Genomic Analysis of Gene Expression in C. elegans

A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, E. L. Brown

Until now, genome-wide transcriptional profiling has been limited to single-cell organisms. The nematode Caenorhabditis elegans is a well-characterized metazoan in which the expression of all genes can be monitored by oligonucleotide arrays. We used such arrays to quantitate the expression of C. elegans genes throughout the development of this organism. The results provide an estimate of the number of expressed genes in the nematode, reveal relations between gene function and gene expression that can guide analysis of uncharacterized worm genes, and demonstrate a shift in expression from evolutionarily conserved genes to worm-specific genes over the course of development.

The nematode C. elegans is a genetically accessible model organism that is widely used to study genetics, development, and other biological processes (1, 2). In 1998, the genome of this organism was completely sequenced, and the presence of 19,099 genes, or open reading frames (ORFs), was predicted. This made it possible to use oligonucleotide arrays (3, 4) to monitor the mRNA levels of 18,791 (98%) of the genes (5). To maximize the number of detected transcripts, we quantitated nematode gene expression in six developmental stages: freshly hatched juveniles, young adult worms, and 2-week life cycle (6). We detected 611 additional transcripts that had a significant increase or decrease in frequency (7) at some point during the life cycle and that were called “present” at least once; 4221 (22%) of the ORFs met these criteria. The expression profile of each of these developmentally modulated genes was normalized to have a mean value of zero and a variance of one, and the normalized profiles were clustered by means of a self-organizing map (SOM) (8). Examples of selected clusters are shown in Figs. 5 and 6.

1Department of Genomics, Genetics Institute, Cambridge, MA 02140, USA. 2Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: ahill@genetics.com (A.A.H.) and hunter@mcb.harvard.edu (C.P.H.)
†Present address: Millennium Predictive Medicine, 700 One Kendall Square, Third Floor, Cambridge, MA 02139, USA.

R E P O R T S

References and Notes

5. V. Borde, A. S. H. Goldman, M. Lichten, data not shown.
25. Supplemental data are available at www.sciencemag.org (version WS6).
26. All strains are diploids of the SK1 background, homozygous for the indicated mutations. Complete genotyping, construction details, and protocols for sporulation and conventional agarose gel analyses are available on Science Online (25).
27. The ars305 ars306 ars307 diploid contains mutations inactivating ARS305, ARS306, and ARS307. The his4::TEL1 diploid contains a reciprocal translocation with breakpoints at HIS4 on chromosome III and at PHO17 on chromosome I. Genotypes, construction details, and protocols for pulsed-field and 2D gel analyses are available on Science Online (25).
29. We thank C. Newlon and D. K. Bishop for strains; T. Allers for valuable advice; K. Smith, A. Nicolas, and C. Newlon for personal communications; and D. Chat toraj, J. Haber, G. Hayhurst, C. Newton, R. Shroff, B. Thomas, and M. Yarmolinsky for comments that improved this manuscript. 2 June 2000; accepted 10 August 2000

www.sciencemag.org SCIENCE VOL 290 27 OCTOBER 2000 809
Nonredundant genes called "present" (% of genes on array) 

<table>
<thead>
<tr>
<th>Developmental time</th>
<th>Array</th>
<th>Total genes ever called &quot;present&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Oocytes</td>
<td>3153</td>
<td>1110</td>
</tr>
<tr>
<td>0 hours</td>
<td>4345</td>
<td>1275</td>
</tr>
<tr>
<td>12 hours</td>
<td>2552</td>
<td>699</td>
</tr>
<tr>
<td>24 hours</td>
<td>1956</td>
<td>786</td>
</tr>
<tr>
<td>36 hours</td>
<td>3547</td>
<td>896</td>
</tr>
<tr>
<td>48 hours</td>
<td>3475</td>
<td>768</td>
</tr>
<tr>
<td>60 hours</td>
<td>3964</td>
<td>373</td>
</tr>
<tr>
<td>2 weeks</td>
<td>579</td>
<td>713</td>
</tr>
</tbody>
</table>

Sensitivity (9) average (ppm) 

- A: 4.1 (2 to 18)
- B: 6.3 (2 to 18)
- C: 4.4 (2 to 18)

Nonredundant genes called "present" by array 

- A: 5189
- B: 2123
- C: 3447

Total genes ever called "present" 

- A: 10,747
- B: 5,230
- C: 4,239

*Numbers in parentheses indicate the range.

To interpret the expression profile clusters, we matched the gene classifications in the Proteome WormPD database (16) to the genes in each cluster. Examination of the gene clusters indicated several expression profiles that were readily interpretable in terms of worm biology. Cluster B contained 104 genes whose expression increased through development to a peak at 60 hours, then declined in 2-week-old worms. This cluster was enriched in markers of metabolic activity, including oxidoreductases (7-fold enrichment as compared to the genome as a whole), amino acid metabolism genes (17-fold enrichment), carbohydrate metabolism genes (15-fold enrichment), and protein synthesis genes (7-fold enrichment). Cluster C contained 80 genes that were up-regulated in eggs and then again later in the reproductively active, egg-laying adult worm. This cluster is enriched in a small set of genes that we classified as being linked to embryogenesis, germ line development or function, and cell-cycle progression. Examples of these genes include F38E1.7 (mom-2), required for polarization of the EMS cell (17), C08B11.1 (zyg-11), required for zygote formation (18), and Y39A1A.12 (putative origin recognition complex subunit).

Other patterns of gene distribution among the clusters were observed, at least one of which emphasizes the influence of assay limitations on the clusters and the need to examine the raw data (19) when interpreting clustered profiles. For example, cluster D is enriched three to four times in known or putative transcription factors and in rare messages [212 out of 263 transcripts in this cluster had frequencies that never exceeded 30 parts per million (ppm); P < 0.001 by hypergeometric statistics]. This cluster exhibited the highest expression in the egg and exhibited low or undetectable expression at later times. However, many of these genes were expressed in the egg just slightly above our limit of detection and sank below our sensitivity at later developmental stages. Thus, the prevalence of transcription factors in this cluster is to some extent driven by the limits of our ability to detect rare messages in older animals, as opposed to the biology of specific transcription factors. This finding appears to have a parallel in reports that many messages that are found in embryonic sea urchin cDNA libraries are no longer detectable in adults (20).

Directed searches for genes that were significantly down-regulated in 2-week-old worms also revealed functionally related transcripts. These include the muscle-related genes T22E5.5 (gene, mup-2; protein, tropominin-T; 15-fold down-regulated at 2 weeks as compared to its mean level between 0 and 60 hours), M03F4.2 (act-4; actin; 17-fold down-regulated), and F07A5.7 (unc-13; paramyosin; 13-fold down-regulated), as well as Y57G11C.12 (similar to ubiquinone oxidoreductase subunit; fivefold down-regulated) and C44B12.2 (ost-1; osteonectin; 22-fold down-regulated). Together, these results suggest impaired muscle function, reduced metabolic activity, and extracellular matrix defects in aged worms, consistent with the aged worm phenotype.

Recent full-genome sequence comparisons (21, 22) have revealed widespread similarities and important differences between the yeast, fly, and worm genomes. Such se-
Table 2. The number of core, animal, and worm genes among detected genes and genes with either high or low expression levels at all developmental time points. Core genes were more likely than animal or worm genes to be detected and were also more likely to be expressed at levels above 30 ppm. Percentages in parentheses are fractions of the total number of genes in each class (core, animal, or worm), f, frequency.

<table>
<thead>
<tr>
<th>Genes</th>
<th>Detected</th>
<th>Never detected</th>
<th>Minimum expression (f &gt; 30 ppm)</th>
<th>Maximum expression (f &lt; 30 ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>2329 (80%)</td>
<td>580 (20%)</td>
<td>116 (4%)</td>
<td>2077 (71%)</td>
</tr>
<tr>
<td>Animal</td>
<td>2114 (67%)</td>
<td>1041 (33%)</td>
<td>18 (0.6%)</td>
<td>2522 (81%)</td>
</tr>
<tr>
<td>Worm</td>
<td>4099 (47%)</td>
<td>4686 (54%)</td>
<td>28 (0.3%)</td>
<td>7662 (67%)</td>
</tr>
</tbody>
</table>

Fig. 2. (A) The fraction of total gene expression among 728 highly expressed (>30 ppm) nonmodulated genes related to the inferred ancestors of the genes. These genes were defined as those that were called "present" 12 or more times and did not change significantly across the worm life cycle (ANOVA, P > 1 x 10^-2). Lines indicate the fraction of total gene expression due to core, animal, and worm genes. Error bars are ±1 SD, based on the variation of subsets of replicated data. Fractions do not add to one because not all genes were classified as core, animal, or worm. (B) Same plot for 4221 modulated genes, defined as those that were detected and changed significantly during the worm life cycle (ANOVA, P < 1 x 10^-2).

Fig. 2: (A) The fraction of total gene expression among 728 highly expressed (>30 ppm) nonmodulated genes related to the inferred ancestors of the genes. These genes were defined as those that were called "present" 12 or more times and did not change significantly across the worm life cycle (ANOVA, P > 1 x 10^-2). Lines indicate the fraction of total gene expression due to core, animal, and worm genes. Error bars are ±1 SD, based on the variation of subsets of replicated data. Fractions do not add to one because not all genes were classified as core, animal, or worm. (B) Same plot for 4221 modulated genes, defined as those that were detected and changed significantly during the worm life cycle (ANOVA, P < 1 x 10^-2).

Reference and Notes

6. Nematodes were grown at 25°C with concentrated OP50 as a food source. For each developmental stage, embryos representing all 12 hours of embryonic development were isolated by hypochlorite treatment of embryos and adults collected from growing mixed-stage populations. For the 0-hour sample, a 1-ml aliquot of embryos was flash-frozen in liquid nitrogen and stored at -80°C. For the larval preparations, the embryos were plated at low density on seeded plates and allowed to develop for the specified time. Additional details on nematode growth conditions, preparation of purified oocytes, and RNA isolation and labeling procedures are available as supplemental information.
7. The Affymetrix GeneChip software provided for each transcript an "absolute decision," which predicted whether the gene was "present" or "absent." The algorithm used to calculate the absolute decision is described in the Affymetrix GeneChip Analysis Suite User Guide (Affymetrix, Santa Clara, CA). Essentially, a gene was called "present" if its specific hybridization intensity was significantly above array background and noise levels. Additional details are available as supplemental information.
9. Eleven control transcripts were spiked into each hybridization. The abundances of these transcripts (stated in terms of control transcripts per total transcripts), ranged from 1:300,000 (or 3 ppm) to 1:1000 (or 1000 ppm). These abundances were calculated by assuming an average worm RNA size of 1000 ribonucleotides [3]. We plotted the absolute decision f as a function of abundance for each of these transcripts, scoring "present calls" as one, and "absent calls" as zero. By logistic regression of these 11 data points, we estimated the likelihood of receiving a present call as a function of transcript abundance. In general, the likelihood of receiving a present call was unity for the highest abundance control transcripts and close to zero for the least abundant control transcript. Using the regression, we defined the sensitivity of detection for each hybridization as the transcript abundance at which the likelihood of receiving a present call fell below 70%. This sensitivity estimated the absolute transcript level at which the hybridization signal became indistinguishable from noise. Because of several sources of variation that affect the quantitation of the control transcripts, we think that our sensitivity estimates and the frequencies (12) derived from them are accurate to within a factor of ~3. Details about the false-positive and false-negative rates of the present call, their impact on our detection limit, and further comments on the uncertainty of our quantitation are available as supplemental information.
10. We isolated 1.2 mg of total RNA from 1 ml of embryos (~7 million embryos). With an average of 3% polyadenylated RNA and an average transcript length of 1 kb, this corresponds to 10 million transcripts per embryo.
12. The GeneChip software provided a specific hybridization intensity value or "average difference" (AD) for each transcript that was proportional to transcript abundance messages also contribute substantially to the increase in worm-specific expression. Thus, the trends in expression during nematode development are consistent with a simple model that envisages the multicellular organism in terms of an ancient cellular core that is organized and regulated by newer organism in terms of an ancient cellular core and further comments on the uncertainty of our quantitation are available as supplemental information.

sequence comparisons provide us with a glimpse of the evolutionary process and are a useful source of information about the function of uncharacterized genes. However, these comparisons do not address the issue of how yeast, fly, and worm compare at the level of expressed transcripts. We formulated a simple "phylogenetic" model for gene expression in the developing worm. In this model, we partitioned nematode genes into three classes on the basis of sequence similarity: "core" genes (shared among yeast, worm, and fly), "animal" genes (shared between worm and fly), and "worm" genes (unique to the worm) [23]. We hypothesized that, during development, the expression of core genes would remain relatively high and constant, reflecting a primary role of these genes in common cellular processes. In contrast, we theorized that animal and worm genes would make up a smaller fraction of the transcriptome and be highly developmentally modulated, reflecting their probable role in defining multicellular processes and worm-specific development.

In agreement with this model, we found that core genes were more likely to be detected and more likely to be highly expressed than animal or worm genes (Table 2). Approximately 80% of core genes were detected, as opposed to only 67% of animal genes and 47% of worm genes. Similar observations have been made from an analysis of EST databases [5]. However, the quantitative nature of our data coupled with the temporal information allows a deeper analysis. For example, 4% of core genes were detected at frequencies greater than 30 ppm at each developmental time point, but less than 1% of animal or worm genes were detected (P < 1 x 10^-3, χ^2 test). Similarly, in terms of the fraction of total transcripts, core genes accounted for most of the transcripts among nondevelopmentally modulated, highly expressed genes (Fig. 2A). In contrast, worm genes accounted for a larger proportion of transcripts among the developmentally modulated genes, and this proportion rose during the course of development (0 to 60 hours), whereas the fraction of expression due to core genes concurrently dropped (Fig. 2B). To determine if the increase in worm-specific expression was driven by a small number of abundant, structural transcripts, we examined the likelihood of worm, animal, and core genes being called "present" at each developmental time and found the same trend, with slightly reduced amplitude. This suggests that, although abundant genes contribute importantly to the trends in Fig. 2B, lower abundance messages also contribute substantially to the increase in worm-specific expression. Thus, the trends in expression during nematode development are consistent with a simple model that envisages the multicellular organism in terms of an ancient cellular core that is organized and regulated by newer genes that evolved from that core. Although this model is too simplistic to be predictive of individual gene functions, it is striking that groups of genes with similar histories are expressed in accord with such a framework.
**Song Replay During Sleep and Computational Rules for Sensorimotor Vocal Learning**

Amish S. Dave and Daniel Margoliash*

Birds learn a correspondence between vocal-motor output and auditory feedback during development. For neurons in a motor cortex analog of adult zebra finches, we show that the timing and structure of activity elicited by the playback of song during sleep matches activity during daytime singing. The motor activity leads syllables, and the matching sensory response depends on a sequence of typically up to three of the preceding syllables. Thus, sensorimotor correspondence is reflected in temporally precise activity patterns of single neurons that use long sensory memories to predict syllable sequences. Additionally, "spontaneous" activity of these neurons during sleep matches their sensorimotor activity, a form of song "replay." These data suggest a model whereby sensorimotor correspondences are stored during singing but do not modify behavior, and off-line comparison (e.g., during sleep) of rehearsed motor output and predicted sensory feedback is used to adaptively shape motor output.

In reinforcement learning, systems learn through interaction with the environment by trying to optimize some measure of performance. Biological systems may experience a substantial delay between premotor activity, a form of song "replay," and off-line comparison of rehearsed motor output and predicted sensory feedback. Thus, a model of how to reward or punish a premotor circuit when that circuit is participating in a sequence of actions (1) has been proposed to the problem of learning sequences of actions (2). Here, we report on neuronal data that represent a solution to the problem of sensorimotor mapping in the bird vocal-motor ("song") system. The physiological properties observed during sleep also suggest a computational implementation for reinforcement learning of song.

Zebra finch songs are organized hierarchically, with one or more notes composing a syllable, and sequences of syllables forming a motif, which is repeated to form song. We investigated neurons in the forebrain nucleus robustus archistriatalis (RA), whose descending projections represent the output of the forebrain song system. During singing, RA neurons exhibit short bursts of activity, whose identity varies with the note that immediately follows the burst (4). In awake birds, outside the context of vocalizations, RA neurons are regularly firing. RA neurons also prominently burst "spontaneously" and respond to sounds, but only during sleep (5). With the goal of comparing motor, auditory, and ongoing bursting activity, we recorded single neurons in the RA of singing male zebra finches, permitted the animals to fall asleep by turning off the lights, and then tested the same neurons' sensory and ongoing discharge properties (6, 7).

The spiking patterns of RA neurons in singing birds consisted of phasic patterns of premotor excitation superimposed over a background of profound inhibition (4) (Fig. 1, B and C). This premotor activity was virtually invariant for multiple occurrences of the same sound. After the lights were turned off, RA auditory responses were initially weak but gained strength with time, reflecting the gradual transition into sleep (5). Responses to playback of the bird's own song (BOS) also consisted of phasic patterns of excitation separated by inhibition that were similar for multiple occurrences of the same sound, differing mainly in the strength of response rather than pattern (8).

The timing of auditory responses to the BOS was very well aligned to the timing of premotor activity (Fig. 1F). The only exceptions were instances of silence following the end of a motif or the end of song, where the auditory response could include an additional burst that corresponded with the syllable that would have followed if the song had continued without pause. To compare motor and auditory activity, we analyzed the singing-related activity surrounding each syllable of song (4, 9). The spike patterns from the response to the BOS playback were then compared with the spike patterns from premotor activity derived from...