Supplementary Text for "Integrative Analysis of Functional Elements in the *Caenorhabditis elegans* Genome by the modENCODE Project"

A. More Details on Author Roles

The authors who were involved with data generation:
Julie Ahringer, Cathleen M. Brdlik, Jennifer Brennan, Ming-Sin Cheung, Luke O. Dannenberg, Abby F. Dernburg, Arshad Desai, Lindsay Dick, Andréa C. Dosé, Thea Egelhofer, Sevinc Ercan, Ghia Euskirchen, Brent Ewing, Reto Gassman, Ting Han, Steven Henikoff, LaDeana W. Hillier, Heather Holster, Tony Hyman, David M. Miller III, Kohta Ikegami, A. Leo Iniguez, Judith Janette, Morten Jensen, Masaomi Kato, Vishal Khivansara, John K. Kim, Stuart K. Kim, Paulina Kolasinska-Zwierz, Mitzi I. Kuroda, Isabel Latorre, Amber Leahey, Jason D. Lieb, Michael MacCoss, Marco Mangone, Gennifer Merrihew, Andrew Muroyama, John I. Murray, Wei Niu, Hoang Pham, Taryn Phippen, Fabio Piano, Elicia A. Preston, Valerie Reinke, Heidi Rosenbaum, Mihail Sarov, Frank J. Slack, Cindie Slightam, Michael Snyder, William C. Spencer, Susan Strome, Teruaki Takasaki, Dionne Vafeados, Anne Vielle, Ksenia Voronina, Guilin Wang, Robert H. Waterston, Christina Whittle, Beijing Wu, Mei Zhong, Xingliang Zhou

The authors who were involved with data analysis:
Ashish Agarwal, Roger P. Alexander, Pedro Alves, Bradley I. Arshinoff, Raymond K. Auerbach, Galt Barber, Adrian Carr, Aurelien Chateigner, Chao Cheng, Hiram Clawson, Sergio Contrino, Jiang Du, Xin Feng, Mark B. Gerstein, Phil Green, Francois Gullier, Kristin C. Gunsalus, Lukas Habegger, Jorja G. Henikoff, Stefan R. Henz, LaDeana W. Hillier, Angie Hinrichs, W. James Kent, Ellen Kephart, Ekta Khurana, Stuart K. Kim, Jing Leng, Suzanna Lewis, Tao Liu, X. Shirley Liu, Paul Lloyd, Lucas Lochovsky, Yaniv Lubling, Zhi John Lu, Rachel Lyne, Sebastian D. Mackowiak, Sheldon McKay, Gos Micklem, Mitzi Morris, Eric L. Van Nostrand, Siew-Loon Ooi, Marc Perry, Nikolaus Rajewsky, Gunnar Ratsch, Andreas Rechtsteiner, Kahn Rhrissorrakrai, Rebecca Robilotto, Joel Rozowsky, Kim Rutherford, Peter Ruzanov, Rajkumar Sasidharan, Andrea Sboner, Eran Segal, Hyunjin Shin, Richard Smith, Lincoln Stein, E.O. Stinson, Scott Taing, Nicole L. Washington, Koon-Kiu Yan, Kevin Yip, Georg Zeller, Zheng Zha

The authors who were lab project coordinators (PIs and co-PIs):
Julie Ahringer, Abby F. Dernburg, Arshad Desai, Mark B. Gerstein, Phil Green, Kristin C. Gunsalus, Steven Henikoff, LaDeana W. Hillier, Tony Hyman, David M. Miller III, A. Leo

Iniguez, W. James Kent, John K. Kim, Stuart K. Kim, Mitzi I. Kuroda, Eric C. Lai, Suzanna Lewis, Jason D. Lieb, X. Shirley Liu, Michael MacCoss, Gos Micklem, Fabio Piano, Nikolaus Rajewsky, Valerie Reinke, Eran Segal, Frank J. Slack, Michael Snyder, Lincoln Stein, Susan Strome, Robert H. Waterston

The authors who were involved in overall project management:
Elise A. Feingold, Peter J. Good, Mark S. Guyer, Rebecca F. Lowdon

B. More Details on the Data Overview

B.1. Comparing and Scaling of Array and Sequencing Data: ChIP-chip vs. ChIP-seq

The modENCODE project began when tiling arrays were still the platform of choice for genome-wide location analysis. Many genome-wide location data sets, especially on histone marks and chromatin factors, were obtained using ChIP-chip on tiling arrays. To ensure the compatibility between ChIP-chip and ChIP-seq data generated by different modENCODE groups, we examined RNA Pol II ChIP data detected by both ChIP-chip (from the Lieb group) and ChIP-seq (from the Snyder group) (Fig. S1).

At 1 kb resolution, the correlation between indvidual Pol II profiles at a given worm stage is 0.75-0.88 within ChIP-seq replicates and 0.77-0.91 within ChIP-chip replicates. The correlation between ChIP-seq and ChIP-chip replicates are 0.56-0.78. Although variations across platform/group are slightly higher than those within platform/group, data across different labs at the same stage are still more correlated than those across different stages by the same lab.

Finally, while ChIP-seq yielded more peaks than did ChIP-chip, the top 3,000 peaks identified by ChIP-chip and ChIP-seq overlap by approximately 2/3, a level of agreement normally observed for ChIP data from different labs on the same platform. These observations not only indicate that the two platforms are comparable, but also attest to the high quality of the respective data sets.

B.2. Tiling Arrays vs. RNA-seq

We also had an opportunity to compare tiling array and RNA-seq technologies for measurement of gene expression, as data sets were generated using both techniques on matched samples. A detailed comparison of these methods was described in (*1*); in addition to presenting some main points from this analysis here, we also repeat this analysis on new unpublished data sets associated with this manuscript. From this comparison, we were able to develop methods of optimally scaling the tiling array measurements to make them best correspond to those from RNA-seq.

Signals from the two platforms agree well (Fig. S2). For a young adult sample, the Pearson correlation is 0.83 between RNA-seq measurements using polyA-selected RNA and tiling array measurements using total RNA. A higher correlation of 0.90 was found when polyA-enrichment was also used for the sample which had been hybridized on tiling arrays. Using the maxgap-minrun algorithm with optimized parameters, we then segmented the signals into transcriptionally active regions (*2, 3*). A ROC curve, parameterized by signal threshold, indicates that RNA-seq consistently outperforms tiling array in its ability to predict known transcribed regions. For instance, at a false positive rate (FPR) of 0.05, the tiling array yields a sensitivity of 0.68, while RNA-seq attains a sensitivity of 0.85. Correspondingly, we also found that the RNA-seq data predicted exon boundaries with greater accuracy, with a median offset of 0 bp (in comparison to 7 bp for the tiling array data). This is to be expected, as the resolution of an array is limited by its probe size, which was 25 bp in this experiment.

Fig. S2 shows several genes in the upper left, indicating they are measured as highly expressed by tiling array but not RNA-seq. We conducted a "nearest neighbor" analysis to investigate the hypothesis that this is due to cross-hybridization effects on the array. For each gene, we computed the expression level from probes lying within that gene, as well as probes similar in sequence, but elsewhere in the genome. For tiling arrays, we found these two values to be similar for many genes, indicating that the suggested expression could arise equally well from true expression or cross-hybridization. These values are similar for fewer genes when using RNA-seq data. Another analysis, using pseudogenes, also confirms cross-hybridization in arrays (*1*). We have used these analyses in formulating our fairly conservative criteria for transcribed pseudogenes (see main text and Fig. 4C).

For determining gene expression values maximally compatible with RNA-seq, we used the following procedure: for 42 of the 44 experiments listed in Fig. 1 (without some of the infection samples), we obtained a signal track by applying pseudomedian smoothing over the three replicates, which provides an expression level for each probe. We then consider all probes overlapping, by at least 50%, the exonic regions of each transcript. We defined the expression level of this transcript as the median of the signal values for all such probes. Gene expression levels were then defined simply as equal to those of the longest isoform. For the inter-sample comparison, we normalized these expression levels by dividing the values by the slide median, i.e. the median of all probes on the array and obtained a large data matrix (42 samples x 20,085 genes). Expression levels for each slide were next centered by subtracting the mean expression value for each slide from all expression values within the slide.

C. More Details on Transcriptome Analysis

C.1. RNA-seq Saturation Analysis

In order to understand the relationship between the robustness of gene expression measurements and the depth of sequencing, we devised the following *in silico* "experiments":
1. We considered an RNA-seq experiment with ~36M mapped reads (mid-L2 25dC 14 hours post L1 - DCCid=2351);
2. We selected fractions of the mapped reads: 1%, 5%, 10%, …, 90%, such that we generated sub-sets of ~300K, 1.6M, 3.3M, …, 30M mapped reads;
3. We computed the expression levels for all 20,051 genes in WormBase190 as reads per kilobase of exonic region per million mapped reads (RPKM). As a gene model, we used the "composite", i.e. the union of exonic nucleotides of all isoforms of a gene.

Fig. S4 reports the density plots at the different sequencing depths. As expected, the low-coverage case shows a higher fraction of non-expressed genes. Interestingly, genes which have a log2(RPKM+1) greater than 2 seem to be less affected by sequencing depth. Fig. S5 reports the comparison between the density plots at different levels of coverage, suggesting that, with a sequencing depth of ~13M mapped reads, most of the expressed genes are captured. This hypothesis is also supported by Fig. S6, which reports the number of non-expressed genes (RPKM=0) as a function of sequencing depth. Indeed, after ~13M mapped reads the number of "genes" with zero expression begins to plateau, although there remain small numbers of lowly-expressed transcripts that can only be identified by further increases in depth.

C.2. RNA-seq Read Mapping and Stage-Specific RNA-seq-Only Genelet Creation Methods

Stage-specific genelets, based solely on stage-specific RNA-seq data were created using methods similar to those described (*4*), but with several additional refinements. Briefly, the Illumina reads were aligned against the genome, and an exhaustive coverage-based splice leader and splice junction database were created for each stage (*4*). Each read was assigned a unique genomic location. Thresholds for read coverage were set for a 0.05 false positive rate, based on a ROC analysis. Transcripts were created by seeding with the highest confidence splice sites and splice leaders in a region, and then extending from those sites and leaders, incorporating coverage and junctions into the model (Fig. 2B). The procedure was iterated until all confirmed splice junctions and leaders were incorporated into models. Instead of producing transcripts containing every possible combination of every splice junction/leader, each splice junction/leader was used in at least one model. We created alternative models, with merged neighboring exons, when above-threshold read coverage suggested the intron had been retained, and when frame was maintained across the merged region. We also generated genelets with alternative start/stop sites

within introns when the entire intron was not retained, but when there were at least 50 bases of above-threshold coverage which extended into the intron initiated by a TSS or terminated by a polyA site.

The stage-specific polyA addition sites (including those generated by this project, as well as those from (*5*)) were clustered (keeping only a single polyA addition site when there are multiple polyA sites within a 10 bp window). While all splice leaders were incorporated into at least one prediction, polyA sites were only incorporated when a genelet model extended to the polyA site. Because overlapping UTRs can cause neighboring same strand predictions to merge if there is no splice leader or no polyA site, whenever a single exon overlapped two separate neighboring WormBase gene predictions, we broke the corresponding transcript into two separate transcripts. We also broke transcripts whenever they overlapped more than one WormBase gene prediction, and three or more neighboring exons were not included in the CDS portion of the transcript. The CDS region was defined by identifying the longest open reading frame. Single exon transcripts from WormBase were incorporated if at least 75 bases had above-threshold coverage. Additionally, single exon transcripts were created when a single block of coverage was at least 75 bases long and extended from an SL to a polyA site, or if it began with an SL and extended at least 250 bases (even if without a polyA site).

C.3. Aggregate Integrated Transcript Set Methods

To create the aggregate integrated transcript set, all of the reads (from all stages) were combined as if they were from a "single project". Splice junctions, splice leaders, and polyA addition sites were identified as they would be in the stage-specific methods. Transcripts were then built in the way described above, seeding with splice junctions and extending using "experimentally confirmed" bases (see below). However, additional evidence from mRNAs/ESTs, WormBase, and modENCODE data was incorporated as described here.

The following splice junctions were included in the aggregate integrated set: (1) splice junctions confirmed in the individual RNA-seq stages or by aggregate read coverage, (2) all mRNA/EST (WormBase209), RT-PCR/RACE(*6*), and mass-spec-confirmed splice junctions (*7*), and (3) WormBase-predicted splice junctions which were supported by RNA-seq data (including those after allowing an RNA-seq read to be placed in all positions at which it had an identical match). Note that, for the splice junction counts in Fig. 2A, we counted any splice junction beginning “before” the 5’ end of an existing WormBase (WS170) transcript prediction as 5’. Similarly, any splice junction extending “past” the 3’ end of an existing WormBase transcript prediction was annotated as 3’. Any splice junction internal to a WormBase transcript prediction was labeled as internal.

In the aggregate transcript set, a base was considered experimentally confirmed when any one of the following criteria were met: (a) above-threshold coverage in the individual stage or aggregate

RNA-seq data set, (b) coverage by an mRNA/EST, RT-PCR/RACE, or mass-spec alignment, (c) coverage by WormBase predictions, as long as the bounding splice junctions are confirmed splice junctions (i.e. holes in coverage within exons which already have evidence based on RT-PCR, RNA-seq, EST/mRNA, etc. can be "filled in" using WormBase coverage), or (d) coverage by a genelet created in the individual stage-specific sets. In addition to the "integrated transcript set," we also created an "integrated genelet set" where evidence "(c)", filling with WormBase predictions, is not included.

For the aggregate set, splice leader and polyA addition site data were included when (a) they were defined by coverage in individual stages (novel splice leaders or polyA sites defined by the RNA-seq-only analyses were required to appear in more than one of the individual stages to be included) and/or in the aggregate set, (b) they were identified by WormBase (WS209) as SLs or polyAs, (c) they were identified in other studies generated from deep 3' RACE sequencing (*5, 8*), or (d) splice leaders were identified by RT-PCR/RACE experiments (*6*).

Transcripts are named after the overlapping WormBase transcript. For instance, the alternative transcripts/isoforms associated with WormBase C10H11.1 would have names such as C10H11.1.T1, C10H11.1.T2, C10H11.1.T3, etc. Those transcripts which do not overlap a WormBase transcript have names beginning with "RIT*" (for RNA-seq Integrated Transcript). The number following "RIT" is the chromosome (1=I, 2=II, etc. 6=X). The number after the first period is a unique number assigned to that transcript. The T1, T2, etc. are used for the alternative versions of that transcript. Currently, the naming does not allow one to know which transcript versions have the same CDS.

For the aggregate transcript set described here, we included all of the 19 stages for which RNA-seq data was available (Fig. S3).

C.4. Differential Splice Junction Usage

We created a non-redundant set of all splice junctions, noting the number of reads which confirmed that intron in each stage. We converted that number into reads per million (RPM) by multiplying 1,000,000, and dividing by the number of aligned reads in that stage. We further tracked the depth of coverage per million reads (DCPM) of a transcript which contained that splice junction. To identify alternative isoforms, we sorted the splice junctions by strand and by intron start position (donor), looked at the coordinates of one intron, and asked if the next intron in the list had a start which was equal to the preceding one (same donor, different acceptor), or if it came after the previous but before the next acceptor, etc. For the splice junctions with alternative forms, we looked at how the ratio of the RPM of the two (or more) forms varies over the stages. In this way the control was internal, the path through the region must use one of the splice junctions, and a change in the ratio means differential splice junction usage. To identify examples we performed pairwise comparisons by stage (e.g. comparing the early embryo to the

young adult) looking for intron pairs where the transcripts involved both had a DCPM of at least one, where one splice junction in the pair was used at least 5 times more frequently in one stage and less frequently in the other stage, and where at least one splice junction in each pair had an RPM of at least 2 (corresponding to ~5 or more reads for the stages with 25M reads aligned). After identifying candidates in this way, we viewed the change in splice junction usage across stages using a normalized read count for each intron in each stage, calculated by dividing the RPM for that intron by the DCPM of a transcript containing that intron.

C.5. Detection of Differentially Expressed Transcripts on Tiling Arrays

This section describes tiling array processing related to detection of differentially expressed genes. More details are in a companion paper (*9*).

RNA was isolated from 25 different embryonic and larval cell types and from all cells derived from 5 selected developmental stages to generate a total of 30 tiling array data sets (*10-12*). Additional 7 tiling array data sets were generated from RNA extracted from synchronized populations of whole animals at 7 different developmental stages. The *C. elegans* Affymetrix 1.0R tiling array was used for all experiments. Non-redundant Transcriptionally Active Regions (nrTARs) were determined by a machine learning approach (*13, 14*). (Note nrTARs were defined slightly different than conventional TARs.) nrTARs with $\geq$ 20 nt overlap with WormBase coding exons or exons of integrated transcript models were counted as hits. For quantification of transcript levels for annotated genes, unique tiling array PM probes wholly contained within exons of gene models were selected to generate a probe set for each gene listed in WormBase version WS199. (obtained from ftp://ftp.wormbase.org as a gff3 format file). Tiling array data sets were quantile normalized and probe sets were median polished using RMA (*15-17*). Significantly expressed ($\leq$ 5% FDR) gene models were determined by comparison to an empirical null model of background expression from intergenic probes for each microarray data set (*18*). The total number of detected genes was calculated from the union of tiling array data sets for cells (30 data sets) stages (7 data sets) and for the combination of cells and stages (37 data sets). As a conservative measure to correct for the accumulation of potential type 1 (false positive) errors, we adjusted the q-value of each detected gene by dividing by the cumulative number of independent samples used for each of these estimates (i.e., 37 for cells and stages, 30 for cells, and 7 for stages). This adjustment applied a similar reasoning as Bonferroni correction of *p*-values by assuming that in the least favorable case, false positives, but not true positives, were independent (*19*). To define genes differentially expressed in cells, tiling array results obtained from specific cell types were compared to corresponding developmentally matched reference data sets obtained from all cells. Similarly, to define genes differentially expressed by stage, the 7 tiling array data sets obtained from staged whole animals were compared to each other. Differentially expressed gene models were estimated with a linear model and moderated t-statistic (*20, 21*). Gene models with a FDR $\leq$ 0.05 and fold change $\geq$ 2 were called significant. Differentially expressed genes detected in cells or stages or in cells and stages were tabulated

from the union of the corresponding comparisons. The estimates were adjusted with a Bonferroni-type correction in which the FDR threshold was divided by the number of comparisons between samples. For differentially expressed genes detected in the 25 cell types, the FDR was corrected by the total number of independent comparisons (total of 25). For stages, the FDR threshold was corrected by the total number of pairwise comparisons between data sets derived from seven stages (total of 21) (see Table S3 footnotes 4, 5 and 6, 7). The fraction of genes differentially expressed was determined by dividing the number of differentially expressed genes for each category by the number of genes detected as expressed in the same category (e.g., 11,229 genes differentially expressed in cells and stages divided by 14,279 genes expressed in cells and stages = 79%).

C.6. Over-Represented Transcripts at Particular Stages

We identified a set of transcripts which are over-expressed in each of the seven main developmental stages (EE, LE, L1, L2, L3, L4, and YA) relative to other stages (Fig. S12 and Fig. S13). The stage specific transcripts were defined as those highly expressed in a particular stage (>90%) but lowly expressed in at least 4 other stages (<70%). Promoter sequences (-1kb to 0 upstream of TSS) for each group were retrieved and searched for enriched motifs using the MEME algorithm (*22, 23*). To remove generic motifs which are present in promoters of all transcripts, we scanned and compared the occurrences of these candidate motifs in specific transcripts of all the 7 stages. As an example, MEME identified 24 candidate motifs that were enriched in EE-specific transcripts, 12 of which were over-represented in the promoters of EE- or LE-specific transcripts but not in other stage-specific transcripts or ubiquitous transcripts (Fig. S14).

C.7. Determining a List of 8,428 Non-overlapping Transcripts to Study the Dynamics of Transcription and Binding

In this section we describe how we derived a high-quality list of non-redundant TSSs for studying expression and binding dynamics in Fig. 3A and 3B. This restrictive list has no transcripts that overlap and for each transcript the closest TSS is farther than 0.5 kb away. To derive the list we started with a list of transcripts obtained from WormBase. For each set of potentially overlapping transcripts at a given locus, we kept the longest one and discarded the rest. Then for each kept transcript, we defined a promoter region as a 1 kb window centered on the TSS. In some cases, promoter regions selected in this manner will overlap with other regulatory regions or transcripts, and cause RNA Pol II signal from potentially unrelated promoter regions to enter the window. To minimize this side effect and to reduce double-counting of signal, we found all TSSs less than 500 bp (i.e. half the window size) apart and picked one from each set. Using this approach, we obtained a final set of 8,428 TSSs (and associated transcripts) used for our analyses.

C.8. Using PCA for Analyzing Expression Changes across Tissues

This section describes how we performed the principal components analysis (PCA) on the tissue samples in Fig. 3C and related it to the overall correlations in the RNA-seq data shown in Fig. 3A. The idea was analyze the overall variation in the RNA-seq and tiling arrays samples in a consistent fashion and then show how the tiling arrays of specific matched embryo-larval pairs related to this overall variation.

First, we used gene-expression values for each of the tiling array samples determined in a way as to maximize compatibility with the RNA-seq DCPM values (see description above in supplement sect. B.2.). This gave rise to a large 42 sample x 20,085 gene data matrix. We then applied PCA to this to reduce dimensionality and to identify sources of variance, generating a 42x42 matrix of principal components. Rows corresponding to matched tissue samples from mixed embryo (MxE) and L2 were then extracted from this matrix and corresponding data points plotted along the first two principal components to produce Fig. 3C. In this manner, we are able to obtain a view of the matched tissue samples in the context of the entire microarray data set across all time points/tissue types queried by the microarray experiments. Note the principal coordinates for this matrix were those of the overall compendium of experiments (and hence were fairly robust to noise). The main component described the overall difference in expression programs between larval and embryo stages. As described in the caption to Fig. 3, this component was particularly enriched in genes having GO terms associated with "nematode larval development, larval development, post-embryonic development, and growth." We compared the overall PCA of all the tiling-array experiments to that of the RNA-seq experiments (obtained from the correlation matrix in Fig. 3A). Both PCAs shared similar top components, with the main axis representing embryo to larval differences.

C.9. Expression Analysis of Alternative Transcripts

We developed two alternative methods to resolve the expression level of individual isoforms for the same gene, using either expectation-maximization (EM) or a Bayesian approach with Gibbs sampling to distribute RNA-seq reads among a set of alternative transcripts in a probabilistic manner. We compared relative and absolute expression of alternative transcripts, as identified by either method, between paired samples and across the entire time course of development. We describe each of these analyses separately below.

C.9.a. IQSeq Analysis

The first method, which we call "IQSeq", uses an expectation-maximization (EM) algorithm to resolve the expression level of individual isoforms. Details on this are available in a companion paper (*24*) and we summarize the main points here. All aligned reads compatible with a transcript cluster are used to build an indicator matrix, in which the entry is 1 if a read is compatible with an isoform or 0 for all other cases. This matrix is then plugged into a likelihood

function, representing the probability of observed read alignment given a set of isoform structures. The EM calculation then produces the most probable expression levels of each transcript in a gene cluster.

Detection of Differential Expression During Development with IQSeq

We applied IQSeq to RNA-seq data of 7 developmental stages (EE, LE, L1, L2, L3, L4, YA) and derived both the relative and absolute RPKMs for all transcripts. These RPKMs values were used to derive an abundance vector □(*i,S,k*) for each gene *i* in stage *S* for isoform *k*. Then for each gene, we computed the average of squared differences for relative RPKMs of the isoforms of each gene between two stages -- i.e. *D(i,R,S)* for gene *i* between stages *R* and *S* (See caption to Fig. 2C for more detail). We computed similar quantities for absolute differences. Genes are then classified based on their scores in these two statistics in pairwise comparisons, revealing the subsets which show only dramatic isoform composition change, only dramatic absolute expression level change, neither, or both. Further analysis on these subsets may reveal key gene players or pathways in dictating worm development.

C.9.b. Deepseq9 Analysis

The second method, which we call "deepseq9", uses a Bayesian approach to estimate the relative expression of alternative transcripts for the same gene. An implementation of the algorithm, including documented source code, is available at SourceForge (*25*). Deepseq9 was developed by B. Carpenter (LingPipe, Inc.) and M. Morris (CGSB, NYU).

Computing transcript-level expression using a joint model of read alignment and expression with deepseq9

Given a data set of sequence reads, our goal is to estimate the expression of each alternative transcript for a gene based on the abundance of reads which map to sequences contained within each isoform. The method effectively distributes all of the observed reads among the possible isoforms using a probabilistic logic. Briefly, expression is inferred from the following data: $K \in N^+$ (the number of variant isoforms), $N \in N^+$ (the number of reads), and $y_1,...,y_N$ (the reads).We assume two model hyperparameters: φ (the expected variation from the reference sequence), and $\alpha_1,...,\alpha_K \in R^+$ (the prior read count per sequence plus one (to avoid zero division errors)). The general-purpose parameter vector φ reflects deviation of the sample sequence from the reference sequence for the given read distribution due to factors such as SNPs, amplification errors during sample preparation, and the sequencing platform's error profile. We infer two model parameters: $t_1,...,t_N \in 1{:}K$ (the mapping of read to splice variant), and $\theta_1,..., \theta_K \in [0,1]$ such that $\sum_{k=1}^{K} \theta_k = 1$ (the read expression probabilities, where expression levels θ are based on prior counts). In sampling notation, the model structure is:

- $\theta \sim$ Dirichlet(α)
- $t_n \sim$ Discrete(θ) for $n \in 1{:}N$
- $y_n \sim$ Channel(t_n,φ) for $n \in 1{:}N$

To estimate expression levels, we must calculate the posterior probability of reads mapping to all possible alternative transcripts. The model uses Gibbs sampling to draw samples from the full posterior distribution $p(\theta,t|y,\alpha,\varphi)$ computed over read mappings t_n and read expression levels θ given the reads y, resulting in a discrete sampling of the mappings t_n onto all annotated isoform variants based on the parameter θ (effectively a beta-binomial model of expression level). The read channel model assigns the probability of a given read y_n being observed, given that it arose from the splice variant t_n under the model parameterized by φ.

Analysis of the aggregate integrated transcript set

The analysis was initiated using pre-computed exon-level coverage for the annotated aggregate integrated transcript models, expressed in DCPM (depth of coverage per base per million reads), and a count of mappable reads for each exon (DCPM_bases), as determined from initial mapping of the RNA-seq data to the *C. elegans* genome (WS190) as described above (see sections above entitled, "RNA-seq Read Mapping and Stage-Specific RNA-seq-Only Genelet Creation Methods" and "Aggregate Integrated Transcript Set Methods"). For each exon, we generated a set of putative alignments to all parent transcripts, and then used our Bayesian model to jointly compute the read assignment and transcript-level expression. The alignment score is the probability of the read given the exon, which is proportional to the exon length (counting only mappable bases): P(read|exon) = log2(ExonLength/TranscriptLength). We multiplied DCPM by 1000 to obtain pseudo-reads which align to the exon, and then generated mappings between each pseudo-read and each possible parent transcript. The average number of mappings to distinct transcripts per read was 3.1 (i.e., on average, reads for each exon could map to one or more of three alternative transcripts). For the deepseq9 expression program, the Gibbs sampler was run for 1000 epochs, with a burn-in parameter of 500 (i.e., the first 500 iterations were discarded to allow the model to reach a stationary distribution); thereafter, we took one sample every 10 epochs (thinning of samples in this way reduces the effect of auto-correlation on samples and produces better variance estimates with fewer samples). Expression was computed as the average number of reads per transcript across all the samples. We compared our estimates with extrapolated transcript DCPM counts from the initial mapping described above, and found good overall correlation between the two approaches (median R^2=0.82 across the 15 samples; data not shown).

Clustering expression by developmental stage using Self-Organizing Maps (SOMs)

We combined the transcript-level expression calculated by deepseq9 for all transcripts in ag1003 across the 15 stages into a single data table. To identify alternative transcripts which show a

relative change in expression (i.e., transcript A > transcript B in stage 1; transcript A < transcript B in stage 2), we applied filtering criteria requiring that: (a) transcripts differ by at least 30% in opposite directions in at least two stages, and (b) the more highly expressed transcript has at least 5 pseudo-reads (corresponding to a DCPM of 0.005). (We note that ~800 of transcript pairs which passed these filters displayed borderline expression levels due to the low minimum read threshold, thus resulting in lower confidence estimates of differential expression.)

The set of transcripts that passed these filters (15,064 transcripts for 3,428 genes) was run through an SOM clustering algorithm (R 2.11 - library(class), function "SOM") that generated 44 clusters, each with at least 5 members. Clusters with similar profiles (based on visual inspection) were merged, resulting in a final set of 25 clusters (Fig. S15).

Identification of alternative transcripts with different developmental profiles

We found that 43% of all genes subjected to clustering showed alternative overlapping transcripts which fell into two or more different SOM clusters (corresponding to 8,077 transcripts for 1,475 genes) (Fig. S16). From a total of 2,107 pairs of clusters containing alternative overlapping transcripts for the same gene, we further examined 1,742 cases in which precisely one isoform fell into a distinct cluster from other isoforms for the same gene. Among these we were able to discern several distinct classes of alterations in features at the 5' end, within the CDS, or at the 3' end of transcripts (Fig. S17and (*26*)).

Individual examples from these different classes are shown in Fig. S18. We observed that while most cluster pairs shared fewer than 4 genes, those pairs with the largest number (proportion) of genes in common also tended to show similar developmental profiles. Thus, for follow-up of individual genes, examples from cluster pairs with fewer genes in common are more likely to reveal alternative transcripts with more obviously divergent developmental expression profiles.

C.10. Pseudogene Identification and Analysis

In order to identify a list of possible *C. elegans* pseudogenes, we looked at a number of features including: amino acid sequence identity, how much the pseudogene covers the parental gene, modifications such as insertions, deletions, and frameshifts, as well as other criteria. This was performed both by using the automated pipeline PseudoPipe as well as by hand annotating the *C. elegans* genome with the help of data available in the WormBase database. Comparing the coordinates from the 2,343 candidate pseudogenes identified by PseudoPipe and 1,541 identified by WormBase, there were 1,025 pseudogenes which had a nucleotide overlap of at least 50 bp between the candidates in each data set. The remaining sequences were reviewed manually, and it was determined that 173 pseudogenes from PseudoPipe, and 95 pseudogenes from WormBase, should also be included in the list, for a final total of 1,293 (Fig. S19). The remaining sequences were found to either overlap with annotated genes, be too small and fragmented to be considered a pseudogene, or should have been curated as part of a functioning gene. To determine if a

pseudogene was abundantly expressed, it had to have a DCPM value of >0.04 in at least one sample. This threshold is 100-fold higher than the minimal DCPM in this set. (DCPM is the Depth of Coverage Per Million reads calculated from the mapped RNA reads).

C.11. Identification of Canonical miRNAs and Mirtrons

Canonical miRNAs are produced by sequential cleavage of inverted repeat transcripts by the Drosha and Dicer RNAse III enzymes. We annotated novel canonical miRNAs using the miRDeep algorithm (*27, 28*), and for confident annotation, required that the cloning of miRNA and star reads mapped to a precursor hairpin with 3' overhangs at both ends of the inferred small RNA duplex. A subset of loci were confirmed to be dependent on the Argonaute encoded by *alg-1* (*29*). In total, 24 confident novel miRNAs were deposited in the miRBase database.

For mirtrons, we built an SVM model based on features of the 14 intially reported *D. melanogaster* mirtrons (*30, 31*) and ran this on the *C. elegans* genome as an independent test of its performance (*32*). Three of the four known nematode mirtrons (*mir-1018*, *mir-62* and *mir-1020*) ranked within the first 15 candidates genomewide; the fourth (*mir-1019*) presents a highly atypical 2:5 hairpin overhang and scored much lower (440th). We validated high-scoring predictions using publicly available small RNA data (*29, 33-41*), yielding 12 novel mitrons that produced at least 5 small RNA reads with a dominant 5' end and extending to the intron terminus; 10 of these also generated star reads with appropriate duplex overhangs. *NM_075944_in2* and *NM_071513_in8* did not have star reads, but the recovery of >40 reads from both loci with precise 5' ends provided strong evidence of specific miRNA production. We also reclassified the previously annotated *mir-2220* as a mirtron and recognized *NM_075943_in1* to produce a mirtron from an unannotated splice site, for a total of 18 confident mirtrons in *C. elegans* at present; Several additional high-scoring predictions yielded <5 intron terminal reads and were classifed as candidates. Full analysis of worm mirtrons is available at (*42*).

C.12. Predicting Long Non-Coding Transcripts from Tiling Array TARs: Building the 21k-set of ncRNAs

We describe below how we construct the 21k-set of ncRNAs. The building of the 7k-set is described in (*43*). The construction of the 21k-set is consistent with this, following similar principles. However, it does not include DNA conservation and RNA secondary structure information.

The tiling array signals were segmented into TARs using the maxgap/minrun algorithm (*2, 3*). Briefly, a contiguous sequence of probes exceeding a signal threshold (selected as described below) was connected to form a TAR. To account for noise, a total of 30 bp (about 1 probe) were allowed to fall below this threshold within a single TAR. Finally, TARs shorter than 100 bp (the total length of 4 probes) were discarded. The signal threshold was optimally selected according

to the criteria of attaining an FPR of 0.05 when compared to a high confidence subset of the annotation. Details are provided in (*1*).

In total, 95,069 TARs (37,026,882 nt in total) were collected from the union of 41 tiling array experiments (Supplement Table 2), of which the minimum length is 100 nt. 1,331 overlap with known ncRNA, and 22,487 include transcribed regions which are not overlapped with any annotated (confirmed or predicted) exons or known ncRNA. The reads from sequencing data from small RNA and polyA-selected RNA were also averaged for each tiling array TAR. Subsequently, different types of expression values were combined to classify each TAR as ncRNA, CDS, or UTR, using machine learning methods. Known ncRNAs, CDSs, and UTRs were selected as a gold-standard set for machine learning (Supplement Table 4 and 5). Before classification, the 95,069 TARs were fragmented into 448,746 small windows (using sliding windows of 150 nt with a 75 nt step size) (Fig. S20). Because of the sample preparation method, the tiling array TAR cannot inform as to which strand the transcript came from.

Although lacking conservation and secondary structure information, the accuracy of the classification model for the gold-standard set in terms of AUC (area under the ROC curve) is still as high as 94.2% for ncRNA prediction from TARs (Supplement Table 6). When applying the classification model to the 49,648 novel transcribed windows (from 22,487 TARs), 45,913 were found to most likely be ncRNA, 3,294 were most likely to be UTR, and 441 were most likely to be CDS (Supplement Table 7). These 45,913 "windows" originated from 21,521 TARs out of the original set of 95,069 TARs. This gave rise to the 21,521 predicted ncRNAs in the 21k-set. Please note that the prediction accuracy of the 21k-set is not as high as the 7k-set, and many of them could come from the UTRs or unprocessed introns. Therefore, we only use the 7k-set for the following novel ncRNA analysis. The genome locations of the 21k-set are available at (*26*).

Subsequently, 1,259 of the predictions in this set were found to overlap the predicted ncRNAs in the 7k-set. The 7k-set includes 7,237 ncRNA fragments (1,045,795 nt) predicted from an integrative method (*43*), where other features, such as RNA secondary structure and protein sequence conservation, are used in addition to the expression features used in generating the 21k-set.

D. More Details on Analysis of Regulation

D.1. Validation of TF Binding Sites

We performed a series of analyses to examine the quality of our ChIP-seq experiments. Much of these are discussed in detail in (*44*) and summarized here. Firstly, we selected several factors (for which primary antibodies are available) to compare our transcription factor (TF) tagging strategy for ChIP to native protein ChIP. We found that: (1) GFP-tagged AMA-1 has the same binding pattern as does native AMA-1 (the correlation coefficient between samples is 0.934, (2) the binding sites of GFP-tagged PHA-4 from embryos and starved L1s are verified by comparing

our list of genes to the list of known pharynx developmental genes (90/238, $p<1.7e^{-13}$), and (3) the binding sites of GFP-tagged HLH-1 are validated by comparing our result to an unpublished data set of binding sites for endogenous HLH-1 (*45*). Overall, these analyses (to date) are consistent with the conclusion that the tagged factor has binding and regulatory properties similar to those of the native protein, and that differences between tagged factor and native protein ChIPs are well within the expected levels of variation which are commonly observed between replicate ChIP samples using the native protein. Secondly, many PHA-4 binding sites from embryos and starved L1s identified by ChIP-seq were verified through an independent method: ChIP-qPCR (76% of the embryonic sites and 74% of starved L1 sites with two-fold or higher enrichments). Lastly, we examined the functional enrichments of protein-coding genes targeted by each of 23 factors. Many Gene Ontology (GO) terms related to developmental processes are found to be enriched for genes bound by many factors, suggesting the general roles of these factors during *C. elegans* developmental processes. More importantly, for factors with known functional roles we identify specific enrichment of GO terms that match these functional roles (*46*). In conclusion, these analyses demonstrate the high quality of our ChIP-seq experiments.

D.2. Identification of Target Coding and Non-Coding Genes for TFs

The details of data sets for 23 factors (22 TFs and one dosage compensation factor) are listed in supplement Table 8. The determination of the target gene associated with a particular TF is described in more detail in (*47*). We used a very simple and straightforward approach: Genes were targeted by TF binding peaks if they were within 500 bp upstream or 300 bp downstream of the TSS. This is a fairly conservative threshold; it is possible to take significantly larger values for the upstream threshold without greatly affecting the results. We obtained the annotations of worm genes from WormBase. Although the TSSes for the majority of worm miRNAs has not been mapped, it has been shown that DNA regions upstream of the pre-miRNA are sufficient to initiate the transcription of miRNAs (*48*). We thereby identified the target miRNAs by examining the existence of TF binding peaks around the start position of pre-miRNA transcripts.

D.3. HOT Regions

Using the 23 factors' ChIP-seq data sets, we determined the number of factors bound at each base in the *C. elegans* genome. Out of the 16,707 genomic regions identified as having significant enrichment in at least one of the ChIP-seq data sets, 304 Highly Occupied Target (HOT) regions were significantly enriched in 15 or more factors (*26*). We combined overlapping peak regions across the 23 factors to annotate each of the 16,707 regions based on the maximum number of factors associated at any point within the region. To determine whether this would be expected by chance, we randomly re-assigned peak regions within the 16,707 regions bound by at least 1

factor. Using 1,000 iterations of random re-assignments, no regions associated with 15 or more factors were observed (Fig. S24).

We used multiple experimental and computational approaches in order to confirm that enrichment for these regions was not simply an artifact of the ChIP-seq procedure. HOT regions were not significantly enriched when IgG antibody was used on transgenic worms, nor when GFP antibody was used on N2 worms lacking a GFP-tagged TF. These negative controls demonstrate that these regions are not simply a chromatin or GFP-antibody artifact (Fig. S25). As an additional negative control, we observed that DPY-27, which is known to bind preferentially to the X chromosome (*49*), is almost exclusively enriched at regions (including HOT regions) on the X chromosome and is not enriched at HOT regions on the autosomes (Fig. S26). As a positive control, we immunoprecipitated endogenous LIN-15B from wild-type worms using anti-LIN-15B antibody, and observed binding peaks in HOT regions similar to those observed using the GFP antibody on *lin-15B*::GFP worms (Fig. S25).

Expression of Genes Associated with HOT Regions

We used a stringent range to associated genes with HOT regions. Genes were associated with peak regions if they were within 1kb upstream or 500nt downstream of the TSS. For staged worm populations, gene expression levels for all *C. elegans* WS190 transcripts were measured by DCPM in RNA-seq data as described previously (*4*), and for genes with multiple annotated alternative transcripts, the average expression level of all transcripts was used. We additionally made use of two different types of tiling array data sets described in (*9*): tissue-specific embryonic expression measurements, performed by expression of GFP under tissue-specific promoters followed by FACS sorting, and tissue-enriched measurements, performed by tissue-specific promoter-driven expression of epitope-tagged polyA binding protein followed by purification of RNA bound by the tagged polyA binding protein (the mRNA tagging method described in (*50*)). Tiling array data were analyzed by first computing the PM - MM value for each probe. Experiments were conducted in triplicate and quantile normalization was used to assure values from the three replicates are comparable. Data from the three replicates were combined using pseudomedian smoothing (*2*) over a window size of 110 bp, and transcript expression levels were calculated as the median signal value for all probes overlapping the transcript's exonic regions by at least 50%. Only the longest isoform was used for genes with multiple transcripts. For inter-sample comparison, we normalized these expression levels by dividing the values by the slide median, i.e. the median of all probes on the array. In the staged population RNA-seq experiments (Fig. S27 for L1 stage worms; other stages not shown) as well as every tissue for both tissue-specific and tissue-enriched tiling array data (Fig. S28), HOT genes had significantly higher levels of expression than genes bound by 1-4 factors (all $p<10^{-15}$ by Kolmogorov-Smirnov test).

Specific vs. HOT Target Comparison

HLH-1 is a muscle-specific TF with a known binding motif (CAGCTG) (*51*). Motif enrichment was calculated by simple hexamer frequency counts, and p-values were calculated using a chi-square test. Genes with muscle-enriched expression were obtained from (*50*). To compare all TFs, we made use of L2 intestine-enriched transcripts (*52*), adult germ-line enriched transcripts (*53*) and embryonic tissue-specific tiling arrays described above (*9*). To identify embryonic tissue-specific genes, each embryonic tissue-specific array was first linearly normalized to the embryonic reference array to correct for array-specific scaling effects. Next, for each gene in each tissue, we calculated a z-score for specificity:

$$z_{tissue} = \frac{x_i - \mu_i}{\sqrt{\frac{1}{N-1}\sum_{j=1,..,i-1,i+1,...,n}(x_j - \mu_i)}}$$, where $\mu_i = \frac{\sum_{j=1,..,i-1,i+1,...,n}(x_j)}{N-1}$ and N=11 tissues (including the reference array).

Genes with $z_{tissue}>2$ were deemed "tissue-specific". In addition to HLH-1, we considered previously identified binding motifs for ELT-3 (GATAA (*54*)), MDL-1 (CACGTG (*51*)), and PHA-4 (T[AG]TT[TG][AG][CT] (*55*)). In all three cases, we observed a similar drop in motif enrichment between specific targets and HOT regions (data not shown).

Essential Genes Comparison

Essential genes were defined as genes having an RNAi phenotype of 100% larval arrest, embryonic lethality, or sterility in a genome-wide screen for RNAi knockdown phenotypes (*56*). Significance was calculated by a chi-square test.

D.4. Identification of Conserved miRNA Binding Sites in 3'UTRs

We used the PicTar algorithm (*57*) to identify conserved microRNA target sites within annotated 3'UTRs from the aggregate transcript models (ag1003) (Supplement Table 11 and (*26*). We applied the version of PicTar described in (*58*) with the slight modification that a perfect seed site if covering the first 5' base of the miRNA was required to match an adenosine at this positon. We used a non-redundant subset of 3'UTRs, considering only those which do not overlap any CDS in an alternative transcript isoform for the same gene, and excluding a small subset of transcripts (~4,500) for which we identified more than one putative ORF in different reading frames. We used 183 miRNAs, either annotated in miRBase14 (*59*) or newly identified from *C. elegans* embryos (*40*) using miRDeep version 2 (*27, 28*), and genome alignments between three (*C. elegans, C. briggsae,* and *C. remanei*) or five (also including *C. brenneri* and *C. japonica*) species. This set of predictions for the ag1003 transcript models are an alternative to our recently published predictions for the worm 3'UTRome (*8*), which use 3'UTRs for AceView (*60*) gene models.

We also independently searched for perfect Watson-Crick complementary seed sites covering the first or second 5' miRNA heptamer which are prefectly conserved. These predictions should be identical to the 'TargetscanS' predictions (*61*) and, by definition, identical to the vast majority of

PicTar predictions. Indeed, a comparison of the results between the two algorithms revealed that PicTar identified 99% of seed sites predicted by TargetScan, and conversely, TargetScan identified 89% of seed sites predicted by PicTar (data not shown). The reasons for the additional PicTar predictions are (a) PicTar uses a more general definition of 'conserved seed site', allowing for evolutionary changes between the different heptamers in the same alignment, (b) PicTar also effectively locally realigns target site candidates to overcome alignment problems, and (c) PicTar also predicts imperfect, conserved seed sites if very significantly compensated by additional basepairings between the remainder of the miRNA and the mRNA. Previous independent comparisons of miRNA target prediction algorithms using other data sets have shown that TargetScan and PicTar are top performers in the field, and generally produce the highest overlap with experimentally determined sites ((*62*); reviewed in (*63, 64*). Compared to our earlier analysis of *C. elegans* 3'UTRs (*58*), our new prediction sets ((*8*) and this study) show a higher signal-to-noise ratio compared to synthetic miRNAs of similar composition (1.8-2.4 and 2.1-3.4 for 3-way and 5-way alignments, respectively, using the method described in (*57*)). We attribute this to a combination of better multi-species genome alignments and exclusion of genomic sequence regions which are not supported by experimental evidence (previous predictions used up to 500nt downstream of the CDS when no annotated 3'UTR was available).

D.5. Calculation of Tissue Specificity Score for TFs in the Hierarchical network

Expression levels of all *C. elegans* genes at 8 different tissues at L2 stage were measured using tiling arrays. The tissues are defined as in Supplement Table 2. Tissue specificity score for a gene is calculate as follows:

$$TSPS = \sum_i f_i log_2(f_i / p_i)$$

,

where f_i is the ratio of the gene expression level in tissue i to its sum total expression level across all tissues, and p_i is 1/8, the fractional expression of a gene under a null model assuming uniform expression across tissues. A higher tissue specificity score suggests more specific expression in a single or multiple tissues, whereas a score of zero suggests uniform expression.

D.6. Calculation of Overrepresented Network Motifs

In order to identify the patterns that are present in the integrated network with significantly higher frequency than expected by chance, we enumerated all the possible patterns with 3 nodes. The frequencies of these patterns in the real network were compared with those in 1,000 random networks. The random networks were generated by rewiring the real network, while keeping its topological statistics constant; i.e., keeping the same number of coding gene targets and the number of miRNA targets for a TF node, the number of regulatory TFs and targets for a miRNA

node, and the number of regulatory TFs and miRNAs for a gene node. For each pattern, a *z*-score was calculated as follows:

$$Z-score=\frac{N_{real}-Mean(N_{rand})}{SD(N_{rand})},$$

where N_{real} and N_{rand} are the number of corresponding patterns in the real network and in the random networks, respectively. A significant positive *z*-score indicates over-representation, whereas a significant negative one indicates under-representation of a pattern in the integrated network. The *p*-value for a *z*-score was calculated by referring to a standard normal distribution. For the network motif analysis, we only used the proximal targets (500bp upstream to 300bp downstream).

E. More Details on Chromatin Organization

E.1. Correlating Chromatin Features and TF Binding Signals

The worm genome was divided into bins of 100 bp. For each bin, the average signal was computed for each chromatin feature and for each TF binding experiment. Consequently, each experiment is associated with a vector of signals. Correlations were computed as the pairwise Pearson correlations between these vectors. We also computed Spearman correlations and normal-score correlations between the vectors. The correlation patterns are similar for the three correlation functions, and we include only the results based on Pearson correlations.

E.2. Machine Learning

For each TF binding experiment, the bins which overlap with the binding peaks form the positive set. The same number of other bins were randomly sampled from the whole genome as the negative set. Half of the bins in the positive and negative sets were used as training examples to train support vector machine (SVM) models using default parameters in Weka (*65*). The other half was used to test the performance of the SVM models. Model accuracy was evaluted using ROCs, as well as the area under the ROC curves (AUROC). We also used precision-recall (PR) curves as a secondary measure, and arrived at the same general conclusions. Different feature sets were used in different configurations. Each of the single-feature models involves only one feature. The integrative model involves all features, and the stage-specific models involve only features from one development stage.

Figure Legends for Supplement

Fig. S1: ChIP-chip and ChIP-seq comparision

A. 2D distributions and pairwise correlations between Pol II ChIP-chip and ChIP-seq replicates with combined profiles at two developmental stages (early embryo, left and L4, right). The sample names are shown on the diagonal. In the lower triangular part of the panel, each blue dot represents the median signal levels of ChIP-chip (MA2C score) and ChIP-seq (sequence read count) within a 1kb-segment on the genome. The upper triangular part provides the correlation coefficient of each pair.
B. The heatmap image represents pairwise correlations between ChIP-chip and ChIP-seq combined profiles at early embryo and L4 stages, and is hierarchically clustered by both rows and columns. It is shown that the variation between the two platforms at the same stage (correlation coefficient of about 0.7) is smaller than that between the two different stages of the same platform (correlation coefficient of 0.4-0.53).
C. The venn diagrams show the overlap of the top 3000 Pol II binding sites identified by ChIP-chip (blue circle) and ChIP-seq (red circle) in early embryo (left) and L4 (right). It can be seen that more than 2/3 of Pol II binding sites were commonly identified by the two platforms.

Fig. S2: Correlation of RNA expression levels for Young Adult between RNA-seq and tiling array platforms

Each point represents a gene. To account for multiple isoforms, a gene is here defined as the union of all exonic nucleotides. RNA-seq expression levels per gene were measured using RPKM, and tiling array levels were measured using the mean intensity of probes falling within exons. The genes in the upper left likely represent cross-hybridization.

Fig. S3: Numbers of RNA-seq Reads

Total reads along with numbers of uniquely and multiply aligned non-rDNA reads for each of the 19 *C. elegans* stages and samples.

Fig. S4: Density plots of 20,051 genes in WormBase190

Each line corresponds to a sequencing depth. The legend reports the number of mapped reads (in millions). The two peaks represent genes not expressed (left) and expressed (right) at each sequencing depth. Note that the number of non-expressed genes drops sharply at first as sequencing depth increases, then reaches a plateau.

Fig. S5: Pair-wise comparison of the density plots

Y-axis reports *p*-values of the Kolmogorov-Smirnov test as a function of depth of sequencing (x-axis). The dotted line shows a *p*-value of 0.01. Higher *p*-values indicate no difference between the distributions. The plot shows that a sequencing depth between 13.4 and 16.8 million reads is sufficient to capture most expressed genes in whole animal samples.

Fig. S6: Rate of gene discovery

The number of genes with RPKM=0 are reported as a function of sequencing coverage. The equation reports the coefficients and the R^2 of the best fitting exponential curve. The fitted curve is: Number of non-expressed genes = 8.5 x (depth of sequencing) $^{-0.88}$ (R^2=0.9044).

Fig. S7: Number of features identified

Number of features identified by stage as compared to features in WormBase (WS170) when the project began. The two right most bars represent the RNA-seq-only aggregate set and the aggregate integrated transcript set created from all available *C. elegans* transcriptome data. All features were clustered when within 25 bases of one another, e.g. if there were three different polyA sites within 25 bases of one another, they were counted as a single polyA site.

Fig. S8: Number of confirmed splice junction over time

This figure indicates the significant contribution of RNA-seq to annotating the worm genome. There were 11,467 splice junctions confirmed when the *C. elegans* full genomic sequence was first published (*66*). The first rise in 2003 was a result of the OST Project (*67*) and the remaining increases were a result of the modENCODE project (e.g. (*4*)).

Fig. S9: Proportion of splice junctions confirmed by various methods

The large overlap in splice junctions confirmed between RNA-seq, RT-PCR/RACE and mass-spec (*7*) provide confidence in the methods used for identifying confirmed junctions by RNA-seq.

Fig. S10: Saturation of discovery of non-coding and coding RNAs with additional RNA-seq data sets

We are presently utilizing a number of approaches to ncRNA discovery, and our initial efforts have revealed thousands of new ncRNAs from the *C. elegans* genome. As assays are performed under additional conditions, and as we develop and refine our computational methods of analysis, we expect to discover many thousands more non-coding RNAs. The saturation plot for novel ncRNAs (left) illustrates this point. In each experimental condition, the total length of ncRNAs expressed in the condition was determined using a combination of experimental and computational methods. When multiple conditions are considered together, the total length of ncRNAs depends on the set of conditions involved. The saturation plot displays that total length (y-axis) at different number of conditions (x-axis). At each point along the x-axis, all possible combinations of conditions are considered, and the distribution of total lengths is summarized by a box plot. The black line shows the slope of the curve connecting the averages at the end of the curve. The steepness of the curve suggests that more ncRNAs are expected to be discovered if additional conditions are considered. We also made the same saturation plot (on the right) for coding exonic regions. The detection of expressed exons tend to be saturated when additional experiments are added.

Fig. S11: Number of stages/samples where a given gene or splice junction is observed

Most genes and splice junctions are represented in all 19 stages, with smaller peaks for those found in only one or two stages and samples. The peak at 2 for stages per splice junction in part results from the requirement that all novel splice junctions (novel is defined as not a part of WormBase170 predictions, which included WormBase, Twinscan and Genefinder predictions) occur in at least two different stages.

Fig. S12: Expression profiles of developmental stage-specific genes

High and low expression levels (normalized DCPMs) are shown in red and blue, respectively. Expression levels of each gene are normalized across the 7 developmental stages by subtracting the mean then dividing the standard deviation.

Fig. S13: Expression profiles of the meta-genes for developmental stage-specific transcripts

The expression level for a meta-gene was calculated by averaging the expression levels of all genes which are specific to a given developmental stage.

Fig. S14: Enrichment of promoter motifs

Enrichment of 24 EE-specific candidate motifs identified by the MEME algorithm in promoters of stage-specific genes. The –log(*p*-value) was calculated by comparing the occurrences of a motif in stage-specific transcripts relative to all the other transcripts, and then color-coded with red (indicating over-representation) or blue (indicating under-representation).

Fig. S15: SOM clusters

Reduced set of 25 SOM clusters displaying different developmental expression profiles, based on probabilistic inference of individual transcript levels using deepseq9. The X-axis is the following 15 stages in order: EmMalesHIM8, EE, LE, L1, LIN35, L2, L3, DauerEntryDAF2, DauerDAF2, DauerExitDAF2, L4, L4MALE, L4JK1107soma, YA, and AdultSPE9. The Y-axis is the log2 of probabilistic read counts from deepSeq9. Solid lines represent mean transcript expression in each of 15 staged samples; dashed lines represent one standard deviation from the mean.

Fig. S16: Number of genes and transcripts shared between pairs of SOM clusters

The size of each cluster is indicated in terms of genes (yellow, g=XX) or transcripts (green, t=XX). Cells are shaded by the Pearson Correlation Coefficient (PCC) between developmental expression profiles for each pair of clusters, calculated from their mean expression across the 15 staged samples.

Fig. S17: Isoform classes

Classes of distinguishable isoform features within alternative transcripts for the same gene that fall into different SOM clusters with distinct developmental expression profiles. Numbers correspond to cases in which a single isoform falls into one SOM cluster, and one or more alternative isoforms fall into another cluster (see text for details).

Fig. S18: Examples of different classes of alternate isoform expression identified from SOM clustering

Ag1003 transcript models are displayed with wiggle plots from relevant stages using the Integrative Genomics Viewer (*68*). These plots represent 36-mer reads aligned without mismatch (trimmed up to 2 bases) and were calculated by the SHRiMP aligner v1.3 (*69*).
A. Unique 5' UTRs of C23H3.7.T3 and C23H3.7.T4. C23H3.7.T3 is absent in early embryo and is co-expressed with C23H3.7.T4 in young adult.
B. An alternative CDS exon is skipped in F26B1.2.T8 and included in F26B1.2.T9. F26B1.2.T9 is more highly expressed in L4 than L2.
C. Overlapping 3' UTR of F28C6.3.T2 and F28C6.3.T4. F28C6.3.T4 is expressed at a much higher level in L4 than in young adult.

Fig. S19: A Breakdown on How the Updated List of Worm Pseudogenes was Created

The figure schematizes the workflow in updating the pseudogenes in WormBase, to arrive at current total of 1293 worm pseudogenes. Pseudogenes came from two sources: those already in WormBase annotations (right) and those identified by Pseudopipe (left). The initial overlap of 1025 pseudogenes from these two sources was kept. The remaining subsets also kept are shown in red. These include 83 additional duplicated (DUP) and 90 additional processed (PSSD) pseudogenes identified by pseudopipe. They also include 95 pre-existing WormBase annotation not found by pseudopipe that were double-checked by the WormBase curators.

Fig. S20: Binning of long TARs built from tiling arrays

The TARs from tiling array data were built from the union of 41 samples (*1*). The minimum length of TAR is 100nt. Since long TARs could cover more than one type of sequence element, such as exons, introns, and UTRs, they were spliced into small windows of at most 150nt each, with adjacent windows having a 75nt overlap. Each bin was defined as intronic TAR, exonic

TAR or UTR depending on which annotation it overlaps (WormBase170 was used). Those small TARs which are less than 150nt are not spliced.

Fig. S21: Transcription factor motif discovery

A. Recovered motifs. Transcription factor ChIP-seq peak data sets were searched for enriched motifs as described in the text . Of the 23 data sets analyzed, enriched motifs were found in 22; however, only 8 transcription factors showed sufficient specificity to be accepted (see panels B and C for example).

[1-] Also enriched in HOT regions. The fact that the CEH-14 motif is enriched in HOT regions either means that this TF binds specifically to HOT regions, or that this TF has a weak motif and that the observed motif is derived from another protein co-binding in HOT regions. Additional experiments will be necessary to decide between these two cases.

[2-] Consistent with a previously published motif for the given TF.

B. Example motif distribution analysis (BLMP-1)- Distribution of BLMP-1 motif "TTTCACTTT" was plotted relative to SPP-point-binding positions (single-base-pair genomic coordinates with highest likelihood for binding (*70*)) for BLMP-1. The motif occurrence distribution is Gaussian-like around BLMP-1 point binding positions (black and yellow lines) while relatively evenly distributed over random upstream regions (red line). Black indicates high confidence peaks with SPP assigned FDR <0.01. Yellow indicates low confidence peaks with SPP assigned FDR >0.01 and <0.05. Red indicates random upstream regions.

C. Example motif density analysis for BLMP-1. 200 base pairs flanking point binding positions for BLMP-1 were analyzed for density of BLMP-1 motif "TTTCACTTT" in occurrences per base pair. BLMP-1 peaks have significantly higher occurrences of the motif than any other transcription factor. Random upstream regions and HOT regions were also analyzed on a motif-per-nucleotide scale and similarly show much lower motif density than what is found in BLMP-1 peaks.

Fig. S22: An example showing GEI-11 binding near three ncRNAs

Four other factors (MAB-5, LIN-39, EGL-27, and PES-1) are also shown as controls. The signal for each TF, as well as for Pol II and input, plots the ChIP-seq raw read counts scaled based on total mapped reads. Pol II and input samples were from N2 worms; TF samples were from worms expressing the factor tagged with GFP. The value of tiling array ChIP-chip signal for H3K27ac and H3K4me are also shown in green. Raw reads of polyA-plus RNA-seq and small RNA-seq, as well as expression (log2 of signal) from total RNA tiling array signal are also shown. The ncRNA annotations and protein annotations are from Refseq.

Fig. S23: Co-occurrence of transcription factors

Co-occurrence is counted if two TFs bind to the promoter region of same gene (2000 bp upstream to 300bp downstream of TSS), without considering the strength of binding. Genes targeted by HOT regions were removed before calculating the co-occurrence. The heat map

reflects the co-bound correlation of each pair of TFs at targeted gene loci, with red indicating more co-bound genes than would be expected by chance and blue, indicating less. TFs have been clustered along both axes based on the similarity of their bound targets with other factors. The same stage is annotated with the same color. The HOX genes are highlighted with orange color.
A. Co-occupancy of transcription factor pairs at targeted coding genes.
B. Co-occupancy of transcription factor pairs at targeted miRNAs.

Fig. S24: Distribution of TF binding numbers

Many regions show overlap of ChIP-seq binding sites for 23 factors. Red indicates the number of regions bound by 1 to 23 factors in ChIP-seq data. There are 16,707 genomic regions bound by at least 1 TF, and 304 regions bound by at least 15 factors. Black indicates the average number of regions bound in 1,000 randomized controls, with error bars indicating standard deviation. In randomized controls, an average of less than 1 region was bound by 12 or more factors, and no regions bound by 15 or more factors were observed.

Fig. S25: Control experiments for HOT regions

The x-axis plots the percentage of the 304 HOT regions which are significantly enriched in the various ChIP-seq controls. In order to verify that antibodies do not bind non-specifically to GFP-tagged proteins, IgG negative control experiments were performed in two different transgenic worm lines expressing LIN-15B::GFP or EGL-27::GFP. In order to verify that GFP-specific antibody does not pull down any other proteins in *C. elegans*, GFP antibody negative controls were performed in wild-type worms at embryonic and L3 stages. As a positive control, LIN-15B antibody was used in wild-type N2 worms to immunoprecipitate endogenous LIN-15B.

Fig. S26: DPY-27 only binds to HOT regions on the X chromosome

The y-axis shows the number of HOT regions found on each chromosome. The set of all 304 HOT regions and the 298 HOT regions that are bound by LIN-15B are evenly distributed across all 6 chromosomes (with chromosomes separated by color). In contrast, all 29 HOT regions bound by DPY-27 are on the X chromosome. DPY-27 regulates gene expression specifically on the X chromosome for dosage compensation (*49*)

Fig. S27: Higher gene expression level in HOT regions in RNA-seq

Genes adjacent to HOT regions have significantly ($p < 10\text{-}30$ by Kolmogorov-Smirnov test) higher expression in late embryonic (LE) worms than do genes located near just 1-4 bound factors. In red, expression level of genes with a HOT region within 1kb upstream or 500nt downstream of the TSS; in blue, expression level of genes proximal only to regions bound by 1-4 factors. The histogram plots the frequency (y-axis) of genes with the listed RNA-seq expression levels in late embryonic worms (x-axis, measured as log10(depth of coverage per million reads)).

Fig. S28: Higher gene expression level in HOT regions in tiling arrays

Genes adjacent to HOT regions have higher expression levels in tissue-enriched tiling arrays across all tissues assayed. In red, expression level of genes with a HOT region within 1kb upstream or 500nt downstream of the transcription start site; in blue, expression level of genes proximal only to regions bound by 1-4 factors. The histogram plots the median of normalized gene expression measurements (y-axis) of genes on the listed tiling array experiment (*x*-axis), with error bars indicating standard error of the mean. Data is further described in (*9*). For embryonic experiments (left), tissue-specific gene expression measurements were obtained from tiling arrays performed on FACS sorted cells expressing a tissue-specific GFP labeling. For post-embryonic experiments (right), gene expression measurements were obtained from tiling arrays performed on samples that were tissue-enriched using the mRNA tagging method. In all experiments shown, genes adjacent to HOT regions are significantly shifted towards higher expression ($p < 10\text{-}15$ by Kolmogorov-Smirnov test).

Fig. S29: HOT regions are broadly expressed

Single-cell gene expression measurement of promoter transcriptional reporter constructs in L1 worms from 3D confocal data stacks (data from (*71*)). The x-axis represents 363 specific cells present in the L1 worm, and the y-axis shows expression of 93 mCherry reporters, with the expression level of the mCherry reporter shown by the red scale bar. Promoters containing HOT regions (bound by 15 or more factors), and even promoters containing regions bound by 10-14 factors, show broad expression across 363 cells in the L1 worm, whereas promoters lacking these regions show a variety of diverse tissue-specific expression patterns. Data is presented identically to Fig. 6C, and gene names are provided in addition to row label codes from Fig. 6C.

Fig. S30: Pair-wise correlations of PHA-4 binding signal across different stages

The union of all PHA-4 binding sites were merged together and then binned into 100nt windows. The raw reads of ChIP-seq data for each window were calculated and normalized over the respective input for each ChIP-seq experiment. The correlation coefficient of each pair of stages was then calculated. HOT regions were removed before merging the binding sites.

Fig. S31: Examples of Pol II binding and expression

Heatmap showing percentage of RNA Pol II binding and expression for *isl-1* and *pgp-2* and C15F1.2, during seven stages of the worm life cycle. For each transcript, RNA Pol II binding levels and gene expression levels increase in concert until the stage where both reach maximum levels. In the following stages, the expression levels tend to drop at a faster rate than RNA Pol II binding. These examples illustrate one scenario in which a change in gene expression in earlier stages can be predictive of a similar change in Pol II binding levels during later stages. The

heatmap is normalized independently along the columns, with the values representing the ratio of signal in a stage to the maximum signal observed.

Fig. S32: Histone marks distribution over repetitive elements

Five repetitive element classes were extracted from WormBase190. The region of the genome underneath each element was subdivided into 10 equal sized bins centered on the element. In addition, the 1 kb regions flanking each element were subdivided into an additional 20 100 bp bins. The mean *z*-scores for ChIP-chip chromatin marks from L3 larvae were then graphed across each bin. The histone marks from top to bottom are: H3K27ac, H3K36me2, H3K36me3, H3K4me2, H3K4me3, H3K27me3, H3K9me1, H3K9me2 and H3K9me3.

Fig. S33: Promoters of chromosome X genes have higher GC content compared to autosomes

A. Average GC content is plotted for chromosome X and autosomal genes centered at their transcription start sites(GC content is calculated within 25 bp upstream and downstream of each coordinate). A region between -250 to -50 shows a spike in GC content on chromsome X.
B. Distribution of average GC content within this region is plotted. Chromosome X gene promoters have significantly higher GC content, as determined by a Wilcoxon rank sum test ($p < 2.2e-16$).

Fig. S34: Correlations between whole-genome transcription factor binding signals and chromatin features

The Pearson correlation between the signals from each of the 27 transcription factor ChIP-seq experiments (rows) and 22 chromatin features (columns) across the whole genome are shown in a heatmap.

Fig. S35: Machine learning procedure for modeling transcription factor binding peaks

The *C. elegans* genome was divided into bins of 100 bp. Histone methylation and binding signals of RNA Pol II were used as features to distinguish between bins which intersect with the binding peaks from those which do not, using the machine learning method support vector machines (SVMs). Models were learned from the training portion of the data sets and evaluated on a separate testing portion.

Fig. S36: Modeling accuracy of integrative models

Each curve represents the accuracy of an integrative model involving all features together used to predict either the binding peaks from a TF binding experiment or HOT regions.

Fig. S37: Modeling accuracy of models involving either all features or individual features

Each column corresponds to the feature(s) involved in constructing statistical models for either the binding peaks of the transcription factors or the HOT regions (represented by the rows). The receiver operator characteristic is a plot of true positive rate against false positive rate of a set of ranked predictions. If all the ground truth positives are ranked higher than the ground true negatives, the curve goes from the origin vertically up to the point (0, 1), and then horizontally to (1, 1). In this case, the area under the curve has the maximum value of 1. If all the ground truth negatives are ranked higher than the ground true positives, the area under the curve has the minimum value of 0. A random ranking has an expected area under the curve of 0.5. In general, a larger area under the curve indicates a higher consistency between the predictions and the ground truth.

Fig. S38: Distinguishing binding peaks of different TFs

Each bar shows the accuracy with which a model distinguishes the binding bins of a TF experiment from random binding bins of other TF experiments. The last column shows that the HOT regions can be accurately separated from other TF binding sites using the chromatin features.

Fig. S39: Developmental stage-specific models

The accuracy of models specific for individual develpmental stages (involving predictors only from that stage) are shown. For each TF, the heights of the three bars correspond to the accuracies of the models, involving predictors measured in (from left to right) embryos only, L3 only, and both stages.

Fig. S40: Combination of chromatin and sequence features

Potential binding sites of HLH-1 were identified by using two known sequence motifs in Jaspar (*72*). Chromatin features were used to model general binding active regions (BAR+) which are not specific to any DNA-binding proteins. The prediction model assigns a probability value for each region to indicate its likelihood of being in BAR+. By varying the probability threshold, different sets of BAR+ regions were identified. At each threshold, three sets of regions were compared: all general binding active regions (BAR+), all regions with high motif scores (PWM+), and binding active regions with high motif scores (BAR+PWM+).

Fig. S41: Coverage of evolutionarily constrained regions by genomic features

From the six-way alignment of *C. elegans*, *C. briggsae, C. brenneri, C. japonica*, and *P. pacificus,* we identified the portion of the genome under evolutionary constraint as described in the main text. From this, we calculated the overlap with pre- and post-modENCODE functional

elements in order to determine the proportion of constrained regions that can be "explained" by known functional elements. The columns indicate the coverage of constrained regions (measured as a proportion of base pairs) for each type of functional element, and the blue line indicates the cumulative coverage. All WormBase annotations are taken from WS190, a database release that preceded importation of modENCODE data. Element sets are as follows: *WB-Conf-CDS:* WormBase CDS annotations that are fully supported by experimental evidence; *WB-3UTR:* WormBase 3' UTRs; *WB-5UTR*: WormBase 5' UTRs; *WB-Partial-CDS*: WormBase CDS annotations that are partially supported by experimental evidence; *ME-CDS*: modENCODE CDS annotations; *ME-3UTR*: modENCODE 3' UTR annotations; *ME-5UTR*: modENCODE 5' UTR annotations; *WB-Pred-CDS*: predicted WormBase CDS (no experimental support); *ncRNA*: modENCODE noncoding RNA annotations; *Pseudogene*: modENCODE pseudogene annotations; *ChIP-TF*: modENCODE binding sites for 23 transcription factors; *ChIP-Other*: the union of modENCODE binding site peaks for chromatin modification factors HCP-3, LEM-2, MES-4 and HRG-1; *ChIP-dosage-comp*: the union of modENCODE binding site peaks for DPY-27, DPY-28, MIX-1, SDC-2 and SDC-3. Because some TF factor binding sites are broad, such peaks were trimmed to be no wider than 250 bp when calculating their coverage.

Fig. S42: Saturation of TF binding

Saturation of the binding sites of 22 *C. elegans* transcription factors (including 6 stages for PHA-4) over the WS190 genome with coding sequence bases and Pol II binding site bases removed. No more than 9% of the bases are covered by these factors. These experiments include: ALR1 L2, BLMP L1, CEH14 L2, CEH30 LE, EGL5 L3, EGL27 L1, ELT3 L1, EOR1 L3, GEI11 L4, HLH1 EMB, LIN11 L2, LIN13 EMB, LIN15B L3, LIN39 L3, MAB5 L3, MDL1 L1, MEP1 EMB, PES1 L4, PHA4 EMB, PHA4 L1, PHA4 L2, PHA4 LE, PHA4 stvL1, PHA4 YA, PQM1 L3, SKN1 L1, and UNC130 L1.

Supplement Files

All the supplement files can be found at
http://www.modencode.org/publications/integrative_worm_2010.

Supplement Tables

Supplement Table 1 : *C. elegans* genes not identified as "transcribed" in 19 polyA RNA-seq samples

Type	Genome total	Not covered	% not found
nuclear hormone receptors	85	21	24.7
7TM/G-protein coupled receptor	1454	323	22.2
math-(meprin-associated Traf homology)	62	9	14.5
F-box	238	12	5
Zinc Finger	236	4	1.7

Stages and strains of worm RNA (polyA) sequenced include: embryonic him-8(e1489) (50% males), early embryos, late embryos, lin-35(n1745), L1, L1, L2, L3, dauer entry daf-2(e1370), dauer daf-2(e1370), dauer exit daf-2(e1370), L4, L4 males, JK1107 L4 (no gonad) glp-1(q224), young adults, aged adults (spe-9(hc88)), adults exposed to Harposporium spp (tentative assignment), and adults exposed to S. marcescens

Supplement Table 2 : Develpmental stages and tissue samples of small RNA-seq and tiling array experiments.

Developmental stages of small RNA-seq experiments
Young adult males (23dC)
Mixed Embryo
mid-L1 20dC 4hrs post-L1 stage larvae
mid-L2 20dC 14hrs post-L1 stage larvae
mid-L3 20dC 25 hrs post-L1 stage larvae
mid-L4 20dC 36hrs post-L1 stage larvae
Young adult 20dC 48hrs post-L1 stage larvae
Young adult (23dC DAY 0 post-L4 molt)
Adult 23dC 12 days post-L4 stage larvae
Adult 23dC 5 days post-L4 stage larvae
Adult spe-9(hc88) 23dC 8 days post-L4 molt
Developmental stages of tiling array experiments
embryo A-class motor neurons
embryo all cells reference
embryo AVA neurons
embryo body wall muscle (v2)
embryo coelomocytes
embryo dopaminergic neurons
embryo GABA motor neurons
embryo germline precursor cells
embryo hypodermal cells
embryo intestine
embryo panneural
gonad from young adult 20dC 42hrs post-L1 N2
L1 20dC 0hrs post-L1 N2
L2 25dC 14hrs post-L1 N2
L2 A-class neuron
L2 body wall muscle
L2 coelomocytes
L2 excretory cell
L2 GABA neurons
L2 glutamate receptor expressing neurons
L2 intestine
L2 panneural
L2 polyA enriched 20dC 14hrs post-L1 N2
L2 reference (mockIP)
L3 25dC 25hrs post-L1 N2
L3-L4 dopaminergic neuron
L3-L4 hypodermal cells
L3-L4 PVD & OLL neurons
L3-L4 reference (mockIP)

L4 25dC 36hrs post-L1 N2
late embryo 20dC 6-12hrs post-fertilization N2
male L4 25dC 36hrs post-L1 CB4689
soma-only mid-L4 25dC 36hrs post-L1 JK1107
young adult 25dC 42hrs post-L1 N2
Young Adult Cephalic sheath (CEPsh)
Young Adult reference (mockIP)
embryo BAG neurons*
embryo PVC neurons*
embryo pharyngeal muscle*
early embryo 20dC 0-4hrs post-fertilization*
pathogen S marcescens 25dC 24hr exposure post-adulthood
pathogen S marcescens 25dC 48hr exposure post-adulthood
pathogen E faecalis 25dC 24hr exposure post-adulthood
non-pathogen control 25dC 24hr exposure post-adulthood
non-pathogen control 25dC 48hr exposure post-adulthood
pathogen P luminscens 25dC 24hr exposure post-adulthood

* Four samples were not included in ncRNA prediction and further analysis of ncRNAs because they were released after the ncRNA companion paper were submitted.

Supplement Table 3 : Summary of cell and stage specific tiling array results

Feature class	FDR	Samples	# of features WS199[1]	% of features WS199[2]
Annotated exons (unique) of coding genes overlapping with nrTARs[3]		cells & stages	119,521 exons	87.1% (137,193)
		cells	116,929 exons	
		stages	100,658 exons	
Annotated coding genes with exons overlapping with nrTARs		cells & stages	18,183 genes	91.3% (19,912)
		cells	18,049 genes	
		stages	15,400 genes	
Exons of integrated transcript models (unique) overlapping with nrTARs		cells & stages	138,433 exons	87.8% (157,612)
		cells	135,654 exons	
		stages	116,799 exons	
Integrated transcript models with exons overlapping with nrTARs		cells & stages	19,325 genes	88.8% (21,774)
		cells	19,173 genes	
		stages	16,152 genes	
Gene models detected[4]	5%	cells & stages	17,452 genes	87.7% (19,912)
	5%	cells	17,075 genes	
	5%	stages	15,822 genes	
Gene models detected (FDR-corrected)[5]	0.14%	cells & stages	14,279 genes	71.7% (19,912)
	0.17%	cells	13,149 genes	
	0.71%	stages	13,713 genes	
Gene models differentially expressed (at least 2 fold)[6]	5%	cells & stages	13,320 genes	66.9 % (19,912)
	5%	cells	10,598 genes	
	5%	stages	9,552 genes	
Gene models differentially expressed (at least 2 fold) (FDR-corrected)[7]	0.11%	cells & stages	11,299 genes	56.7 % (19,912)
	0.20%	cells	7,983 genes	
	0.24%	stages	8,606 genes	

[1] Protein-coding gene models are as described in WS199. Overlapping features were merged to produce a total of 19,912 gene models.
[2] Experimental results were calculated for 19,181 genes on the Affymetrix *C. elegans* 1.0R Tiling Array with $\geq$ 3 nonrepetitve exon probes. % of features is based on the total # of genes in WS199 (19,912) which is substantially similar to WS190 (20,121).
[3] non-redundant Transcriptionally Active Regions (nrTARs): Contiguous stretch of nucleotides all of which are inclusive to a TAR detected in $\geq$ 1 of the samples.
[4] The False Discovery Rate (FDR) of 5% was calculated for each sample independently and the total number of genes tabulated from the union of these results.
[5] Correction for potential accumulation of false positives arising from multiple testing. The FDR of each sample (5%) was divided by the cumulative number of samples for each category considered: cells & stages = 37; cells = 30; stages = 7.
[6] The False Discovery Rate (FDR) of 5% was calculated for each independent comparison and the total number of genes tabulated from the union of these results.
[7] Correction for accumulation of false positives arising from multiple testing. The FDR of each sample (5%) was divided by the cumulative number of comparisons for each category considered: cells & stages = 46; cells = 25; stages = 21. (see supplemental methods for this table)

Supplement Table 4 : Different types of known ncRNAs in the gold standard set

Type	Number
rRNA	19
scRNA	1
snRNA	94
snlRNA	4
snoRNA	139
tRNA*	630
miRNA	174
Total	**1061**

The miRNAs are collected from miRBase 14, and other ncRNAs are collected from WormBase 200.
*24 tRNAs are from Mitochondria.

Supplement Table 5 : Annotated regions used for the training of machine learning methods – tiling array TARs

	Transcribed regions overlapped with confirmed annotations (Training Set)		
	CDS	UTR	known ncRNA[d]
	97.4%[a]	86.7%[a]	58.9%[a]
Total number of bases	5,117,511 (9,714,480)[b]	2,682,448 (7,498,856)[b]	51,928 (181,034)[b]
Total number of windows	51,721 (14,230)[b]	27,084 (9,854)[b]	489 (225)[b]
Number of windows with known 2" structure[c]	318 (183)[b]	320 (201)[b]	305 (160)[b]

[a] Fraction of annotated elements overlapped with tiling array TARs

[b] Values in the parenthesis are counted for the TARs, from which the fragmented windows are derived.

[c] Predicted with RNA secondary structure models from Rfam

[d] This is just the gold standard set and doesn't include any unconfirmed ones. Only 10% of known ncRNA were sampled because of large number of annotated tRNAs in the gold standard set.

Supplement Table 6 : Performance of our integrated method on tiling array TARs[a] with three different ways to define element classes in the gold-standard set

Class definition 1		Class definition 2		Class definition 3	
Element class	AUC	Element class	AUC	Element class	AUC
ncRNA	0.9718	ncRNA	0.9246	ncRNA	0.9418
Coding exon	0.9718	Coding exon	0.7485	Coding exon	0.7361
		3' UTR	0.7448	5' and 3' UTR	0.7315

[a]The minimum length of a TAR is 100nt. Large TARs are binned into 100nt windows with a step size of 75nt.

Supplement Table 7 : Annotated and novel tiling array TARs going into 21K-set of ncRNAs

	Transcribed regions overlapped with annotated exons or known ncRNA (Confirmed and predicted)		Novel transcribed regions		
	Exon (81.3%) [a]	known ncRNA (13.1%) [a]	CDS-like [b]	UTR-like [b]	ncRNA-like [b]
Total number of bases	32,744,074 (33,532,732) [c]	265,250 (640,719) [c]	45,208 (134,041) [c]	368,771 (1,048,017) [c]	4,352,048 (6,503,326) [c]
Total number of windows	396,551 (77,131) [c]	2,547 (1,331) [c]	441 (194) [c]	3,294 (1,983) [c]	45,913 **(21,521)** [c]
Number of windows with known secondary structure) [d]	7,314 (3,988) [c]	961 (519) [c]	26 (19) [c]	152 (113) [c]	3,537 (2,083) [c]

[a] Fraction of annotated elements overlapped with tiling array TARs.

[b] In the prediction, if the probability of being ncRNA is larger than 0.009 but less than 0.297, it is ncRNA-like; if the probability of being a UTR is larger than 0.297 or less than 0.692, it is UTR-like; otherwise, if the probability of being CDS is larger than 0.692, it is CDS-like. The cut-offs are determined from the ROC curves.

[c] The long TARs are fragmented into small windows, and values in the parenthesis are counted for the original TARs.

[d] Secondary structure is predicted from Rfam/INFERNAL.

Supplement Table 8 : Total mapped reads, numbers of peaks bound by each of 23 factors (22 TFs and one dosage compensation factor, 28 experiments in total) from ChIP-seq, and numbers of targeted coding and non-coding genes.

	# of Peaks			# of Targeted Coding Genes		# of Targeted ncRNAs		# of Total Mapped Reads	
	Proximal to TSS	Distal to TSS	Unassigned	Targets	Targets (extended)	Targets	Targets (extended)	GFP	Input
ALR-1_L2	737	470	940	908	468	23	21	3,746,542	2,506,542
BLMP-1_L1	2254	1485	2361	2493	1218	80	34	13,699,035	7,832,710
CEH-14_L2	719	211	240	852	234	56	9	4,369,374	1,124,270
CEH-30_LE	895	320	325	989	358	153	12	6,915,024	6,288,570
EGL-27_L1	401	222	302	394	231	92	6	3,402,816	2,862,812
EGL-5_L3	607	375	735	698	392	58	10	2,970,537	1,861,526
ELT-3_L1	1158	629	953	1309	580	49	21	5,558,439	6,612,443
EOR-1_L3	1956	834	1095	2210	767	205	22	2,386,942	3,327,484
GEI1-1_L4	230	75	184	167	73	116	6	2,744,559	4,498,845
HLH-1_MxE	436	329	449	449	343	61	15	4,052,296	2,488,302
LIN1-1_L2	365	110	117	423	117	52	6	2,942,539	4,563,448
LIN1-3_MxE	1018	306	285	1193	324	132	10	5,108,200	9,056,899
LIN-15B_L3	1827	506	519	2236	514	156	20	2,024,367	6,045,335
LIN-39_L3	1197	709	1362	1290	656	147	19	3,399,898	1,993,494
MAB-5_L3	675	307	478	762	333	74	8	3,517,148	3,568,848
MDL-1_L1	2501	1058	1393	2686	918	330	18	4,134,371	4,264,998
MEP-1_MxE	1716	519	629	1985	525	210	17	4,239,180	5,082,534
PES-1_L4	1718	628	914	1990	607	217	28	3,417,784	2,630,081
PHA-4_MxE	2185	920	1259	2439	824	253	32	7,719,682	9,994,939
PHA-4_L1	2333	1264	1866	2592	1111	179	39	15,222,883	11,556,011
PHA-4_L2	1927	924	1300	2210	823	139	30	4,593,131	2,284,558
PHA-4_LE	2609	1169	1655	2907	1003	278	45	4,574,629	6,295,331
PHA-4_StvL1	1263	726	1112	1438	684	78	28	17,845,198	26,819,222
PHA-4_YA	233	111	207	259	114	37	9	5,123,545	10,555,219
PQM-1_L3	1242	724	1001	1253	616	86	28	2,626,971	6,505,184
SKN-1_L1	1751	586	719	1987	598	232	19	4,517,511	2,474,805
UNC-130_L1	284	117	206	333	125	36	3	3,174,776	5,312,775
DPY-27_MxE*	123	-		-	-	-	-	2,074,238	7,578,449

The peaks are overlapped from two replicas, and determined by Peakseq (q value <= 0.001). A proximal peak is defined as 500 bp upstream or 300 bp downstream of transcription start site (TSS) of a gene. This is used by default for the following analysis of TF targets. If a binding peak is located within 501-2000 bp upstream of the TSS sites, it is a distal peak. The numbers of extended targets bound by distal peaks are also reported here. The rest of the peaks are not assigned to any genes, because it is more than 2000 nt away from any TSS sites.The promoter and start regions of 20,069 protein-coding genes (derived from 27,310 coding transcripts) were used to map the targeted genes. 1,061 annotated ncRNAs were collected from Wormbase 200 and miRBase 14.

*Dosage compensation factor. No distance to TSS is calculated.

MxE: mixed embryo; LE: late embryo; StvL1: starved L1 YA: young adult

Supplement Table 9 : GO analysis of genes associated with HOT regions

GO ID	Name	P-value	Sample frequency	Background frequency	Genes
0040007	growth	1.49E-19	82/153 (53.6%)	2845/15340 (18.5%)	rps-22 rps-12 Y87G2A.1 hsp-1 rps-25 lin-54 rps-28 F36A2.7 rps-24 pfd-1 dpy-23 Y82E9BR.3 glit-1 vps-32.1xbx-5 sys-1 rpl-5 wrt-5 rpl-43 R11D1.9 Y65B4BR.5 K12H4.5 Y49A3A.1 Y71H2AM.5 rps-1 K10D2.5 puf-9 taf-4rpl-3 hsr-9 eif-3.B epc-1 F17C11.9 W04A4.5 mei-2 K04G7.1 rps-30 ash-2 wip-1 H28O16.1 Y48A6B.3 ZK550.3cco-2 rpl-13 his-37 mbk-1 set-16 vha-8 kbp-4 cap-2 nipi-3 C34C12.2 F48C1.4 pbs-2 sor-1 dpm-3 T23F11.1eft-4 rpl-6 rpl-7 ekl-4 rpl-32 rpl-22 E02D9.1 emo-1 atad-3 nuo-1 LLC1.3 cct-1 Y51H4A.15 cco-1 rpn-3 rps-26rpl-24.1 rpl-14 prp-8 mdt-19 rpl-35 rfc-4 mdt-26 htz-1 eft-2
0009792	embryo development ending in birth or egg hatching	1.67E-17	82/153 (53.6%)	3054/15340 (19.9%)	rps-22 rps-12 cbp-1 atx-2 hsp-1 rps-25 lin-54 F36A2.7 F40F11.2 pfd-1 dpy-23 Y82E9BR.3 vps-32.1 sys-1 rpl-5rpl-43 R11D1.9 Y65B4BR.5 K12H4.5 Y49A3A.1 T08B2.11 Y71H2AM.5 rps-1 taf-4 rpl-3 eif-3.B epc-1 F17C11.9W04A4.5 mei-2 K04G7.1 ile-2 daf-21 wwp-1 ash-2 wip-1 H28O16.1 hsp-60 ZK550.3 klc-1 mdl-1 vig-1 cco-2rpl-13 his-37 pqn-51 set-16 cls-2 tre-1 vha-8 kbp-4 cap-2 F25E2.2 cpt-2 nipi-3 F48C1.4 let-268 pbs-2 eft-4rpl-6 rpl-7 ekl-4 rpl-22 dnj-11 emo-1 atad-3 nuo-1 LLC1.3 cct-1 Y51H4A.15 cco-1 rpn-3 rps-26 rpl-24.1 rpl-14 prp-8 mdt-19 rpl-35 rfc-4 mdt-26 htz-1 eft-2
0005737	cytoplasm	7.43E-15	47/153 (30.7%)	1130/15340 (7.4%)	rps-22 rps-12 cbp-1 atx-2 hsp-1 egl-30 trap-3 rps-28 rps-24 ddp-1 pfd-1 dpy-23 sys-1 rpl-5 rpl-43 R11D1.9Y71H2AM.5 rps-1 puf-9 rpl-3 ain-1 F17C11.9 rps-30 daf-21 wwp-1 hsp-60 eat-16 deb-1 rpl-13 cls-2 vha-8 tra-4cap-2 unc-108 let-268 eft-4 rpl-6 rpl-7 rpl-32 rpl-22 nuo-1 LLC1.3 cco-1 rps-26 rpl-24.1 rpl-14 rpl-35
0005840	ribosome	1.65E-13	19/153 (12.4%)	141/15340 (0.9%)	rps-22 rps-12 rps-28 rps-24 rpl-5 rpl-43 R11D1.9 rps-1 rpl-3 rps-30 rpl-13 rpl-6 rpl-7 rpl-32 rpl-22 rps-26 rpl-24.1 rpl-14 rpl-35

Supplement Table 10 : Expression correlation of tanscription factors with target genes and non-target genes.

TF	Target	non-Target	Z-score	P-value
ALR-1	-0.007457	-0.008889	0.129717	0.896799
BLMP-1	-0.066699	-0.041505	-2.489284	0.012836
CEH-14	-0.096922	0.006595	-11.395537	0
EGL-27	0.128903	-0.034187	12.274861	0
EGL-5	0.0441	0.007088	5.185788	0
ELT-3	0.014695	-0.019066	3.251677	0.001165
EOR-1	0.133573	-0.042516	21.571089	0
GEI-11	0.068925	-0.005121	2.314903	0.021352
LIN-11	-0.081399	-0.048996	-2.416469	0.015807
LIN-15B	0.132488	-0.047239	21.460631	0
LIN-39	0.072205	-0.013786	11.372508	0
MAB-5	0.026751	0.009732	1.416596	0.156752
MDL-1	-0.0611	0.037063	-17.003577	0
PES-1	0.195083	-0.037515	20.789856	0
PHA-4	0.016003	-0.052655	16.739494	0
PQM-1	0.312876	0.016214	32.681676	0
SKN-1	0.090011	-0.017028	29.322469	0
UNC-130	-0.080614	-0.028788	-1.698375	0.090283

For each TF, the Pearson correlation coefficents of the expression level of the TF with those of its target genes and non-target genes were calculated across the 7 developmental stage time aourse. The significance of difference between target and non-target genes were calculated using t-test.

Supplement Table 11 : Overview of PicTar-predicted miRNA target sites within 3'UTRs of the ag1003 transcript set (see text for details).

	3 species conservation	**5 species conservation**
Number of miRNAs analyzed	183	183
Number of 3'UTRs analyzed	25,539	25,539
Number of genes analyzed	14,519	14,519
Number of target sites detected	20,427	8,810
Number of 3'UTRs with target sites	4,866	2,406
Number of genes with target sites	2,349	1,162
Number of miRNAs that target a 3'UTR	182	178

Supplement Table 12 : Overlap of 3.4 Mb of residual constrained blocks with various genomic elements.

Genomic elements	observed base pair overlap	expected by GSC simulation	p-value
Introns	0.49	0.35	1e-34
Intra-genic regions	0.26	0.16	1e-34
1000 bp upstream of gene TSS	0.19	0.18	1e-6
1000 bp downstream genic regions	0.19	0.25	1e-15

Supplement References

1. A. Agarwal *et al.*, Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**, 383 (2010).
2. D. Kampa *et al.*, Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331-342 (2004).
3. T. E. Royce *et al.*, Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **21**, 466-475 (2005).
4. L. W. Hillier *et al.*, Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.* **19**, 657-666 (2009).
5. C. H. Jan, R. C. Friedman, J. G. Ruby, C. B. Burge, D. P. Bartel, Formation and regulation of 3′ untranslated regions in Caenorhabditis elegans. *Manuscript in preparation*, (2010 (manuscript in preparation)).
6. B. Ewing, A. Leahey, L. Hillier, C. Davis, P. Green, Targeted closure of the C. elegans transcriptome (Green companion). *In preparation*, (in preparation).
7. G. E. Merrihew *et al.*, Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.* **18**, 1660-1669 (2008).
8. M. Mangone *et al.*, The Landscape of C. elegans 3'UTRs. *Science* **329**, 432-435 (2010).
9. W. C. Spencer *et al.*, A spatial and temporal map of C. elegans gene expression. *In preparation*, (in preparation).
10. M. Christensen *et al.*, A primary culture system for functional analysis of C-elegans neurons and muscle cells. *Neuron* **33**, 503-514 (2002).
11. R. M. Fox *et al.*, A gene expression fingerprint of C-elegans embryonic motor neurons. *BMC Genomics* **6**, 42 (2005).
12. S. E. Von Stetina *et al.*, Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C-elegans nervous system. *Genome Biol.* **8**, R135 (2007).
13. G. Zeller, S. R. Henz, S. Laubinger, D. Weigel, G. Ratsch, Transcript normalization and segmentation of tiling array data. *Pac Symp Biocomput*, 527-538 (2008).
14. S. Laubinger *et al.*, At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biol.* **9**, R112 (2008).
15. R. A. Irizarry *et al.*, Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, E15 (2003).
16. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
17. L. Gautier, L. Cope, B. M. Bolstad, R. A. Irizarry, affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004).
18. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).
19. C. E. Bonferroni, Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, 13–60 (1935).
20. G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, 3 (2004).
21. G. K. Smyth, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor,* R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber, Eds. (Springer, New York, 2005), pp. 397-420.
22. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
23. T. L. Bailey, N. Williams, C. Misleh, W. W. Li, MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369-W373 (2006).
24. J. Du, ..., M. B. Gerstein, IQSeq: Integrated Isoform Quantification Analysis Based on Next-generation Sequencing Data. *In preparation*, (in preparation).
25. documented source code for Deepseq9 algorithm at SourceForge, http://deepseq9.sourceforge.net.
26. Supplemental files are available at http://www.modencode.org/publications/integrative_worm_2010/.

27. M. R. Friedlander *et al.*, Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407-415 (2008).
28. M. Friedlander, S. Mackowiak, N. Rajewsky, miRDeep 2.0. (in preparation).
29. M. Kato, A. de Lencastre, Z. Pincus, F. J. Slack, Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol.* **10**, R54 (2009).
30. J. G. Ruby, C. H. Jan, D. P. Bartel, Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).
31. K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, E. C. Lai, The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**, 89-100 (2007).
32. Chung, E. C. Lai, modENCODE mirtron companion paper. *Genome Res.*, (submitted).
33. P. J. Batista *et al.*, PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C-elegans. *Mol. Cell* **31**, 67-78 (2008).
34. J. M. Claycomb *et al.*, The Argonaute CSR-1 and Its 22G-RNA Cofactors Are Required for Holocentric Chromosome Segregation. *Cell* **139**, 123-134 (2009).
35. C. C. Conine *et al.*, Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3588-3593 (2010).
36. E. de Wit, S. E. V. Linsen, E. Cuppen, E. Berezikov, Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res.* **19**, 2064-2074 (2009).
37. J. I. Gent *et al.*, Distinct Phases of siRNA Synthesis in an Endogenous RNAi Pathway in C. elegans Soma. *Mol. Cell* **37**, 679-689 (2010).
38. J. I. Gent *et al.*, A Caenorhabditis elegans RNA-Directed RNA Polymerase in Sperm Development and Endogenous RNA Interference. *Genetics* **183**, 1297-1314 (2009).
39. J. G. Ruby *et al.*, Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C-elegans. *Cell* **127**, 1193-1207 (2006).
40. M. Stoeckius *et al.*, Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat. Methods* **6**, 745-U716 (2009).
41. J. C. van Wolfswinkel *et al.*, CDE-1 Affects Chromosome Segregation through Uridylation of CSR-1-Bound siRNAs. *Cell* **139**, 135-148 (2009).
42. Full analysis of worm mirtrons, http://cbio.mskcc.org/leslielab/mirtrons/ce6/mirtron.reports.ce.20100719.html.
43. Z. J. Lu *et al.*, Prediction and characterization of non-coding RNAs in C. elegans by integrating conservation, secondary structure and high throughput sequencing and array data. *Submitted to Genome Research*, (in preparation).
44. M. Zhong *et al.*, Genome-Wide Identification of Binding Sites Defines Distinct Functions for Caenorhabditis elegans PHA-4/FOXA in Development and Environmental Response. *PLoS Genet.* **6**, e1000848 (2010).
45. H. Lei *et al.*, A widespread distribution of genomic CeMyoD binding sites revealed and cross-validated by ChIP-chip and ChIP-Seq techniques. *PLoS One*, (submitted).
46. W. Niu *et al.*, Systematic dissection of regulatory networks dictated by *C. elegans* sequence-specific transcription factors. *In preparation*, (in preparation).
47. C. Cheng, K. Yan, ... Integrated regulatory network analysis in C. elegans. *to be submitted*, (2010).
48. S. M. Johnson, S. Y. Lin, F. J. Slack, The time of appearance of the C-elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259**, 364-379 (2003).
49. S. Ercan *et al.*, X chromosome repression by localization of the C-elegans dosage compensation machinery to sites of transcription initiation. *Nature Genet.* **39**, 403-408 (2007).
50. P. J. Roy, J. M. Stuart, J. Lund, S. K. Kim, Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* **418**, 975-979 (2002).
51. C. A. Grove *et al.*, A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. *Cell* **138**, 314-327 (2009).
52. F. Pauli, Y. Y. Liu, Y. A. Kim, P. J. Chen, S. K. Kim, Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C-elegans. *Development* **133**, 287-295 (2006).
53. V. Reinke, I. S. Gil, S. Ward, K. Kazmer, Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. *Development* **131**, 311-323 (2004).

54. J. S. Gilleard, Y. Shafi, J. D. Barry, J. D. McGhee, ELT-3: A Caenorhabditis elegans GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.* **208**, 265-280 (1999).
55. J. Gaudet, S. E. Mango, Regulation of organogenesis by the Caenorhabditis elegans, FoxA protein PHA-41. *Science* **295**, 821-825 (2002).
56. R. S. Kamath *et al.*, Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-237 (2003).
57. A. Krek *et al.*, Combinatorial microRNA target predictions. *Nature Genet.* **37**, 495-500 (2005).
58. S. Lall *et al.*, A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.* **16**, 460-471 (2006).
59. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-D158 (2008).
60. AceView, http://www.aceview.org.
61. B. P. Lewis, C. B. Burge, D. P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).
62. A. Stark, J. Brennecke, N. Bushati, R. B. Russell, S. M. Cohen, Animal microRNAs confer robustness to gene expression and have a significant impact on 3 ' UTR evolution. *Cell* **123**, 1133-1146 (2005).
63. N. Rajewsky, microRNA target predictions in animals. *Nature Genet.* **38**, S8-S13 (2006).
64. M. Selbach *et al.*, Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63 (2008).
65. M. Hall *et al.*, The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**, 10-18 (2009).
66. C. elegans Sequencing Consortium, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
67. P. Vaglio *et al.*, WorfDB: the Caenorhabditis elegans ORFeome database. *Nucleic Acids Res.* **31**, 237-240 (2003).
68. Integrative Genomics Viewer, http://www.broadinstitute.org/igv/v1.2.
69. SHRiMP aligner v1.3, http://compbio.cs.toronto.edu/shrimp.
70. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351-1359 (2008).
71. X. Liu *et al.*, Analysis of Cell Fate from Single-Cell Gene Expression Profiles in C. elegans. *Cell* **139**, 623-633 (2009).
72. E. Portales-Casamar *et al.*, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105-D110 (2010).