mRNA tagging line, S. Schultheiss for advice with the motif analysis, J. Tap for helpful discussions, D. Anastassiou for support, V. Varadan for advice, G. Seydoux for encouragement, D. Hall and Z. Altun for use of images from Wormatlas, N. Kurn, S. Wang, and J.D. Heath (NuGEN) for help with developing amplification and labeling methods for tiling array hybridization, the Vanderbilt Functional Genomics Shared Resource for microarray processing, the Veterans Administration Medical Center Flow Cytometry Core and the Vanderbilt University Medical Center Flow Cytometry Core for FACS. Some of the strains used in this work were provided by the *C. elegans* Genetic Center, which is supported by NIH NCRR. Some GFP expression data was obtained from the Genome BC *C. elegans* Gene Expression Consortium, which is funded by Genome Canada and Genome British Columbia. This work was supported by

the Max Planck Society and DFG RA1894/1-1 (GR, SRH, GZ), an EMBL postdoctoral fellowship (GZ), NIH grants HG004263 (DMM, VJR), NS49743 (JDW), GM34059 (S. Strome), GM83548 (LP), R01 NS064273 (S. Shaham), NS26115 (DMM), MH077302 (DMM), P50 DK44757 (A. Fogo) and by NIH grants to Vanderbilt University: P30 CA68485, P60 DK20593, P30 DK58404, HD15052, P30 EY08126 and PO1 HL6744.

FIGURE LEGENDS

Fig. 1. Strategies for generating tiling array data sets from specific *C. elegans* cells in embryos and larva and from whole animals at defined developmental stages.

(A) In the MAPCeL (Micro-array Profiling of *C. elegans* Cells) method, embryos are isolated from gravid adults and blastomeres released by treatment with chitinase. Dissociated embryonic cells are either sorted immediately or cultured for 24 hrs before FACS. Total RNA is amplified for tiling array analysis.

(B) The mRNA-tagging strategy was used to isolate RNA from specific larval and adult cells. The epitope-tagged (FLAG) polyA-binding protein (PAB-1) is expressed under the control of cell-specific promoters. The PAB-1:RNA complex is immunoprecipitated and RNA is amplified for tiling array analysis.

(C) Total RNA is isolated from synchronized populations of embryonic, larval and adult animals for tiling array analysis.

(D) Tiling array data (middle) is shown in a region around the annotated transcript *C15A7.1* (top). Each vertical bar corresponds to the signal of one probe feature. A transcript identified by mSTAD using only the tiling array signal is shown at bottom.

Fig. 2. *De novo* transcript identification with mSTAD and overlap of TARs with annotated and experimentally defined gene models.

(A) Transcript normalization (red) improved exon probe recognition over raw data (black) and compared to normalization using genomic DNA hybridization as reference (blue). Sensitivity and precision were estimated after thresholding the intensity data with increasing cutoffs in a fivefold cross-validation. Sensitivity is defined as the percentage of exon probes with signal

above the threshold among all annotated exon probes. Precision is defined as the percentage of annotated exon probes among those with signal above the threshold (see Methods). Values in parentheses indicate area under the curve. Based on data from LE-ref (see **Table 1**).

(B) Cross-validation accuracy of mSTAD for probes (green), for exons (blue) and for exons with independently confirmed expression (brown). For exons, sensitivity is defined as the percentage of annotated exons for which all corresponding tiling probes were predicted as exonic by mSTAD. Precision is defined as the percentage of predicted exons for which all probes are annotated as such. Definitions for probes are as in (A) but with respect to predictions by mSTAD. Exon-level evaluation was repeated with the subset of predicted exons also detected as expressed by a statistical test (see Expressed exon level). Enlarged crosses correspond to predictions used for subsequent analysis. Based on data from LE-ref (**Table 1**).

(C) Accuracy of exon and intron recognition increased with gene expression. Colored bars correspond to equally sized expression bins. Here exon overlap sensitivity equals the percentage of predicted exons, which overlap by at least 75% of their length with annotated exons. Exon overlap precision equals the percentage of exon predictions overlapping with annotated exons (by $\geq$ 75%) among all predicted exons (Intron overlap sensitivity and precision are defined analogously with respect to predicted and annotated introns). Based on data from LE-ref (**Table 1**).

(D) Overlap between non-redundant TARs (nrTARs), the portion detected as expressed and annotated coding gene models. ~45% of expressed nrTAR bases do not overlap with annotated coding gene models.

(E) Overlap between TARs and the modENCODE integrated transcript model (Hillier et al. 2009) Gerstein et al., *in press*). ~41% of expressed nrTAR bases do not overlap with the integrated transcript model.

(F) Unannotated and novel TARs and their overlap with TARs expressed above array background. Unannotated TARs are defined as TARs without significant overlap (≥ 20bp) with exons of annotated coding genes, pseudogenes and non-coding RNAs. Novel TARs are defined as the subset of unannotated TARs without significant overlap (≥ 20bp) with exons in the integrated transcript model (see main text for details).

Fig. 3. mSTAD detects TARs corresponding to protein-coding genes and to novel transcribed regions.

(A) Novel TARs detected in larval L2 intestine. Enlarged region shows location of primers and predicted RT-PCR amplicon from two TARs, L2-int-1 and L2-int-2.

(B) RT-PCR detects novel TARs expressed in specific cell types. TARs L2-int-1, 2, 3 are detected in RNA isolated from the larval L2 intestine (L2-int) but are not amplified from RNA in the absence of reverse transcriptase (L2-int-RT).

(C) qPCR validates enrichment of novel TARs in specific cell types. Log2 ratio of enrichment in specific tissue *vs*. corresponding reference samples (Table 1).

Fig. 4. Transcripts enriched or depleted in certain cell types.

(A) Genes differentially expressed between a given cell type and the corresponding reference sample (FDR ≤ 0.05). Bars pointing up and down indicate the number of enriched and depleted genes, respectively, relative to reference. Expression fold change is color-coded (see key).

(B) Log2-expression fold change relative to reference shown as gray lines for genes selectively enriched in LE dopaminergic neurons (highlighted in yellow). Four selectively enriched genes (*ast-1, dat-1, cat-2, cat-4*) with known function in these neurons are plotted in color (see key).

(C) Coverage of the genome by expressed transcripts at bp-resolution. Nucleotides in non-redundant TARs (nrTARs) (for 25 cell-type samples, **Table 1**) were binned according to the number of samples for which a TAR was detected at the given position. Bars pointing upward correspond to expressed TARs overlapping with exons of annotated coding genes and those defined by the integrated transcript model. Bars pointing downward correspond to nucleotides in expressed novel TARs (see main text for definition) organized into subgroups according to their location relative to annotated protein coding gene models (see key). Intergenic positions were classified as proximal if within 500 bp of any annotated gene and otherwise as distal.

(D) 7TM genes are selectively expressed in a specific neuron. Two members (*sra-32* and *sra-36*) of a tandem array (yellow highlights) of 7TM-encoding genes are selectively enriched in the A-type motor neuron data set derived from L2 larvae.

Fig. 5. Global characteristics of Z2/Z3-enriched gene expression.

(A) Z2/Z3-expressed genes are under-represented on the X chromosome. The chromosomal locations of genes expressed in the adult germ line ("all germline")(Reinke et al. 2004), genes identified from tiling array profiles of Z2/Z3 minus hypodermis-expressed genes ("Z2/Z3 core", see text) and genes identified from tiling array profiles of adult hermaphrodite somatic cells ("soma-only") were determined. Observed/expected was calculated by comparing the number of genes per chromosome from each data set to the entire number of protein-coding genes on that chromosome.

(B) Meiotic proteins are expressed in Z2/Z3. Antibodies to HTP-3 and REC-8 were used to stain mixed stage embryos, along with PGL-1 or PGL-3 to mark P granules in Z2/Z3. Stage of germline development is indicated on the left, with the earliest stage shown (P2/P3) likely showing persistence of maternal protein, the P4 stage showing decreased protein in the P4 cell, and the Z2/Z3 stage showing an increase in protein level in Z2 and Z3.

Fig. 6. Expression patterns during *C. elegans* development.

(A) Component planes of a self-organizing map (SOM) fitted to the developmental stage data set. Each component plane visualizes mean-centered gene expression (log2-scale) in one stage as a color gradient from blue to red indicating low and high expression, respectively (see color bar): EE - early embryos, LE - late embryos, L1 - larvae stage 1, L2 - larvae stage 2, L3 - larvae stage 3, L4 - larvae stage 4, YA - young adults.

(B) Eight regions (CS1 to CS8) of the SOM, which robustly clustered together, are color-coded (see main text for details).

(C) - (F) Mean-centered log2-expression values of genes corresponding to four of the clusters in (B) are plotted for the 50% of best-fitting genes (additional clusters in **Supplemental Fig. S20**). Colored lines indicate the expression of a selected subset of genes (see key). *mec-17* and *nlp-8* encode neuron-enriched transcripts; *chn-1* and *spo-11* are highly expressed in the adult hermaphrodite gonad; *puf-8* is highly expressed in embryonic and adult germline and *ssq-2* encodes a sperm-specific transcript. See results for other labeled genes.

Fig. 7. SOM clustering of tissue- and cell-type data

(A) Regions SOM for cell-type data as defined by *k*-means clustering.

(B) - (F) Expression patterns of genes from selected clusters in the SOM. Cell types are indicated at bottom and reference samples shaded in gray. Box plots were generated from the mean-centered log2-expression values of prototypical genes for a given cell type with horizontal lines indicating the median, boxes delineating the interquartile range and whiskers extending to the most extreme values within 1.5 times the median-quartile range; outliers are depicted as black crosses. Some of the SOM clusters correspond to peaked expression in a subset of cell types and/or developmental stages: (B) higher expression in larval stages and YA compared to embryo with a prominent peak for intestine, (C) elevated expression in L3/L4 reference, (D) high expression in LE neurons, (E) most neurons, (F) intestine (see **Supplemental Fig. S21** for SOM component planes and **Supplemental Fig. S22** for additional clusters).

Fig. 8. Selected regulatory elements discovered in stage and cell-type expression clusters.

FIRE analysis identifies motifs over- and under-represented in (A) developmental profile clusters and (B) cell-type profile clusters. Complete results are shown in **Supplemental Fig. S23**. A heat map indicates whether each motif is over-represented (yellow) or under-represented (blue) in each cluster. Motifs are arranged in rows and clusters in columns. Significant over-representation is indicated by red box outlines and under-representation is indicated by blue outlines ($p \leq 0.05$, Bonferroni-corrected). The optimized motif logo, location of the motif (5' upstream promoter or 3' UTR), mutual information with the genes in the cluster, and matching transcription factors and miRNAs listed in public data bases for indicated motifs are shown alongside the heat map.

Tables

Table 1. Samples used for expression profiling

Name	Stage*	Description	RNA
Cell types and tissues			
emb-0hr-ref	EE	all (EE) embryonic cells	total RNA
emb-GLP	EE	Germ-line precursors	total RNA
emb-BAG	EE	BAG neurons	total RNA
emb-reference	LE	all (LE) embryonic cells	total RNA
emb-panneural	LE	all neurons	total RNA
emb-AVA	LE	AVA neurons	total RNA
emb-AVE	LE	AVE neurons	total RNA
emb-A-class	LE	A-class motor neurons	total RNA
emb-bwm	LE	Body muscle	total RNA
emb-coelomocytes	LE	Coelomocytes	total RNA
emb-dop	LE	Dopaminergic neurons	total RNA
emb-GABA	LE	GABAergic motor neurons	total RNA
emb-hypodermis	LE	Hypodermal cells	total RNA
emb-intestine	LE	Intestine	total RNA
emb-PhM	LE	Pharyngeal muscle	total RNA
L2-reference	L2	all (L2) cells	poly A+/total RNA
L2-panneural	L2	all neurons	poly A+/total RNA
L2-A-class	L2	A-class motor neurons	poly A+/total RNA
L2-bwm	L2	Body muscle	poly A+/total RNA
L2-coelomocytes	L2	Coelomocytes	poly A+/total RNA
L2-excretory_cell	L2	Excretory cell	poly A+/total RNA
L2-GABA_neurons	L2	GABA neurons	poly A+/total RNA
L2-glr	L2	Glutamate receptor (*glr-1*+) neurons	poly A+/total RNA
L2-intestine	L2	Intestine	poly A+/total RNA
L3-L4-reference	L3-L4	all (L3-L4) cells	poly A+/total RNA
L3-L4-dop	L3-L4	Dopaminergic neurons	poly A+/total RNA
L3-L4-hypodermis	L3-L4	Hypodermis	poly A+/total RNA
L3-L4-PVD_OLL	L3-L4	PVD and OLL neurons	poly A+/total RNA
YA-ref	YA	all (YA) cells	poly A+/total RNA
YA-CEPsh	YA	CEP sheath cells	poly A+/total RNA
Whole Animal			
N2EE	EE	Early embryos	total RNA
N2LE	LE	Late embryos	total RNA
L1	L1	L1 animals	total RNA
L2	L2	L2 animals	total RNA
L3	L3	L3 animals	total RNA
L4	L4	L4 animals	total RNA
YA	YA	Young adult animals	total RNA
soma-only	L4	L4 hermaphrodite somatic cells only	total RNA
male	L4	L4 males	total RNA
Gonad	YA	Hermaphrodite gonad	total RNA

* EE (Early Embryo), LE (Late Embryo), L1 (L1 larva), L2 (L2 Larva), L3-L4 (L3-L4 larvae), L4 (L4 larva), YA (Young Adult)

(For more details see also **Supplemental Table S1**).

Table 2. Gene models detected as expressed above background and with differential expression between cell types and references or between developmental stages.

	Cell types	Dev. stages	Both data sets
Expressed genes (5% FDR)	17,075	15,822	17,452 87.7%
Average # expressed genes per data set (5% FDR)	12,228[a]	12,252	12,232 61.4%
Expressed genes (stringent FDR)	13,149	13,713	14,279 71.7%
Differentially expressed genes (5% FDR, FC ≥2)	10,598	9,552	13,320 66.9 %
Differentially expressed genes (stringent FDR, FC ≥2)	7,983	8,606	11,827 59.4%
Differentially expressed genes (5% FDR, FC ≥5)	1,596	1,981	3,218 16.2%
Differentially expressed genes (stringent FDR, FC ≥5)	1,586	1,974	3,206 16.1%
Differentially expressed genes (5% FDR, FC ≥10)	270	620	873 4.4%
Differentially expressed genes (stringent FDR, FC ≥10)	270	620	873 4.4%

[a] On average per cell-type, 5,698 genes with expression higher than in the reference sample were detected (see also **Supplemental Fig. S5**). This transformation is designed to remove transcripts that are highly expressed in other cell types but detected as background in a given cell specific sample (see Results).

REFERENCES

Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. 2010. Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res* **38**(3): 868-877.

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

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M. et al. 2009. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**(9): 647-649.

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

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

Baltus, A.E., Menke, D.B., Hu, Y.C., Goodheart, M.L., Carpenter, A.E., de Rooij, D.G., and Page, D.C. 2006. In germ cells of mouse embryonic ovaries, the decision to enter meiosis precedes premeiotic DNA replication. *Nat Genet* **38**(12): 1430-1434.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A. et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37**(Database issue): D885-890.

Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. 2003. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* **130**(5): 889-900.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**(1): 289-300.

Birney, E. Stamatoyannopoulos, J.A. Dutta, A. Guigo, R. Gingeras, T.R. Margulies, E.H. Weng, Z. Snyder, M. Dermitzakis, E.T. Thurman, R.E. et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blacque, O.E., Perens, E.A., Boroevich, K.A., Inglis, P.N., Li, C., Warner, A., Khattra, J., Holt, R.A., Ou, G., Mah, A.K. et al. 2005. Functional genomics of the cilium, a sensory organelle. *Curr Biol* **15**(10): 935-941.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* **19**(2): 185-193.

Bond, A.M., Vangompel, M.J., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., and Kohtz, J.D. 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. *Nat Neurosci* **12**(8): 1020-1027.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.-s., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E., and Chory, J. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome research* **13**(3): 513-523.

Bradnam, K.R. and Korf, I. 2008. Longer first introns are a general property of eukaryotic gene structure. *PLoS One* **3**(8): e3093.

Brenner, S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77**(1): 71-94.

Budovskaya, Y.V., Wu, K., Southworth, L.K., Jiang, M., Tedesco, P., Johnson, T.E., and Kim, S.K. 2008. An elt-3/elt-5/elt-6 GATA transcription circuit guides aging in *C. elegans*. *Cell* **134**(2): 291-303.

Calella, A.M., Nerlov, C., Lopez, R.G., Sciarretta, C., von Bohlen und Halbach, O., Bereshchenko, O., and Minichiello, L. 2007. Neurotrophin/Trk receptor signaling mediates C/EBPalpha, -beta and NeuroD recruitment to immediate-early gene promoters in neuronal cells and requires C/EBPs to induce immediate-early gene transcription. *Neural Dev* **2**: 4.

Chatzigeorgiou, M., Yoo, S., Watson, J.D., Lee, W.H., Spencer, W.C., Kindt, K.S., Hwang, S.W., Miller, D.M., 3rd, Treinin, M., Driscoll, M. et al. 2010. Specific roles for DEG/ENaC and

TRP channels in touch and thermosensation in *C. elegans* nociceptors. *Nat Neurosci* **13**(7): 861-868.

Chen, N., Pai, S., Zhao, Z., Mah, A., Newbury, R., Johnsen, R.C., Altun, Z., Moerman, D.G., Baillie, D.L., and Stein, L.D. 2005. Identification of a nematode chemosensory gene family. *Proc Natl Acad Sci U S A* **102**(1): 146-151.

Chi, W. and Reinke, V. 2006. Promotion of oogenesis and embryogenesis in the *C. elegans* gonad by EFL-1/DPL-1 (E2F) does not require LIN-35 (pRB). *Development* **133**(16): 3147-3157.

Chikina, M.D., Huttenhower, C., Murphy, C.T., and Troyanskaya, O.G. 2009. Global prediction of tissue-specific gene expression and context-dependent gene networks in *Caenorhabditis elegans*. *PLoS Comput Biol* **5**(6): e1000417.

Christensen, M., Estevez, A., Yin, X., Fox, R., Morrison, R., McDonnell, M., Gleason, C., Miller, D.M., 3rd, and Strange, K. 2002. A primary culture system for functional analysis of *C. elegans* neurons and muscle cells. *Neuron* **33**(4): 503-514.

Cinar, H., Keles, S., and Jin, Y. 2005. Expression profiling of GABAergic motor neurons in *Caenorhabditis elegans*. *Curr Biol* **15**(4): 340-346.

Colosimo, M.E., Brown, A., Mukhopadhyay, S., Gabel, C., Lanjuin, A.E., Samuel, A.D., and Sengupta, P. 2004. Identification of thermosensory and olfactory neuron-specific genes via expression profiling of single neuron types. *Curr Biol* **14**(24): 2245-2251.

del Castillo-Olivares, A., Kulkarni, M., and Smith, H.E. 2009. Regulation of sperm gene expression by the GATA factor ELT-1. *Dev Biol* **333**(2): 397-408.

Deshpande, G., Calhoun, G., Yanowitz, J.L., and Schedl, P.D. 1999. Novel functions of nanos in downregulating mitosis and transcription during the development of the Drosophila germline. *Cell* **99**(3): 271-281.

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

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

Fay, D.S., Stanley, H.M., Han, M., and Wood, W.B. 1999. A *Caenorhabditis elegans* homologue of hunchback is required for late stages of development but not early embryonic patterning. *Dev Biol* **205**(2): 240-253.

Flames, N. and Hobert, O. 2009. Gene regulatory logic of dopamine neuron differentiation. *Nature* **458**(7240): 885-889.

Fox, R.M., Von Stetina, S.E., Barlow, S.J., Shaffer, C., Olszewski, K.L., Moore, J.H., Dupuy, D., Vidal, M., and Miller, D.M., 3rd. 2005. A gene expression fingerprint of *C. elegans* embryonic motor neurons. *BMC Genomics* **6**(1): 42.

Fox, R.M., Watson, J.D., Von Stetina, S.E., McDermott, J., Brodigan, T.M., Fukushige, T., Krause, M., and Miller, D.M., 3rd. 2007. The embryonic muscle transcriptome of *Caenorhabditis elegans*. *Genome Biol* **8**(9): R188.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.a. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)* **20**(3): 307-315.

Goodyer, W., Kaitna, S., Couteau, F., Ward, J.D., Boulton, S.J., and Zetka, M. 2008. HTP-3 links DSB formation with homolog pairing and crossing over during *C. elegans* meiosis. *Dev Cell* **14**(2): 263-274.

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

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**(7308): 835-840.

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

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., and Waterston, R.H. 2005. Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res* **15**(12): 1651-1660.

Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H. 2009. Massively parallel sequencing of the poly-adenylated transcriptome of *C. elegans*. *Genome Res*.

Hirota, K., Miyoshi, T., Kugou, K., Hoffman, C.S., Shibata, T., and Ohta, K. 2008. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**(7218): 130-134.

Hubbard, E.J.A., and Greenstein, D. 2005. Introduction to the germ line. In *WormBook*, (ed. T.*C.e.*R. Community).

Huber, W., Toedling, J., and Steinmetz, L.M. 2006. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics (Oxford, England)* **22**(16): 1963-1970.

Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A. et al. 2007. High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol* **5**(9): e237.

Hwang, S.B. and Lee, J. 2003. Neuron cell type-specific SNAP-25 expression driven by multiple regulatory elements in the nematode *Caenorhabditis elegans*. *J Mol Biol* **333**(2): 237-247.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**(4): e15.

Isik, M., Korswagen, H.C., and Berezikov, E. 2010. Expression patterns of intronic microRNAs in *Caenorhabditis elegans*. *Silence* **1**(1): 5.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S.K. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **98**(1): 218-223.

Kato, M., de Lencastre, A., Pincus, Z., and Slack, F.J. 2009. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol* **10**(5): R54.

Kawasaki, I., Amiri, A., Fan, Y., Meyer, N., Dunkelbarger, S., Motohashi, T., Karashima, T., Bossinger, O., and Strome, S. 2004. The PGL family proteins associate with germ granules and function redundantly in *Caenorhabditis elegans* germline development. *Genetics* **167**(2): 645-661.

Kawasaki, I., Shim, Y.H., Kirchner, J., Kaminker, J., Wood, W.B., and Strome, S. 1998. PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell* **94**(5): 635-645.

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**(5537): 2087-2092.

Kimble, J. 1981. Alterations in cell lineage following laser ablation of cells in the somatic gonad of *Caenorhabditis elegans*. *Dev Biol* **87**: 286-300.

Kimble, J.a.C., S.L. 2005. Germline proliferation and its control. In *WormBook*, (ed. T.C.e.R. Community).

Koh, K. and Rothman, J.H. 2001. ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in *C. elegans*. *Development* **128**(15): 2867-2880.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**(1): 59-69.

Krause, M., Park, M., Zhang, J.M., Yuan, J., Harfe, B., Xu, S.Q., Greenwald, I., Cole, M., Paterson, B., and Fire, A. 1997. A *C. elegans* E/Daughterless bHLH protein marks neuronal but not striated muscle development. *Development* **124**(11): 2179-2189.

Laubinger, S., Zeller, G., Henz, S.R., Sachsenberg, T., Widmer, C.K., Naouar, N., Vuylsteke, M., Scholkopf, B., Rätsch, G., and Weigel, D. 2008. At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biol* **9**(7): R112.

Li, X., Panea, C., Wiggins, C.H., Reinke, V., and Leslie, C. 2010. Learning "graph-mer" motifs that predict gene expression trajectories in development. *PLoS Comput Biol* **6**(4): e1000761.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17**(8): 991-1008.

Liu, X., Long, F., Peng, H., Aerni, S.J., Jiang, M., Sanchez-Blanco, A., Murray, J.I., Preston, E., Mericle, B., Batzoglou, S. et al. 2009. Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell* **139**(3): 623-633.

Martinez, N.J., Ow, M.C., Reece-Hoyes, J.S., Barrasa, M.I., Ambros, V.R., and Walhout, A.J. 2008. Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res* **18**(12): 2005-2015.

McGhee, J.D., Fukushige, T., Krause, M.W., Minnema, S.E., Goszczynski, B., Gaudet, J., Kohara, Y., Bossinger, O., Zhao, Y., Khattra, J. et al. 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine, from embryo to adult. *Dev Biol* **327**(2): 551-565.

McGhee, J.D., Sleumer, M.C., Bilenky, M., Wong, K., McKay, S.J., Goszczynski, B., Tian, H., Krich, N.D., Khattra, J., Holt, R.A. et al. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Dev Biol* **302**(2): 627-645.

McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E. et al. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb Symp Quant Biol* **68**: 159-169.

Meissner, B., Warner, A., Wong, K., Dube, N., Lorch, A., McKay, S.J., Khattra, J., Rogalski, T., Somasiri, A., Chaudhry, I. et al. 2009. An integrated strategy to study muscle development and myofilament structure in *Caenorhabditis elegans*. *PLoS Genet* **5**(6): e1000537.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**(2): 716-721.

Mukai, M., Kitadate, Y., Arita, K., Shigenobu, S., and Kobayashi, S. 2006. Expression of meiotic genes in the germline progenitors of Drosophila embryos. *Gene Expr Patterns* **6**(3): 256-266.

Murray, J.I., Bao, Z., Boyle, T.J., Boeck, M.E., Mericle, B.L., Nicholas, T.J., Zhao, Z., Sandel, M.J., and Waterston, R.H. 2008. Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat Methods* **5**(8): 703-709.

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**: 385-404.

Ow, M.C., Martinez, N.J., Olsen, P.H., Silverman, H.S., Barrasa, M.I., Conradt, B., Walhout, A.J., and Ambros, V. 2008. The FLYWCH transcription factors FLH-1, FLH-2, and FLH-3 repress embryonic expression of microRNA genes in *C. elegans*. *Genes Dev* **22**(18): 2520-2534.

Pasierbek, P., Jantsch, M., Melcher, M., Schleiffer, A., Schweizer, D., and Loidl, J. 2001. A *Caenorhabditis elegans* cohesion protein with functions in meiotic chromosome pairing and disjunction. *Genes Dev* **15**(11): 1349-1360.

Pauli, F., Liu, Y., Kim, Y.A., Chen, P.J., and Kim, S.K. 2005. Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in *C. elegans*. *Development*.

Reinke, V., Gil, I.S., Ward, S., and Kazmer, K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131**(2): 311-323.

Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J., Davis, E.B., Scherer, S., Ward, S. et al. 2000. A global profile of germline gene expression in *C. elegans*. *Mol Cell* **6**(3): 605-616.

Reynolds, N.K., Schade, M.A., and Miller, K.G. 2005. Convergent, RIC-8-dependent Galpha signaling pathways in the *Caenorhabditis elegans* synaptic signaling network. *Genetics* **169**(2): 651-670.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E. et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7): 1311-1323.

Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W.J., Davis, P., Fernandes, J. et al. 2008. WormBase 2007. *Nucleic acids research* **36**(Database issue): D612-617.

Rousseeuw, P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**(1): 53-65.

Roy, P.J., Stuart, J.M., Lund, J., and Kim, S.K. 2002. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418**(6901): 975-979.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**(6): 1193-1207.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**(Database issue): D91-94.

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

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

Shakes, D.C., Wu, J.C., Sadler, P.L., Laprade, K., Moore, L.L., Noritake, A., and Chu, D.S. 2009. Spermatogenesis-specific features of the meiotic program in *Caenorhabditis elegans*. *PLoS Genet* **5**(8): e1000611.

Smith, C.J., Watson, J.D., Spencer, W.C., O'Brien, T., Cha, B., Albeg, A., Treinin, M., and Miller, D.M., 3rd. 2010. Time-lapse imaging and cell-specific expression profiling reveal dynamic branching and molecular determinants of a multi-dendritic nociceptor in *C. elegans*. *Dev Biol* **345**(1): 18-33.

Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**(1): Article3.

Strome, S. and Wood, W.B. 1983. Generation of asymmetry and segregation of germ-line granules in early *C. elegans* embryos. *Cell* **35**(1): 15-25.

Subramaniam, K. and Seydoux, G. 1999. nos-1 and nos-2, two genes related to Drosophila nanos, regulate primordial germ cell development and survival in *Caenorhabditis elegans*. *Development* **126**(21): 4861-4871.

Sulston, J.E. and Horvitz, H.R. 1977. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev Biol* **56**(1): 110-156.

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**(1): 64-119.

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

Touroutine, D., Fox, R.M., Von Stetina, S.E., Burdina, A., Miller, D.M., 3rd, and Richmond, J.E. 2005. acr-16 encodes an essential subunit of the levamisole-resistant nicotinic receptor at the *Caenorhabditis elegans* neuromuscular junction. *J Biol Chem* **280**(29): 27013-27021.

Troemel, E.R., Chou, J.H., Dwyer, N.D., Colbert, H.A., and Bargmann, C.I. 1995. Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. *Cell* **83**(2): 207-218.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**(5992): 689-693.

van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**(5): e1000371.

Von Stetina, S.E., Watson, J.D., Fox, R.M., Olszewski, K.L., Spencer, W.C., Roy, P.J., and Miller, D.M., 3rd. 2007. Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the *C. elegans* nervous system. *Genome Biol* **8**(7): R135.

Wang, X., Zhao, Y., Wong, K., Ehlers, P., Kohara, Y., Jones, S.J., Marra, M.A., Holt, R.A., Moerman, D.G., and Hansen, D. 2009. Identification of genes expressed in the hermaphrodite germ line of *C. elegans* using SAGE. *BMC Genomics* **10**: 213.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23**(13): 1494-1504.

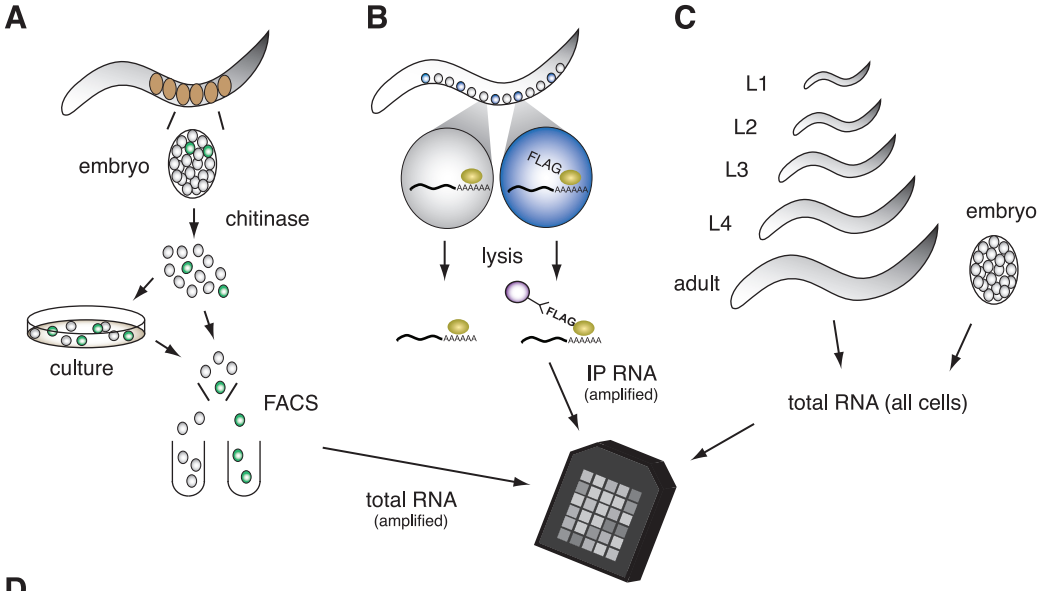
Yang, Y., Xu, S., Xia, L., Wang, J., Wen, S., Jin, P., and Chen, D. 2009. The bantam microRNA is associated with drosophila fragile X mental retardation protein and regulates the fate of germline stem cells. *PLoS Genet* **5**(4): e1000444.

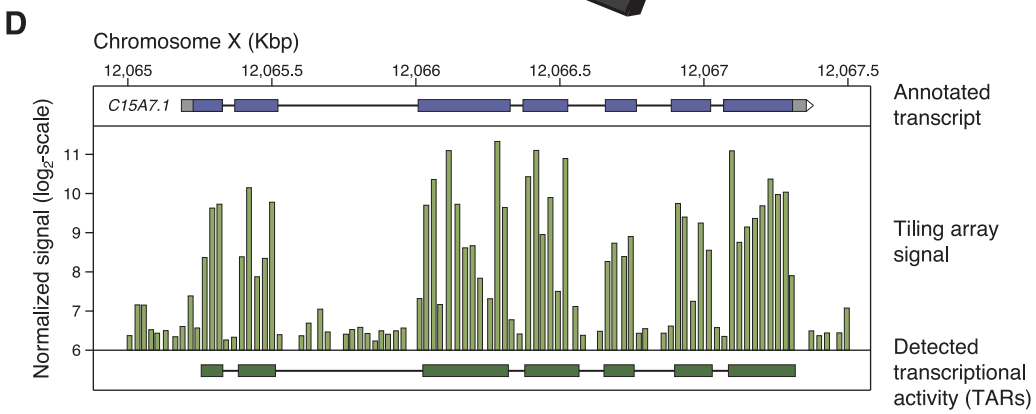
You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., and Anderson, O.D. 2008. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**: 253.

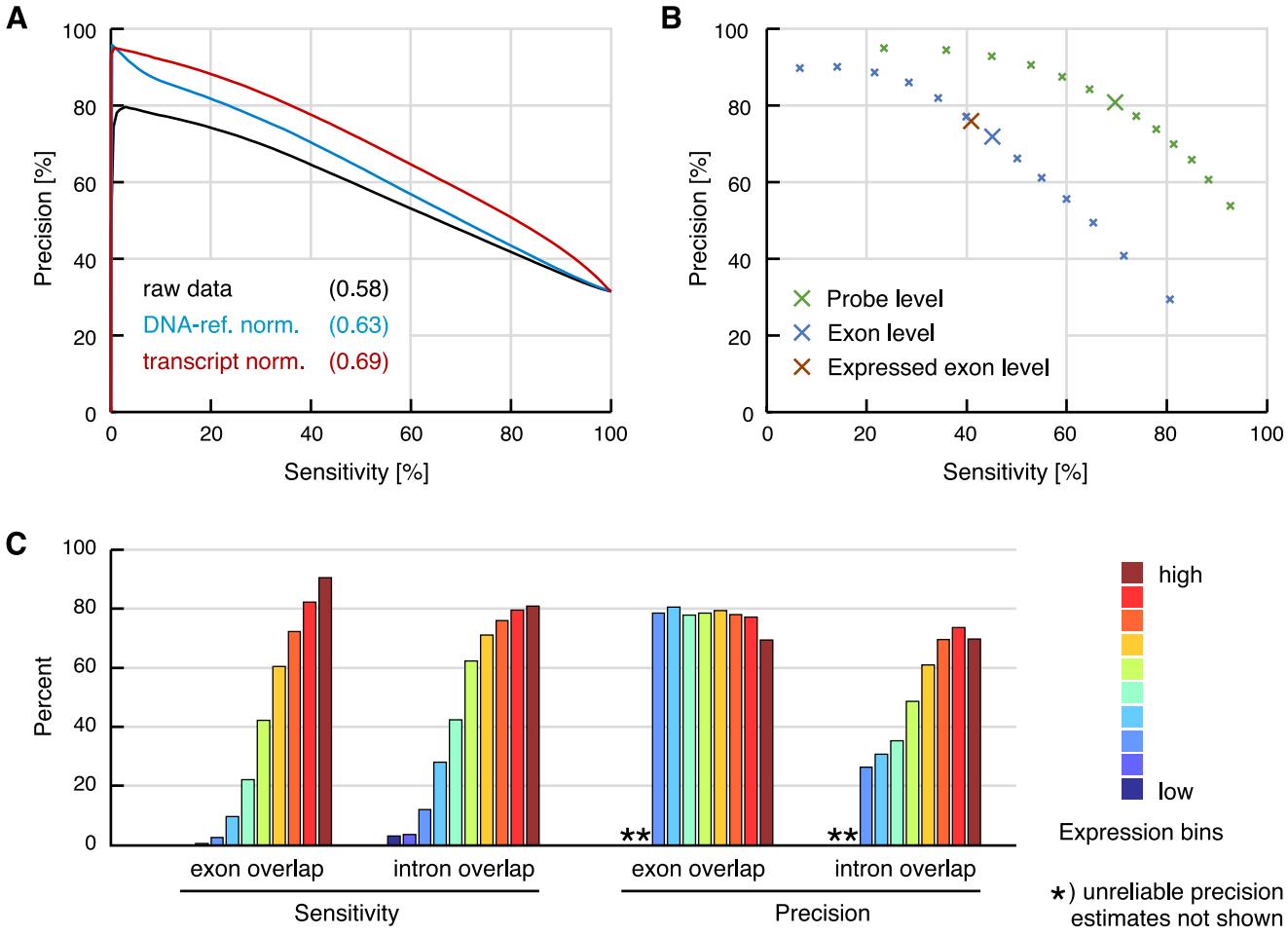
Zeller, G., Henz, S.R., Laubinger, S., Weigel, D., and Rätsch, G. 2008. Transcript normalization and segmentation of tiling array data. *Pac Symp Biocomput*: 527-538.

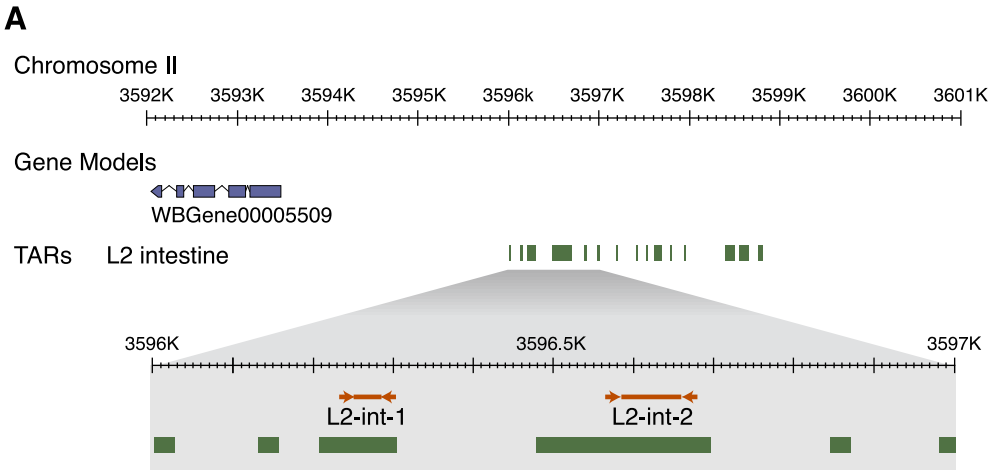
Zeller, G., Henz, S.R., Widmer, C.K., Sachsenberg, T., Rätsch, G., Weigel, D., and Laubinger, S. 2009. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. *Plant J*.

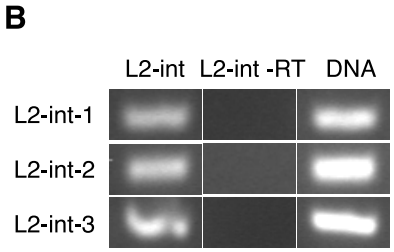
Zhang, Y., Ma, C., Delohery, T., Nasipak, B., Foat, B.C., Bounoutas, A., Bussemaker, H.J., Kim, S.K., and Chalfie, M. 2002. Identification of genes expressed in *C. elegans* touch receptor neurons. *Nature* **418**(6895): 331-335.

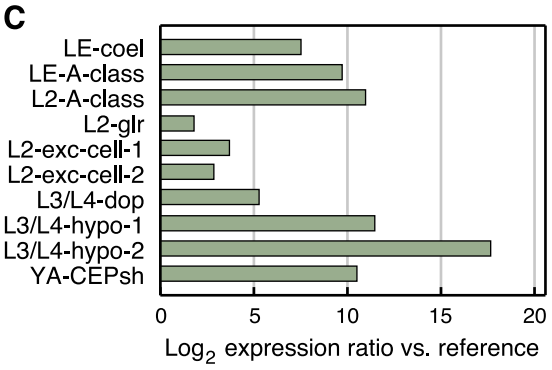


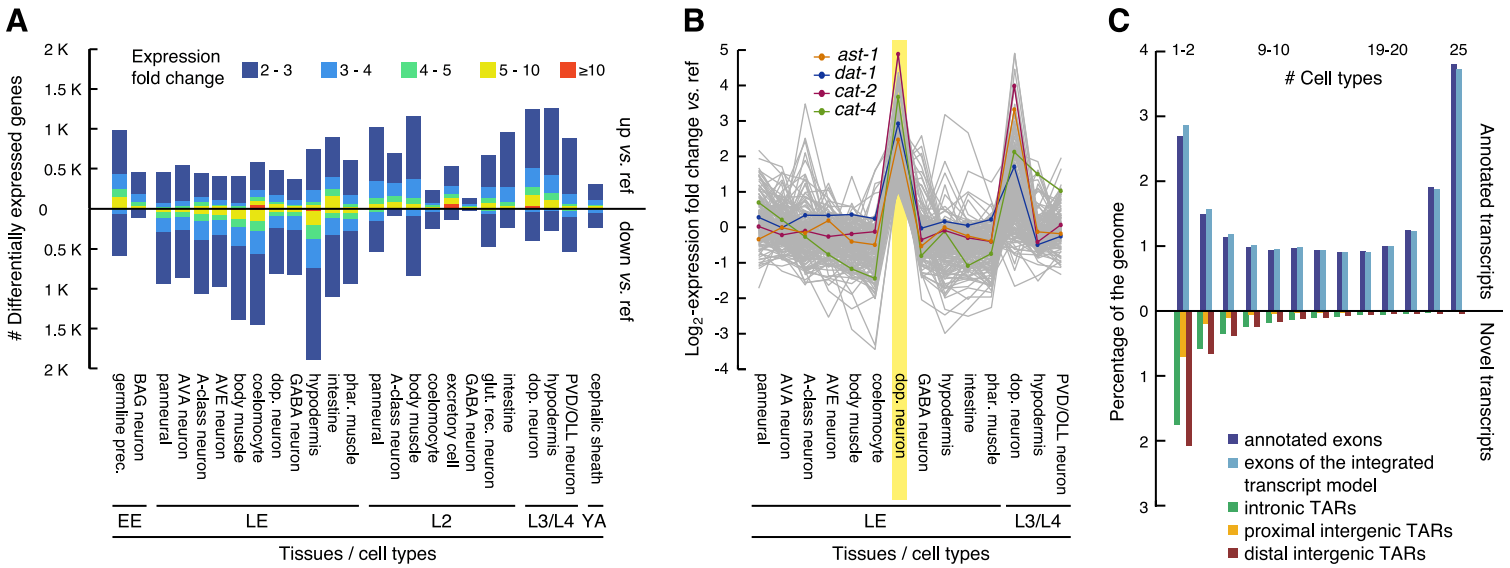


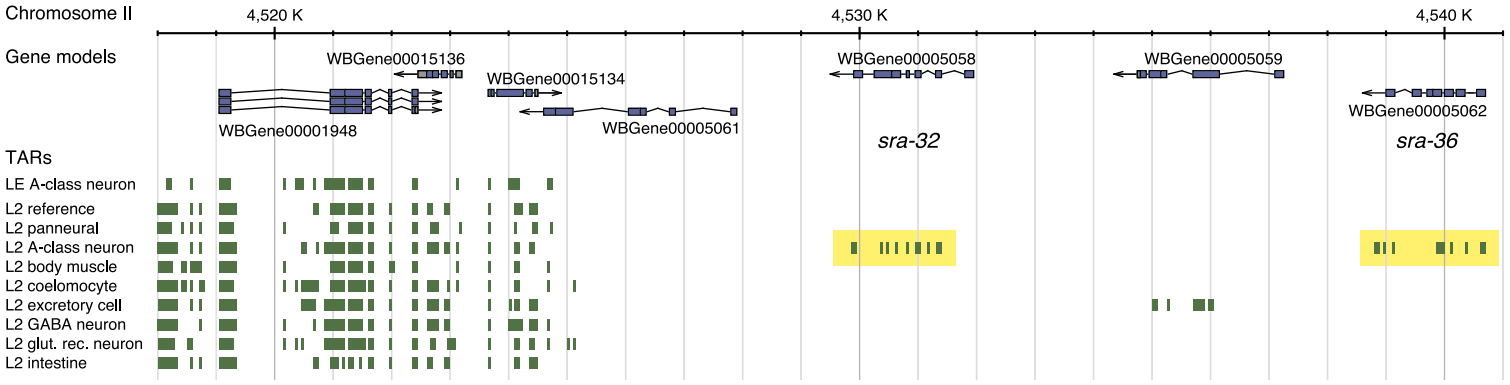


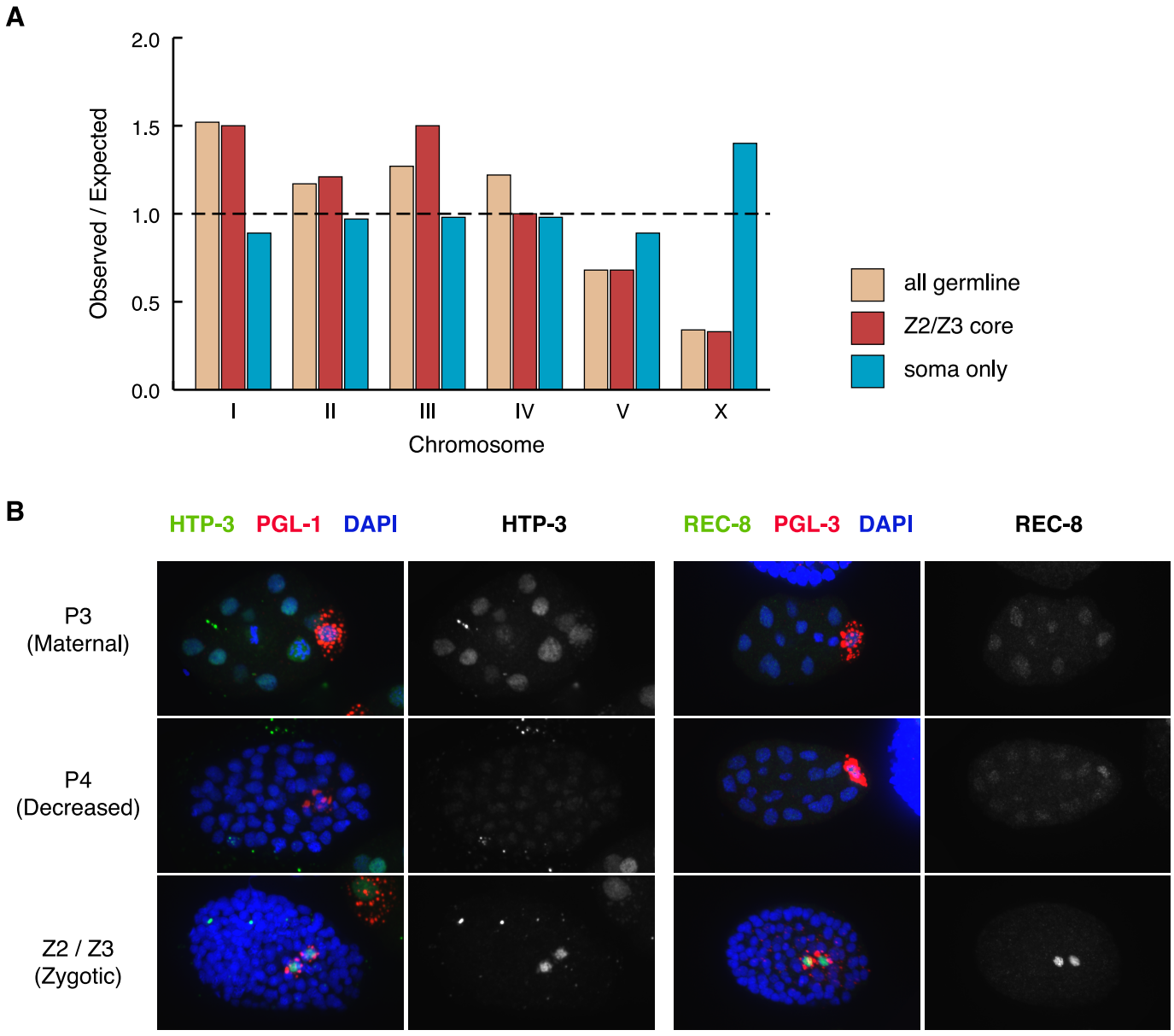


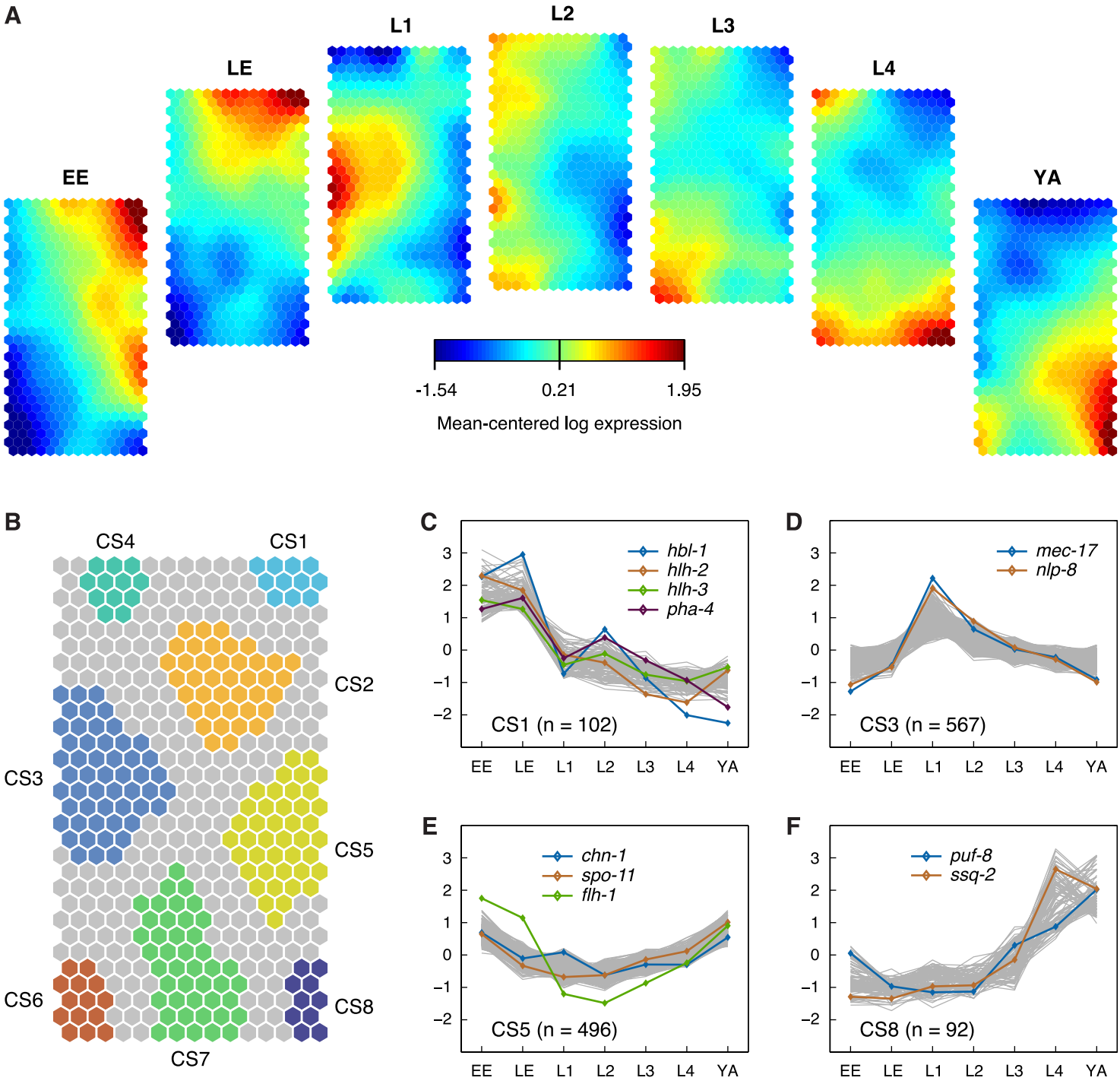


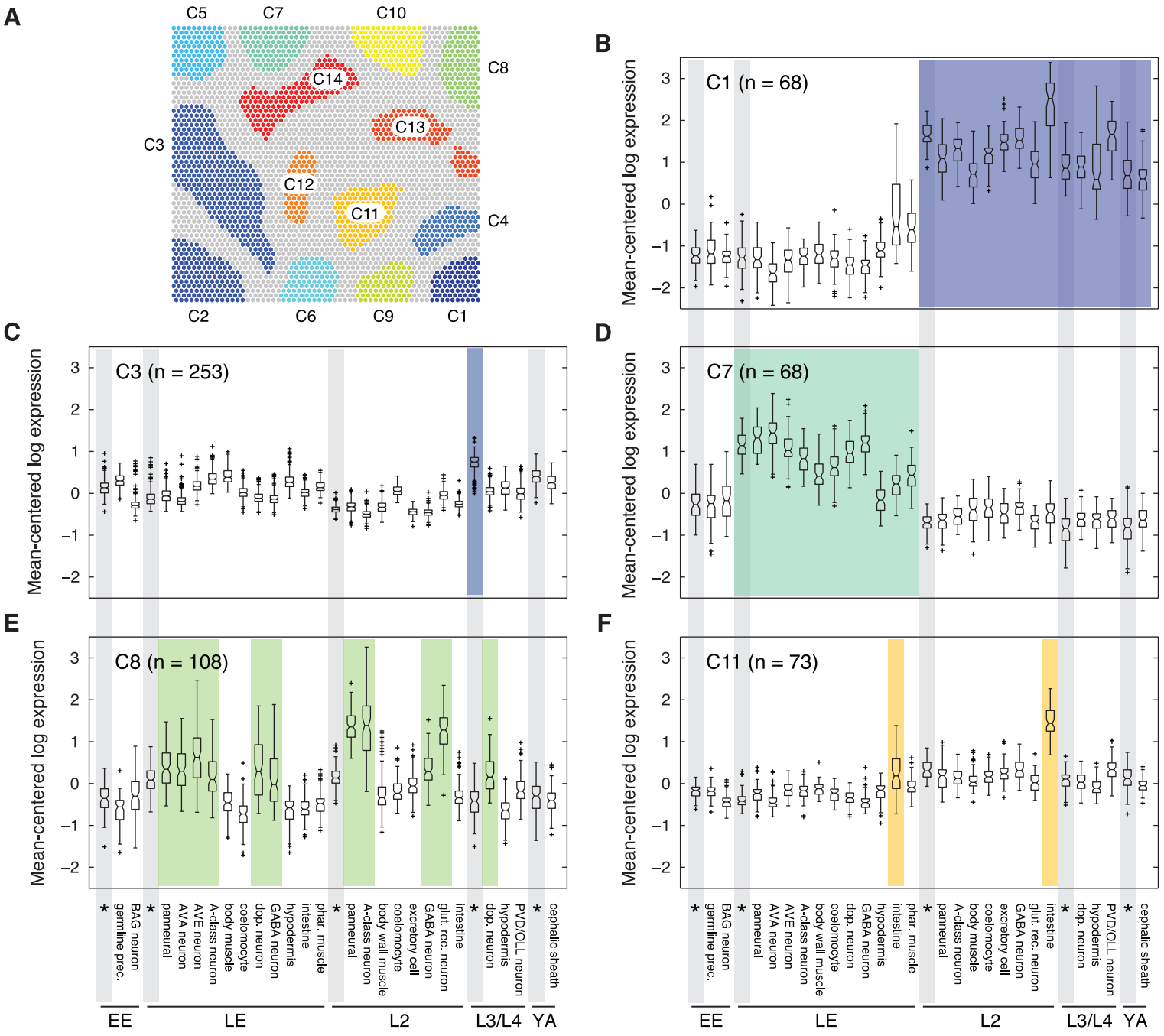




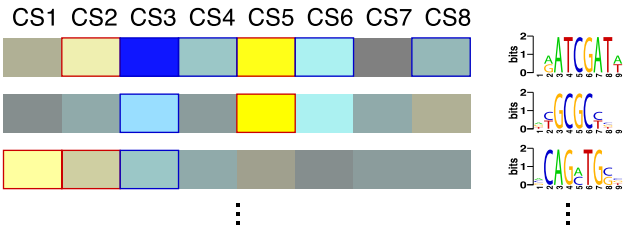






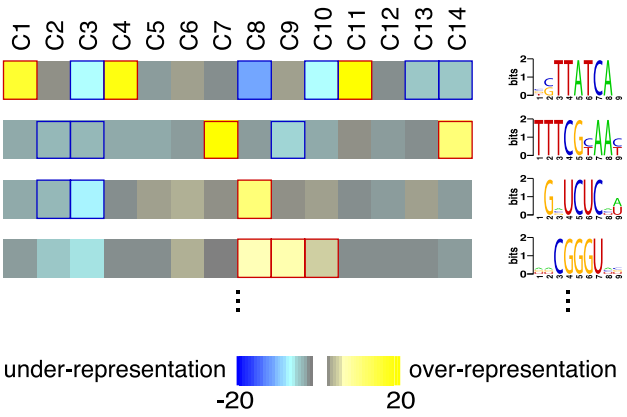


A



opt. motif	location	mut. inf.	motif name
	5'	0.036	Homez, GATA2, CDP_CR3+HD ...
	5'	0.035	RSC3, E2F3_1, E2F2_1
	5'	0.010	E47, MyoD, Sn ...
⋮	⋮	⋮	⋮

B



opt. motif	location	mut. inf.	motif name
	5'	0.046	Gata3_1, GATA2, GATA-3 ...
	5'	0.033	C_EBPalpha, C_EBPbeta, Cebpa
	3'UTR	0.012	miR-47*, miR-46*, miR-80 ...
	3'UTR	0.011	miR-54, miR-56, miR-55 ...
⋮	⋮	⋮	⋮