Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*

LaDeana W. Hillier, Valerie Reinke, Philip Green, et al.

Genome Res. 2009 19: 657-666 originally published online January 30, 2009
Access the most recent version at doi:10.1101/gr.088112.108

Resource

Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*

LaDeana W. Hillier,[1] Valerie Reinke,[2] Philip Green,[1,3] Martin Hirst,[4] Marco A. Marra,[4] and Robert H. Waterston[1,5]

[1]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195-5065, USA;* [2]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520-8005, USA;* [3]*Howard Hughes Medical Institute, Seattle, Washington 98195, USA;* [4]*British Columbia Cancer Agency (BCCA), Genome Sciences Centre, Vancouver, British Columbia V5Z 456, Canada*

Using massively parallel sequencing by synthesis methods, we have surveyed the polyA+ transcripts from four stages of the nematode *Caenorhabditis elegans* to an unprecedented depth. Using novel statistical approaches, we evaluated the coverage of annotated features of the genome and of candidate processed transcripts, including splice junctions, *trans*-spliced leader sequences, and polyadenylation tracts. The data provide experimental support for >85% of the annotated protein-coding transcripts in WormBase (WS170) and confirm additional details of processing. For example, the total number of confirmed splice junctions was raised from 70,911 to over 98,000. The data also suggest thousands of modifications to WormBase annotations and identify new spliced junctions and genes not part of any WormBase annotation, including at least 80 putative genes not found in any of three predicted gene sets. The quantitative nature of the data also suggests that mRNA levels may be measured by this approach with unparalleled precision. Although most sequences align with protein-coding genes, a small fraction falls in introns and intergenic regions. One notable region on the X chromosome encodes a noncoding transcript of >10 kb localized to somatic nuclei.

[Supplemental material is available online at www.genome.org. These short-read sequence data have been submitted to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA003622.7. Alignments, confirmed sequence features, and relevant data have been submitted to the modENCODE Data Coordinating Center/WormBase.]

The protein-coding genes of a genome are one of its fundamental attributes, yet for *Caenorhabditis elegans* and metazoans more generally, the large introns, complicated alternative splice forms, and numerous pseudogenes have made an accurate, comprehensive annotation of protein-coding transcripts a challenging task. For example, for *C. elegans*, with its highly accurate genome sequence (The *C. elegans* Genome Sequencing Consortium 1998), systematic efforts to collect expressed sequence tag (EST), cDNA, and open-reading-frame sequence tag (OST) sequences (Waterston et al. 1992; Kohara 1996; Reboul et al. 2001, 2003) and other expression data sets (Wei et al. 2005; Merrihew et al. 2008; Rogers et al. 2008) over the past decade and a half have confirmed many of the initial gene predictions and modified others. Yet as of January, 2007, only 7825 genes (34%) had full experimental support for all splice junctions and resultant open reading frames, whereas one-fifth of the genes lacked any experimental support (Table 1).

Many of the unsupported genes have probably escaped previous efforts because the transcripts are expressed at low levels and/or only in very specific stages or conditions. Because biases in base order and composition used in gene prediction are often weaker in poorly expressed genes, even prediction of these genes can be problematic. These unsupported gene models may contain fused or split genes, incorrect splice junctions, and missing alternative splice forms, as suggested by the results from OST sequencing (Reboul et al. 2001, 2003). In addition, shotgun proteomics data suggest some genes or exons are missing altogether from the current WormBase predictions (Merrihew et al. 2008).

To remedy this situation, one of us (P. Green) has for several years undertaken a directed approach utilizing RT-PCR and RACE to test the remaining unconfirmed splice junctions and to define the 5′ and 3′ ends of transcripts. More recently, as part of the modENCODE Consortium, we have been able to expand this effort to examine RNA from specific stages, cells, and tissues to uncover previously unknown transcripts and to refine extant predictions (modENCODE Consortium, in prep.). Although initially the intent was to use tiling microarrays to help direct more targeted experiments, the emergence of powerful new sequencing technology offered the opportunity to explore the transcriptome by sequencing more deeply than had been possible by Sanger-based EST sequencing and more completely than possible through serial analysis of gene expression (SAGE) (Velculescu et al. 1995). For example, a single L2 larval stage animal has just 687 nuclei. If each cell contains ~25,000–50,000 mRNA molecules, based on interpolation from estimates for yeast and mammalian cells and supported by estimates of total RNA per isolated embryonic cell for *C. elegans* (see Methods for details; Hardy 1976; Hereford and Rosbash 1977; Coupar et al. 1978), an L2 animal contains 17–35 million mRNA molecules. With the Illumina 1G, as many as 30 million aligned sequence tags can be obtained from a single flow cell, thus sampling an average-sized RNA molecule reproducibly expressed in that stage once on average, although of course there will be statistical variation of the level of sampling.

We report here the results of applying this sequencing technology to polyadenylation-enriched mRNA from hermaphrodite

[5]Corresponding author.
E-mail waterston@gs.washington.edu; fax (206) 685-7301.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.088112.108.

Table 1. WormBase annotations

Exp.			Splice junctions		SL1	SL2			
Support	CDS[a]	Transcripts[b]	Predicted	Confirmed	Confirmed	Confirmed	5′ UTR[c]	3′ UTR	PolyA[d]
Full	7825	11,037	1772	31,397	3050	1071	9802	8135	880
Partial	10,746	11,620	18,720	41,578	1996	509	4460	4811	181
None	4653	4653	17,904	3	167	23	2	2	0

[a]CDS represents distinct protein-coding sequences from the initiator methionine to the termination codon, including alternative splice forms.
[b]Transcripts include alternative 5′ and 3′ UTRs, with counts of each based on whether the underlying CDS has full support, partial support, or no support.
[c]Number of transcripts in each class that have a 5′ UTR.
[d]Number of transcripts that have a polyA_site or signal sequence ≤25 bp from transcript end.

populations of L2, L3, L4, and young adult stages. Despite the limitations imposed by short, error-prone sequence reads, the data have been a rich source of new insight into the transcriptome. Not only have we been able to provide experimental support for a large number of splice junctions and transcripts that previously were simply predicted in WormBase, but we also have been able to identify new splice junctions by comparing the data to other gene predictions and through a computational search for novel splice junctions. The result is a substantially improved view of the worm transcriptome.

While this work was in progress, other studies employing related methods have also appeared documenting the power of deep RNA sequencing to studies of transcriptomes (Cloonan et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008). Our own goal of improving the annotation of gene models lacking experimental support led us to develop statistical approaches to evaluate coverage and to identify splice junctions, spliced leaders, polyadenylation sites, and transcription initiation and termination sites. With about one flow cell's worth of data from each of three larval stages and young adults we have, for example, found support for 20,121 WormBase splice junctions that lacked prior experimental evidence. The data confirm another 6137 splice junctions not represented in WormBase at all, including 983 not included in gene predictions from any of three different prediction programs.

Results

RNA-seq data sets

Using polyA+ RNA isolated from L2, L3, and L4 larval stages as well as young adults, we prepared double-stranded cDNA using random hexamer primers. After shearing, a ~200-bp fraction was isolated and used to generate libraries for massively parallel sequencing, using the Illumina 1G instrument. The resulting 36-base reads were aligned using MAQ (Li et al. 2008) and cross_match (P. Green, unpubl.) to the genome and to several databases representing processed RNA products.

For each of the four stages, 15–33 million reads aligned, of which ~91% were designated as high mapping quality by MAQ (Supplemental Table S1; Li et al. 2008). With ~25 million bases (Mb) of the genome transcribed as polyA+ RNA (WormBase WS170), these aligned reads achieve an average depth of coverage of 20- to 40-fold per base for each stage. Viewed another way, for L2 animals containing 687 nuclei and for young adults containing ~2000 nuclei (including germline nuclei), with each nucleus representing 25,000–50,000 mRNA molecules of an average length of 1.5 kb, each molecule of average length is sampled with a range of averages from 0.2 to 1.6 times (see Methods for details of these estimates).

Coverage of WormBase features in high-scoring bases

Preliminary inspection of the aligned reads with respect to features annotated in WormBase indicated that although most reads aligned to protein-coding regions, some reads were scattered in intergenic and intronic regions. The reads aligning to intronic and intergenic regions, often with only 24–26 bases of the read aligning rather than the longer alignments found in known coding regions, might have derived from previously unrecognized transcripts, but also might have resulted from artifacts associated with using such a large number of short, sometimes error-prone reads or from very low levels of contaminating genomic DNA. To reduce the contribution to coverage of such potential artifacts, we developed a scoring system that reflects read density in a window around each base, reasoning that reads arising from a transcript would be clustered, thereby increasing the score of a base, whereas artifacts would be more randomly distributed with the score, simply reflecting the isolated read. In addition, by providing a score for each base, further manipulations are simplified. To evaluate possible window sizes and the relationship of the scores to a false-positive rate, we used a set of confirmed WormBase exons as a positive control and intronic and intergenic regions depleted of other probable transcribed regions as a negative control (see Methods). Based on this ROC-like analysis (Supplemental Fig. S1), we selected a window size of 51 bases and established thresholds that yielded a cumulative false-positive rate of 5% for each stage as well as the aggregate data set.

Using these thresholds, we calculated coverage of WormBase features by high-scoring bases, that is, bases scoring above the threshold set for each stage (Table 2A). About 17 million high-scoring bases in each of the stages and almost 19 million bases in the combined data sets fell in exonic sequence (including untranslated regions [UTRs]). In contrast, 1.7–2.4 million high-scoring bases fell in introns and 4.3–5.0 million high-scoring bases were in intergenic regions. The number of high-scoring bases is remarkably similar across the stages and is only marginally greater in the aggregate data set, suggesting that the bulk of the transcribed genome is similar across these larval stages.

WormBase intergenic regions likely contain additional coding bases

Even with the scoring system, some 6–7 million high-scoring bases fall outside of WormBase exonic sequence, or about double what might be expected with a 5% false-positive rate. Others (He et al. 2007) using tiling array analysis also noted that a small fraction of the polyA+ signal arose from portions of the genome not annotated in WormBase. To investigate whether the fraction observed here might include polyA+ transcripts missing in WormBase annotations,

Table 2. Coverage of annotated features

Coverage of WormBase features

	Exonic		Intronic		Intergenic		Total	
	Bases	Percent	Bases	Percent	Bases	Percent	Bases	Percent
WB TOT[a]	25,196,984	100	31,126,299	100	42,287,639	100	98,610,922	100
Stage								
L2	17,097,242	67.9	2,440,886	7.8	4,843,187	11.5	24,381,315	24.7
L3	17,210,508	68.3	1,953,376	6.3	4,960,692	11.7	24,124,576	24.5
L4	17,025,950	67.6	1,796,841	5.8	4,273,535	10.1	23,096,326	23.4
YA	17,090,725	67.8	1,759,002	5.7	4,637,344	11	23,487,071	23.8
AG	18,957,102	75.2	2,770,106	8.9	5,779,481	13.7	27,506,689	27.9

Coverage of combined gene models

	Exonic		Intronic		Intergenic		Total	
	Bases	Percent	Bases	Percent	Bases	Percent	Bases	Percent
Com TOT[b]	44,482,639	100	36,857,949	100	17,322,420	100	98,663,008	100
Stage								
L2	20,624,497	46.4	2,220,499	6	879,045	5.1	23,724,041	24
L3	20,859,564	46.9	1,699,835	4.6	763,336	4.4	23,322,735	23.6
L4	20,615,125	46.3	1,585,744	4.3	585,246	3.4	22,786,115	23.1
YA	20,751,614	46.7	1,507,604	4.1	655,769	3.8	22,914,987	23.2
AG	23,123,854	52	2,439,352	6.6	936,917	5.4	26,500,123	26.9

[a]The total number of WormBase genes per category.
[b]The total bases per category in the combined predicted gene set, based on WormBase, Twinscan, and a revised Genefinder.
(YA) Young adult; (AG) aggregate data set containing L2 + L3 + L4 + YA.

we incorporated gene models derived from Twinscan (Korf et al. 2001) and revised Genefinder (P. Green, unpubl.) programs, where parameters for the latter were set to reduce the fraction of false negatives, albeit at the expense of an increased number of false positives. As a result, ~3.5 Mb of covered sequence shifted to the exonic category, and coverage of intergenic bases dropped to less than 1 million bases (Table 2B). This fivefold drop is much greater than the 60% decrease in total intergenic bases resulting from the increased predictions, suggesting that WormBase may be missing an important segment of protein-coding regions. For these revised intronic and intergenic regions, the number of high-scoring bases in each stage is ~5% or less. Thus, the vast bulk of the polyA+ transcriptome as sampled by the RNA-seq data appears to be confined to bases within protein-coding genes.

Coverage of individual genes

With the very high average sequence redundancy in our data sets, we evaluated the representation of individual WormBase transcripts. WormBase (WS170) annotates 20,069 protein-coding genes. Of these, many produce multiple transcripts through alternative splicing and alternative start and stops, yielding a total of 23,224 different protein-coding sequences (CDS). In addition, some genes have alternative polyA-addition sites or other alternative UTRs outside of coding sequence; these additional alternative forms lead to 27,310 distinct annotated protein-coding transcripts. In our aggregate data set, 13,869 of these transcripts had 100% of their bases represented in high-scoring bases, 19,737 transcripts (73%) had at least 90% of the bases covered, and 22,268 (82%) had at least 50% coverage (Supplemental Fig. S2). We also looked specifically at genes expressed in only one or few cells. The gene *daf-7*, which encodes a TGF-beta signaling molecule critical for control of entry into the dauer stage, is expressed in just the two ASI chemosensory neurons (Ren et al. 1996); it has 100% coverage in the L2 data set. The guanylyl cyclase *gcy-7*, expressed only in the ASEL sensory neuron, has 466 of 3138 bases above threshold, and its counterpart, *gcy-5*, in ASER shows 312 of its 3369 transcribed bases above threshold (Chang et al. 2003). Thus, even for genes expressed at no more than moderate levels, in just one of the 700–2000 cells of the animal we could detect their product with this level of sampling. Complete coverage of genes such as *gcy-5* and *gcy-7*, however, would require substantially greater depth of sequencing or normalization approaches.

To determine if the coverage seen in the ASE-specific genes and other weakly expressed genes was meaningful, we established a false-positive rate as a function of both the number of high-scoring bases and the number of independent high-scoring blocks using artificial transcripts (see Methods). From this analysis, the expression of the ASE genes, for example, is well above threshold; altogether, 21,600 WormBase transcripts from L2 are above threshold (see Supplemental Table S2 for summaries and complete lists for each stage).

In contrast, 1871 WormBase models had no bases scoring above threshold in the aggregate data set. We were interested to learn what classes of genes were represented in this set and if they were known to be poorly expressed in the stages we used. Some were genes with few or no unique 24-mers, the minimum length we used for alignment, and had thus been excluded from consideration. If we allowed coverage to include those ambiguous bases, the number of untouched gene models dropped to 1336 candidate transcripts. Of these, preliminary evidence suggests that 480 are represented in early or late embryos, so sampling more stages and populations enriched for males should reduce this number further. Table 3 shows the breakdown of the major classes of remaining putative transcripts based on their WormBase annotation. Seven-transmembrane/serpentine receptors dominate

Table 3. Genes with no high-scoring bases

Type	Genome total	Not covered	% not found
7TM/G-protein coupled receptor	1454	331	22.76
F-box	238	3	1.26
Zinc finger	236	3	1.27
Nuclear hormone receptors	85	3	3.53
Math (meprin-associated Traf homology)	62	2	3.23
Testis-specific protein TPX-1 like	13	4	30.77

the set. 81% of these unsampled transcripts have no experimental support in WormBase.

Depth of coverage along the length of genes

In inspecting coverage plots of individual genes, we noted that depth of coverage varied along the length of genes, as might be expected from chance, from context-specific effects on polymerase extension, and perhaps for other reasons, but we also noted a consistently increased depth at the 5′ ends and decreased depth near the 3′ ends. Because these distributions might provide useful signatures of the start and stop of transcripts, we investigated the distribution of reads in more detail.

We first plotted depth of coverage for genes normalized for length and for overall depth of coverage (Fig. 1A). Across the body of the gene, depth of coverage is quite uniform on average, but depth rose sharply at the 5′ end and fell at the 3′ end.

We then looked at the strand bias of the sequence reads in genes normalized only for overall depth of coverage. These plots reveal a strong strand bias, with almost no reads coming from the minus strand within 100 bases of the annotated start and a 3.5- to fourfold greater depth of coverage on the positive strand (Fig. 1B). There was a weaker, but reproducible bias at the 3′ ends (Fig. 1C). The bias at the ends might be explained by the library construction and sequencing methods: With 36-base reads from 200-base fragments, the bulk of the gene can be represented by reads from either end of the 200-base fragments and thus from either strand, but near the ends, the sequence can originate only from one strand. The difference in overall depth of coverage is consistent with random priming, where at the 5′ end multiple primers all end at the same place (and all the fragments produced by fragmentation end at the same place), whereas at the 3′ end, the number of initiated and extended molecules drops. Regardless of the underlying mechanism, the patterns, particularly at the 5′ end, should aid in recognizing the ends of the transcripts.

Expression levels of individual WormBase genes

Although our primary goal in generating the RNA-seq data was to improve the annotation of the *C. elegans* genome, we also explored the ability of the data to provide a digital measure of the expression levels of genes. Using RNA fragmentation to obtain a relatively random distribution of reads along the transcript length, others (Mortazavi et al. 2008) have shown that RNA-seq data can accurately reflect mRNA levels over seven orders of magnitude. Beyond the relatively random distribution of reads along the length of the transcripts, this requires that the results be normalized for transcript length to compare genes within the same sample and for the number of aligned reads in each sample to compare genes between samples. Since our analysis in the preceding section indicated our random-primed cDNA preparation and sequencing also produced a relatively random coverage for the body of the gene, we adapted the normalization methods of Mortazavi et al. (2008) to utilize high-scoring bases, and from the normalized coverage values of each base we calculated the expression levels for each transcript and exon across all four stages (Supplemental Table S3; see Methods for details).

In accord with microarray studies (Reinke et al. 2000; Jiang et al. 2001; Kim et al. 2001; Baugh et al. 2003, 2005), the expression levels of individual transcripts varied widely within each stage, spanning more than five orders of magnitude. For example, in the L2 data set, 3644 WormBase transcripts had an average depth of coverage per base per million reads (dcpm) of <0.05 (equivalent to onefold coverage for 20 million aligned reads),

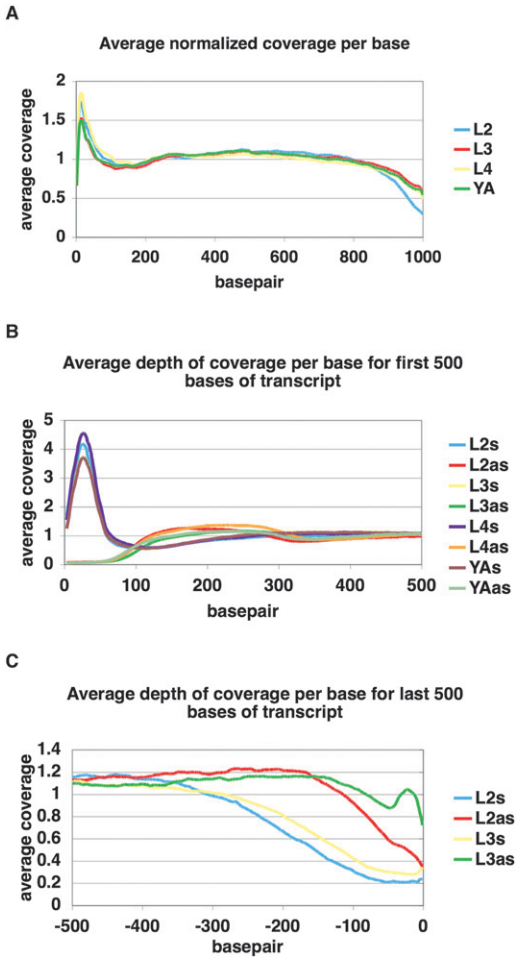

Figure 1. Relative coverage of WormBase genes by position and strand. (*A*) Results of analyzing 7571 full-length transcripts, normalized for length and depth of coverage, show almost uniform coverage along the body of the gene, with a pronounced overrepresentation at the 5′ end and decreasing depth of coverage at the 3′ end. (YA) Young adult. (*B*) Average depth of coverage at the 5′ end by strand, measured in bases from the start of the transcript. This plot confirms the overrepresentation at the 5′ end and demonstrates that it is almost entirely restricted to the coding strand. (*C*) Average depth of coverage at the 3′ end, by strand, measured from the transcript end. Coverage on both strands falls approaching the 3′ end, with loss of the sense (s) strand coverage followed by decrease of the antisense (as) to less than half the average depth of coverage. Only data for L2 and L3 are shown for clarity.

whereas a few genes, primarily those involved in protein synthesis, were very highly expressed in all stages with dcpm > 100. Inspection of the transcript length and GC content of the transcripts at the extremes did not reveal any systematic bias in dcpm values. (The overall distribution of expression levels for L2 is shown in Supplemental Fig. S3.)

We also compared transcript levels between stages. Many genes as measured by dcpm were expressed at similar levels across all four stages, with high Pearson correlation coefficients between the different stages, especially between the larval stages (Table 4; see also Supplemental Table S4 for a full list). However, again in accord with previous studies (Reinke et al. 2000; Jiang et al. 2001; Kim et al. 2001; Baugh et al. 2003, 2005), the dcpm measure also revealed genes that were highly differentially expressed between stages, including genes with known stage-specific expression patterns. For example, *vit-6* had no significant expression in the L2 stage in the RNA-seq data but rose to a dcpm of 94 in young adults, making it one of the most highly expressed genes. On the other hand, the *daf-7* gene described above is abundantly expressed in L2 (dcpm = 1.20) but falls almost 15-fold in young adults. The collagen genes also change dramatically over time, consistent with previous results (J. Kramer, pers. comm.). For example, *rol-1* levels increase 11,000-fold from the L2 to L4 stage and *col-19* rises from undetectable in L2 to a dcpm of 36 in young adults. In contrast, *col-17*, *rol-6*, and *sqt-1* levels fall by about 80-fold from L2 to young adult. As can be done with microarray data, the dcpm for transcripts across stages can be clustered by their patterns of gene expression, thereby revealing genes with potentially related transcriptional control (Fig. 2). Although biological replicates will be required to establish which differences are truly stage-specific as opposed to differences arising from particular growth conditions and technical variation, and more extensive comparisons with microarray and other data will be important, these initial observations establish that RNA-seq data as reflected in the dcpm metric can be a useful and precise measure of gene expression.

Representation of pseudogenes and noncoding genes

In addition to protein-coding genes, WormBase annotates pseudogenes and non-protein-coding transcripts. Inspection of our data shows that 272 of 954 annotated pseudogenes have high-scoring bases and 188 of these have expression for the pseudogene above threshold in the aggregate data set, suggesting that they were transcribed. Only ~15% of noncoding RNAs are sampled, probably reflecting the polyA enrichment used in making the library. Also, ribosomal RNAs were sampled, which because of their abundance and perhaps the presence of A-rich tracts within their sequences are not fully depleted by the polyA+ selection. Excluding the ribosomal components, only 15,225 known noncoding transcript bases are sampled. Other approaches will be required to recover these and novel noncoding transcripts.

Table 4. Pearson correlation coefficients for expression levels between stages

Stage	L2	L3	L4
L3	93.0	—	—
L4	92.5	94.6	—
Young adult	87.0	92.7	91.6

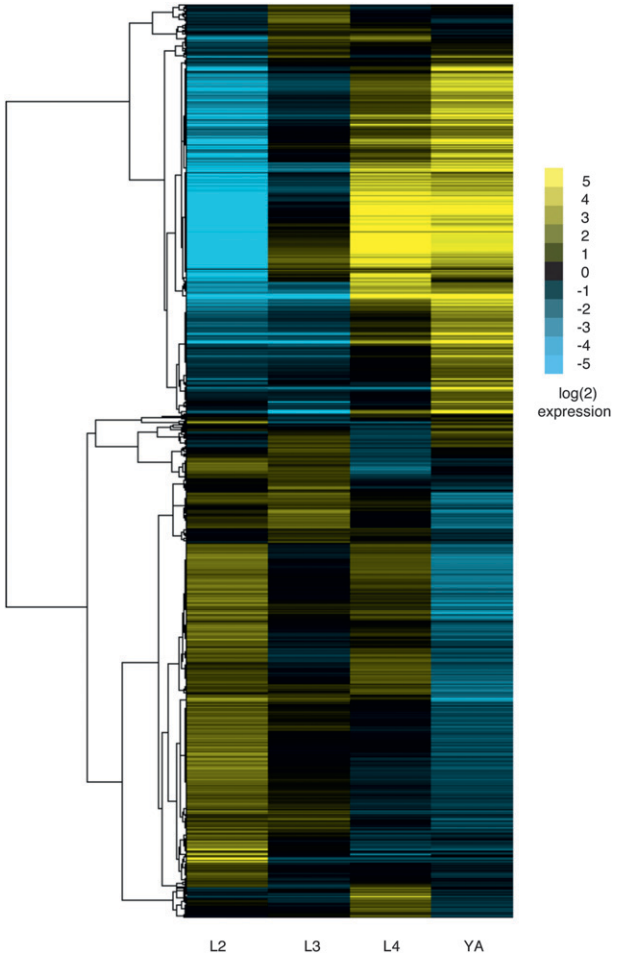

Figure 2. Heat map of the 2000 most variably expressed transcripts in the four *C. elegans* stages. The 15,000 most abundantly expressed transcripts, based on their mean expression, were evaluated for their variation in expression, based on the ratio of the variance to the mean. We arbitrarily selected the 2000 most variable transcripts (Supplemental Table S8), which were then clustered based on their coverage levels as reflected in the dcpm statistic. The columns represent the four stages, L2, L3, L4 and young adult (YA), respectively, and each row represents a different transcript. Transcripts absent in L2 and abundant in L4 and young adult include the major sperm protein (*msp*-), vitellogenin (*vit*-), and other transcripts related to the germline.

Post-transcriptional processing

In addition to evaluating representation of the genome in high-scoring bases to find evidence of transcription, we also looked for sequence reads that derived from post-transcriptional processing events, including formation of splice junctions, transplicing, and polyadenylation. Reads matching such sequences provide clear evidence of transcription and also can define the transcript with base-level precision. They also inherently define the strand of the transcript, something lacking in reads matching the genome. Importantly, finding such events may provide evidence for new transcripts not presently represented in any of the gene models.

To find splice junctions using these short reads, we created databases consisting of the 35 bases on either side of non-redundant splice junctions from WormBase, Genefinder, and Twinscan, noting those already confirmed and those lacking any experimental confirmation. We also combined splice sites from

existing models in novel combinations to look for alternative splice events as well as weak splice junctions not part of any gene model (see Methods for details). The large number of possible splice junctions considered, the very large number of imperfect short reads, and the frequent similarity of flanking intronic bases to the junction sequences all contribute to the likelihood of false-positive matches. To establish a false-discovery rate for each class of overlap of reads with a splice junction (length and number of reads), we created a set of false junctions composed of splice sites combined across chromosomes (see Methods for details). Table 5 shows the number of supported splice junctions, broken down by class and by the source of the junction. In addition to matching almost 91% (64,360/70,911) of previously confirmed junctions, the data provide support for 26,024 predicted junctions previously lacking any experimental support. Although the bulk of these represent splice junctions annotated in WormBase, importantly, more than 6137 of the confirmed junctions derive from either Twinscan or Genefinder gene models and are not represented in WormBase (Table 5B). We also note that the increase in newly supported junctions falls with each additional sample. The young adult sample, the last sample to be analyzed, added just 1239 new junctions, compared with 3721 when L3 is added to the initial L2 set.

In addition to these splice junctions that are part of existing gene models, we found support for a small but significant number of novel junctions identified by looking for splice junctions predicted using minimal constraints and by looking for new combinations of splice sites in existing models (Table 5A). Inspection of the novel junctions suggests that some represent new 5′ exons and others define new alternative exons or extensions to exons within currently predicted introns. A few appear to suggest new genes, including one in the intron of an existing gene transcribed in the opposite direction. The new combinations predominantly appear to be new alternative splice forms, but in some instances merge exons from different gene predictions.

For transplicing events and polyadenylation, we applied similar methods to find likely sites (Supplemental Tables S5, S6). About 60%–70% of worm genes are expected to receive a 22-base leader RNA sequence from transplicing. Although the functional importance of transplicing is unclear for the worm, when SL2 transplicing is the dominant event at a site, it generally (but not always) marks downstream genes in operons. WormBase lists 7635 confirmed SL1 sites and 2043 confirmed SL2 sites for a total of 8725 different confirmed SL sites (some sites have evidence for both SL1 and SL2 splicing). Our analysis detects 6673 confirmed SL1 acceptor sites and provides support for 4105 additional candidate SL1 sites. For SL2, 556 potential new sites were found in addition to 1959 previously confirmed sites. The 556 includes some sites only previously confirmed in WormBase as SL1 sites and supports the notion that SL2 and SL1 may be used at a few sites. These new sites are potentially valuable in assigning the start sites of transcripts, and the new SL2 sites where SL2 predominates could suggest additional operons.

Discovering new polyadenylation sites was less successful, perhaps because of the drop in representation near the 3′ end as well as the low information content of the polyT tract at the beginning of the reads. We found support for 434 of 1587 sites annotated as "polyA_site" in WormBase. However, the data suggest an additional 6212 previously unannotated transcript termination sites, the vast bulk of which lie near the end of a predicted transcript.

Table 5. Supported splice junctions

All supported junctions by category

Splice junction source	L2	L3	L4	YA	Total splice junctions
Confirmed[a]	56,410	56,367	53,788	54,515	64,360
Predicted_not_confirmed	18,885	18,415	17,389	17,417	26,258
Novel combinations[b]	277	271	218	206	360
Novel[b]	521	583	473	532	700
TOTAL	76,093	75,636	71,868	72,670	91,678

Source of supported, predicted-not-confirmed splice junctions

Model source	L2	L3	L4	YA	Total
WormBase[c]	14,859	14,537	13,733	13,648	20,121
Genefinder[d]	3384	3284	3131	3191	5154
Twinscan	642	594	525	578	983

[a]WormBase170 lists 70,911 different confirmed splice junctions of a total of 108,671 annotated splice junctions.
[b]Present in more than one stage.
[c]Includes junctions also predicted by Genefinder and Twinscan.
[d]Includes junctions also predicted by Twinscan.
(YA) Young adult.

Validation of supported processing sites

To obtain an independent estimate of the validity of our supported splice junctions and transplicing sites, we utilized data that have been collected in a parallel project targeting specific splice junctions for testing by RT-PCR and 5′ and 3′ RACE (P. Green, unpubl.). Primer pairs that spanned newly supported junctions were identified and the resultant sequences examined for their representation of the possible sites. In short, of the 1532 junctions from all stages that were spanned by the RT-PCR sequence reads, <3% were not supported by sequence alignments across the site. Inspection of the traces suggested that many of these failed to confirm the junction simply because of poor read quality (Supplemental Table S7). The RT-PCR-confirmed junctions included both 633 predicted but not confirmed junctions as well as 11 novel junctions. These failure rates are well within the predicted false-discovery rate of this data set.

Integrating the data sets

The coverage in high-scoring bases, splice junctions, transplicing sites, and polyadenylation sites can combine with the depth of coverage and strand biases at the start and stops of transcripts to give an improved picture of the polyadenylated transcriptome of *C. elegans*. One can imagine a program that would exploit the statistical measures we can provide for each of the features above and combine that with intrinsic signals and other experimental evidence to provide an automated, accurate view of much of the worm genome, along with likely levels of expression at different stages.

In the absence of such a sophisticated program, we nonetheless wanted to evaluate the extent to which the combined RNA-seq data sets modified and supplemented existing WormBase models. To facilitate this endeavor, we built a tool that, using only the data sets described here, begins with the highest-scoring splice junction in a region and extends from there based on coverage data and additional splice sites to create gene models. When coverage drops to zero, the model is terminated. The tool then evaluates the ends to see if the termination points correspond to transplice or polyadenylation sites and regions of strand bias. By

using all splice junctions in a region it is also able to generate alternative splice forms, but of course the short reads do not disambiguate complicated splicing patterns for genes with multiple alternative exons. Because these models are truncated when coverage falls to zero and only use the RNA-seq data, we have dubbed these models *genelets*. Nonetheless, where coverage is high the genelets might be expected to correspond well to known genes. Indeed, for the 50% of WormBase gene models with 100% coverage in high-scoring bases, there is a genelet model that incorporates all of the WormBase model in >95% of cases.

The genelets are also able to incorporate splice junctions and covered regions not part of any predicted transcripts, resulting in a wealth of putative revised gene models. For example, in the L2 data set, of 25,190 potentially protein-coding genelets, 4039 genelets incorporate splice junctions from Genefinder or Twinscan that are not represented in WormBase. For each stage there are around 800 alternative splice junctions incorporating alternative 5′ or 3′ exons, new included exons, or skipped exons in about equal number. There are 82 genelets composed only of novel splice junctions that are not part of WormBase models or represented in the gene models from either Twinscan or Genefinder.

We also looked for signatures of transcription starts and stops in the genelets. In young adults, for example, 7257 protein-coding genelets had a high-scoring SL1 or SL2 within 25 bases of the start and 7501 had a biased distribution of scores by strand suggestive of a transcript start. These measures putatively identify 10,206 different transcription start sites. By similar methods we identify 7181 ends of transcripts compared with 1587 in WormBase.

Blocks of high-scoring bases in introns and intergenic regions

Having incorporated gene models both from other gene prediction tools and from genelet formation into the total coding potential of the genome, we returned to examine the residual blocks of high-scoring bases not part of any annotated protein-coding transcript, pseudogene, or noncoding transcript. We focused our attention on high-scoring segments longer than 20 bases and then clustered proximate segments into single blocks. We looked at both those that fell within introns and those lying in unannotated (intergenic) regions. These blocks presumably derive from as yet unrecognized transcripts and could represent noncoding RNAs. Alternatively, they might also represent exons of protein-coding genes, particularly new terminal and alternative exons or even whole new genes that escaped algorithmic detection even with the permissive parameters used. Of course, some might be incompletely processed transcripts and yet others could be false-positive signals. We examined high-scoring blocks from both intronic and intergenic regions to glean some hint of their origins.

Of the 8707 high-scoring blocks within introns, almost three-quarters were small (<50 bases) and had low depth of coverage (dcpm < 0.05). These small, low-coverage blocks were distributed widely, with 4425 blocks falling into introns of 2334 distinct genes. Many of these isolated blocks seem likely to be false positives. However, some large introns contained multiple blocks, often spanning most of the intron, whereas other introns in the same gene were devoid of high-scoring bases. Intronic dcpm was in these cases generally much lower than for the exons of the gene. These aspects were all consistent with the hypothesis that these intronic blocks or clusters of blocks arose from incompletely processed transcripts, perhaps because of inefficiently processed splice sites. Occasionally, multiple introns from the same gene contained multiple small blocks. For example, Y46G5A.1, a gene spanning >31 kb of genomic sequence, contains 35 blocks covering 2976 intronic bases scattered across most of its 15 introns.

Some intronic blocks spanned larger regions and were represented with greater depth. Inspection of the characteristics of a sample of these blocks suggests they include previously missed transcripts, new terminal exons, and possible alternative skipped exons. For example, one block extended the 3′ end of the *asd-2* gene (but lay in the intron of a Genefinder prediction) and had more than one-fourth depth of coverage of the flanking exons, suggesting that the block may represent an alternative 3′ UTR. In another example, in an intron of *ajm-1* the depth of coverage of a block covering the first 1100 bases of intronic sequence approached that of the upstream exons in some stages, suggesting that this too might be an important alternate carboxyl terminus and 3′ UTR. In another case, although the dcpm of the intron block was substantial (1.67), the dcpm of the flanking exons of the ribosomal protein gene, *rps-17*, was some 50-fold higher, perhaps suggesting that the splicing was inefficient here.

In contrast to the intronic blocks, there were only 2465 high-scoring blocks in intergenic (unannotated) regions. Moreover, only 912 contained more than 50 bases above threshold, and few of these had more than modest expression levels (475 with dcpm >0.05). Manual review of several blocks suggests they represent a variety of different features, including 5′ or 3′ extensions or alternate terminal exons of existing genes as well as altogether new transcripts. Some of the latter had the biased distribution of reads suggestive of the end of a transcript, further supporting this speculation. The extent of conservation varied but EST representation was generally absent.

One particularly intriguing region lies on chromosome X, extending for almost 14 kb from the start of a gene, C30E1.9, annotated in WormBase as a probable noncoding RNA gene. The region includes roughly 50 copies of a degenerate tandem repeat of unit length 151. There are no large open reading frames, and the region is poorly conserved with other nematodes. The putative transcript was found to be abundantly expressed at all stages. Northern blots confirmed the presence of both the small transcript annotated in WormBase as well as a transcript larger than 10 kb, corresponding to the region covered in the RNA-seq data sets (Supplemental Fig. S4). Experiments by one of us (V. Reinke, unpubl.) show that the transcript is nuclear localized and confined to two spots in hermaphrodites and one spot in males. No phenotype for this locus was detected with RNAi knockdown.

Discussion

By applying next-generation sequencing we have been able to sample RNA far more deeply than has been previously possible, and as a result provide evidence for a large set of protein-coding genes and exons in *C. elegans* that had no previous experimental support. The results incorporate predictions from WormBase, Twinscan, and Genefinder and suggest novel splice junctions and exons, including alternative exons that are not currently part of any of the sets of gene models used here. The data also provide a more complete definition of the UTR regions, features that are traditionally very difficult to predict computationally. The result is a substantially altered view of the *C. elegans* protein-coding potential that will be important to integrate into WormBase annotations (Table 6).

Table 6. Summary of features identified

	Splice junctions confirmed	SL1	SL2	Start	Stop
WormBase	70,911	7635	2043	—[a]	1587[b]
RNA-seq	91,678	10,778	2515	14,261	12,455
WormBase+RNA-seq	98,308	11,740	2599	14,261	13,151

[a]WormBase does not specifically annotate transcript "starts."
[b]Features annotated in WormBase as polyA_site.

Development of rigorous statistical approaches to these short, error-prone reads has been key to the analysis. For genome and transcript coverage, splice junctions, splice leaders, and polyadenylation sites, we derived estimates of false-positive rates or false-discovery rates. In each case the thresholds have been set based on conservative assumptions. The data analyses also intrinsically provide a confidence measure for each feature detected. These individual estimates can provide valuable input into automated methods for gene model construction as well as for users investigating specific features or even individual genes. For example, in the absence of ancillary information, a user might want to use quite stringent thresholds for SL2 detection, but in the context of related information such as genome coverage and the proximity of an upstream gene, a more lax threshold might suffice. For our purposes in improving annotation of the *C. elegans* genome, we can use the analysis to prioritize validation experiments.

We were also able to exploit the patterns of strand representation and bias to predict the start of transcripts and the end of transcripts. Other studies using next-generation sequencing have not exploited this bias. Some studies sheared the RNA before priming, presumably eliminating the signal. Shearing reportedly has the benefit of providing more uniform coverage of some RNAs. Investigators will have to weigh the trade-offs of the two methods for their own applications. We would speculate as well that the signal depends strongly on the quality of the RNA; even slight degradation is likely to diminish the signal. Even in the preparations here we noticed slight variations in the shapes of the curves.

In addition to the strand bias, we were able to use transplicing events as a second signal of the start of the processed transcript. In many cases, these sites corroborated the strand bias signals above, but in other cases they provide unique evidence of the transcript start. Also, for those sites in which the SL2 splice leader sequence predominates, the data provide evidence for previously unrecognized operons. These will have to be confirmed in further studies. Indeed, more detailed investigations of the wealth of information about SL1 and SL2 usage on this large number of transcripts could provide new insights into the transplicing process. The spliced leader data combined with the strand bias analysis in the young adult data, for example, point to a total of 10,206 transcript starts, 5274 of which are supported by both approaches.

The digital nature of the data provides ready quantification of the signals. We have calculated average depth of coverage for each of the transcripts normalized to the number of aligning reads in each data set in a modification of the procedure proposed by Mortazavi et al. (2008). Depth of coverage is perhaps a more intuitive measure and automatically adjusts for read length. We calculated this value based only on nonredundant bases in a transcript, thereby avoiding ambiguous read placements. Of course, these expression values will need to be replicated for each of the stages studied here before they can be used with confidence. However, given the high correlation coefficients between stages, the sources of variation are likely to be biological, not technical.

The very high sampling of the transcriptome of each of the stages facilitated by next-generation sequencing provides evidence of transcription from genes reported to be expressed only in a single cell of the 700–2000 cells of the animals. This result suggests that almost all protein-coding genes of an average size expressed consistently in these stages have been sampled at this point. Of course, genes expressed in only a few copies per animal and genes more stochastically expressed have probably not been detected. Nonetheless, the results suggest that the genes with no evidence of expression are used in other stages or conditions, such as embryos, males, response to environmental signals, etc. Future experiments could explore these other stages and conditions to complete the catalog. Another possibility is that these predicted genes are no longer active in the N2 strain.

We also detected a modest amount of polyadenylated transcription outside of predicted gene models. Although some of it may represent noncoding RNAs, much of what we detected could be 5′ or 3′ UTR/exons. These features are notoriously difficult to predict, especially when alternative start and stop exons are present. Still other regions appear to be inefficiently spliced transcripts. Further experimental data will be required to distinguish conclusively which of these possibilities obtains in each instance. There may be additional noncoding transcripts that are not polyadenylated and have thus escaped detection here. Similarly, because our methods do not retain strand information from the transcript, we cannot recognize antisense transcripts that are not processed. Again, other experimental approaches will be required to detect such transcripts.

The rapid advances in next-generation sequencing will only make RNA sequencing more powerful and cost-effective in the annotation of genomes. Already, the technology allows detection of transcripts regularly present in the population. Advances now being implemented such as paired-end reads and longer reads will provide more reliable mapping of the reads and will begin to add information about the contiguity of transcripts. But even currently, the approach provides a much-improved view of the protein-coding potential of a genome than can be obtained from the best prediction programs, especially for organisms where the various stages and tissues are readily available. Undoubtedly, RNA sequencing will become a standard part of any genome analysis project.

Methods

Using the short reads produced by the Illumina 1G sequencer required the development of a variety of new methods to provide statistically defined views of the polyA+ transcribed genome. These and other methods are presented in detail in the supplemental material. A synopsis of those methods is presented below.

Growth of synchronized animal populations, RNA isolation, library preparation, Illumina sequencing, and RT-PCR validation of predicted splice junctions

These methods followed published procedures with minor modifications as detailed in the Supplemental Methods section.

Estimating transcript number

Yeast, with an average cell volume of 35–60 μm^3 depending on bud scar number (Johnston et al. 1979), has been estimated to

Table 7. Composition of splice junction and splice leader databases

Stage	Threshold	Confirmed	Predicted not confirmed	Novel alternative	Novel	Splice leaders
L2	57	69,514	68,062	598,377	7,855,027	674,526
L3	26	69,514	68,062	598,377	8,899,527	699,899
L4	26	69,514	68,062	598,377	7,397,890	633,119
YA	26	69,514	68,062	598,377	8,025,180	659,399

(YA) Young adult.

contain 15,000 mRNA molecules (Hereford and Rosbash 1977), whereas mammalian hepatic cells with a volume of about 30–40 times that of yeast are estimated to contain 300,000–500,000 mRNA molecules (Coupar et al. 1978). Both yeast and mammalian cells then have about 300–350 mRNA molecules per μm^3. With 558 cells in the $30 \times 30 \times 60$-μm egg, the average embryonic cell volume can be estimated to be $\sim 60\ \mu m^3$, yielding an estimate of about 20,000 mRNA molecules per embryonic cell. David M. Miller, III, independently has estimated that embryonic cells contain 1 pg of total RNA. If 1%–2% of the total RNA is polyA+ RNA and the average mRNA is 1500 bases, cells on average contain 12,500–25,000 mRNA molecules, in good agreement with the interpolation from yeast and mammalian cells. If cells have increased in size by two- to threefold by L2, with the remaining increase in volume accounted for by an increase in cell number, the average L2 cell contains then about 25,000–50,000 mRNA molecules.

Read placements

To determine the origin of the reads, we aligned the reads against databases of (1) *C. elegans* genomic DNA (WS170); (2) splice junctions derived from WormBase, predicted genes from Genefinder and Twinscan, novel combinations of those splice sites, and novel splice sites not part of any gene model in either WormBase annotations or the gene predictions (Table 7); and (3) all of the known splice leader sequences fused to initial splice acceptors as well as internal acceptor sites. We also trimmed leading polyT sequences from reads to find reads initiating in the polyA sequence. All reads were aligned to all the databases using either MAQ or cross_match or both, retaining only a single best match for each read, using a variety of criteria to define the best match.

False-discovery and false-positive rates

Estimates of false-discovery rates and false-positive rates were obtained by comparing match rates between confirmed or likely transcripts and constructed false or unlikely data sets. For genome alignments, each base was assigned a score based on read coverage of that base and bases within a surrounding window centered on that base. To allow for a sharp transition from covered to uncovered bases, any base which itself is not covered is scored as zero, regardless of the coverage of flanking bases. Scores of bases within confirmed genes were compared against probable introns and intergenic regions in a ROC-like analysis to threshold scores for specific false-positive rates. For splice junctions, false junctions were created between acceptors and donors from different chromosomes. For splice leaders, the transpliced leader sequences were appended to splice acceptor sites on the opposite strand of confirmed WormBase annotated exons with above-threshold coverage. For polyA tails, reads with four or more matching As were trimmed and matched to the genome. For each case, the experimental and false matches were compared in detail to establish false-discovery rates. For the gene level analysis, we constructed false genes from 200-base segments outside of annotated regions and combined them randomly across the genome.

Transcript coverage

To normalize for both transcript length and read number aligned in each stage, we calculate the average depth of coverage per million reads (dcpm). The score of all bases of a transcript were summed, converted to a depth of coverage estimate by dividing by the window size, and then divided by the number of aligned reads in millions after removing rDNA matching reads. To look at normalized coverage along the length of a gene, each gene was normalized for length and for its overall average coverage.

Integrating data sets

Coverage, splice junctions, splice leaders, and other signals were used to create simple gene models by beginning with the highest confidence splice sites in a region and extending from those incorporating coverage and junctions into the model. This procedure was iterated until all confirmed splice junctions were incorporated into models. Potential transcription start and stop sites were identified by examining the ends of the models for features of the 5′ and 3′ start and stop sites, including splice leader sequences, biased strand coverage, and polyA tails.

Noncoding blocks

Blocks of putative transcription outside of protein-coding genes were formed by merging regions of 20 bp or more and scoring above threshold that were within 100 bases of one another.

Acknowledgments

We thank David Miller and Mark Gerstein for helpful advice throughout the project and for comments on the manuscript; John Murray for carrying out the cluster analysis of the differentially expressed genes; Tom Blumenthal for evaluating an early list of possible transpliced leader sites; Darin Blasiar for his assistance with WormBase questions; Jim Thomas for his insights into the 7TM/serpentine receptor class of genes; James Kramer for sharing his expertise in collagen gene expression; the BCCA Genome Sciences Centre Functional Genomics Group for expert technical assistance in library construction and sequencing; the Howard Hughes Medical Institute for funding (P.G.); and NIH/NHGRI for their funding through the modENCODE project, 5U01HG004263.

References

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

Baugh, L.R., Wen, J.C., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. 2005. Synthetic lethal analysis of *Caenorhabditis elegans* posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol.* **6:** R45. doi: 10.1186/gb-2005-6-5-r45.

The *C. elegans* Genome Sequencing Consortium.1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Chang, S., Johnston Jr., R.J., and Hobert, O. 2003. A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of *C. elegans*. *Genes & Dev.* **17:** 2123–2137.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5:** 613–619.

Coupar, B.E., Davies, J.A., and Chesterton, C.J. 1978. Quantification of hepatic transcribing RNA polymerase molecules, polyribonucleotide elongation rates and messenger RNA complexity in fed and fasted rats. *Eur. J. Biochem.* **84:** 611–623.

Hardy, H. 1976. Letter: Correction on the number of presumed beryllium-induced osteosarcomas in human beings. *N. Engl. J. Med.* **295:** 624.

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

Hereford, L.M. and Rosbash, M. 1977. Number and distribution of polyadenylated RNA sequences in yeast. *Cell* **10:** 453–462.

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

Johnston, G.C., Ehrhardt, C.W., Lorincz, A., and Carter, B.L. 1979. Regulation of cell size in the yeast *Saccharomyces cerevisiae*. *J. Bacteriol.* **137:** 1–5.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.

Kohara, Y. 1996. [Large scale analysis of *C. elegans* cDNA]. *Tanpakushitsu Kakusan Koso* **41:** 715–720.

Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* (Suppl. 1) **17:** S140–S148.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18:** 1851–1858.

Merrihew, G.E., Davis, C., Ewing, B., Williams, G., Kall, L., Frewen, B.E., Noble, W.S., Green, P., Thomas, J.H., and Maccoss, M.J. 2008. Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. *Genome Res.* **18:** 1660–1669.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45:** 81–94.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5:** 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J., et al. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet.* **27:** 332–336.

Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., et al. 2003. *C. elegans* ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34:** 35–41.

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

Ren, P., Lim, C.S., Johnsen, R., Albert, P.S., Pilgrim, D., and Riddle, D.L. 1996. Control of *C. elegans* larval development by neuronal expression of a TGF-beta homolog. *Science* **274:** 1389–1391.

Rogers, A., Antoshechkin, I., Bieri, T., Blasiar, D., Bastiani, C., Canaran, P., Chan, J., Chen, W.J., Davis, P., Fernandes, J., et al. 2008. WormBase 2007. *Nucleic Acids Res.* **36:** D612–D617.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., et al. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nat. Genet.* **1:** 114–123.

Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. 2005. Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15:** 577–582.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453:** 1239–1243.

Received October 21, 2008; accepted in revised form January 16, 2009.