Published in final edited form as:
Nat Methods. 2008 August ; 5(8): 703–709. doi:10.1038/nmeth.1228.

Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*

John Isaac Murray[1,2], **Zhirong Bao**[1,2,3], **Thomas J. Boyle**[1], **Max E. Boeck**[1], **Barbara L. Mericle**[1], **Thomas J. Nicholas**[1], **Zhongying Zhao**[1], **Matthew J. Sandel**[1], and **Robert H. Waterston**[1,4]

1 *Department of Genome Sciences University of Washington School of Medicine 1705 NE Pacific Street, Seattle, WA 98195*

Abstract

We describe a system that permits the automated analysis of reporter gene expression in *Caenorhabditis elegans* with cellular resolution continuously during embryogenesis and demonstrate its utility by defining the expression patterns of reporters for several embryonically expressed transcription factors. The invariant cell lineage permits the automated alignment of multiple expression profiles, allowing the direct comparison of the expression of different genes' reporters. We have also used the system to monitor perturbations to normal development involving changes both in cell division timing and in cell fate. Systematic application could reveal the gene activity of each cell throughout development.

Introduction

A major goal of current biological research is to understand how the genome directs the process by which a single celled zygote gives rise to the complexity of a multicellular organism and in turn a new zygote. Simply knowing the full complement of transcriptionally active genes for each cell throughout development would be a major advance, and would provide the molecular framework for understanding the network of interactions with which development proceeds. For example, the description of the expression pattern of pair-rule genes such as *even skipped* in *Drosophila melanogaster* at cellular resolution with high temporal resolution was critical in understanding how these genes can cause expressing cells to adopt a fate different from adjacent non-expressing cells[1].

Several different methods have attempted to capture expression information with high temporal and spatial resolution on a broad scale. Hybridization to microarrays has provided valuable information for bulk samples, including for time courses and some specific cell types [2-7]. However, the method is unwieldy and technically challenging when trying to achieve both cellular and high temporal resolution. Microscopy of an organism more readily provides single-cell resolution by using a visible transgenic reporter or through direct labeling of transcripts or proteins. However, typically only one or a few genes are assayed at once, often only at specific time points. Integrating such data sets across multiple genes and multiple specimens

[2] These authors contributed equally to this work
[3] Present address: Developmental Biology Program Sloan-Kettering Institute 1275 York Avenue, New York, NY 10065
[4] Corresponding Author Email: waterston@gs.washington.edu FAX: 206-685-7301 PHONE: 206-685-7347

Author Contributions
RHW, ZB and JIM conceived the experiments. JIM, ZB, MEB , TJN, BLM, ZZ and MJS performed the experiments. TJB, ZB and JIM wrote software for data analysis. JIM, ZB, TJB and JIM analyzed the data. JIM and RHW wrote the paper with input from the other authors.

throughout development generally requires expert anatomists and the expression profiles of individual cells are often lost.

The nematode *C. elegans* presents the possibility of new approaches to the comprehensive description of expression patterns. Because it is transparent from the zygote to the adult and with just 959 somatic cells as an adult, every cell can be visualized throughout the life cycle in living animals. Because it develops through an invariant cell lineage, knowledge of an imaged animal's lineage allows the identity of each cell to be assigned unambiguously[8]. This potentially allows the alignment of expression patterns from individual animals onto a reference lineage, providing an integrated view of gene expression for each cell.

To exploit these possibilities systematically we previously developed methods that allow automated computational tracing of the *C. elegans* lineage [9-11]. These methods use custom software to identify nuclei and track them over time in 3D movies of worm embryos ubiquitously expressing nuclear-localized histone-GFP (green fluorescent protein) fusion proteins. This system generates highly accurate lineages through the 350-cell stage, which comprises all but the last round of embryonic cell division. Here we extend the system to provide detailed spatiotemporal characterization of reporter gene expression in both wild type and genetically perturbed embryos.

Results

Annotating reporter expression with the lineage

To monitor gene expression we generated promoter::fluorescent protein::histone reporter constructs, using a gene's 5' intergenic sequence to drive expression of an mCherry[12,13] red fluorescent reporter. Fusing the reporter to a *C. elegans* histone H1.1 coding sequence directs the reporter to the nucleus to facilitate quantification and assignment of the signal to specific cells. These constructs, although lacking post-transcriptional control signals and perhaps even some transcriptional controls, should drive temporal-spatial expression of the reporter and provide a test of our ability to describe expression patterns over time at the single cell level. We used microparticle bombardment to generate stably integrated transgenes in an attempt to avoid high-copy number, mosaicism, silencing and other artifacts associated with extrachromosomal arrays. The resultant transgenes were then placed in a background ubiquitously expressing GFP-histone to follow the lineage. We adapted our previous imaging methods[9] to collect two-color 3D movies from embryos by time lapse confocal microscopy and traced the embryonic lineages and visualized the results with the programs StarryNite[10] and AceTree[11].

We selected four developmentally important and well-studied transcription factors for initial testing (Table 1). *pha-4* encodes a FoxO transcription factor homologous to *Drosophila* forkhead required for specification of the pharynx[14-16]. *cnd-1* encodes a NeuroD ortholog required throughout metazoans for proper neuron development[17]. *hlh-1* encodes a MyoD homolog important for muscle development[18]. Finally, *end-3* encodes a GATA binding transcription factor important in specifying the intestine[19]. We used manual review of movies collected of reporter strains for each gene to identify expressing cells through the 350-cell stage and compared our annotations to descriptions from the literature. We describe expression using the conventional *C. elegans* naming scheme (Fig. 1).

Our results generally agree with those in the literature (Table 2), and the minor differences are consistent with known differences in the specifics of the experiments. For example, for *end-3* the earliest visible fluorescence appeared during the E2 stage (daughters of the E cell), while *end-3* transcript is detectable by *in situ* hybridization 15-30 minutes earlier in the late E1 stage[19]. This difference likely reflects the time required for translation and folding of the mCherry protein (15 minute half-maturation time in *E. coli*). Timing of fluorescence onset for the other reporters described here are also consistent with a 15-30 minute lag. For *hlh-1*, reporter signal was detected not only in the myoblast lineages (D, Cap and Cpp), but also in the lineage (MS) which gives rise to pharyngeal cells and glia as well as body wall muscle cells. Transient *hlh-1* expression was previously detected in all descendents of the MS cell using protein::reporter fusions[18]; the more persistent expression found here likely reflects the use of the promoter::reporter constructs and the stability of the histone::reporter fusion. For *pha-4* the brighter expression we find in the E lineage relative to the pharyngeal precursors in the AB and MS lineages is similar to other reporter patterns and differs from antibody staining patterns of native protein[15,16], which reveal more abundant protein in the pharyngeal lineages. Presumably, these differences reflect regulatory sequences missing in the reporter constructs.

Importantly, our annotations extend the expression pattern analysis to cellular resolution and in so doing reveal novel spatial and temporal details. Most notably, for *cnd-1* we detected low signal consistently above background throughout the AB lineage from the 24-cell stage onward, with much brighter expression in a subset of AB cells starting at the 100-cell stage, as was reported previously[17]. However, the previous work did not specifically identify these brightly expressing cells other than to state that a few were likely to be ventral cord motor neurons and that most did not express UNC-86 protein. We identified the brightly expressing sublineages, which produce ring and other neurons in addition to motor neurons (Table 2). Our broader description of *cnd-1* promoter activity (114 neurons and glia vs. 13 non-nervous system cells) suggests this gene may play a more complex role in neurogenesis than previously suggested.

Quantitative description of expression

Because manual annotation of expressing cells is both subjective and time-consuming, we sought to develop methods that would automate quantitative measurements of expressing cells. Such automated quantification also permits a more systematic and rigorous evaluation of reproducibility within and between independently derived strains.

We calculated the mCherry fluorescence intensity in arbitrary units within each nucleus at each time point using the nuclear positions and diameters estimated by StarryNite (roughly 20,000 measurements per embryo). Because in some cases out of focus light from nearby expressing nuclei can increase signal within non-expressing nuclei, leading to false positives, we subtracted locally calculated background from each expression measurement (see methods for details). The resulting high-resolution expression data can be displayed as a color-coded lineage tree or in a spatial representation of the lineage (Fig. 1). To distinguish non-expressing cells from expressing cells we developed methods to estimate the statistical significance and time of onset of expression for each cell and also estimated the time of onset of expression for the lineage of each terminal cell in each recording (Supplementary dataset, see Methods for details).

The measured intensity values vary continuously from near zero ($-0.07 +/- 0.69$ (arbitrary) units per pixel) for non-expressing cells in negative control embryos to very high values (> 100) for brightly expressing cells (Supplementary Table 3). Even among expressing cells in a single embryo we observed a large dynamic range. For example, peak expression levels of a *cnd-1* reporter in the mostly neuron-producing ABarapp lineage ranged from 58-116 units ($P < 10^{-22}$), while its pharynx-producing sister lineage ABarapa was similarly significant ($P < 10^{-19}$) but reached only 6-7 units.

Analysis of two negative control embryos with no RFP transgene gave no cells with $P < 3 \times 10^{-6}$ (Supplementary Fig. 1). Cells in reporter embryos with expression significantly below this threshold corresponded well to the cells identified as expressing by manual inspection.

Several novel sublineages not annotated manually also pass this cutoff. A few are likely false positives, while others represent previously overlooked expression. For example, all three *cnd-1* reporter strains have significant expression throughout the EMS lineage, which generates mostly endoderm and mesoderm. This expression is dimmer than the AB expression (2-10 units compared with 10-100 units for the AB expressing cells), which may be why it was not identified in previous studies.

Automated estimates of the time of onset agreed with the manual review of images as well as with inspection of the plots of brightness versus time for each terminal cell. For example, the automated estimate of the time of *end-3* reporter onset in the E lineage was in the E2 stage (daughters of E), the same as was annotated manually.

Reproducibility of expression patterns—We compared replicate image series for strains expressing reporter constructs for each of the four genes to assess the reproducibility of the measured expression levels, of the time of onset and of the identities of significantly expressing cells in the reporter strains using our automated methods.

Relative expression levels for each cell (normalized to other cells in the same embryo) were quite reproducible from one series or strain to the next, despite the many possible sources of variation (Supplementary Figs. 4-6, Supplementary Table 1). Comparing the average intensity of each cell across four embryos of a *pha-4* reporter strain yielded high reproducibility (mean r = 0.96) as did comparing replicate expression patterns for *cnd*-1 (r = 0.92), *hlh-1* (r = 0.94) and *end-3*(r = 0.95) reporter strains. As additional variability might be introduced by the strain construction process, two independent *hlh-1* reporter strains made with the same construct were compared and were also highly correlated (mean r = 0.95) as were three independent strains containing the same *cnd-1* reporter construct (mean r = 0.86).

In contrast, a second *pha-4* reporter strain made with a different fluorescent reporter (unoptimized DsRed.T1 instead of worm-optimized mCherry) and transformation marker (*pha-1* instead of *unc-119*) showed more variable expression (mean r = 0.84, range 0.68-0.95). Examining the lineage trees and expression levels for this strain (Supplementary Fig. 5) identified substantial differences in the identities of expressing cells: some series had bright expression in the ABalpp, ABarap and MSxp lineages. These lineages give rise predominantly to nonpharyngeal cells and are the sister lineages of predominantly pharyngeal lineages that express the *pha-4* reporter in all embryos. While determining the cause and biological relevance of this observation will require additional experiments, our ability to detect the differences in pattern demonstrates the utility of lineage analysis in identifying and characterizing variability in reporter expression.

The identities of expressing cells were also highly reproducible between embryos of the same strain. Of highly significant expressing cells ($P < 10^{-12}$), 94.3% (8192/8683) were significant at $P < 10^{-6}$ in a second embryo containing the same reporter construct (38 comparisons involving 14 embryos for four genes, see Methods for details). Even for cells with intermediate significance ($10^{-12} < P < 10^{-6}$), a substantial fraction, 68%, were significant in the second embryo (1054/1550) (Supplementary Fig. 2). It is not surprising that the confirmation rate is lower for these cells because they have much lower expression levels (average 3.3 units) than cells in the more significant set (average 25 units) and are thus closer to the detection limit of the system. Supporting this, the confirmation rate rises to > 97% if only cells reaching an arbitrary threshold of ten units are considered, and over half of all unconfirmed cells have $P <$ 0.01 in the second embryo (compared with 10% of cells in the negative control embryos).

The majority of cells with expression brighter than five units that failed to confirm in the second series represent real differences in expression between embryos based on manual inspection

(13 of 15 cases tested). For cells below this cutoff, 34% (113 of 333 tested) represent clear false positives in one series caused by imaging artifacts such as dust flecks and coverslip reflections. The remaining cells are difficult to score – cells missed at this level could be either not expressing or expressing below our sensitivity level. The relatively small number of differences suggests that it should be possible to reliably identify all brightly expressing cells with 2-3 replicate image series, while 4-5 or more replicates would allow higher sensitivity for dimly or occasionally expressing cells.

The estimated onset times for the four genes ranged from 48 minutes to 154 minutes after the ABa (4-cell stage) division. Across all expressing cells, the median standard deviation of onset time in replicate embryos was 11 minutes and 90% of cells had a standard deviation of less than 20 minutes (Supplementary Fig. 3). By comparison, the mean cell cycle length through the 200-cell stage is 32 minutes.

Integrating expression patterns—The invariance of the wild type allows mapping of multiple expression patterns onto a single reference lineage tree to examine their relationships. For example, the expression patterns of the *pha-4*, *cnd-1* and *hlh-1* reporters are largely orthogonal, while the *end-3* reporter is coexpressed in the E lineage with the *pha-4* reporter and has an earlier onset (Fig. 2). These relationships can be expressed in terms of correlation coefficients: the mean correlation for *pha-4* and *end-3* reporters was 0.68 compared with −0.002 for all other inter-gene comparisons. As a result, relationships between expression patterns can be systematically explored by hierarchical clustering (Supplementary Fig. 7). Replicates for a given reporter form tight clusters, and the relationship between reporters with similar expression, notably *end-3* and *pha-4*, are evident in the clustering patterns, as are the differences in expression in *pha-4*::H1-DsRed embryos. With a larger dataset and more sophisticated comparison algorithms, this type of analysis may be useful in identifying the modules that make up the embryonic transcriptional regulatory network.

Identifying altered expression in mutants

Detecting quantitative changes in reporter expression—While expression patterns are well correlated between embryos expressing the same reporter, we observed variability in absolute expression levels (acting multiplicatively on all cells of a given embryo) over a two-fold range. This is not surprising due to the many technical and biological factors that can influence the reporter signal. However, with sufficient replicates and controls, we reasoned it might be possible to use the magnitude of expression as a quantitative readout of developmental pathways.

We tested the effect of the GATA transcription factor encoded by the gene *elt-7* on the *pha-4*::H1-mCherry reporter in the E lineage. ELT-*7* is thought to regulate early gut differentiation redundantly with the GATA factor encoded by *elt*-2[20]. The *pha-4* promoter is known to contain sites bound *in vitro* by ELT-2 and ectopic expression of ELT-*2* can lead to ectopic expression of *pha-4* reporters[15]. However, the effect of ELT-*7* on *pha-4* reporter expression is unknown. We compared *pha-4* reporter intensity in the gut cells of comma-stage wild type embryos ($n = 28$) with intensity in embryos homozygous for a deletion allele of *elt-7* ($n = 56$). With these large numbers and a careful protocol (see Methods for details) we observed a mean 23% decrease in *pha-4* reporter intensity ($P = 3.9 \times 10^{-7}$). This difference cannot be explained by differences in the z position of the expressing cells (see methods). A similar effect was noted when *elt-7* was targeted by RNAi (data not shown), indicating that genetic background differences are not responsible for the effect. These data suggest that ELT-*7* is required for full *pha-4* reporter expression in the E lineage. The residual expression may be due to activation by ELT-*2*, for which RNAi inactivation shows a similar effect (data not shown), and possibly by other factors acting at this promoter.

Characterizing expression in altered lineages—Our quantitative image analysis methods can also detect changes in both expression pattern and expression level at cellular resolution in altered lineages. Some of the cells that express the *pha-4* reporters derive from the E and MS founder cells. E and MS are sister cells with different fates because of a WNT signal received by E but not MS[21,22]. Loss of function of the gene *lit-1*, which encodes a NEMO-like kinase in the WNT pathway, causes E to adopt the MS fate[23]. Disrupting *lit-1* function by RNAi leads to a change in the spatial expression of the *pha-4* reporter (Fig. 3). The embryo fails to gastrulate, altering the anatomy of the embryo dramatically and making it difficult to determine the identity of the expressing cells anatomically. The lineage patterns combined with the *pha-4* expression pattern, however, reveal that the central misplaced expressing cells are generated from the E founder cell, which has adopted an MS-like fate. Some of the expressing cells in the anterior of the embryo have the same lineage identity (descended from ABalpa, ABaraa) as expressing cells in wild type embryos, and are simply misplaced because of gastrulation defects. However, many cells (most cells in the ABalpp and ABarap lineages) show robust ectopic expression not seen in wild type. These cells normally give rise to non-pharyngeal cells and these data suggest that WNT signaling acts to prevent them from adopting a pharyngeal fate and expressing *pha-4*.

Discussion

The methods described here allow integration of expression patterns from different genes in different animals, providing the potential to obtain a comprehensive picture of gene expression in every cell. For example, using this methodology it should be possible to describe the patterns of activity of the promoters for all of the transcription factor genes active during embryogenesis. Combined with emerging knowledge of transcription factor binding sites, these expression patterns would begin to reveal the network of regulatory control. Predicted networks could then be tested by quantitative analysis of reporters after RNAi or genetic depletion of predicted regulators. With appropriate controls for integration site and copy number, promoter dissection could allow the identification of the specific DNA sequences required for each regulator's effects.

The ability to trace altered lineages and generate a quantitative readout of cell-type specific reporters extends the phenotypic analysis of mutants and RNAi treatments that perturb development. As the number of cell fate markers increases, the method will become increasingly powerful in phenotypic analysis that will be required to gain a functional understanding of regulatory networks.

Future developments could considerably enhance the already powerful information obtained. Extending automated lineaging to include the last round of embryonic cell division will facilitate describing the expression of those genes only expressed in the terminally differentiated cells. This remains a challenging goal because of the active cell migrations and close packing of nuclei in the late embryo. Another improvement would be the development of an RFP-based lineaging system, which would allow the embryonic expression patterns for existing genomically integrated GFP reporters[24-26] to be characterized. Whether the many other non-integrated strains available would be useful for systematic analysis is questionable, because their mosaicism would be expected to necessitate additional replicates to ensure identification of all expressing cells and because a fraction of embryos imaged would not contain the transgene at all.

Our results emphasize that promoter::reporter constructs only partially capture the complexity of regulation, emphasizing the need for more faithful transgenic strategies. We intentionally used transcriptional fusions to histones for this study to maximize sensitivity for weakly expressed genes. Protein fusions could also be generated to reveal subcellular localization of the proteins and to contrast transcriptional and translational controls, at the potential expense of reduced sensitivity for proteins with faster turnover than HIS-24. Improved cloning and transformation methods such as recombineering[27] would allow the use of larger genomic segments so that the observed patterns are likely to more faithfully reflect the native pattern. Faster-folding, brighter reporters would ensure that the system detects the earliest expression of even weakly expressed genes. Extension of the system to multiple colors could increase throughput and provide kinetic information about colocalization.

In vivo single-cell analysis of gene expression is an important step towards a comprehensive molecular understanding of development. The transparency of the worm and its invariant lineage make it ideal for such analyses, but with continued progress in non-invasive imaging to track cells and ever-expanding sets of cell fate markers to substitute for the lineage, we envision equivalent analyses for more complex organisms, including mammals.

Methods

Constructs, Strains and RNAi treatment

See Supplementary Methods for details.

Imaging

Confocal imaging with a Zeiss LSM510 was performed as described[9] with the addition of parallel acquisition of mCherry (or DsRed) signal. For this we used a second track with excitation by a 5mW 543nm HeNe laser attenuated linearly between 5% (top plane) and 25% (bottom plane) by the acousto-optical tunable filter, and collected emitted light with a 560nm long pass filter and PMT with the gain varied linearly between 1100 (top plane) to 1150 (bottom plane), except for strain RW10062 the laser was attenuated between 8% (top plane) and 40% (bottom plane). We collected time points once per minute between the four-cell and comma stages (5-6 hours). The same RFP imaging settings were used throughout each recording. All expression patterns reported are from animals whose development proceeded normally. When scored, these animals hatched into L1 larvae with normal morphology and the lineages were identical to wild type through the 350-cell stage. While fluctuations in laser output over time would be predicted to alter the absolute magnitude of expression in an embryo, we measured this output regularly (about once per month) during the project and found that intensity varied within a modest range (< 10%).

Lineage analysis

We generated and edited lineages with the programs StarryNite[9,10] and AceTree[11] as described.

Quantitative intensity analysis

We compared three different strategies to quantify the RFP signal. In each case, we began by summing the signal within each nucleus. For one strategy, we did no background subtraction. We observed that on occasion this led to signal from strongly expressing cells interfering with the measurements of nearby cells, resulting in false positives. To address this problem, we tested two strategies in which we subtracted locally computed background signal either from the raw images (unmasked) or from images in which nearby cells were excluded from the background calculation (masked) (See Supplementary methods for details), and chose the more conservative unmasked strategy.

To generate the expression intensity values along the lineage, the local background, $B_{n,t}$, for each nucleus in the red channel images at each time point was defined as the average pixel intensity of pixels between 1.2r and 2r from the nuclear centroid, where r is the radius of the nucleus. The raw intensity, $R_{n,t}$, was defined as the average intensity of pixels within the bounds of the predicted nucleus. The corrected intensity, $I_{n,t}$, was calculated as $I_{n,t} = R_{n,t} - B_{n,t}$ and was used for all subsequent analysis. This algorithm was implemented as a stand-alone extension to AceTree, available by request. Visualization of expression-coded trees and projections was performed using AceTree[11].

An important imaging issue that could obscure subtle expression differences between lineages is depth – typically intensity decreases with depth in the specimen. The variable excitation light settings with depth were designed to reduce this effect. See Supplementary Methods for a further discussion of the residual impact of depth on quantification.

For multiple series comparisons requiring alignment to a reference lineage, a single series (102405_pha4red) was selected as the reference series and other series were scaled to match the branch lengths from this series. The scaling was done by interpolating expression values when the number of time points on a branch did not match. To illustrate, imagine a branch in the reference series with 20 time points. If the series being aligned had 19 time points in this branch, the 19 expression measurements would be mapped linearly onto the 20 time points of the reference lineage. For terminal branches, a 1:1 mapping was used. For hierarchical clustering, the aligned data were linearized into a single vector for each series, clustered using the program Cluster[28], and visualized in Java TreeView[29].

For the *elt-7* expression analysis, we collected single Z-stacks for wild type ($n = 28$) or *elt-7 (ok835)* ($n = 56$) animals using the same image settings as were used for image series collection. The images were acquired in 3 sessions; in each one roughly comparable numbers of wild type and mutant embryos were imaged to ensure variation in laser power over time or other factors were not confounded with mutant status. StarryNite was used to assign nuclei based on RFP expression, nuclei with locations and sizes compatible with E-derived nuclei were selected using a script, and the average raw RFP expression per nucleus was calculated for each embryo. The observed differences cannot be explained by differences in z position of expressing cells: we measured the effect of this factor to be approximately 3% per plane (see supplementary Methods). The positions of the E cells in wild type and *elt-7* embryos were very similar, with wild type lower by 1.5 planes. Given that our analysis suggests lower z plane causes reduced expression, the actual expression difference is likely slightly larger than was measured.

Statistical analysis

We evaluated the significance of the expression for each terminal cell (cells that did not divide during the recording) as follows: First, we calculated the trajectory of expression values of each cell and its parents, back to the first time point for that embryo. For each time point in each trajectory, we calculated the significance for expression beyond that time point being greater than local background by using a Wilcoxon signed rank sum test. This led to a time series of *P*-values for each cell trajectory. For expressing cells, a global minimum could be identified – in 10/10 cases examined in detail, the last time point that *P* was within 100-fold of this minimum was within five time points of the point subjectively identified as the onset of fluorescence. This method was more robust for weakly expressing genes than choosing an arbitrary intensity threshold as the onset time. Examining the significance levels determined this way in two image series with no RFP transgene identified $P < 10^{-6}$ as a threshold that would lead to less than one false positive per series. The distributions of red expression per embryo in wild type and *elt-7* mutants were compared by a Wilcoxon rank sum test with continuity correction.

Quantitative comparisons of replicates

Although we generated multiple strains for both the *cnd-1* and *hlh-1* reporters, we grouped these strains together for the analysis of reproducibility because expression was not substantially more different between strains than it was between embryos of the same strain. To identify the rates at which significant cells were confirmed in a second series, we generated a list of all cells for all 38 pairwise combinations of replicate embryos containing the same reporter construct and filtered based on various criteria (as specified in Results). To reduce the effect of outliers, peak expression level was defined as the second highest measured expression value in a cell's history. We then assessed the fraction of cells that had $P < 0.01$ in the second embryo. For comparison < 10% of cells in the negative control embryos had $P < 0.01$. To describe variability in time of onset, we identified terminal cells identified as expressing ($P < 10^{-6}$) in each replicate for a given gene, then calculated the standard deviation of the time of onset for each cell. Because the onset was usually several cell cycles prior to the terminal stage, related terminal cells frequently share a common point of onset. To avoid overcounting these cells, we limited further analysis to unique onset events. Quantitative comparisons of expression level were performed by first calculating the average expression level for each cell and then comparing these average expression levels between two embryos.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Thomas, J. Priess and members of their labs for helpful discussions. We also thank J. Thomas for critical comments on the manuscript. We thank A. Audhya (University of Wisconsin, Madison) for providing worm-optimized mCherry prior to publication and J. Waddle (Southern Methodist University) for providing unpublished HIS-24 protein fusion vectors. Some nematode strains used in this work were provided by the Caenorhabditis Genetics Center, which is funded by the NIH National Center for Research Resources. This work was partially supported by funding from the National Institutes of Health. J.I.M. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research. Z.B. is a Damon Runyon Fellow supported by Damon Runyon Cancer Research Fellowship DRG-1813-04.

References

1. Lawrence PA, Johnston P, Macdonald P, Struhl G. Borders of parasegments in Drosophila embryos are delimited by the fushi tarazu and even-skipped genes. Nature 1987;328:440–2. [PubMed: 2886916]
2. Fox RM, et al. A gene expression fingerprint of C. elegans embryonic motor neurons. BMC Genomics 2005;6:42. [PubMed: 15780142]
3. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development 2003;130:889–900. [PubMed: 12538516]
4. Baugh LR, et al. The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development 2005;132:1843–54. [PubMed: 15772128]
5. Furlong EE, Andersen EC, Null B, White KP, Scott MP. Patterns of gene expression during Drosophila mesoderm development. Science 2001;293:1629–33. [PubMed: 11486054]
6. Arbeitman MN, et al. Gene expression during the life cycle of Drosophila melanogaster. Science 2002;297:2270–5. [PubMed: 12351791]
7. Zhang Y, et al. Identification of genes expressed in C. elegans touch receptor neurons. Nature 2002;418:331–5. [PubMed: 12124626]
8. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode Caenorhabditis elegans. Developmental Biology 1983;100:64–119. [PubMed: 6684600]
9. Murray JI, Bao Z, Boyle T, Waterston RH. The lineaging of fluorescently- labeled Caenorhabditis elegans embryos with StarryNite and AceTree. Nature Protocols 2006;1:1468–1476.

10. Bao Z, et al. Automated cell lineage tracing in Caenorhabditis elegans. Proc Natl Acad Sci U S A. 2006
11. Boyle TJ, Bao Z, Murray JI, Araya CL, Waterston RH. AceTree: a tool for visual analysis of Caenorhabditis elegans embryogenesis. BMC Bioinformatics 2006;7:275. [PubMed: 16740163]
12. McNally K, Audhya A, Oegema K, McNally FJ. Katanin controls mitotic and meiotic spindle length. J Cell Biol 2006;175:881–91. [PubMed: 17178907]
13. Shaner NC, et al. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nat Biotechnol 2004;22:1567–72. [PubMed: 15558047]
14. Azzaria M, Goszczynski B, Chung MA, Kalb JM, McGhee JD. A fork head/HNF-3 homolog expressed in the pharynx and intestine of the Caenorhabditis elegans embryo. Dev Biol 1996;178:289–303. [PubMed: 8812130]
15. Kalb JM, et al. pha-4 is Ce-fkh-1, a fork head/HNF-3alpha,beta,gamma homolog that functions in organogenesis of the C. elegans pharynx. Development 1998;125:2171–80. [PubMed: 9584117]
16. Horner MA, et al. pha-4, an HNF-3 homolog, specifies pharyngeal organ identity in Caenorhabditis elegans. Genes Dev 1998;12:1947–52. [PubMed: 9649499]
17. Hallam S, Singer E, Waring D, Jin Y. The C. elegans NeuroD homolog cnd-1 functions in multiple aspects of motor neuron fate specification. Development 2000;127:4239–52. [PubMed: 10976055]
18. Krause M. MyoD and myogenesis in C. elegans. Bioessays 1995;17:219–28. [PubMed: 7748176]
19. Maduro MF, et al. Genetic redundancy in endoderm specification within the genus Caenorhabditis. Dev Biol 2005;284:509–22. [PubMed: 15979606]
20. Maduro MF, Rothman JH. Making worm guts: the gene regulatory network of the Caenorhabditis elegans endoderm. Dev Biol 2002;246:68–85. [PubMed: 12027435]
21. Lin R, Hill RJ, Priess JR. POP-1 and anterior-posterior fate decisions in C. elegans embryos. Cell 1998;92:229–39. [PubMed: 9458047]
22. Lin R, Thompson S, Priess JR. pop-1 encodes an HMG box protein required for the specification of a mesoderm precursor in early C. elegans embryos. Cell 1995;83:599–609. [PubMed: 7585963]
23. Kaletta T, Schnabel H, Schnabel R. Binary specification of the embryonic lineage in Caenorhabditis elegans. Nature 1997;390:294–8. [PubMed: 9384382]
24. Bieri T, et al. WormBase: new content and better access. Nucleic Acids Res 2007;35:D506–10. [PubMed: 17099234]
25. Hunt-Newbury R, et al. High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. PLoS Biol 2007;5:e237. [PubMed: 17850180]
26. Reece-Hoyes JS, et al. Insight into transcription factor gene duplication from Caenorhabditis elegans Promoterome-driven expression patterns. BMC Genomics 2007;8:27. [PubMed: 17244357]
27. Sarov M, et al. A recombineering pipeline for functional genomics applied to Caenorhabditis elegans. Nat Methods 2006;3:839–44. [PubMed: 16990816]
28. Sherlock G. Clustering Software 1999;2003
29. Saldanha AJ. Java Treeview--extensible visualization of microarray data. Bioinformatics 2004;20:3246–8. [PubMed: 15180930]

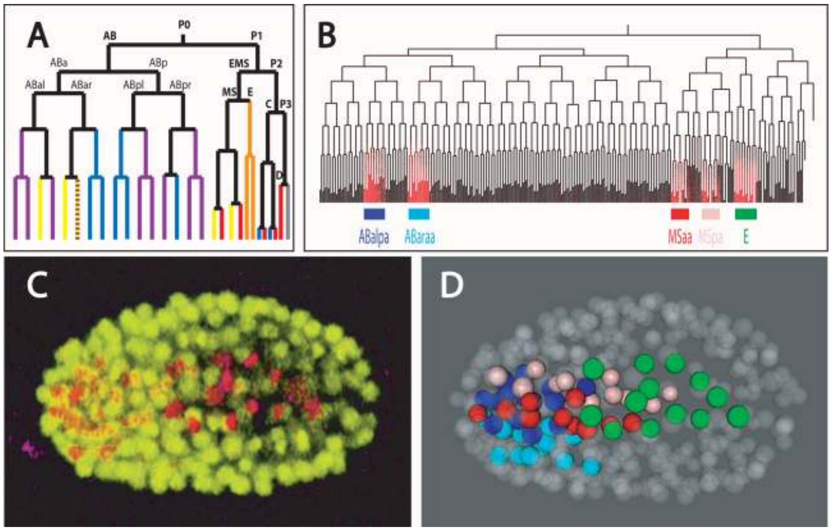

Figure 1. Displaying lineage-based expression data

(**a**) Overview of the *C. elegans* lineage, displayed using the conventional naming scheme[8]. In this naming, initial divisions lead to a set of founder cells cells (AB, C, D, E, MS and P4), with subsequent daughters named based on their orientation at birth relative to the primary embryonic axes (anterior-posterior, dorsal-ventral and left-right). For example ABala is the anterior daughter of ABal, which is itself the left daughter of ABa. Branch color indicates the predominant fate of terminal cells within each lineage (purple = neurons and neuronal support cells, yellow = pharynx, blue = epidermis, red = muscle, orange = intestine, grey = germ line) (**b**) Embryonic lineage tree color-coded by *pha-4* reporter expression. Branches are ordered as in Fig. 1a. The pharynx is formed from most of the cells in the lineages ABalpa, ABaraa, MSaa and MSpa. E forms the intestine. (**c**) 3D projection of a 350-cell stage RW10007 embryo with *pha-4* reporter expression (red) and ubiquitously expressed histone-GFP fusion proteins (yellow). (**d**) 3D model of the same embryo as in Fig. 1c generated from automated lineaging data. Expressing cells (> 5 units) are color coded by lineage identity, as labeled below the lineage tree in Fig. 1b. Nonexpressing cells are semitransparent.

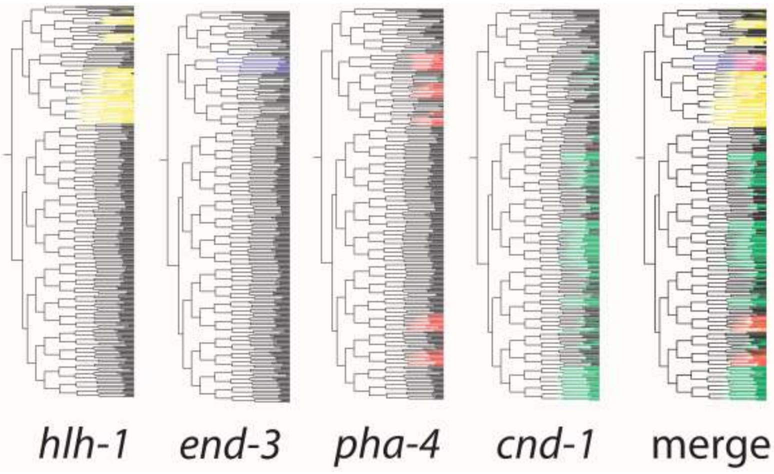

Figure 2. Aligning and comparing expression patterns

Expression-coded lineage trees for RW10055(*cnd-1*$^{3.2kb}$::HIS-24::mCherry) (green), RW10064(*end-3*$^{1.0kb}$::HIS-24::mCherry) (blue), RW10007(*pha-4*$^{4.1kb}$::HIS-24::DsRed) (red) and RW10097(*hlh-1*$^{3.3kb}$::HIS-24::mCherry) (yellow) reporters. Branch order is the same as in Figure 1.

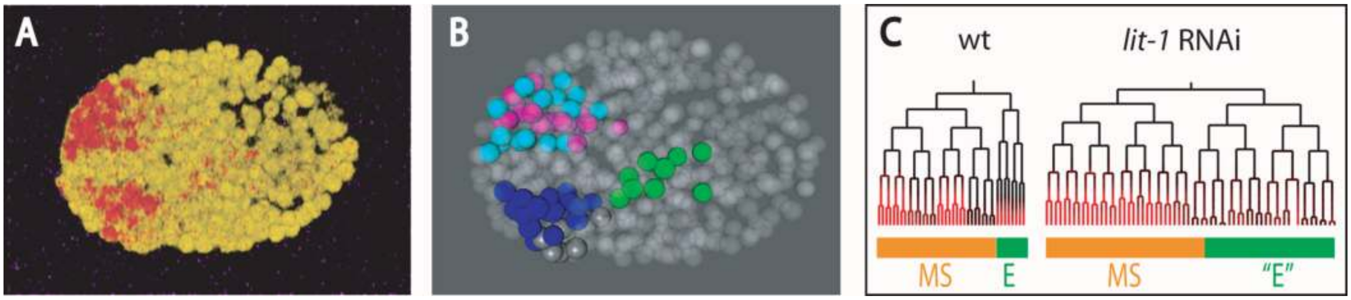

Figure 3. Identifying expression changes after RNAi treatment

(**a**) 3D projection of a *pha-4::*DsRed.T1-expressing embryo (red = DsRed; yellow = ubiquitously expressed histone-GFP fusion) from a *lit-1* RNAi-treated mother (RW10007). Note multiple isolated regions of expression while wildtype (Fig. 1a) has a single cohesive expression domain. (**b**) 3D model of the embryo in (**a**) with expressing cells colored as in Fig. 1c with the addition of the nonpharyngeal lineages ABarap (gray) and ABalpp (magenta). (**c**) EMS sublineage tree with *pha-4* reporter expression in embryos from wildtype (left) and *lit-1* RNAi-treated (right) mothers, showing extra cells and reduced *pha-4* expression in *lit-1* E lineage.

Table 1

Strains and replicates

Promoter	Reporter	Strain	Number of embryos
pha-4(F38A6.1a(4.1 kb))	HIS-24::mCherry	RW10062	4
cnd-1(3.2 kb)	HIS-24::mCherry	RW10055	2
cnd-1(3.2 kb)	HIS-24::mCherry	RW10060	2
cnd-1(3.2 kb)	HIS-24::mCherry	RW10083	1
hlh-1(3.3 kb)	HIS-24::mCherry	RW10097	2
hlh-1(3.3 kb)	HIS-24::mCherry	RW10112	1
end-3(1.0 kb)	HIS-24::mCherry	RW10064	2
pha-4(F38A6.1a(4.1 kb))	HIS-24::DsRed	RW10007	6

Table 2

Comparing lineage-based expression patterns to descriptions from literature

gene	literature cells	literature onset	observed cells (all terminal fates in parentheses)	observed onset
pha-4	Pharynx[14]	> 500 cells	ABalpaxxx, ABaraaxxx, ABarapaxxx (57 pharyngeal cells, 4 neurons, 4 hypodermal cells) MSaaxxx, MSpaxxx (37 pharyngeal cells, 2 neurons, 10 muscle cells)	< 200 cells < 200 cells
	Rectal precursors[14]	> 500 cells	ABprpapppxx, ABprpppppax, ABplpppppax, ABplpapppxx (7 rectal and digestive muscle cells, 2 neurons, 1 muscle cell	> 350 cells
	Intestine (E)[14]	100-200 cells	Exxx (20 intestine cells)	< 200 cells
end-3	Intestine blast cell (E)[19]	28 cells	Both E daughters (20 intestine cells)	2E(< 50 cells)
hlh-1	Muscle precursors and transiently in MS[18]	12-24 cells(MS) 90+ cells (C	MSxx (Muscle and pharynx) Cxpx (Muscle) Dxx (Muscle)	24 cells (MS) 90 cells (C) 180 cells (D)
cnd-1	15 of 16 initial AB descendents[17]	24 cells	Descendents of all 16 initial AB descendents with highly patterned expression	50-200 cells
	Non-hypodermal AB descendents not expressing UNC-86[17]	by > 500 cells	ABalaaxx, ABalpapax, ABalppapx, ABarappx, ABplaapax, ABplapax, ABplpaax, ABplppapx, ABpraapax, ABprapppx, ABprpaax, ABprppapx (97 neurons, 3 hypodermal cells, 1 arcade cell, 17 glia, 2 excretory system cells, 7 postembryonic blast cells)	100-200 cells