Integrative Analysis of Functional Elements in the *Caenorhabditis elegans* Genome by the modENCODE Project

Mark B. Gerstein[1,2,3,*,#], Zhi John Lu[1,2,*], Eric L. Van Nostrand[4,*], Chao Cheng[1,2,*], Bradley I. Arshinoff[5,6,*], Tao Liu[7,8,*], Kevin Yip[1,2,*], Rebecca Robilotto[1,*], Andreas Rechtsteiner[9,*], Kohta Ikegami[10,*], Pedro Alves[1,*], Aurelien Chateigner[11,*], Marc Perry[5,*], Mitzi Morris[12,*], Raymond K. Auerbach[1,*], Anne Vielle[13,*], Wei Niu[14,15,*], Kahn Rhrissorrakrai[12,*], Ashish Agarwal[2,3], Julie Ahringer[13], Roger P. Alexander[1,2], Galt Barber[16], Cathleen M. Brdlik[4], Jennifer Brennan[10], Adrian Carr[11], Ming-Sin Cheung[13], Hiram Clawson[16], Sergio Contrino[11], Luke O. Dannenberg[17], Abby F. Dernburg[18], Arshad Desai[19], Lindsay Dick[10], Andréa C. Dosé[18], Jiang Du[3], Thea Egelhofer[9], Sevinc Ercan[10], Ghia Euskirchen[14], Brent Ewing[20], Elise A. Feingold[21], Xin Feng[5,22], Reto Gassman[19], Peter J. Good[21], Phil Green[20], Francois Gullier[11], Mark S. Guyer[21], Lukas Habegger[1], Ting Han[23], Jorja G. Henikoff[24], Stefan R. Henz[25], Angie Hinrichs[16], Heather Holster[17], Tony Hyman[26], David M. Miller III[27], A. Leo Iniguez[17], Judith Janette[15], Morten Jensen[10], Masaomi Kato[28], W. James Kent[16], Ellen Kephart[5], Vishal Khivansara[23], Ekta Khurana[1,2], John K. Kim[23], Paulina Kolasinska-Zwierz[13], Mitzi I. Kuroda[29], Eric C. Lai[30], Isabel Latorre[13], Amber Leahey[20], Jing Leng[1], Suzanna Lewis[31], Paul Lloyd[5], Lucas Lochovsky[1], Rebecca F. Lowdon[21], Yaniv Lubling[32], Rachel Lyne[11], Michael MacCoss[20], Sebastian D. Mackowiak[33], Marco Mangone[12], Sheldon McKay[34], Gennifer Merrihew[20], Andrew Muroyama[19], John I. Murray[20], Siew-Loon Ooi[24], Hoang Pham[18], Taryn Phippen[9], Elicia A. Preston[20], Nikolaus Rajewsky[33], Gunnar Ratsch[25], Heidi Rosenbaum[17], Joel Rozowsky[1,2], Kim Rutherford[11], Peter Ruzanov[5], Mihail Sarov[26], Rajkumar Sasidharan[2], Andrea Sboner[1,2], Eran Segal[32], Hyunjin Shin[7,8], Frank J. Slack[28], Cindie Slightam[35], Richard Smith[11], William C. Spencer[27], E.O. Stinson[31], Scott Taing[7], Teruaki Takasaki[9], Dionne Vafeados[20], Ksenia Voronina[19], Guilin Wang[15], Nicole L. Washington[31], Christina Whittle[10], Beijing Wu[35], Koon-Kiu Yan[1,2], Georg Zeller[25], Zheng Zha[5], Mei Zhong[14], Xingliang Zhou[10], modENCODE Consortium, Susan Strome[9,#], Kristin C. Gunsalus[12,36,#], Gos Micklem[11,#], X. Shirley Liu[7,8,#], Valerie Reinke[15,#], Stuart K. Kim[35,4,#], LaDeana W. Hillier[20,#], Steven Henikoff[24,#], Fabio Piano[12,36,#], Michael Snyder[4,14,#], Lincoln Stein[34,5,6,#], Jason D. Lieb[10,#], Robert H. Waterston[20,#]

[1] Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA.
[2] Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA.
[3] Department of Computer Science, Yale University, 51 Prospect St, New Haven, Connecticut 06511, USA.
[4] Department of Genetics, Stanford University Medical Center, Stanford CA 94305, USA.
[5] Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto Ontario M5G 0A3, Canada.

[6] Department of Molecular Genetics, University of Toronto, 27 King's College Circle,
Toronto, Ontario M5S 1A1, Canada.
[7] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
44 Binney St, Boston, 02115, USA.
[8] Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave,
Boston, 02115, USA.
[9] Molecular, Cell, and Developmental Biology, University of California Santa Cruz,
Santa Cruz, California 95064, USA.
[10] Department of Biology and Carolina Center for Genome Sciences, The University of
North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
[11] Department of Genetics, University of Cambridge, CB2 3EH, UK; Cambridge Systems
Biology Centre, Tennis Court Road, Cambridge CB2 1QR, UK.
[12] Center for Genomics and Systems Biology, Department of Biology, New York
University, 1009 Silver Center, 100 Washington Square East, New York, New York
10003-6688, USA.
[13] Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge,
Tennis Court Road, Cambridge CB2 1QN, UK.
[14] Department of Molecular, Cellular and Developmental Biology, Yale University, New
Haven, Connecticut 06824, USA.
[15] Department of Genetics, Yale University School of Medicine, New Haven, Connecticut
06520-8005, USA.
[16] Department of Biomolecular Engineering, University of California, Santa Cruz,
California 95064 USA.
[17] Roche NimbleGen, Inc., 500 South Rosa Road, Madison, Wisconsin 53719, USA.
[18] Howard Hughes Medical Institute; Department of Molecular and Cell Biology,
University of California, Berkeley; Life Sciences Division, Lawrence Berkeley National
Laboratory, USA.
[19] Ludwig Inst. Cncr. Res/Dept. of Cellular & Molecular Medicine, University of
California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0653, USA.
[20] Department of Genome Sciences and University of Washington School of Medicine,
William H. Foege Bldg. S350D, 1705 N.E. Pacific Street, Box 355065 Seattle ,
Washington 98195-5065, USA.
[21] Division of Extramural Research, National Human Genome Research Institute,
National Institutes of Health, 5635 Fishers Lane, Suite 4076, Bethesda, Maryland 20892-
9305, USA.
[22] Department of Biomedical Engineering, State University of New York at Stonybrook,
Stonybrook, NY 11794, USA.
[23] Life Sciences Institute, Department of Human Genetics, University of Michigan, 210
Washtenaw Avenue, Ann Arbor, Michigan 48109-2216, USA.
[24] Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview
Avenue North, Seattle, Washington 98109, USA.
[25] Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany.
[26] Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108,
01307 Dresden, Germany.
[27] Department of Cell and Developmental Biology, Vanderbilt University, 465 21st
Avenue South, Nashville, Tennessee 37232-8240, USA.

[28] Department of Molecular, Cellular and Developmental Biology, PO Box 208103, Yale University, New Haven, Connecticut 06520, USA.
[29] Department of Medicine, Brigham & Women's Hospital, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.
[30] Sloan-Kettering Institute, 1275 York Avenue, Box 252, New York, New York 10065, USA.
[31] Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mailstop 64-121, Berkeley, California 94720 USA.
[32] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel.
[33] Max-Delbrück-Centrum für Molekulare Medizin (MDC), Division of Systems Biology, Robert-Rössle-Str. 10, D-13125 Berlin-Buch, Germany.
[34] Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11542 USA.
[35] Department of Developmental Biology, Stanford University Medical Center, 279 Campus Drive, Stanford, California 94305-5329, USA.
[36] New York University, Abu Dhabi, United Arab Emirates.

[*] Joint First Authors

[#] Corresponding Authors

Abstract

We describe the systematic acquisition of genome-wide data sets that have led to a much improved annotation of the *C. elegans* genome. These include a full transcriptome analysis over a developmental time course, genome-wide identification of transcription-factor binding sites, and high-resolution maps of chromatin organization. Integrative analysis of these data allowed us to provide more complete and accurate gene models, including alternative splicing, and to find new noncoding RNAs. We identified chromosomal locations bound by an unusually large number of transcription factors and described hierarchical networks of transcription-factor-binding and microRNA interactions. We found striking differences in chromatin composition and histone modification between the arms and centers of chromosomes, and similarly prominent differences between the autosomes and the X chromosome. We integrated these data to build statistical models relating chromatin structure to transcription-factor binding and gene expression. Finally, our data can be used to ascribe a putative function to most of the evolutionarily conserved DNA in the genome. The data, materials and associated analyses are publicly available through modencode.org.

Introduction

The ability to completely sequence genomes marked a major advance in biological research, providing for the first time a view of the complete instruction set required to direct the development and behavior of an organism. However, the ability to comprehend the functional content of a genome by DNA sequence alone is limited. Accurate and comprehensive annotation requires direct experimental evidence. To address this need, in 2003 the United States National Human Genome Research Institute (NHGRI) initiated a pilot project (called ENCODE, for ENcyclopedia of DNA Elements) focusing initially on 1% of the human genome and now the whole genome (*1*). Recognizing the importance of well-annotated genomes for experimental systems and the need to subject candidate functional elements to rigorous testing, in 2007 the NHGRI initiated the modENCODE project (*mod*el organism ENCODE) on *Caenorhabditis elegans* and *Drosophila melanogaster*. The project aimed to systematically annotate the functional genomic elements in these key model organisms (*2*).

The nematode *C. elegans* offers a critical perspective on genome organization and function, given its intermediate complexity between single-cell eukaryotes, such as yeast, and highly complex organisms, such as humans. Following Brenner's pioneering work describing its genetics (*3*), *C. elegans* became the first multicellular animal with a fully defined cell lineage, the first with a fully reconstructed nervous system by serial electron microscopy, and the first with a fully sequenced genome (*4-6*). It has also contributed to the discoveries of apoptosis (*7*), RNA interference (RNAi) (*8*), and gene regulation by microRNAs (miRNAs)(*9, 10*). As biology moves toward a more complete description of the molecular basis of development and behavior, the relevance of *C. elegans* to understanding biology continues. Its 100.3 Mb genome is only 8 times larger than that of the single-celled yeast *S. cerevisiae*, and yet it contains almost as many genes as human and all of the information necessary to specify the major tissues and cell types of metazoans.

Despite the fact that the *C. elegans* genome is considered fairly well annotated, many key genomic features remain poorly defined. In particular, before the beginning of the project, there was a lack of experimentally verified information about protein-coding gene organization, including precise starts, stops, intron/exon boundaries, and alternative splicing events. The universe of noncoding RNAs (ncRNAs) had only been partially explored. As is the case in most organisms, even less was known about regulatory and structural elements that control gene expression and chromosomal organization. To address this, starting in 2007, five worm modENCODE groups began collecting genome-wide data (*2*). Here, we present our progress to date. In particular, our analysis reveals:

* An extensive set of directly supported protein-coding genes containing alternative splice junctions, 5′ and 3′ ends.

* An initial set of non-coding RNAs, which includes new RNAs belonging to both known classes and novel types.

* The dynamics of gene expression and transcription factor binding across many developmental stages, showing coordinated changes between binding and expression.

* The discovery of genomic locations bound by many of the transcription factors analyzed, which we call “HOT” (Highly Occupied Target) regions.

*A hierarchy of candidate regulatory interactions amongst the transcription factors analyzed, which could, in turn, be related to the network of connections between microRNAs and their targets.

* Striking differences in histone modifications between the centers and arms of autosomes in somatic cells that correlate with nuclear envelope interactions and germline function.

* Specific histone modifications (including H4K20me1) and other features of chromatin organization associated with the X chromosome.

* Evidence for trans-generational transmission of the pattern of germline gene expression from mother to daughter through a chromatin-mediated mechanism.

* Models of the chromatin state around genes that provide statistical predictions of both transcription-factor binding and gene expression. Models developed for protein-coding genes are directly applicable to microRNA encoding genes.

The summation of features annotated through these functional data sets now provides a plausible explanation for most of the conserved sequences in the *C. elegans* genome. The genome-wide view of critical gene and chromosome features revealed by the project lays the foundation for the study of how the genome of this multicellular organism accurately directs development and maintains homeostasis.

Data Overview

The *C. elegans* modENCODE data sets span the domains of gene structure, RNA expression profiling, chromatin structure and regulation, and evolutionary conservation. Many data sets were collected across a standardized developmental time course consisting of all the major stages of the life cycle (embryos, adults, and the four larval stages: L1, L2, L3, and L4) to facilitate integrated analysis (Fig. 1). Experiments were generally carried out using the standard laboratory strain N2 under normal conditions (*3*). Mutants were used when needed for specific experiments, for example to obtain populations enriched in dauers or males. In total, over 50 distinct combinations of stages, tissues, cells, genetic backgrounds, and environmental conditions have been assessed.

To enable the analyses presented here, the data from completed projects were 'frozen' in February 2010, at which time we had released 237 *C. elegans* data sets (Fig. 1A) . Over 4.2 billion sequencing reads covering 109 billion bases from 65 experiments were used for gene structure determination, RNA profiling, and mapping protein-genome interactions. Whole-genome tiling microarrays were used to detect the results of 128 chromatin immunoprecipitation (ChIP) and 44 RNA profiling experiments. To facilitate the integration of results from both sequencing and tiling arrays, we standardized ways in which the interpreted results of these technologies are reported, thereby allowing them to be used as merged data sets (see supplement B, Fig. S1 and Fig. S2) (*11*). Also part of the modENCODE corpus are per-base evolutionary constraint scores, generated from a six-way alignment between *C. elegans* and the related nematodes *C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica*, and *P. pacificus*. These conservation data assist with the identification of functional elements and potentially help to distinguish essential, highly-conserved elements from those that are either recently acquired by *C. elegans*, or are under weak selective pressure.

To ensure the completeness and standardization of modENCODE data, all data sets were, and continue to be, submitted to the modENCODE Data Coordinating Center . Each data set is hand curated to include extensive structured metadata, which is validated for completeness and checked for consistency before public release (*12*). The metadata describe the overall design, reagents and protocols for each experiment. All raw and analyzed data, metadata, and interpreted results can be searched, displayed, and downloaded at modencode.org (*13, 14*). Raw sequencing reads and microarray data are also archived at the Short Read Archive (*15, 16*) and the Gene Expression Omnibus (*16, 17*). All of the results are being incorporated into WormBase (*18, 19*).

Transcriptome Analysis

Accurate and comprehensive annotation of all RNA transcripts (the transcriptome) produced from the *C. elegans* genome is one of the main objectives of the modENCODE project. In addition to their functional importance, transcripts provide a framework for interpreting other genomic features, such as transcription-factor (TF) binding sites and chromatin marks. At the outset of the project, the worm genome lacked direct experimental support for more than one third of predicted splice junctions, and many transcription start sites and most polyA addition sites were not annotated at all. Furthermore, systematic tests of the predicted gene models revealed that many existing annotations were erroneous or incomplete (*20, 21*). Here, we sought cDNA-based evidence from high-throughput sequencing (RNA-seq), RT-PCR/RACE, mass-spectrometry (mass-spec) and tiling arrays to discover previously unrecognized protein-coding genes, precisely define known protein-coding genes, examine the dynamics of expression and alternative splicing, find evidence of transcription of pseudogenes, and begin to define non-coding RNA genes. We also used mass-spectrometry to verify the existence of predicted proteins and to distinguish short single-exon protein-coding transcripts from ncRNAs.

Detecting Features of Protein-Coding Genes

To detect protein-coding transcripts, we relied primarily on RNA-seq (*22-24*). We generated >1 billion uniquely aligned RNA-seq reads from 19 different worm populations, representing all the major life stages and both embryonic and late L4 males, as well as worms exposed to pathogens (Fig. S3). In addition, data sets targeting the 3′ ends of polyA-plus transcripts were collected (*25*), and sequence tags representing polyadenylated 3' ends, which had been acquired using 3P-Seq (polyA-position profiling by sequencing) were made available to the consortium (*26*).

By exhaustively mapping the RNA-seq reads, we were able to detect with nucleotide resolution transcribed exons, 5′ and 3′ UTRs, splice junctions, transpliced leaders and polyA addition sites independent of the pre-existing WormBase models. Relative to WormBase annotations available in January 2007, our results significantly increase the number of experimentally supported transcribed features of all types (Fig. 2A and Fig. S7). For example, the number of supported splice junctions rose from 70,028 in January 2007 to 98,308 with the analysis of the first four stages (*24*) and then to 111,786 with 15 additional data sets (Fig. S8). Importantly, these include 8,174 splice junctions not represented in WormBase gene models at the start of the project, with 2,126 deriving from genes not previously represented in WormBase. Similarly, the number of genes with a transpliced leader (either Splice Leader 1 or Splice Leader 2) at the 5′ end rose from 6,012 in Jan 2007 to 12,413, with those genes showing a total of 20,515 different transplice sites (*27*) and the number of polyA sites associated with genes rose from 1,330 defined in WormBase in January 2007 to 28,199 sites distributed across 15,531 genes.

In addition to RNA-seq, we sought to define protein coding gene features through directed approaches using RT-PCR/RACE targeting regions with predicted but unsupported splice junctions, and mass-spectrometry targeting small single exon genes as well as stage-specific proteins. RT-PCR provided direct support for 37,797 splice junctions and mass-spec proved that >75 single exon genes produced protein products. About 95% of the splice junctions detected by RT-PCR/RACE and mass-spec overlapped with those detected by RNA-seq, thus validating 37,830 of these features (Fig. S9). They also detected features not previously found in the RNA-seq data set. The intersection of these data sets and previous WormBase data indicate that perhaps only 2,000-3,000 exons (2-3%) remain undetected (Fig. S10, Right).

Other lines of evidence also suggest that only a small number of protein coding genes remain to be discovered. For embryonic and early larval stages, we have obtained a number of RNA-seq reads that approaches, if not exceeds, the number of mRNA molecules estimated to be in a single animal (*24*). Further, we have shown that we can detect genes expressed at low levels in single cells, such as *gcy-5* (*24*). For the determination of conventional gene expression, our evidence indicates we are able to sequence to saturation, and consequently have identified most of the reproducibly expressed transcripts.

The yield of new features with each additional RNA-seq sample is diminishing. For example (Fig. 2A), the L4 male and late embryo samples were among the richest sources of splice junctions, but respectively only supported 3,029 and 3,564 additional splice junctions beyond those detected in the previously published four data sets. The dauer exit was the least complex sample and only added 721 newly supported junctions. By evaluating the cumulative increase in features after randomizing the addition of new samples, we can show that many features discovered are approaching saturation (Fig. S10). However, we continue to detect rare events, particularly those associated with more abundantly expressed genes. For example, we find unusual splice junctions, some of which disrupt the reading frame. While these could represent isoforms with important biological function, they also might represent aberrant splice products that occur below the limits of evolutionary selective forces. Transplicing provides another example where, occasionally, splice leader sequences are detected at splice acceptor sites downstream of longer introns (*27*). These types of events complicate assessments of the comprehensiveness of the gene catalog.

What genes remain to be supported? Of the 478 WormBase-predicted genes with known domains for which we did not detect any evidence with RNA-seq, 369 are members of rapidly evolving gene families that have been implicated in environmental responses (Supplement Table 1). Although some of these models may represent pseudogenes (see below), sampling of additional pathogen-treated populations or other environmental exposures may provide support for them.

Building Gene Models

We built likely gene models based solely on the evidence produced by transcript sequencing, allowing for multiple transcripts (isoforms) from a region (*24*). We called these models genelets, because they could include just fragments of full genes. Genelets were initiated with the most highly represented splice junction in a region, and extended in each direction to incorporate regions covered by above-threshold sequence reads and splice junctions, terminating the model when either a transcript start or stop signal was encountered, or when coverage was interrupted (Fig. 2B). By iterating the process, we generated alternative gene isoforms. We then inspected the various genelet transcripts for the longest open reading frame to annotate protein-coding sequences (CDSs) and 5′ and 3′ untranslated regions (UTRs).

For each of the 19 stages and conditions, we built a stage-specific transcript set based solely on the RNA-seq data (stage-specific RNAseq-only genelets). In addition, we built three aggregate sets: (1) aggregate RNAseq-only genelets, based only on the total RNA-seq data; (2) aggregate integrated genelets, which combine the RNA-seq data with available ESTs (expressed sequence tags), cDNAs, and OSTs (open reading-frame sequence tags) from the community (*20, 21, 25*), as well as the RT-PCR/RACE and mass spectrometry data produced in this project; and (3) aggregate integrated transcripts, a set where we allowed WormBase predictions to fill small coverage gaps within exons. The aggregate integrated transcripts incorporate all the splice junctions and all the splice leader sites, as well as multiple polyA addition sites, and thus often contain multiple different isoforms. Altogether, we generated 64,824 transcripts from 21,733 regions (genes), compared to 23,710 transcripts from 20,082 genes in *C. elegans*. Our gene models, based only on direct experimental evidence, match the internal splice junction pattern exactly for 10,123 previous gene models, but even for many of these genes we provide revised 5′ or 3′ ends. For 6,418 models, we have the same internal splice junction pattern as WormBase but add 5′ or 3′ exons and associated splice junctions. For the remaining models, we overlap WormBase models but differ in splice junctions (3,292) or fail to cover all of the splice junctions (2,235).

Expression Dynamics

We sought to determine the dynamics of gene expression over development and in specific cell types. In addition to the RNA-seq data sets, we used tiling array data because arrays allowed us to assay very small amounts of starting RNA (<10 ng) in biological triplicate. We analyzed 44 biological samples, comprising 17 different combinations of growth stages and conditions, and 25 different cell and tissue types (Supplement Table 2). Importantly, for many of the stages, the RNA-seq data were obtained from samples that were verified by the tiling array data sets.

We find that transcripts for more than 95% of genes are detected in more than one stage and that almost half the genes are expressed in every stage (Fig. S11). In contrast, only a small number of genes (~100/stage) have strong stage specific expression (Fig. S12, Fig. S13, and Fig. S14), suggesting that the differences between stages might be more related to modulation in expression levels than the existence of discrete stage-specific genes. We find that ~75% of genes show at least a two-fold difference between a pair of stages or between tissues/cells and the reference (Supplement Table 3).

To further investigate the relationship of gene expression between different stages, we correlated all RNA-seq expression values for each gene at a given stage with all other stages. To simplify this calculation, we focused on a set of 8428 non-overlapping transcripts (see supplement C.7). The resultant correlation matrix (Fig. 3A, Left) clearly shows that the time course subdivides into distinct embryo and larval phases. We also performed principal-components analysis on tiling array data generated from several matched worm tissues sampled during the embryo and L2 stages (see supplement C.8). The embryonic cells were isolated by Fluorescence-Activated Cell Sorting (FACS) of fluorescently tagged cells that had been extracted from dissociated embryos and cultured for 24 hours to allow further differentiation (*28*). The matched tissues from the L2 stage were isolated by precipitation of PolyA Binding Protein. Points representing stages were plotted along the first two principal components (Fig. 3C), representing the linear combination of genes explaining the two greatest sources of variance within the data. The tight clustering of embryo tissues contrasts with that of the more dispersed organization of the matched L2 tissues, which form multiple clusters that are generally distinct from their embryonic counterparts. GABA motor neurons, coelomocytes, and A-class motor neurons differentiate along the same axis, whereas intestine and body-wall muscle differentiate along a different path. Overall, our results suggest the presence of different gene-expression programs in the L2 tissues compared to embryo, consistent with expectations for cell differentiation and specialization the worm must undergo as it develops from an embryo into an adult. However, the clustering could also be related to differences in RNA collection protocols.

Alternative Splicing Dynamics

Alternative splicing provides another mechanism for differential transcript usage. To discover the strongest stage-specific examples of alternative splice forms, we identified sites with two or more splice forms across the same region where the ratio of abundance of the two forms differed by more than five-fold from the beginning to the end of the timecourse (i.e. between embryo and L3 or YA). Differential splice junction usage ranged from simply having alternative exons to more complicated examples, including those where a series of introns spliced out in one stage are retained entirely in another (Fig. 2D and 2E).

To look more broadly for evidence of differential isoform usage, we developed algorithms to infer the quantitative expression of alternative transcripts and analyzed samples from a selection of stages with respect to the aggregate transcript models. Using the experimental evidence for exon and splice-junction usage, these algorithms distribute sequence reads among a set of distinct alternative transcripts in a probabilistic manner by applying either expectation maximization (EM) or Gibbs sampling (see supplement C.9). We found that alternative transcription generally does not change dramatically between stages (Fig. 2C). However, in systematic pairwise comparisons of the 7 observed stages , on average, ~300 of the ~13,000 genes with multiple isoforms show isoform switching between pairs of stages (see supplement C.9.a). Moreover, to identify alternative isoforms with divergent expression patterns across the developmental timecourse, we used expression profiles across 15 stages (including dauer) to cluster transcripts into 25 distinct expression profiles (Fig. S15 and supplement C.9.b). We found nearly 1,500 genes for which different isoforms fell into different expression-profile clusters (Fig. S16) and used these to distinguish several classes of variation in terminal or internal exons (Fig. S17; see Fig. S18 for examples). These genes provide candidates for possible stage-specific functions.

Pseudogenes

We noted several examples of gene models derived from RNA-seq that fell in regions previously annotated as pseudogenes. Pseudogenes are DNA sequences that are similar to protein-coding genes, but are thought to be non-functional in a conventional sense, producing no protein products (*29*). They are usually identified by the presence of disablements such as premature stop codons. Many were created from existing protein-coding genes, either by duplication followed by disablement or from reverse transcription of processed transcripts. Although considered to be non-functional, in some instances, pseudogenes have been found to be transcribed, potentially acting as endo-siRNA (endogenous-small-interfering RNA) regulators of their parent genes (*30-32*).

We began with a reanalysis of worm pseudogenes using the automated pipeline PseudoPipe (*33*) (see supplement C.10). Predicted pseudogenes were then manually reviewed with the help of the WormBase curators in order to identify a total of 1,293 likely pseudogenes in the worm genome (Fig. S19), adding 173 new annotations and removing 213 others. We also established the probable source (parent) gene for 1,198 pseudogenes.

We investigated the pseudogenes for evidence of transcription using the RNA-seq and tiling-array data. For the pseudogenes with identified parents, we found evidence of transcription for 789. We further characterized 323 of these as abundantly expressed, based solely on the RNA-seq data (see supplement C.10). To address the possibility that the reads were derived from the parent gene and not the pseudogene, we classified the pseudogenes into three subcategories. The first includes pseudogenes with expression levels at least two-fold higher than the parent gene. The second subclass contains pseudogenes for which the expression patterns of the pseudogene and parent are discordant across samples (see Fig. 4C for an example). Both of these cases indicate independent transcription of pseudogene and parent, arguing against mapping artifacts. The last subclass includes instances where the expression pattern of the pseudogene is concordant with the parent gene across multiple samples, which by itself would not exclude mapping artifacts. Altogether 191 of the 323 candidates fell into the first two subclasses (87 and 104, respectively) and are thus likely transcribed independently from their parents. The transcripts from these pseudogenes may potentially function in a variety of ways, from creating endo-siRNAs that regulate the parent gene to producing small peptides (*30, 31, 34*). Intriguingly, 17 of the potentially transcribed pseudogenes have a mass-spec peptide match, where the peptide does not match any protein already annotated in WormBase. This implies that these 17 pseudogenes could potentially give rise to novel, short peptides.

Non-Coding RNAs

In addition to protein-coding genes, the genome produces a variety of transcripts that do not code for proteins and function directly as RNA (ncRNAs). These include snoRNAs and well-known RNAs involved in mRNA translation and splicing (e.g., rRNAs, tRNAs and snRNPs). In addition, diverse classes of small regulatory RNAs that program Argonaute effector complexes have been characterized, such as: miRNAs, 21-24 bp RNAs produced by Dicer cleavage of short hairpins that direct post-transcriptional repression (*35, 36*); piRNAs (21U-RNAs in *C. elegans*) that may control transposon activity in the germline (*37*); and multiple classes of endo siRNAs (*38, 39*).

Because the catalog of ncRNA types is still being defined (*40*), we sought to provide a more comprehensive annotation of small RNA transcripts, and profiled miRNA gene expression using RNA-seq on size-fractionated total RNA. We obtained a total of 81 million aligned reads from small RNA fractions from 11 different stages. This enabled us to identify 154 previously annotated miRNA genes and their relative expression levels (*39, 41*). Most of these are products of the canonical Drosha-Dicer cleavage pathway. However, 4 are mirtrons whose precursor hairpins are independent of Drosha but are instead generated directly by intron splicing (*42, 43*). Our small RNA data defined 102 additional candidate canonical miRNAs (*39*), of which 42 have evidence of star sequences (the complementary strand sequence, providing confirmation of the miRNA). Of these, 20 were recently incorporated into miRBase (*44*). We also developed a computational model for prediction of mirtrons which identified 13 such loci that potentially produce endogenous miRNAs (*45*) (see supplement C.11). Finally, the small RNA data revealed thousands of piwi-interacting RNAs (21U-RNAs), although all of these were from previously identified loci (*39, 46*).

To identify other candidate ncRNA genes, particularly longer ones than those discussed above, we integrated all of our transcriptome data sets, along with RNA structure and sequence conservation. Looking at individual data sets, we found that, in comparison to other genomic "elements" (i.e. well curated CDSs, UTRs, or intergenic regions), known ncRNAs (Supplement Table 4) tended to have different values for characterized "features" (e.g. a higher small RNA-seq signal, more stable and conserved RNA secondary structures, and very little polyA-plus RNA-seq signal). However, no single feature was able to reliably distinguish the known worm ncRNAs from the other elements (Fig. 4A, Left). By using pairs of the features, discrimination improved, but was still incomplete (Fig. 4A, Right).

To make further improvements, we combined many features together in the framework of a machine-learning model (*47*). Initially, we focused only on features derived from expression data. Because the tiling array data were obtained from total RNA samples, we began by looking outside of coding exons and known ncRNAs for novel transcriptionally active regions >100 nt in length (Fig. S20) . By then integrating them with the additional expression data in a machine-learning model, we found support for 21,521 ncRNAs (4,352,048 bp). These tiling array-based predictions, which we call the 21k-set of ncRNAs, were characterized by a lower polyA RNA-seq signal and a higher small RNA-seq signal than other genomic elements (see Supplement Table 7).

Because the identification of potential ncRNAs from tiling arrays can be problematic due to cross-hybridization and other array issues (*11, 48, 49*), we sought to define a more confident set, incorporating both conservation and RNA secondary structure. The requirement for conservation restricts the search space to only ~15% of the genome. Nonetheless, the potential for greater specificity warranted these trade-offs. Incorporating these additional features into our machine-learning model (*47*), we predicted 7,237 novel ncRNAs (1,045,795 nt), with an estimated positive-predictive value of 91%, based on testing against the known ncRNAs. We call this the 7k-set. In this set, 1,678 ncRNAs (181,552 bp) fall in intergenic regions, with the remainder in introns, pseudogenes, or antisense to coding exons. Using RT-PCR, we tested 15 novel ncRNAs located in intergenic regions, and detected RNA products for 14 of them (*47*). We found many RNA structural motifs among the ncRNAs, many of which were not found in the known RNA secondary structure families from the Rfam database (*50*). In contrast to many known ncRNAs, such as rRNAs and tRNAs, (but similar to miRNAs (*51*)) our novel ncRNAs tended to be differentially expressed across developmental stages (*47*).

In comparing the 7k and 21k sets of candidate ncRNAs, the overlap is small, with just 1,259 of the 7,237 predicted ncRNAs in the former overlapping with the latter. Thus, the additional constraints of conservation and structure allowed us to detect candidate ncRNAs not found in the expression data sets alone. On the other hand, the additional constraints may well have omitted non-conserved ncRNAs or those with less structure.

Regulatory Sites and Interactions

TF Binding Sites and Targets

Accurate annotation of sites regulating gene expression and the transcription factors that bind to these sites is central to understanding the regulatory networks underlying development and homeostasis. To date, large-scale projects mapping TF binding sites have been performed either in cell culture or in single-celled organisms, and fail to link the identified regulatory elements to developmental events. We investigated binding sites within the whole animal using high-throughput sequencing ChIP (ChIP-seq) to map 23 GFP-tagged fusion proteins and RNA polymerase II (RNA Pol II, Supplement Table 8 and (*52*)). Generally, the factors were mapped at the developmental stages during which they have their highest expression levels, as deduced using Green Fluorescent Protein (GFP) fusion proteins. However, both PHA-4 and RNA Pol II were analyzed at six developmental stages to examine how TF and RNA Pol II binding sites change over the life cycle. A number of the factors have expression patterns limited to ~10% of cells in the whole animal, indicating that we can successfully identify binding sites for factors expressed in a fraction of somatic cells. Control experiments using antibodies directed against native proteins demonstrate that tagged protein binding sites correlate strongly with those from native protein. Also, TF binding sites identified through ChIP-seq have been verified through an independent method, ChIP-qPCR (see supplement D.1 and (*52, 53*)). At least two independent ChIP-seq experiments were performed for each factor. The binding peaks of each factor were scored using PeakSeq with a stringent threshold (*54*). We only kept those peaks reproduced in both replicates. For comparison we also analyzed the binding sites scored with SPP (*55*). The numbers of total mapped reads and binding sites for 23 factors are shown in Supplement Table 8.

DNA Motifs Bound by TFs

A major characterization for each TF and its associated sites is a sequence motif. These motifs are typically short, inexact sequences ranging in size from 8 to 12 bp (*56*). We developed a technique to identify high-likelihood cis-regulatory motifs from the modENCODE ChIP-seq TF binding data sets. We combined information from both PeakSeq (*54*) and SPP (*55*) with information from the six-way nematode alignment (see Conservation section, below). For these calculations we excluded the HOT regions (described below). We weighted sequences under peaks for each TF by their degree of evolutionary constraint and distance from the peak center. To discover motifs, weighted sequences were presented to a standard sequence-pattern discovery algorithm (*57, 58*), along with background sequence generated by a fourth-order Markov model from peak flanking regions. We then performed a specificity analysis on each recovered motif by measuring the frequency of the motif occurrences in peak regions relative to random upstream sequences and peaks from other TF data sets (e.g. Fig. S21C). We also performed localization tests for each motif relative to point binding positions (e.g. Fig. S21B). We recovered statistically enriched motifs for 21 of the 22 TFs' binding profiles, only 8 of which remained after specificity testing (Fig. S21A). Of the three TFs with previously described putative binding site motifs, we recovered the previously described motif in two cases.

The Distribution of TF Binding Sites

As shown in Fig. 5B, most TF binding sites defined by ChIP-seq peaks lie within 500 bp upstream of transcript start sites (TSS). In comparison to coding genes, binding sites assigned to known ncRNAs are even closer to the 5′ end of the transcript (Supplement Table 8). Consequently, binding sites could be readily assigned to specific protein-coding or known-ncRNA genes, based on proximity to the TSS. Most binding sites were assignable to annotated loci (see supplement D.2), but a subset remains unassigned for each factor. Although most factors bind sites near both protein coding and known ncRNA genes, GEI-11 mostly binds to ncRNAs (Fig. 5C; see Fig. S22 for example). We also examined whether any TF binding sites were adjacent to our novel predicted ncRNAs (intergenic ncRNAs from the 7k-set above). About 59% are potential targets of these 23 TFs, providing additional evidence for their activity (*47*). (An example is in Fig. 4B; 59% is significantly more than would be expected by chance (P<0.001).)

Pairwise correlation analysis of target genes reveals that factors with related functions often show substantial overlap in the target genes to which they bind (Fig. S23A). Three Hox genes involved in establishing the body plan (*MAB-5*, *LIN-39*, and *EGL-5*) provide a particularly striking example (*59*). They are more strongly correlated with each other in terms of targets than with the four other HOX genes analyzed, which have more diverse developmental roles. In contrast, pairwise correlation of miRNA targets shows that the factors bound to them tend to cluster together more by stage than by factor type (Fig. S23B). For example, one group of 4 different TFs analyzed in embryos target similar miRNAs, whereas a different group of six disparate TFs analyzed at L3 target another set of miRNAs. Integrated regulation by multiple TFs at a given developmental stage may be connected to the fact that the expression of miRNAs tend to show strong stage-enrichment (*39, 51*).

In sum, the binding sites cover a total of 11,831,636 base pairs and target 8,859 protein-coding genes as well as 652 known ncRNA genes. The large fraction of the genome associated with sites and the high number of genes targeted from the relatively small set of TFs we analyzed (from >900 candidate TFs in the worm) suggests that each gene may have sites for many factors.

Clustered Binding in HOT Regions

We identified 304 short DNA regions (avg. ~400 bp) that were significantly enriched in most TF ChIP-seq experiments despite the fact that the 23 analyzed TFs have diverse functions and expression patterns. These regions were bound by 15 or more factors at a q-value cutoff of 1e-5; we term these Highly Occupied Target (HOT) regions (Fig. 5A, 6A and Fig. S24 (*60*)). Control ChIP-seq experiments, using either Immunoglobulin G (IgG) antibodies in transgenic worms or GFP antibodies on N2 worms lacking a GFP-tagged TF revealed that apparent enrichment of these regions was not the result of generally open chromatin or the GFP-antibody binding to secondary targets (Fig. S25; see supplement D.3 and Fig. S26 for further discussion of control experiments).

Most TFs also cross-link to factor-specific DNA regions (bound by 0-3 additional factors) in addition to the HOT regions (Fig. 5A). We compared these different classes of sites to look for functional differences. For example, the HLH-1 TF drives muscle development in *C. elegans* (*61*) and is associated with 598 specific regions and 165 HOT regions. The specific HLH-1 ChIP-seq regions were over four-fold enriched for the HLH-1 binding motif (*62*) and were more than seven-fold enriched for genes with muscle-enriched expression ("muscle genes") (*63*)(Fig. 6B). By contrast, the 165 HOT regions associated with HLH-1 were less than two-fold enriched for the motif and were not enriched for muscle genes (Fig. 6B). For 3 other factors with identified binding motifs, we observed that motif enrichment was higher in specific targets than in HOT regions (see supplement D.3). Also, like HLH-1, 13 other TFs have targets whose expression is highly enriched for specific tissues. In every case, target genes bound specifically by the TF were enriched for expression in specific tissues, but target genes associated with HOT regions were not enriched (data not shown). These results suggest that there are functional differences between HOT regions and factor-specific sites.

To look for additional differences, we examined the expression patterns and functional classifications of genes associated with HOT regions. Initially, we looked at data from a single-cell gene expression atlas for L1 worms (*64*), in which single-cell expression levels were extracted from confocal microscope data stacks of 93 genes in 363 individual cells. Promoter regions that contained HOT regions often drove expression in most or all cell types (Fig. 6C), whereas most other genes show tissue-specific expression. As might be expected given this ubiquitous expression, we found that genes associated with HOT regions were associated with higher expression levels in whole worm RNA-seq measurements (Fig. S27) as well as tissue-specific tiling arrays (Fig. S28).

Using the results of a genome-wide RNA interference (RNAi) screen (*65*), we also found that genes associated with HOT regions are much more likely to be essential than other genes (Fig. 6D and supplement D.3). Specifically, 21% of genes associated with HOT regions are essential, compared to 3% of genes associated with binding to 1-4 TFs (8.7-fold enrichment; p < 1e-40) (*65*). Gene Ontology (GO)(*66*) analysis reveals a variety of biological processes highly represented in genes associated with HOT regions, including growth, larval and embryonic development, and reproduction (p < 1e-15), as well as 19 ribosomal protein genes (a more than 12-fold enrichment, p < 1e-12) (Supplement Table 9). In comparison, GO analysis of the remaining (non-HOT) targeted genes identifies functional terms consistent with the known tissue-specificity and function of the TFs (*59*).

Overall, these results suggest that many TFs cross-linked to HOT regions are not directly associated with DNA via specific binding, consistent with findings for highly occupied regions in *Drosophila* (*67, 68*). Rather, they suggest that association with HOT regions may be driven by protein-protein interactions to a currently unknown set of HOT-region-associated DNA-binding factors. HOT regions are significantly enriched for a few sequence motifs (Fig. S21), but additional experiments will be needed to discover which protein factors bind directly to them.

Building a TF Hierarchy

With the assignment of binding sites to target genes, the results of a ChIP-seq experiment can be represented by a series of potential regulatory interactions in a "binding network". Such a regulatory network graph has commonly been used in yeast (*69*) and *E. coli* (*70*), but has not previously been employed in metazoans because of the scarcity of the binding-site data for multiple TFs. A section of the worm network focusing primarily on interactions between TFs for the larval stages after removing the HOT regions is shown in Fig. 7A. Given only 18 larval factors, this is a relatively dense network, with each TF regulating an average of 828 genes including both TF genes and other gene targets. The amount of auto-regulation among the factors is notable, including the known example of LIN-39 (*59, 71*)

The expression of a TF tends to be more strongly correlated with the expression of its targets (over the time course) than its non-targets (Supplement Table 10). For example, transcript levels for the *pqm-1* gene have an average correlation coefficient of 0.31 with PQM-1 target genes, whereas the average correlation coefficient for non-target genes is 0.02 (p < 1e-200). The correlation is positive for potential activators and negative for repressors. We can then prune the network to reflect this correlation, keeping only strong relationships (Fig. 7B). Interestingly, we find more putative activators, with the only two identified repressors targeting GEI-11, specifically.

Within the network, TFs were organized hierarchically based on the extent to which they “target” other TFs (most targeting on top rows) or are themselves targets for other TFs (most targeted on bottom rows). This layout is motivated by the fact that TFs are often thought to act in regulatory hierarchies carrying out spatial and temporal control over developmental events (*72, 73*). Moreover, there are a number of clear differences that can be observed between the TFs at each level of the hierarchy (Fig. 7). First, we examined the expression of the TFs in 8 tissues in L2 and found that TFs at the lower layers tended to be more uniformly expressed across multiple tissues (see Fig. 7C caption for numbers). Next, we found that TFs at the bottom level tended to be essential more often than those at the top. This is, in a sense, consistent with results for tissue-specificity, with lower level TFs always being “necessary”. In contrast, TFs of the Hox family were more often at the top of the hierarchy -- among the six Hox TFs, four were at the top layer of nine TFs -- perhaps reflecting their role in globally modulating specific developmental processes in different tissues. Finally, we examined the TFs for their connectivity in the existing *C. elegans* protein-protein interaction network (*74*), and found that the TFs at the top of the hierarchy tended to have significantly fewer protein-protein interactions than those below. This result suggests that TFs in the middle and bottom layers act as “mediators” or "effectors," which are more likely to exchange information with other proteins. While the larval network here is obviously small and one cannot make strong statistical statements, these conclusions follow a pattern consistent with the more studied regulatory hierarchies in yeast and *E. coli*, where essential and highly connected, "workhorse" regulators tend to be at lower levels while overall modulators are on the top of the hierarchy (*72*).

An Integrated miRNA-TF Network and its Motifs

miRNAs can mediate post-transcriptional regulation of mRNAs, including those that encode TF proteins. We therefore endeavored to build a combined TF and miRNA network to identify potential interactions between these two types of regulators. We first made new predictions of candidate miRNA binding sites in *C. elegans* mRNAs using the integrated transcript models described above (*25, 75*). We identified a total of 20,427 predicted target sites within 4,866 3′UTRs for 2,244 genes that are conserved in *C. briggsae* (see supplement D.4).

We then used the miRNA data in combination with the TF network to generate an integrated network (Fig. 7C). We focused on miRNAs expressed during larval stages, as determined by the small RNA sequencing data, to match the selection of larval TFs in Fig. 7A. Each miRNA is placed on a hierarchical level depending on the highest-level TF it regulates or, if it does not regulate a TF, the lowest level TF that regulates it. Even with only 18 TFs and only larval regulation, the hierarchy reveals impressive complexity, with miRNAs clearly falling into several distinct levels, paralleling the TF arrangement. Moreover, the different levels effectively create different classes of miRNAs -- i.e. those that are more strongly regulated by TFs (at bottom right) in contrast to those that predominantly regulate TFs (top left).

In actual biological networks, some recurring regulatory patterns, or network motifs, are over- or under- represented relative to randomly wired graphs (*76, 77*). We identified over-represented network motifs, involving at least 3 members, in the integrated miRNA-TF network (Fig. 7D). Specifically, we compared the number of network patterns in the real network to 1000 random networks generated by re-wiring the real network, while keeping its rough topological statistics constant (see supplement D.6). As an example, the feed-forward loop, in which a TF regulates another TF, and both jointly regulate a target coding gene or a miRNA, is highly over-represented in the integrated network. Previous studies suggest that this motif can reduce the time required to turn on the expression of a target (*76*). Another interesting motif involves one in which a miRNA binds to a transcript encoding a TF, as well as a target gene of that TF. This motif could represent a way in which down regulation of a target gene is ensured by inhibiting both it and its activator. Finally, in reviewing two-member motifs, we observed many instances of specific miRNA-TF loops, where in which a miRNA regulates a TF, and the same TF regulates the miRNA. While these individual examples are interesting and fit a pattern reported earlier(*78*), overall, the occurrence of miRNA-TF loops was not significantly enriched relative to that in random network rewirings.

Dynamics of RNA Polymerase II Binding & Expression

To explore regulatory dynamics, we profiled RNA Pol II and one specific factor (PHA-4) in each of the main stages of *C. elegans* development and compared their binding profiles to the corresponding RNA-seq data. For RNA Pol II, we analyzed the aggregate ChIP-seq signal over promotors for the set of 8,428 non-overlapping transcripts (see expression dynamics discussion above and in supplement C.7). The binding profiles for each stage were then compared to each other and to the average expression profiles across each of the corresponding genes, giving rise to the correlation matrices in Fig. 3A and 3B.The most evident finding is that the embryonic stages form a distinct cluster from the larval stages (Fig. 3A, Right). This is broadly similar to what was observed for the gene expression correlation described earlier (Fig. 3A, Left). The embryonic-larval division is also observed for PHA-4 binding across different stages (Fig. S30). It presumably reflects the different transcriptional programs at work in embryos compared to larvae and adults.

In comparing the RNA Pol II profiles to gene expression, the correlation between RNA Pol II binding and expression profiles within the same stage is relatively high (0.64 to 0.70), as would be expected (Fig. 3B). However, the intra-stage correlation is lower for embryonic stages than larval ones, perhaps reflecting the fact that worm embryos have a large number of transcripts that are inherited from the parent which are not transcribed in the embryo(*79, 80*).

Fig. 3B also shows that expression in earlier developmental stages is more tightly correlated with binding at later stages, rather than a relationship in which RNA Pol II binding precedes RNA production. The correlation structure follows a consistent trend across rows: it is low initially, peaks at the matching stage and then remains high for later stages. This can be interpreted as RNA Pol II binding to genes at the same developmental stage where they are initially expressed, with RNA Pol II then remaining bound in later stages, even if expression drops. The initial round of transcription may affect the accessibility of the promoter, which may remain unaltered in later stages for these non-dividing cells. Moreover, this result may potentially reflect the presence of paused RNA polymerase at genes with reduced expression at later stages. Indeed, we have found several examples of genes where RNA Pol II binding remains high in later stages where gene expression falls (e.g. *isl-1* and *pgp-2*, Fig. S31).

Chromatin Organization and its Implications

The modENCODE project aims to identify functional elements that control chromatin and chromosome behavior and to identify chromatin features that control the function of associated DNA elements. *C. elegans* has several unusual features that offer an opportunity to study diverse aspects of chromatin behavior. Its holocentric chromosomes have microtubule attachment sites distributed along the length of each chromosome, rather than being embedded in the long, highly repeated sequences of mammalian chromosomes. Gene expression from the two sex chromosomes present in hermaphrodites (XX) is down regulated in somatic cells by a dosage compensation mechanism to better match the output from the single X chromosome in males (X0) (*81*). A distinct mechanism silences almost the entire X chromosome in the germline cells of both hermaphrodites and males (*82*). Finally, *C. elegans* autosomes have distinct domains – a central region flanked by two “arms”, where the two arms together comprise more than half the chromosome. Compared to the centers, the arms have higher meiotic recombination rates, lower gene density and higher repeat content (*6, 83, 84*). These domains are less distinct on the X. Overall, these features provide fertile ground for the investigation of chromosome-level mechanisms of regulation.

To discover elements that control chromatin and chromosome behavior, we have performed 68 sets of experiments to map the distribution of chromatin proteins and histone modifications, most examined in at least two developmental stages. This information has been integrated with transcriptional and regulatory data described above to determine how chromatin organization is related to the specification of TF-binding sites and the levels of gene expression.

Chromosome-Scale Domains of Histone Modification

Using ChIP, we examined the distribution of 19 histone modifications and three key histone variants (*C. elegans* homologs of H2A.z, CENP-A, and H3.3) (*85-88*). Several of these histone marks revealed striking, broad domains of histone modification states on the autosomes, with relatively sharp boundaries between the central region of each autosome and the distal chromosomal regions (Fig. 8A, 8C and 8D). We found that modifications traditionally associated with gene activity and euchromatin such as acetylation and H3K4 and H3K36 methylation are enriched in the central regions of the chromosomes. In contrast, H3K9 mono-, di-, and tri-methylation, histone modifications traditionally associated with transcriptional repression and heterochromatin formation, were very strongly enriched on the arms of the autosomes, and relatively depleted from the central regions (Fig. 8A). Despite the biased distribution of these repressive marks, the terminal regions of the chromosomes do not appear heterochromatic by DAPI staining or classical banding techniques (*89*). The chromosome-scale domains of histone modification do not vary significantly in composition or position between embryos and L3 larvae. Even though these animals contain only a small fraction of germline cells, the broad domains of histone modifications correspond to regions defined by differences in recombination rate, with the boundaries located at the recombination rate inflection point (Fig. 8A) (*6, 83, 84*). We also note that these megabase-scale chromosomal domains are far from homogeneous, with small zones of repressive marks occurring within the generally active central regions, and active marks occurring within the generally repressed arms.

On each chromosome, one arm contains a specialized region known as a homolog recognition region, or pairing center, which mediates homologous pairing and synapsis (*89, 90*). We found that H3K9me3 is more highly enriched on the chromosomal arm that contains the meiotic pairing center than the opposite arm (Fig. 8A)*,* consistent with previous observations (*91*). However, the methylation is not particularly enriched within the pairing center regions themselves (*92*).

The pattern of H3K9 methylation on the autosome arms also mirrors the genomic distribution of repetitive DNA elements. We therefore examined the histone modification patterns associated with each of five DNA repeat classes: tandem repeats, inverted repeats, transcribed mobile elements, non-transcribed mobile elements, and inactive mobile remnants. H3K9me1, H3K9me2, and H3K9me3 are all enriched over repeat elements, while transcribed mobile elements are additionally marked with H3K36me2 and H3K36me3 (Fig. S32).

The X Chromosome Exhibits Several Unique Chromatin Features, Including Enrichment of H4K20me1

The organization of the X chromosome differs from the five autosomes in several key respects, with gene density, recombination rates, and repeat content more uniformly distributed along the length of the X (*6*). Consistent with this, we find that the chromatin marks on X are more uniformly distributed. A high density of repressive marks, similar to that seen throughout the autosome arms, is only associated with two ~300 kb regions at the left end of X. These more autosome-like regions are on the same end as the X-chromosome pairing center (Fig. 8B); they flank the pairing center, which has been mapped to between 0.5 and 1.5 Mb from the left telomere (*92*).

To further investigate the differences between X and the autosomes, we determined the genomic distribution of proteins mediating dosage compensation, which comprise the dosage-compensation complex, and found that all of them were highly enriched on the X chromosomes of XX animals (Fig 8B)(*93-96*). In addition to the dosage-compensation complex, we also observed a prominent enrichment of H4K20me1 on the X. The X-enrichment of this mark is detected in early embryo populations, which contain a mixture of embryos that have not initiated dosage compensation, and those that have. However, the X enrichment is more pronounced at L3, by which time all somatic cells are thought to have established dosage compensation. Interestingly, H4K20me1 has been linked to mammalian chromatin maturation during development (*97, 98*) and to mammalian X chromosome inactivation (*99*) but has not previously been associated with the X chromosome or dosage compensation in *C. elegans*.

Interaction Between Chromosomes and the Nuclear Envelope

We examined the interactions between the genome and the nuclear envelope by ChIP of LEM-2, a transmembrane protein associated with the nuclear lamina (*100*). We found that in embryos LEM-2 interacts strongly with the repeat-rich, H3K9 methylated arms of the autosomes, and does not interact with the autosome centers (Fig. 8A). The transition between the LEM-2 associated and free chromosomal regions is rather sharp, and coincides with the transition between regions of low and high meiotic recombination rate. In addition to this large-scale organization, regions associated with the nuclear envelope exhibit a complex underlying subdomain structure (*100*). LEM-2 also shows strong interaction with the small regions on the left end of the X chromosome that have an autosome-like distribution of repressive chromatin marks, as described above (Fig. 8B). These findings suggest that the three-dimensional organization of the X chromosome is subject to different constraints than those of the autosomes.

X-linked Genes are Enriched in Several Mono-Methylated Histone Marks

To measure how histone modifications were distributed on a typical gene, we plotted the distribution of each chromatin mark relative to gene features. We further subdivided these plots by the expression level of the associated gene and by X-linkage (i.e. autosome vs X) (Fig. 9). Overall, the results are consistent with the known distributions and functions of chromatin marks in other eukaryotes (*101*). However, a striking exception is the distribution of several mono-methyl marks that are more strongly associated with highly transcribed genes on the X than with similarly expressed genes on autosomes. In addition to the H4K20me1 modification mentioned above, H3K36me1, H3K9me1, H3K27me1 are all more highly associated with the bodies of highly expressed X-linked genes than with autosomal ones. H3K36me1 was also associated with highly transcribed genes on autosomes, but was confined to gene bodies only on the X (Fig. 9). Conversely, H3K36me3 and H3K36me2 were more strongly associated with autosomal than with X-linked genes. The causes and consequences of these differential histone-modification patterns remain unclear, but we speculate that they reflect dosage-compensation mediated repression of X-linked gene expression.

Nucleosome Organization

Nucleosome positioning and occupancy were determined by paired-end Illumina sequencing of isolated mononucleosomal DNA generated by micrococcal nuclease digestion of chromatin in embryos and adults (*102*). Overall, the results are consistent with previously published maps of *C. elegans* nucleosome organization (*91, 103, 104*). We observed a typical nucleosome-depleted region upstream of TSSs, and at the 3′ ends of genes. Also consistent with previous data (*104*), our nucleosome maps were highly concordant with computational predictions derived from yeast data or *in vitro* nucleosome affinities (*105*), suggesting a prominent role for DNA sequence in nucleosome organization in *C. elegans* (*102*).]] Nucleosome positioning was not markedly dependent on developmental stage.

We noted a striking difference in the average GC content and nucleosome occupancy of X vs. autosomal gene promoters (Fig. S33). X-linked promoters have a higher GC content, which is predictive of high nucleosome occupancy in vitro (*105-107*). Accordingly, *in silico* models predicted that X-linked genes would have higher nucleosome occupancy at their promoters than autosomes (*102, 105*). Our experimental nucleosome mapping data from both embryos and *glp-1* germlineless adults were consistent with this. While both X and autosomal promoters exhibit nucleosome depletion at promoters and a well-positioned +1 nucleosome, average nucleosome occupancy at 5′ ends of genes on the X is 1.6-fold higher than that of genes on autosomes (as measured between -300 to +200 bp relative to the TSS, $p < 2.2e\text{-}16$ by Wilcoxon rank sum test). The increased nucleosome occupancy we observe between X and autosomes appears to be specific to the region immediately upstream of the +1 nucleosome (*102*).

Notably, we observed a similar difference between X and autosomal promoters when naked DNA was digested with micrococcal nuclease (MNase). This result was not unexpected, because MNase has an intrinsic bias to cleave at sequences that define linker DNAs in chromatin assembled either *in vitro* or *in vivo* (*108-115*). Moreover, a wealth of data validates the correspondence between MNase cleavage patterns and nucleosome positions *in vivo* (*104, 111, 112, 116, 117*).

DNA sequences that exclude nucleosomes or that support higher nucleosome occupancy have been shown to evolve according to the expression requirements of the downstream gene (*114, 118*), and the single most important feature of DNA sequences in supporting higher nucleosome occupancy is GC content(*106*). Therefore, higher GC content may have evolved to support higher nucleosome occupancy on X-linked promoters as part of the mechanism of somatic dosage compensation or silencing of the X in the germline.

Evidence for Epigenetic Transmission of Chromatin State to Progeny

Our data also provide evidence for chromatin-mediated transmission of the pattern (i.e. "memory") of germline gene expression from mother to progeny. This phenomenon is illustrated by the activity of the *C. elegans* protein MES-4, a histone H3K36 methyltransferase that is required for the survival of nascent germ cells in developing animals. ChIP analysis in early *C. elegans* embryos revealed that MES-4 binding sites are concentrated on the autosomes and on the leftmost ~2% (300 kb) of the X chromosome (Fig. 8A and 8B), consistent with cytological evidence (*119, 120*). Also, like other methyltransferases, MES-4 is associated with gene bodies. However, in contrast to previously studied H3K36 methyltransferases, which are targeted to genes by association with RNA Pol II (*121-123*), our data show that MES-4 can continue to associate with genes in an RNA Pol II-independent manner. In embryos, MES-4 binds preferentially to genes that were highly expressed in the maternal germ line, many of which are no longer expressed in embryos. Conversely, MES-4 was not associated with genes expressed specifically in early embryos, despite recruitment of RNA Pol II to those genes. Therefore, RNA Pol II association with genes is neither necessary nor sufficient to recruit MES-4 in embryos (*95*). These and other findings suggest that MES-4 serves as a maintenance methyltransferase, perhaps by recognizing H3K36 methylation itself, to propagate a memory of gene expression from the parental germ line to the cells of the next generation (*95*). This propagation is required for the viability of the next generation of germ cells, since offspring of *mes-4* mutants are sterile.

Another case of epigenetic transmission was observed for the histone H3 variant HCP-3, an ortholog of CENP-A/CenH3, which defines centromeres in both monocentric organisms, including yeast, flies, and mammals, and holocentric organisms such as *C. elegans* (*124*). We found that HCP-3 was generally excluded from regions of transcriptional activity in early embryos, and that RNA Pol II occupancy was inversely correlated with the presence of HCP-3. Unexpectedly, we found that HCP-3 was also absent from regions that are silent in embryos (and lack RNA Pol II) but were previously transcribed in the maternal germline. Furthermore, HCP-3 distribution does not change between early and late-stage embryos, even on genes whose transcription levels change. This suggests that the memory of the maternal germline transcriptional state, rather than the transcriptional state in the embryo, defines genomic regions permissive for HCP-3 incorporation (*86*).

Statistical Models for TF Binding and Gene Expression from Integrating Chromatin Features

We have shown how the signals of chromatin features vary substantially at different positions around genes and with levels of gene expression (Fig. 9), observations consistent with previous findings (*125*). This suggests that the various chromatin marks may be used to statistically "predict" the locations of promotors and enhancers, as well as gene expression levels. We show here by integration of the data sets this, in fact, is possible. Moreover, the models themselves reveal which specific chromatin marks are most strongly associated with the binding of particular TFs and the amount of gene expression. They also reveal the relative importance of different chromosomal locations (e.g. positions relative to the TSS) in relation to binding or expression.

TF-Binding Model

Since TF binding is influenced by chromatin features (*126, 127*), we used our data sets to explore relationships between TF binding and chromatin patterns. In pairwise combinations, we observed that there are only weak global correlations across the whole genome between histone marks and TF-binding signals (Fig. S34). However, the relationship between these two types of genomic features potentially involves more complex, non-linear relationships than is ascertained by straight correlation. To probe these complexities, we built machine-learning classifiers that predict the location of TF binding from chromatin features, considered individually and in combination (Fig. S35). The prediction success of a classifier can be interpreted as a measure of the degree of association between specific histone marks and particular TFs (Fig. 10A).

Using such an approach on features individually, we found that H3K4me2 and H3K4me3, which are usually enriched in promoters, identified the binding peaks of most TFs with reasonable accuracy (Area Under the Receiver Operating Characteristic curve, AUROC, up to .92 in Fig. 10A). In contrast, the repressive marks H3K9me3 and H3K27me3 were found to be discriminative between different TFs by being significantly depleted at the binding peaks of some TFs. Further, some of the TFs were clearly more strongly associated with individual marks than others. For instance, CEH-14, LIN-13, and LIN-15B were more strongly correlated with promoter-specific marks H3K4me2 and H3K4me3 than the other TFs. In fact, the association of chromatin marks can be used to cluster the TFs into distinct groups (e.g. promoter associated). The HOT regions were also strongly related to the individual chromatin features. As expected, the binding of RNA Pol II was also strongly indicative of HOT regions, suggesting strong active transcription at them.

Chromatin features are thought to work in combination to influence binding site selection (*126, 127*). Indeed, when we combined all the histone marks together, the resulting integrative models were more accurate in identifying binding sites in the genome (Fig. S36) than any of the models involving single histone marks (Fig. S37), illustrating the complex interactions between different chromatin features. Although RNA Pol II peaks alone provide reasonable predictions of TF binding, the integration shows that adding in the histone marks provides quantitative improvement in prediction accuracy. The chromatin features are also effective in distinguishing the specific binding regions of some groups of TFs from those of others, as well as the HOT regions from other TF binding sites (Fig. S38). In fact, the chromatin features more readily identify the HOT regions than individual TF targets, perhaps reflecting the broad, high expression of HOT-region associated genes. We also observed that in order to construct an accurate model for the binding sites of a TF measured at a certain developmental stage, we usually must include chromatin features measured at the same stage (Fig. S39). This finding suggests a dynamic relationship between chromatin features and protein binding sites across developmental stages.

Although the models constructed from chromatin features generally discriminate strongly between TF binding peaks and other regions, they are not always sufficient for specifically discriminating between the binding sites of individual TFs. Additional discriminating information undoubtedly comes from the exact binding motif of a given TF, usually summarized in terms of a position-weight matrix (PWM). On the other hand, direct searching of a genomic sequence with just a PWM will usually result in large numbers of false positives (*128*). However, we find that when the static sequence information from the PWM and the dynamic chromatin data are combined, the resulting models are more accurate than the models from either type of information alone, as seen, for example, with HLH-1 (Fig. 10B and Fig. S40 with motif from Fig. S21). One interpretation of this finding is that chromatin features enable one to predict TF-accessible regions and broad classes of binding sites but motifs provide information on the exact sites of particular factors, chosen from these broad classes. These findings are consistent with those reported in previous work (*129-131*).

Gene-Expression Model

Next, we developed a model to relate the levels of gene expression to the chromatin marks near the TSS and transcript termination site (TTS) of a gene. To understand the spatial effect of chromatin features, we divided the DNA regions around the TSS and TTS of each transcript into small 100-bp bins, and calculated the average signal of each chromatin feature and RNA Pol II (13 total different features) in a set of 160 bins up to 4 kb upstream and downstream of these two anchors to include even long-range effects. Then, as shown in Fig. 10C, we constructed a matrix, whose elements were the correlation of signals in each of the 100-bp bins with the stage-matched gene expression value. As shown, the elements form two categories, those positively correlated with expression and those negatively correlated. The first category shows substantial spatial variation across the different bins, i.e., the effect of activating marks appears to be more sensitive to their exact positioning relative to the TSS and TTS than the repressive ones.

We then combined all features and constructed a statistical model for gene expression at each of the 160 bins. In particular, we predicted the quantitative expression levels of transcripts using support vector regression (SVR). We found, for instance, that the predicted expression levels based on the bin closest to the TSS are highly correlated with the real expression levels, with a Pearson correlation coefficient of 0.75 (cross-validation result). As an overall benchmark of the model performance we compared it to a model based on the level of RNA Pol II binding upstream of a gene. Our statistical model based on chromatin features achieves more accurate prediction of expression levels than the one based on RNA Pol II binding alone (Fig. 10D).

Next, we evaluated the relative importance for gene expression of the 160 bin locations upstream and downstream of the TSS and TTS. To simplify the model, we divided transcripts into high and low expression classes using the median as a divider, and predicted the class of a given gene from the bin values using a support vector machine (SVM). The prediction accuracy of each bin was estimated by cross-validation. The best predictions were obtained for the bins immediately after the TSS and just prior to the TTS. Moreover, with increasing distance upstream of the TSS, predictive power fell off smoothly. Intriguingly, the predictive capability of chromatin features extends as much as 4kb upstream of the TSS and 4kb down-stream of the TTS. These results were seen even when we restricted the analysis to widely separated genes with distant upstream neighbors and may indicate long-range influences of chromatin features on gene expression.

In contrast to protein-coding genes, the association of histone modifications with miRNA gene expression is largely unknown. Since protein-coding and miRNA genes are both transcribed by RNA Pol II, we investigated the effectiveness of the above chromatin model for predicting miRNA expression and histone association. Because precise TSSs are not generally available for worm miRNAs, we calculated the signals of chromatin features in the genomic region corresponding to candidate pre-miRNAs, and used them as the input features for our chromatin model trained solely on data for protein-coding genes. The predicted expression levels of 162 worm microRNAs for which genomic locations are provided by miRBASE (*44*) were compared with the experimental results measured by the modENCODE small RNA-seq data set (*39*). Our predictions showed remarkable agreement with the experimental results, with a correlation coefficient of 0.6 (Fig. 10D). Since some miRNAs are located within or near protein-coding gene loci, a fact that may confound the prediction of microRNA expression, we checked the prediction accuracy using the miRNAs that are isolated from any known gene; a similar prediction accuracy was achieved (Pearson correlation coefficient is 0.62). The fact that the expression of miRNAs may be accurately predicted using a chromatin model trained on protein-coding genes suggests that miRNAs and protein-coding genes share similar mechanisms of transcriptional modulation by histone marks.

Conservation Analysis

Because purifying selection slows the rate of divergence of functional sequence relative to the "neutral model" (*132*), the knowledge of evolutionarily constrained regions can assist in identifying true functional elements. For example, TF-binding sites that are active in regulating a nearby transcriptional promoter could be distinguished from those that may be biochemically real, but have no effect on gene expression (*133*). While there may be many functional sequences that are not conserved, are conserved in a way that we are unable to detect, or are under strong positive selection, our ability to account for conserved sequences with annotations provides one measure of the completeness of our annotations. For these reasons, we characterized regions of the *C. elegans* genome under evolutionary constraint by constructing a six-way multiple alignment among the nematodes *C. elegans*, *C. remanei*, *C. briggsae*, *C. brenneri*, *C. japonica*, and *P. pacificus* using the previously described methods (*1*). Evolutionary conservation scores were then calculated using PhastCons (*134*).

Considering only regions under purifying selection, there are 59,504 constrained blocks covering 29.6% of the *C. elegans* genome. Coverage ranges from a low of 27.4% on chromosome IV, to a high of 31.9% on chromosome X. The single largest constrained block is 3558 bp on chromosome V (spanning a portion of *unc-70,* spectrin beta-chain), but conserved blocks are typically much smaller (mean 49 +/- 58.6 bp). These conserved regions are highly correlated with functional elements. This finding applies to previously known elements as well as to those discovered by modENCODE.

First, we asked what proportion of evolutionarily constrained regions could be explained by experimentally annotated regions on the genome (Fig. 11A and Fig. S41). Prior to modENCODE, 58% of constrained blocks were covered by genomic annotations (primarily CDS and UTR) but only 51% were covered by annotations supported by direct experimental evidence. When new protein-coding transcribed regions discovered by modENCODE are included, the proportion of the genome covered by experimentally-supported CDS and UTR annotations increases to 59%. When ncRNAs and TF binding sites are included, coverage increases to 71%. Sites bound by dosage compensation and other chromatin modifying factors increases coverage further to 82%. Hence, modENCODE provides an explanation for an additional 31% of the constrained regions of the genome, whereas the amount of sequence covered by the new data would be predicted to cover only an additional 18% by chance (p-value < 1e-34 by GSC, see below).

Next we looked at the converse question: What fraction of the annotated functional elements is constrained? First, in Fig. 11B we graph the distribution of PhastCons conservation scores of a variety of annotations. Non-coding RNAs (both known and in the 7k-set) are the most constrained. However, this likely reflects the fact conservation was one of the features used in their identification. Then come coding elements followed by transcribed UTRs and miRNAs. Among the elements annotated by modENCODE, the TF binding sites and other chromatin regulatory factor binding sites have intermediate levels of conservation, falling between CDSs and UTRs.

Fig. 11C shows the distribution in another form more suitable for testing whether the degree of constraint in a functional element is significantly different from the genomic background. We first show the degree of constraint after normalizing against a background model drawn from sections of the genome that have not been annotated. The background distribution (expectation) of constrained bases is represented as a horizontal line at 0; this was derived independently for each site group examined. Values higher than 0 represent a larger fraction of bases at that conservation level than background, whereas values lower than 0 represent a smaller fraction of bases at that level than background. To perform significance testing on the evolutionary constraint we used the Genome Structure Correction (GSC) statistic (*1*) to calculate confidence intervals on the degree of overlap between evolutionarily constrained blocks with functional elements defined by modENCODE and other annotations. This confirms that coding regions, UTRs, and TF-binding sites are all significantly more constrained than would be expected by chance; in contrast, the overlap of pseudogenes with constrained blocks is not significantly different from chance.

Roughly 18% of the constrained genome remains uncovered by current functional annotations. However, some of this sequence partially overlaps known functional elements, suggesting that it might be covered if the borders of transcribed regions or binding sites were modestly extended. This leaves ~3.4 Mb (11%) of constrained blocks that do not overlap with functional genomic elements at all. Interestingly, these residual constrained blocks are markedly enriched in introns and intra-genic regions (see Supplement Table 12). The blocks are slightly overrepresented in the 1kb regions upstream of genes and markedly under-represented in the 1kb downstream regions. Together, these observations suggest that the residual constrained blocks might contain a population of intronic regulatory sites, such as splice enhancers, or possibly alternative exons that are expressed only under unusual circumstances. Finally, plots of the increase in coverage of constrained sequence by each additional TF experiment shows the beginning of saturation even with the relatively small numbers of TFs studied here (Fig. S42). This implies that doing more TF experiments may not account for all the remaining constrained regions and that the population of unexplained constrained blocks in intragenic regions is particulary mysterious and may represent novel classes of functional elements that are yet to be discovered.

Discussion

Model organisms such as *C. elegans* have long served as key experimental systems for developing technological advances and providing fundamental insights into human biology. As part of the modENCODE project, we have generated multiple data sets aimed at functional annotation of the *C. elegans* genome for two main purposes. First, decoding the genome of this powerful model organism will directly enhance and facilitate future studies in the worm. Second, these data sets will provide insights that can be useful for understanding general principles of genome organization and function, which will ultimately aid in deciphering the function of the human genome.

Our analysis illustrates general patterns evident at multiple genomic scales: individual gene, chromosomal domain, and whole-chromosome. At the single-gene level, we have annotated thousands of individual gene features, substantially improving gene models and increasing the number of new splice variants and transcript ends. In addition to improving annotation of protein-coding genes, we have found transcribed pseudogenes and identified many candidate noncoding RNAs, including additional miRNA genes. More accurate and reliable knowledge of gene models will simplify and expedite in-depth functional analysis, in both global and gene-by-gene studies. For instance, improved annotation has already been useful for mapping binding sites of TFs, allowing us to more accurately assign sites to specific target genes and to build regulatory networks. Finally, we found the relationship between histone marks and both TF-binding locations and gene-expression levels to be strong enough for individual genes so that we could build statistical models predicting binding location and expression level with reasonable accuracy.

On the level of chromosomal domains, we have found large-scale patterns characterized by repressive marks and interactions with the nuclear envelope on the arms of autosomes, with activating marks and a lack of nuclear envelope interaction at the centers of autosomes. The boundaries of these domains correlated with many known chromosomal features, including the shift from low to high recombination frequency between chromosome centers and ends. Quite strikingly, this relationship between recombination frequency, histone modifications, and nuclear envelope components is apparent on the autosomes in somatic nuclei, even after multiple mitotic divisions. The persistence of this chromosomal organization throughout development suggests that the events occurring in the germline are one of the strongest and most lasting influences on chromatin and chromosomal organization in the soma. Additionally, although it has long been appreciated that the X chromosome has unique features and regulation, our data sets have identified several additional X-specific properties, including the preferential accumulation of multiple mono-methylated histone marks. These features provide the opportunity to develop and ultimately test new hypotheses about mechanistic differences in X chromosome vs. autosome expression in both the soma and the germline.

Overall, one major advantage of our large-scale approach is the ability to discover unexpected biological phenomena that could not be discovered by single-gene studies. In particular, upon profiling the binding sites for only 23 different TFs, we identified obvious regions of clustered TF binding, which we have termed HOT regions. Since these HOT regions were apparent after analyzing a relatively small fraction of TFs, we may assume that HOT regions can be cross-linked to hundreds of different TFs. Further characterization of these HOT sites will improve our ability to predict the functional consequences of individual TF binding events on gene regulation from genome-wide binding site data in the future. Mechanistically, HOT regions may represent a particular 3-D nuclear organization, resulting in previously uncharacterized regions of dense TF co-localization within the nucleus.

We believe that these data provide an important foundation for functional annotation of the *C. elegans* genome, and that persistent coordinated collection of these types of data will continue to provide new global insights into genome organization and function. More TF profiles will provide important functional information about individual TFs and will also expand regulatory networks. One current limitation of the data is that almost all experiments were performed in whole animals composed of multiple tissues. A primary focus in the future will be to increase the tissue-specific resolution of the data. Such efforts will lead us ever closer to unraveling the complexity and elegance of the *C. elegans* genome.

Exhibit Legends

Figure 1: Overview of the modENCODE Worm Data Sets

Part (A) gives an overview of the amount of raw experimental data (e.g. reads and replicates) present in the February 2010 data freeze, and Part (B) shows some of the derived quantities (e.g. peaks and transcripts) for key developmental stages. For brevity, in part (A) developmental substages, isolated tissues, and several mutant strains have been collapsed into single columns, although not every experiment type may have been performed on every substage. The following list enumerates all stages possibly represented within each column, given numerically in the column header. Background strain is N2, unless otherwise stated: *Embryo*: early embryo, late embryo, mixed-stage embryo, one-cell stage embryo, post-gastrulation embryo, two-to-four cell embryo; *L1 larva*: N2, N2 starved, *lin-35*; *L4 larva*: hermaphrodite, JK1107 soma, L3-L4; *Dauer*: *daf-2* dauer larva (entry, mid, exit), *daf-3, daf-7, daf-9, daf-11*; *Adult hermaphrodite*: adult (includes controls for pathogen assays), young adult, *spe-9* adult (0, 5, 8, 12 days), JK1107 soma, L4-YA; *Male*: *him-8* embryo, *dpy28(y1);him-8(e1489)* L4 male, *him-8* adult male; *Isolated tissues*: GABA neurons, A-class motor neurons, AVA neurons, body wall muscle, coelomocytes, dopaminergic neurons, GABA motor neurons, germline precursor, hypodermal cells, intestine, panneural, BAG neurons, pharyngeal muscle, PVC neurons, excretory cell, glutamate receptor neurons, PVD & OLL neurons, cephalic sheath cells (CEPsh), spermatids, oocytes, gonad; *Infection (3 pathogens)*: *E. faecalis, P. luminscens, S. marcescens.* In Part (B) we summarize genomic elements that have been inferred for each major element type across the developmental series. For simplicity, we have chosen a single representative subcondition for each stage. *Embryo:* early N2 embryo for all experiments except for the miRNA and other ncRNA experiments, which were performed on mixed embryonic stages from N2; *L1-L4:* L1 through L4 larva in the N2 strain; *YA:* Young adult N2 hermaphrodites

Figure 2: Transcriptome Features, Gene Models & Alternative Splicing

(A) The histogram indicates the extent to which the modENCODE project has increased our knowledge of splice junctions in *C. elegans*. On the left is the number of directly supported splice junctions annotated as of January 2007 in WormBase (WS170). Columns L2, L3, L4 and YA (*24*) indicate the number of splice junctions identified by RNA-seq experiments in the indicated stage, either confirming known junctions (blue) or finding novel splice junctions compared to WormBase transcript predictions in the 5' UTR (yellow), internal to the CDS (green), in the 3' UTR (purple), or in novel transcript sequences (orange). For stages and samples to the right of L2-YA, the red box represents splice junctions already confirmed by the L2-YA RNA-seq data, while newly supported splice junctions are colored as indicated above.

(B) This diagram illustrates the process of gene model construction. The top half shows the various features identified through RNA-seq and the bottom half shows the resultant models. To build gene models in regions across the genome we search for the most abundantly represented splice junction, indicated by “(1)”, and then move away in both directions until another feature is encountered. Moving to the right in this example, coverage continues until a second splice junction is encountered, so the model incorporates this junction and continues through the next area of coverage until the end of coverage is encountered. Here, this position corresponds to a polyA site, indicating a transcript stop signal (black line). Moving to the left of the initiating splice junction, a splice junction is again encountered and incorporated. The first gene model is completed when the end of coverage is encountered. A splice junction indicated by “(2)” that was not incorporated into the first model is then used to initiate a second gene model. Moving to the right, this gene model is the same as model 1. Moving to the left, it encounters the end of coverage, with an associated start site (either a spliced leader junction or a strand bias signal) and the model is complete. Orientation is implicit in the sequences of the splice junctions and the start and stop sites.

(C) This histogram shows the distribution of differences in isoform composition for all genes with multiple isoforms in 21 pairwise comparisons across 7 developmental stages (EE, LE, L1, L2, L3, L4, YA). (Note that the Y-axis is on log scale). Isoform composition for gene *i* in stage *S* is represented by a vector □(*i,S,k*) where the kth component is the relative abundance of isoform *k* in relation to the other isoforms. Between two stages *R* and *S* for a given gene i the difference in abundance vectors gives a measure of the change in isoform usage for a gene. This is represented as D(*i,R,S*) = $\sum_{/k}$ ((□(*i,R,k*) - □(*i,S,k*))^2)/k. The difference *D* is a fractional number between 0 and 1; scores close to 1 indicate dramatic differences in the relative composition of different isoforms of the gene. The histogram plots the distribution of *D* values for all genes *i.* It is averaged over all pairs of stages *R* and *S*. The error bars represent the range of number of genes in every histogram across the 21 pairwise comparisons. Overall, the histogram shows that most genes have minor differences in their isoforms, but a small fraction (~300) have major and minor isoform switching between stages. (The minor isoform is defined as that with the lowest expression and account for less than 15% of the total expression of a given gene, while the major isoform account for more than 85% of of total expression. This corresponds to a cutoff value 0.5 of the fractional difference).

(D) This region of the gene F01G12.5 (*let-2*) is a simple example illustrating alternative exon usage and the dynamics of expression across stages. The alternative integrated transcript models are shown (black) followed by raw read counts per base across stages. The first alternative exon is used almost exclusively by stages L1 through young adult. The second alternative exon is used primarily in early and late embryo, with decreasing usage in later stages. Note that the raw read counts for each exon may come from multiple isoforms.

(E) This region of the transcript ZK783.1, homologous to human fibrillin-1, illustrates that alternative splicing in worm can be quite complex. The current WormBase model is shown at the top with our aggregate integrated transcript models shown below. Raw read counts per base for early embryo (orange) reveal clearly evident splice junctions, whereas in L3 (blue gray), a series of introns are apparently read through without splicing until apparently splicing to either the penultimate exon in the region or skipping this to the final exon shown.

Figure 3: Expression & Binding Dynamics

(A) (LEFT) Spearman correlation of expression of 8,428 genes across seven different stages of the *C. elegans* life cycle. Gene expression in early and late embryo are highly correlated with each other, while gene expression between larval and young adult stages are also highly correlated. (RIGHT) Cluster patterns similar to those in the left panel also become evident when correlating RNA Pol II binding levels across the same 8,428 genes. Values in each cell represent Spearman's ρ. For both RIGHT and LEFT panels correlations were done over 8,428 genes with simply defined TSSs (see supplement C.7).

(B) Spearman correlation of RNA Pol II binding levels and gene expression for 8,428 genes across seven different stages of the *C. elegans* life cycle. Correlation is generally high between RNA Pol II binding levels and gene expression within the same stage, with correlation the highest for the larval stages. Additionally, in considering correlation between stages, embryonic stages and larval/young adult stages form an symmetric pattern. RNA Pol II binding in the embryonic stages shows poor correlation with gene expression in the larval and young adult stages but expression in the embryo stages correlates moderately well with RNA Pol II binding in later stages. Values in each cell represent Spearman's ρ.

(C) Results of principal components analysis (PCA) of 6 matched tissue samples from the MxE and L2 stages. Tissues/cell types from embryo cluster together along both principal components. Tissues from L2 are differentiated with respect to their embryonic counterparts along principal components 1 and 2, spreading out with respect to the embryonic cluster. Component 1 is particularly enriched in genes associated with the GO terms nematode larval development, larval development, post embryonic development, and growth.

Figure 4: Non-protein-coding RNA

(A) The two panels illustrate the increased power achieved by combining features to discriminate between ncRNAs and other regions of the genome. These graphs show the distribution of expression feature values (e.g. from small RNA-seq) for genomic regions in the worm genome corresponding to ncRNAs and other types of sequence elements. The two panels show that while each feature alone cannot discriminate among different types of genomic elements, combining features into an integrated model can. The left panel shows the distributions of expression values for four representative features of the nine features examined using the gold-standard set of annotated regions (see (*47*)for the definition of the gold-standard set). The gold standard consists of four types of genomic elements: the known non-coding RNA, coding sequences (CDSs), untranslated regions (UTRs), and intergenic regions. A scatter plot of individual regions with values normalized to the same scale shows that the known ncRNAs are not readily distinguished from other regions, particularly using the bottom two features. At right, the maximum signal of polyA RNA on a tiling array is plotted in a two-dimensional scatter plot against predicted secondary structure conservation. Even using just two features, the ncRNAs begin to separate from the other regions. Expression values in the right panel are log-transformed normalized read counts (DCPM). Where multiple experimental data sets exist, the maximum value is used. The data used in the plots are from gold standard bins defined in (*47*).

(B) Example of a novel ncRNA with support from multiple sources of information in embryos. Track labels are PHA-4, HLH-1, RNA Pol II: ChIP-seq reads from the indicated protein, where signal heights are normalized by their total mapped reads; H3K27ac (histone 3 lysine 27 acetylation), H3K4me (H3K4 methylation): log-transformed values of the ChIP-chip data for two chromatin features normally associated with active genes; PolyA and Small RNA-seq: reads from polyA-selected and small RNA sequencing; Total RNA tiling arrays: log-transformed values of transcription on the tiling array in embryo; TARs: Transcriptionally Active Regions called from the tiling array signal track; Refseq: annotated genes in the region. The grey box at center shows a novel non-coding RNA ~160 nt in length captured only by the tiling array, indicating that it is not polyadenylated and is longer than the 30 nt size cutoff of the small RNA-seq experiment.

(C) Example of a differentially transcribed pseudogene presumably creating a ncRNA. Rows represent normalized signal tracks derived from uniquely mapped reads for the developmental stages indicated. The left panel denotes the expression pattern of the parent gene (*T01B11.7.1*; shown in orange), whereas the right panel shows the expression profile of the associated pseudogene (duplicated pseudogene derived from PseudoPipe; *PP00501*; shown in green). The dashed vertical lines demarcate the exon boundaries of the parent gene and pseudogene. Note that the scales of the y-axis (normalized signal tracks) are set independently for the parent gene and the pseudogene. The expression patterns of the pseudogene and parent gene are discordant. For instance, the parent gene is expressed in mid-L2, while the pseudogene does not appear to be expressed in that development stage. On the other hand, the pseudogene is expressed in the dauer entry and dauer stages, whereas the parent gene is not expressed during these stages. The discordant expression patterns indicate that the pseudogene is expressed independently of the parent gene.

Figure 5: Transcription Factor Binding

(A) Examples of TF binding peaks at a HOT region (blue box) and two specific regions (orange boxes), integrated with chromatin features and expression data. The top tracks (red labels) show raw reads from ChIP-seq experiments where a GFP antibody precipitated each of the 22 TFs indicated, from strains carrying a GFP-tagged form of that TF. The remaining tracks are: polyA RNA-seq: the raw reads from early embryos; H3K4me and H3K27ac: the logarithm of ChIP-chip signals for histone methylation and acylation in early embryos; Pol II: raw reads of ChIP-seq data in early embryos; N2 Input track: raw reads from genomic DNA of the N2 strain; N2 GFP: ChIP-seq of genomic DNA from the N2 strain with pull-down by a GFP antibody; EGL-27 IgG: ChIP-seq of genomic DNA from the EGL-27 strain with pull-down by IgG antibody; EGL-27 and HLH-1 Input: raw reads from genomic DNA of the EGL-27 and HLH-1 strains, respectively. All tracks of TFs are scaled based on the total mapped reads of each experiment (To fit all the tracks on a consistent scale, the range of each track is ~20 to ~200 reads). The RNA-seq range is between 1 to 20 reads. High peaks exceeding the maximum range are truncated, as indicated by a dotted line. The displayed region is chrIII:7,193,967-7,224,48, where the HOT peak (blue box) is adjacent to the *rpl-6* (ribosomal large subunit) gene; downstream, there are specific peaks for the HLH-1 and UNC-130 TFs (orange boxes) upstream of *R151.1*, which is a muscle-enriched gene. The two specific binding peaks are enlarged at right. (Note, ChIP-chip is chromatin immunoprecipitation ("ChIP") using microarray technology ("chip").)

(B) Average ChIP-seq signal around the transcript start site (TSS) of target coding (red) and non-coding (blue) transcripts for four representative TFs. The signal is the normalized mapped reads over input at each position (window size is 100nt).

(C) Enrichment of binding targets and signal of TFs in non-coding vs. coding genes. Max signal value represents the ratio of maximum binding signal of a TF around its target non-coding genes to that of its target coding genes. Target fraction represents the ratio of target percentage in non-coding genes to that in coding genes. Only TFs at the larval stages are shown. Some factors such as GEI-11 clearly bind more to ncRNA than others (e.g. PHA-4).

Figure 6: Highly Occupied Target (HOT) Regions

(A) Compares the TF binding signals 304 HOT regions (left) across the *C. elegans* genome are bound by 15 or more TFs (out of 23 ChIP-seqs) and 150 regions (right) randomly chosen from the rest of the genome that indicate little to no TF binding. The results show that the HOT regions are highly enriched with TFs. Each row represents a TF ChIP-seq dataset, and rows are ordered by the number of HOT regions bound. Columns are separated by chromosome, and then ordered from left to right by increasing numbers of TF's bound. ChIP-seq experiments were performed in worms synchronized for a specific developmental stage as indicated. Black indicates regions without significant TF binding, while colored regions have significant binding (maximum *q*-value 1e-5) colored based on the enrichment *q*-value (with increasing significance from red to white).

(B) (Left) HOT regions containing HLH-1 binding show a relative lack of HLH-1 binding motifs. In black, the frequency of the *in vitro* HLH-1 binding motif (hexamer CAGCTG) is greater in HLH-1-specific regions than in HLH-1 binding sites within HOT regions. The sequences in HLH-1 peak regions were randomized using the Fisher-Yates shuffling algorithm, and motif density was calculated for these shuffled regions (grey bar, error bars indicate standard deviation). 598 HLH-1 specific targets are defined as regions with 1-4 factors (including HLH-1); 165 HOT regions are bound by 15 or more factors (requiring inclusion of HLH-1).
(Right) HLH-1 binding does not correlate with muscle expression in HOT regions.

Genes associated with specific peaks for HLH-1, a muscle-specific TF, are over 7-fold more likely to be muscle-specific genes (*61, 63*) than genes located near HLH-1-containing HOT regions. For each dataset, the frequency of muscle-specific genes is shown in black, and the frequency in random gene sets of equal size is shown in grey (error bars indicate standard deviation).

(C) 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 (*64*)). 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. (See Fig. S29 for row annotations.)

(D) Genes near HOT regions are enriched for essential function. Genes were separated based upon the presence of ChIP-seq peaks within 1kb of the TSS. The y-axis shows the percent of genes bound only by 1-4 factors (“specific targets”) or genes bound by 15 or more factors (“HOT regions”) that serve essential functions, as indicated by RNAi knockdown. The dotted line signifies the percentage of all genes that are essential. By Chi-square test, genes nearby HOT regions are significantly more likely to be essential (9-fold; p < 10-40), whereas genes that only had specific peaks were not.

Figure 7: Integrated Regulatory Network

The TFs (blue triangles) in panels **A**, **B**, and **C** are organized hierarchically: the top layer has TFs that are not regulated by any other TF in the network, the second layer contains TFs that are regulated by and also regulate other TFs in the network, and the third layer contains TFs that are regulated by but do not regulate any other TFs. In addition a sampling of TF interactions with other genes (purple) is shown in the fourth row. The TFs are as follows. In the first row: LIN-11, MDL-1, PQM-1, LIN-15B, MAB-5, ALR-1, ELT-3, PES-1, CEH-14; in the second row: PHA-4, EOR-1, LIN-39, BLMP-1; in the third row: GEI-11, UNC-130, EGL-5, SKN-1, EGL-27. The totals interactions associated with each type of edge are shown in parentheses.

(A) shows all of the TF-TF interactions in the larval stages based on the TF binding experiments with the HOT regions already removed. The HOX genes are highlighted in the figure.

(B) shows the interactions of the same genes after filtering by gene expression correlation, calculated using Pearson's correlation and allowing a stage shift of +/- one stage, with the maximum correlation being used. Correlations with absolute values less than 0.75 were filtered out. Essential TFs are highlighted red in this panel. The key to the colors used in the edges in **B** and **C** is shown to the right.

(C) shows interactions between miRNA (red circles) and TFs. miRNAs that bind to 3'

UTRs of TFs are shown on the left whereas those that fail to bind to the 18 TFs but do have associated TF binding sites upstream are shown to the right. miRNAs that had their max larval expression level lower than 10% of their maximum expression level in all stages were filtered out. The miRNAs are also arranged hierarchically based on their interactions with the TFs. At left is an indication for the average value of the number of protein-protein interaction and tissue specificity for each level (see Supplement D.6 for detail on tissue specificity calculation). The star indicates that the difference the top and lower (bottom and middle) layers is statistically significant. (Specifically, it is P=0.002 for degree in protein-protein interaction network and P=0.003 for tissue specificity score). The differences between the levels for HOX genes and essential genes, while notable, are not statistically significant given the small size of the network.

(D) shows significantly enriched network motifs involving all of the interactions in the figure.

Figure 8: Chromosome-scale domains of chromatin organization

(A, B) Data from whole-genome ChIP-chip experiments for various histone modifications and chromatin-associated proteins, along with other relevant genome annotations, were normalized, placed into 10 kb bins, and displayed as a heatmap. Red indicates a higher value, whereas blue signifies a weaker signal. The continuous black line plots the relationship between physical and genetic distance (*135*). Three major groups of data were identified by hierarchical clustering. Group 1 contains H3K9 methylation marks and LEM-2, which tend to be enriched at distal autosomal regions, and correlate with repetitive DNA and a high recombination rate. Group 2 contains the dosage compensation complex members and H4K20me1, which are highly enriched on the X chromosome. Group 3 contains regulatory element and gene-body marks for active chromatin. In general, the signal for active marks is much weaker on the X chromosome than on autosomes. This megabase-scale chromatin organization persists throughout all stages of development and adulthood. Chromosome III **(A)** is representative of all autosomes, whereas the X **(B)** has a distinct chromatin configuration. **(C, D)** The H3K9me1/2/3 signals gradually decrease at the boundaries between the central and distal domains, whereas the boundary defined by LEM-2 is relatively sharp.

Figure 9: Chromatin Patterns around Genes

Average gene profiles around the TSS and TTS of various histone marks displayed for the X chromosome (red), and autosomes (blue). Genes were further stratified according to their expression level, with the top 20% of expressed genes shown in darker shade, and the bottom 20% of expressed genes shown in lighter color. The top two panels show that histone variant H3.3 marks regions of active chromatin on both autosomes and the X chromosome. _Marks typically associated with active or repressed transcription are labeled on the left. Plots from L3 worms (bottom row), highlight some of the differences in histone mark patterns between the EE and L3 stages. For example, H3K27me1 and H3K27me3 show stronger enrichment on expressed genes on the X in EE, whereas H4K20me1 is more strongly enriched on the X in L3.

Figure 10: Integrated Statistical Models Predicting Regulation and Expression from Chromatin Features

The figure shows the results of developing statistical models to predict TF binding and gene expression from chromatin features.

(A) Modeling binding peaks of TFs using chromatin features. The color of each cell represents the modeling accuracy (quantified by the area under the receiver-operating characteristics curves, AUROC) of a statistical model where a single chromatin feature acts as a predictor for the binding peaks from a TF binding experiment. The last row shows the prediction accuracy for the HOT regions.

(B) An example of combining chromatin and sequence features. Potential binding sites of HLH-1 were identified by the MEME algorithm (*57, 58*). Chromatin features were used to model general potential binding active regions (BAR+) that are not specific to any DNA-binding proteins. 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+). Clearly the last combination does the best.

(C) Correlation pattern of each chromatin feature in 160 100-nt bins around the TSS and TTS (from 4kb upstream to 4kb downstream from each) of worm transcripts at the EE stage. The Spearman correlation coefficient of each chromatin feature with gene expression levels was calculated for each bin. Ab1 and Ab2 represent experimental results using different antibodies for the same chromatin feature.

(D) Chromatin models enhance the accuracy of predicting expression levels for both protein-coding genes and miRNAs. The x-axis shows the expression levels of protein-coding genes (upper and middle figures) and miRNAs (lower figure), measured by RNA-seq and small RNA-seq, respectively. The y-axis shows the RNA Pol II binding signals (upper figure) and predicted expression levels for protein-coding genes (middle figure) or microRNAs (lower figure) by chromatin models. For the latter miRNA prediction we used model is just trained on the chromatin features of protein coding genes.

Figure 11: Relative Proportion of Different Annotations Among Constrained Sequences

(A) shows the proportion of constrained and unconstrained blocks in the *C. elegans* genome as a pie chart. The bar shows the cumulative proportion of the constrained region covered by various types of annotated functional elements, starting with CDS at the bottom. The indicated percentage shows the increase in the total proportion of the constrained region covered by the addition of that element type

(B) Kernel density estimate of distribution of element conservation. Enrichment of constrained bases among several types of annotated elements, relative to a background model drawn from a random distribution of untranscribed regions of the genome are plotted across the entire conservation spectrum (0 having no alignment, 1 having an alignment perfectly generated by the conserved state of the PhastCons phylo-HMM (*134*)). Points above and below the dotted horizontal line are enriched and depleted, respectively, relative to expectation. Higher conservation values represent subsets of bases within each annotation class that are more evolutionarily constrained.

(C) shows the fraction of constrained and unconstrained bases underneath annotated elements. 95% confidence intervals for random placement of elements are indicated and were calculated using the GSC algorithm (*1*). If the ends of the columns are outside the confidence interval, then it is unlikely that the fraction of constrained or unconstrained bases underneath the annotation could have occurred by chance. Thus CDSs cover a higher fraction of constrained bases and a lower fraction of unconstrained bases than expected by chance, and the opposite is true for pseudogenes. Annotation types: *CDS:* all predicted and confirmed coding regions from WormBase; *5' UTR, 3' UTR:* WormBase 5'- and 3'-UTRs confirmed by EST alignments; *pseudogenes:* WormBase-designated pseudogenes; *small RNA*: Small RNAs identified by modENCODE; *TF sites:* Transcription factor binding sites identified by modENCODE; *Dosage compensation:* the union of binding site peaks for the factors DPY-27, DPY-28, MIX-1, SDC-2 and SDC-3; *ChIP-Other*: the union of binding site peaks for the factors HCP-3, LEM-2, MES-4 and MRG-1; HOT: ChIP target regions occupied by at least 15 TFs.

References

1. The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
2. S. E. Celniker *et al.*, Unlocking the secrets of the genome. *Nature* **459**, 927-930 (2009).
3. S. Brenner, Genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94 (1974).
4. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic-cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64-119 (1983).
5. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* **314**, 1-340 (1986).
6. C. elegans Sequencing Consortium, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
7. M. M. Metzstein, G. M. Stanfield, H. R. Horvitz, Genetics of programmed cell death in C-elegans: past, present and future. *Trends Genet.* **14**, 410-416 (1998).
8. A. Fire *et al.*, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811 (1998).
9. V. Ambros, microRNAs: Tiny regulators with great potential. *Cell* **107**, 823-826 (2001).
10. G. Ruvkun, Molecular biology - Glimpses of a tiny RNA world. *Science* **294**, 797-799 (2001).
11. A. Agarwal *et al.*, Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**, 383 (2010).
12. N. L. Washington *et al.*, The modENCODE Data Coordination Center: Lessons in harvesting comprehensive experimental details. *Genome Res.*, (2010).
13. S. Contrino *et al.*, Accessing modENCODE data. *Genome Res.*, (in preparation).
14. modMine, http://intermine.modencode.org.
15. Short Read Archive, http://www.ncbi.nlm.nih.gov/sra.
16. D. L. Wheeler *et al.*, Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13-D21 (2008).
17. Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/.
18. WormBase, http://www.wormbase.org.
19. T. W. Harris *et al.*, WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38**, D463-D467 (2010).
20. J. Reboul *et al.*, C-elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet.* **34**, 35-41 (2003).
21. P. Lamesch *et al.*, C-elegans ORFeome version 3.1: Increasing the coverage of 2064 ORFeome resources with improved gene predictions. *Genome Res.* **14**, 2064-2069 (2004).
22. U. Nagalakshmi *et al.*, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
23. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621-628 (2008).
24. L. W. Hillier *et al.*, Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.* **19**, 657-666 (2009).
25. M. Mangone *et al.*, The Landscape of C. elegans 3'UTRs. *Science* **329**, 432-435 (2010).
26. 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)).
27. M. A. Allen, L. W. Hillier, R. H. Waterston, T. Blumenthal, A global analysis of trans-splicing in C. elegans. *Genome Res.*, (submitted).
28. W. C. Spencer *et al.*, A spatial and temporal map of C. elegans gene expression. *In preparation*, (in preparation).
29. P. M. Harrison, M. Gerstein, Studying Genomes through the aeons: Protein families, pseudogenes and proteome evolution. *J Mol Biol* **318**, 1155-1174 (2002).
30. O. H. Tam *et al.*, Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534-538 (2008).

31. D. Y. Zheng, M. B. Gerstein, The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* **23**, 219-224 (2007).

32. R. Sasidharan, M. Gerstein, Protein fossils live on as RNA. *Nature* **453**, 729-731 (2008).

33. Z. L. Zhang *et al.*, PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-1439 (2006).

34. T. Kondo *et al.*, Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis. *Science* **329**, 336-339 (2010).

35. M. Kato, F. J. Slack, microRNAs: small molecules with big roles - C. elegans to human cancer. *Biol. Cell* **100**, 71-81 (2008).

36. M. R. Fabian, N. Sonenberg, W. Filipowicz, Regulation of mRNA Translation and Stability by microRNAs. *Annu. Rev. Biochem.* **79**, 351-379 (2010).

37. V. Reinke, Gene Regulation: A Tale of Germline mRNA Tails. *Curr. Biol.* **18**, R915-R916 (2008).

38. J. Brennecke *et al.*, Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007).

39. 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).

40. R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, M. B. Gerstein, Annotating non-coding regions of the genome. *Nature Reviews Genetics* **11**, 559-571 (2010).

41. M. Stoeckius *et al.*, Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat. Methods* **6**, 745-U716 (2009).

42. J. G. Ruby, C. H. Jan, D. P. Bartel, Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).

43. 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).

44. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-D158 (2008).

45. Chung, E. C. Lai, modENCODE mirtron companion paper. *Genome Res.*, (submitted).

46. 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).

47. 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).

48. T. E. Royce, J. S. Rozowsky, M. B. Gerstein, Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res.* **35**, e99 (2007).

49. H. van Bakel, C. Nislow, B. J. Blencowe, T. R. Hughes, Most Dark Matter Transcripts Are Associated With Known Genes. *PLoS. Biol.* **8**, e1000371 (2010).

50. P. P. Gardner *et al.*, Rfam: updates to the RNA families database. *Nucleic Acids Res.* **37**, D136-D140 (2009).

51. M. Frasch, A matter of timing: microRNA-controlled temporal identities in worms and flies. *Genes Dev.* **22**, 1572-1576 (2008).

52. W. Niu, ..., V. Reinke, Diverse transcription factor binding features revealed by genome-wide ChIP-Seq in C. elegans. *In preparation*, (in preparation).

53. 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).

54. J. Rozowsky *et al.*, PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66-75 (2009).

55. 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).

56. G. D. Stormo, DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).

57. 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).

58. 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).

59. W. Niu *et al.*, Systematic dissection of regulatory networks dictated by *C. elegans* sequence-specific transcription factors. *In preparation*, (in preparation).
60. Supplemental files are available at http://www.modencode.org/publications/integrative_worm_2010/.
61. T. Fukushige, M. Krause, The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early C-elegans embryos. *Development* **132**, 1795-1805 (2005).
62. C. A. Grove *et al.*, A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. *Cell* **138**, 314-327 (2009).
63. 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).
64. X. Liu *et al.*, Analysis of Cell Fate from Single-Cell Gene Expression Profiles in C. elegans. *Cell* **139**, 623-633 (2009).
65. R. S. Kamath *et al.*, Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-237 (2003).
66. The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25-29 (2000).
67. C. Moorman *et al.*, Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12027-12032 (2006).
68. S. MacArthur *et al.*, Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol.* **10**, R80 (2009).
69. T. I. Lee *et al.*, Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
70. G. Balazsi, A. L. Barabasi, Z. N. Oltvai, Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7841-7846 (2005).
71. J. A. Wagmaister *et al.*, Identification of cis-regulatory elements from the C-elegans Hox gene lin-39 required for embryonic expression and for regulation by the transcription factors LIN-1, LIN-31 and LIN-39. *Dev. Biol.* **297**, 550-565 (2006).
72. H. Y. Yu, M. Gerstein, Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14724-14731 (2006).
73. N. Bhardwaj, K.-K. Yan, M. B. Gerstein, Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proceedings of the National Academy of Sciences* **107**, 6841-6846 (2010).
74. N. Simonis *et al.*, Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **6**, 47-54 (2009).
75. S. Lall *et al.*, A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.* **16**, 460-471 (2006).
76. S. Mangan, U. Alon, Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11980-11985 (2003).
77. U. Alon, Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**, 450-461 (2007).
78. N. J. Martinez *et al.*, A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev.* **22**, 2535-2549 (2008).
79. L. G. Edgar, N. Wolf, W. B. Wood, Early transcription in Caenorhabditis elegans embryos. *Development* **120**, 443-451 (1994).
80. G. Seydoux, A. Fire, Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. *Development* **120**, 2823-2834 (1994).
81. B. J. Meyer, X-Chromosome dosage compensation. *WormBook*, 1-14 (2005).
82. W. G. Kelly *et al.*, X chromosome silencing in the germline of C. elegans. *Development* **129**, 479-492 (2002).
83. T. M. Barnes, Y. Kohara, A. Coulson, S. Hekimi, Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics* **141**, 159-179 (1995).
84. M. V. Rockman, L. Kruglyak, Recombinational Landscape and Population Genomics of Caenorhabditis elegans. *PLoS Genet.* **5**, e1000419 (2009).
85. C. M. Whittle *et al.*, The Genomic Distribution and Function of Histone Variant HTZ-1 during C-elegans Embryogenesis. *PLoS Genet.* **4**, e1000187 (2008).

86. R. Gassmann *et al.*, Centromeric Chromatin Domains are Defined by an Inverse Relationship to Germline Transcriptional Memory in C. elegans. (To be Submitted).

87. T. Liu, A. Rechsteiner, J. Ahringer, S. Strome, X. Liu, C. elegans chomosome organization revealed by histone modificaiton mapping. *Genome Res*, (To Be Submitted).

88. S. L. Ooi, J. G. Henikoff, S. Henikoff, A native chromatin purification system for epigenomic profiling in Caenorhabditis elegans. *Nucleic Acids Res.* **38**, e26 (2010).

89. D. G. Albertson, A. M. Rose, A. M. Villeneuve, in *C. elegans II,* D. L. Riddle, T. Blumenthal, B. J. Meyer, J. R. Preiss, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1997), pp. 47–78.

90. A. J. Macqueen *et al.*, Chromosome sites play dual roles to establish homologous synapsis during meiosis in *C. elegans*. *Cell* **123**, 1037-1050 (2005).

91. S. Guoping, A. Fire, Partitioning the C. elegans genome by nucleosome modification, occupancy, and positioning. *Chromosoma* **119**, 73-87 (2010).

92. C. M. Phillips *et al.*, Identification of chromosome sequence motifs that mediate meiotic pairing and synapsis in C. elegans. *Nat. Cell Biol.* **11**, 934-U966 (2009).

93. 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).

94. S. Ercan, L. L. Dick, J. D. Lieb, The C. elegans Dosage Compensation Complex Propagates Dynamically and Independently of X Chromosome Sequence. *Curr. Biol.* **19**, 1777-1787 (2009).

95. A. Rechtsteiner *et al.*, The histone H3K36 methyltransferase MES-4 acts epigenetically to transmit the memory of germline gene expression to progeny. *PLoS Genet.*, (2010 (In Press)).

96. J. Jans *et al.*, A condensin-like dosage compensation complex acts at a distance to control expression throughout the genome. *Genes Dev.* **23**, 602-618 (2009).

97. A. N. D. Scharf *et al.*, Monomethylation of Lysine 20 on Histone H4 Facilitates Chromatin Maturation. *Mol. Cell. Biol.* **29**, 57-67 (2009).

98. H. Oda *et al.*, Monomethylation of Histone H4-Lysine 20 Is Involved in Chromosome Structure and Stability and Is Essential for Mouse Development. *Mol. Cell. Biol.* **29**, 2278-2295 (2009).

99. A. Kohlmaier *et al.*, A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS. Biol.* **2**, e171 (2004).

100. K. Ikegami, J. D. Lieb, Interactions between chromosomes and the nuclear envelope in C. elegans. *Genome Res*, (To be submitted).

101. T. Kouzarides, Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

102. S. Ercan, Y. Lubling, E. Segal, J. D. Lieb, DNA-encoded differences in nucleosome organization between the C. elegans X chromosome and autosomes. *Genome Res*, (To be Submitted).

103. S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, A. Z. Fire, Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.* **16**, 1505-1516 (2006).

104. A. Valouev *et al.*, A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051-1063 (2008).

105. N. Kaplan *et al.*, The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366 (2009).

106. D. Tillo, T. R. Hughes, G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**, 442 (2009).

107. D. Tillo *et al.*, High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. *PLoS One* **5**, e9129 (2010).

108. W. Horz, W. Altenburger, Sequence Specific Cleavage of DNA by Micrococcal Nuclease. *Nucleic Acids Res.* **9**, 2643-2658 (1981).

109. C. Dingwall, G. P. Lomonossoff, R. A. Laskey, High Sequence Specificity of Micrococcal Nuclease. *Nucleic Acids Res.* **9**, 2659-2673 (1981).

110. J. T. Flick, J. C. Eissenberg, S. C. R. Elgin, Micrococcal Nuclease as a DNA Structural Probe - Its Recognition Sequences, Their Genomic Distribution and Correlation with DNA-Structure Determinants. *J Mol Biol* **190**, 619-633 (1986).

111. G. C. Yuan *et al.*, Genome-scale identification of nucleosome positions in S-cerevisiae. *Science* **309**, 626-630 (2005).

112. I. Albert *et al.*, Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572-576 (2007).

113. W. Lee *et al.*, A high- resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235-1244 (2007).
114. Y. Field *et al.*, Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
115. E. Segal, J. Widom, Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struc Biol* **19**, 65-71 (2009).
116. M. A. Keene, S. C. R. Elgin, Micrococcal Nuclease as a Probe of DNA-Sequence Organization and Chromatin Structure. *Cell* **27**, 57-64 (1981).
117. D. J. Clark, Nucleosome Positioning, Nucleosome Spacing and the Nucleosome Code. *J Biomol Struct Dyn* **27**, 781-793 (2010).
118. A. M. Tsankov, D. A. Thompson, A. Socha, A. Regev, O. J. Rando, The Role of Nucleosome Positioning in the Evolution of Gene Regulation. *PLoS. Biol.* **8**, e1000414 (2010).
119. Y. Y. Fong, L. Bender, W. C. Wang, S. Strome, Regulation of the different chromatin states of autosomes and X chromosomes in the germ line of C-elegans. *Science* **296**, 2235-2238 (2002).
120. L. B. Bender *et al.*, MES-4: an autosome-associated histone methyltransferase that participates in silencing the X chromosomes in the C-elegans germ line. *Development* **133**, 3907-3917 (2006).
121. K. O. Kizer *et al.*, A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3K36 methylation with transcript elongation. *Mol. Cell. Biol.* **25**, 3305-3316 (2005).
122. J. X. Li, D. Moazed, S. P. Gygi, Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. *J. Biol. Chem.* **277**, 49383-49388 (2002).
123. N. J. Krogan *et al.*, Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **23**, 4207-4218 (2003).
124. P. S. Maddox, K. Oegema, A. Desai, I. M. Cheeseman, "Holo"er than thou: Chromosome segregation and kinetochore function in C. elegans. *Chromosome Res.* **12**, 641-653 (2004).
125. B. M. Lee, L. C. Mahadevan, Stability of Histone Modifications Across Mammalian Genomes: Implications for 'Epigenetic' Marking. *J. Cell. Biochem.* **108**, 22-34 (2009).
126. S. L. Berger, The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412 (2007).
127. K. A. Gelato, W. Fischle, Role of histone modifications in defining chromatin structure and function. *Biol Chem* **389**, 353-363 (2008).
128. S. Sinha, A. S. Adler, Y. Field, H. Y. Chang, E. Segal, Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res.* **18**, 477-488 (2008).
129. Y. Zhang *et al.*, Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucleic Acids Res.* **37**, 7024-7038 (2009).
130. K. J. Won, B. Ren, W. Wang, Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.* **11**, R7 (2010).
131. J. Ernst, H. L. Plasterer, I. Simon, Z. Bar-Joseph, Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* **20**, 526-536 (2010).
132. G. M. Cooper *et al.*, Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14**, 539-548 (2004).
133. A. M. Moses *et al.*, Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.* **2**, e130 (2006).
134. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050 (2005).
135. C. Rezvoy, D. Charif, L. Gueguen, G. A. B. Marais, MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* **23**, 2188-2189 (2007).

Acknowledgements

The authors acknowledge funding from the National Human Genome Research Institute of the NIH as part of the modENCODE project. They also acknowledge funding from NIH (grant R01GM088565), Muscular Dystrophy Association and the Pew Charitable Trusts to John K. Kim; the Helmholtz-Alliance on Systems Biology (Max Delbrück Centrum Systems Biology Network) to Sebastian D. Mackowiak; and the AL Williams Professorship funds to Mark Gerstein. They acknowledge help from Sarah Anthony, Kim Bell, Colleen Davis, Yair Field, Ann S. Hammonds, Jeanyoung Jo, Noam Kaplan, Arjun Manrai, Bernard Mathey-Prevot, Rebecca McWhirter, Stephanie Mohr, Stephen Von Stetina, Joseph Watson, Kathie Watkins, Chenghai Xue, and Yong Zhang. They thank Bob Carpenter for co-development of the deepseq9 algorithm. We thank Calvin Jan and David Bartel sharing their data on polyA sites with us prior to publication.