Resource

Cell

Analysis of Cell Fate from Single-Cell Gene Expression Profiles in *C. elegans*

Xiao Liu,[1] Fuhui Long,[2] Hanchuan Peng,[2] Sarah J. Aerni,[3] Min Jiang,[1] Adolfo Sánchez-Blanco,[1] John I. Murray,[4] Elicia Preston,[4] Barbara Mericle,[4] Serafim Batzoglou,[3] Eugene W. Myers,[2] and Stuart K. Kim[1,*]

[1]Department of Developmental Biology, Stanford University Medical Center, Stanford, CA 94305, USA
[2]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA
[3]Department of Computer Science, Stanford University Medical Center, Stanford, CA 94305, USA
[4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
*Correspondence: kim@cmgm.stanford.edu
DOI 10.1016/j.cell.2009.08.044

SUMMARY

The *C. elegans* cell lineage provides a unique opportunity to look at how cell lineage affects patterns of gene expression. We developed an automatic cell lineage analyzer that converts high-resolution images of worms into a data table showing fluorescence expression with single-cell resolution. We generated expression profiles of 93 genes in 363 specific cells from L1 stage larvae and found that cells with identical fates can be formed by different gene regulatory pathways. Molecular signatures identified repeating cell fate modules within the cell lineage and enabled the generation of a molecular differentiation map that reveals points in the cell lineage when developmental fates of daughter cells begin to diverge. These results demonstrate insights that become possible using computational approaches to analyze quantitative expression from many genes in parallel using a digital gene expression atlas.

INTRODUCTION

A powerful approach to dissect apart cellular phenotypes is to use molecular expression signatures. This is typically accomplished by using DNA microarrays to measure changes in expression of all or nearly all of the genes in the genome associated with an experiment or a condition. The combination of all of the expression changes in a cell generates a molecular phenotype for the state of the cell that has very high resolution. For cancer, expression signatures provide a powerful method to classify tumors and predict clinical outcomes (Potti and Nevins, 2008). For pharmacological drugs, one can generate a connectivity map showing molecular responses to different drugs (Lamb et al., 2006). For aging, molecular signatures can inform about the physiological age of tissues, apart from their chronological age (Rodwell et al., 2004).

Since molecular signatures are typically generated with DNA microarrays, the resulting data are noisy and reveal average expression from the entire sample. Thus an attractive alternative is to use libraries of images of GFP reporters or RNA in situ hybridizations. Images of GFP reporter expression or RNA in situ hybridizations have very high resolution, showing differential expression in different tissues or cells within a sample (Lecuyer and Tomancak, 2008).

Of all of the GFP expression data sets, images for *C. elegans* are particularly appealing because one can identify expression in specific individual cells. *C. elegans* is nearly unique among model organisms in that it has an essentially invariant cell lineage that gives rise to 558 cell nuclei in the newly hatched larva and 959 somatic cell nuclei in the adult hermaphrodite (Kimble and Hirsh, 1979; Sulston and Horvitz, 1977; Sulston et al., 1983). For worms expressing a fluorescence reporter, one can identify each nucleus, measure levels of fluorescence expressed in that nucleus, and thus analyze gene expression patterns at the level of single cells.

However, a major limitation for all of the GFP reporter and RNA in situ expression data is that the images must be manually browsed. The images show general patterns of expression but do not reveal quantitative levels of expression. Thus, the GFP expression data are not suitable for computational analysis, which is necessary to analyze all of the genes in parallel or to extract molecular signatures. To go beyond manual browsing, a key step is to automatically extract quantitative expression data from high-resolution images. This is analogous to converting images of DNA microarrays to data files showing expression of genes, except with single-cell resolution and more precise measurement of expression levels.

In *Drosophila* and zebrafish, digital atlases have been constructed that allow one to examine patterns of expression of multiple genes in a virtual embryo (Fowlkes et al., 2008; Keller et al., 2008). However, *Drosophila* and zebrafish do not have a fixed cell lineage, and hence it is not possible to precisely line up specific cells in different individuals as in *C. elegans*. In *C. elegans*, computational algorithms allow one to follow gene expression in the embryonic lineage from the one-celled zygote to the ~350-celled stage embryo (Murray et al., 2008).

In this work, we develop an automated method to extract quantitative expression data from single cells in postembryonic *C. elegans* (Long et al., 2009). This approach combines the advantages of high-resolution confocal microscopy and the ability to computationally analyze the data similar to analysis of

DNA microarray data. This combined approach provides a powerful new way to investigate patterns of gene expression and molecular signatures of cell fates in *C. elegans*.

RESULTS

A Gene Expression Database with Single-Cell Resolution

We developed an experimental pipeline to create a gene expression data set using images of worms carrying fluorescence protein reporters as a proof of principle to demonstrate that important biological insights can be extracted from single-cell gene expression data. To generate mCherry reporter constructs in a systematic way, we inserted the upstream regulatory region of a gene of interest into an expression vector using a library of cloned upstream regions (Dupuy et al., 2004). In *C. elegans*, upstream regions contain most of the regulatory information, and the promoter library has been previously shown to be sufficient to recapitulate patterns of gene expression (Dupuy et al., 2007; Dupuy et al., 2004). The expression vector contains mCherry fused to the coding region of histone H1, which produces a stable fluorescent protein localized to the nucleus. Transgenic *C. elegans* strains carrying integrated copies of the reporter construct were generated by biolistic transformation. To aid in identification of nuclei, we crossed in a GFP reporter that is expressed in the body wall muscle cells and the anal depressor muscle (from the *myo-3* promoter). Newly hatched first larval stage worms (L1) were stained with DAPI, and then worms were scanned by confocal microscopy in three fluorescence channels. The mCherry channel revealed expression from the regulatory region of interest, the GFP channel labeled body muscle and anal depressor muscle nuclei as landmarks, and the DAPI channel revealed all 558 nuclei (Figure 1A).

We used knowledge of the cell number, morphology of the cell nuclei, and their relative position with respect to each other to develop an automatic method to first identify specific cells in confocal images of worms expressing a fluorescent reporter, and then measure expression in specific cell nuclei. This approach captures high-resolution expression information available from confocal images of worms and converts the information into quantitative expression data suitable for computational analysis similar to output from DNA microarray experiments. We first computationally straightened the three-dimensional worm images and then registered them by aligning each image into a canonical rod shape that has the same precise orientation and size (Figure 1B) (Peng et al., 2008). Next, we developed segmentation software to automatically identify nuclei as bright objects in the foreground of dark, surrounding cytoplasm (Figure 1C). Third, we automatically named the nuclei in the confocal image stacks. GFP labeling of the 81 body wall muscle cells and the anal depressor muscle cell from the *myo-3* reporter aided us in identifying surrounding cell nuclei. Currently, the software can recognize and name 357 nuclei with 86% accuracy (Long et al., 2009). In addition to these 357 nuclei, an additional six nuclei were named manually. We have thus annotated 363 of the 558 nuclei in newly hatched L1 larvae (64%). These nuclei include all of the cell nuclei in the trunk, tail, and pharynx, representing nearly all tissue types in the worm. The only region that has not been well annotated is the nerve ring, which contains nuclei that are clustered too tightly to be reliably recognized at this time. Finally, we extracted values for mCherry expression for each identified nucleus (see the Experimental Procedures).

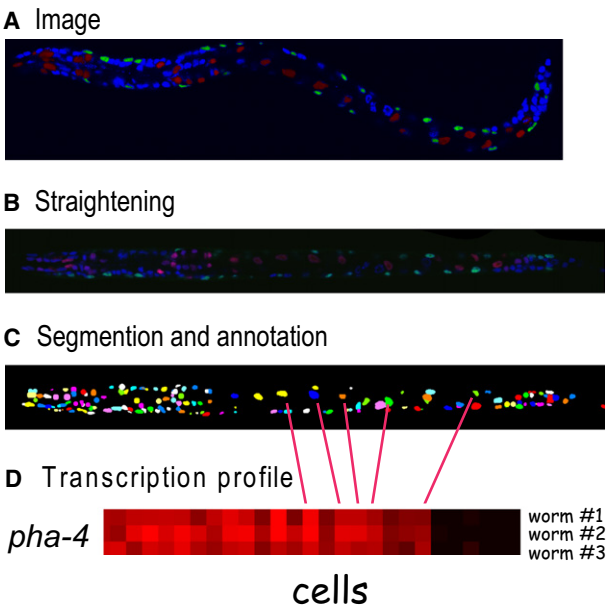

Figure 1. Generation of Cell Lineage Expression Profiles

(A) Three-dimensional image of a worm expressing mCherry from a promoter of interest, GFP in muscle cells from the *myo-3* promoter, and stained with DAPI.

(B) The confocal image is computationally straightened and set to a standard three-dimensional size.

(C) Nuclei stained with DAPI are automatically identified (Peng et al., 2009). Nuclei were labeled by pseudo-colors for visualization.

(D) Expression of the mCherry reporter in each of the 363 identified nuclei is calculated and displayed as a heat map.

Each of the steps in the pipeline can be scaled up, enabling one to generate much larger gene expression data sets in the future. The expression dataset currently contains 324 images from 93 reporter genes, including 60 that encode transcription factors (Table S1 available online). To control for differences in fluorescence intensity due to sample thickness, we normalized mCherry expression to DAPI fluorescence because the DNA content of every nucleus is constant. By plotting the expression values in a heat map, we converted the complex expression information embedded in fluorescence images into a form that is suitable for computational analysis (Figures 1D and 2). Each row in these expression profiles shows the pattern of expression of a mCherry reporter gene in a highly quantitative manner with single-cell resolution. The full data set can be queried with wormDB (http://www.computationalbio.com/Stanford/KimLab/wormDB/) and downloaded from Tables S2–S4.

We performed several tests to evaluate the reproducibility of our system to measure mCherry expression levels. First, we reannotated three images to determine the reproducibility of the annotation procedure, and we found that 98% of the nuclei were assigned the same cell name. Second, we found that expression values from different images of the same worm are

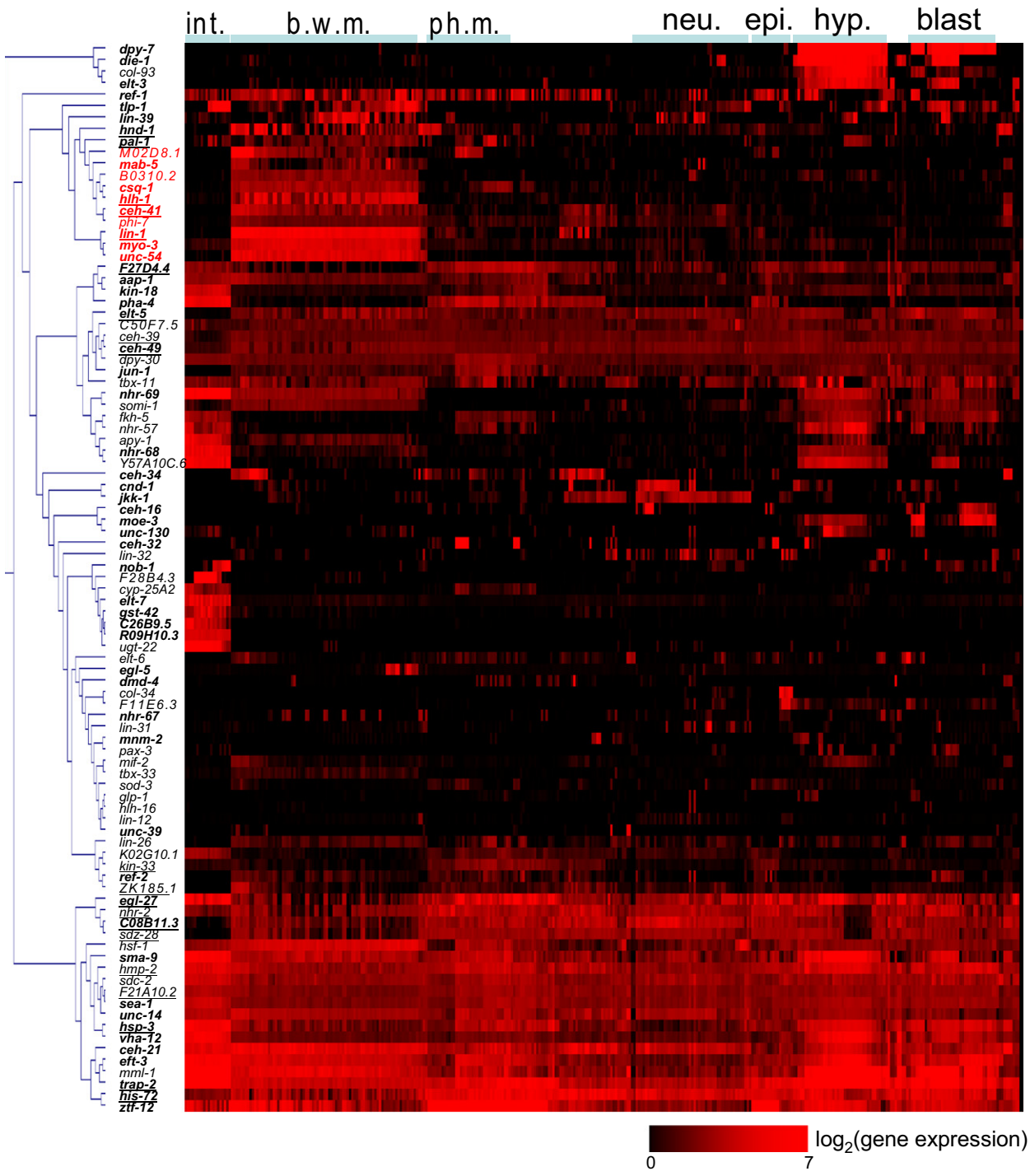

Figure 2. A Profile of Gene Expression at Single-Cell Resolution

Shown is the mCherry expression level for 93 reporter genes in 363 cells (out of 558 total, 64%) in newly hatched L1 larvae. We adjusted the gene expression level by calculating (gene expression level + 500)/500. The scale bar shows $\log_2$ (adjusted gene expression level). Genes were clustered according to their expression profile. Cells were manually arranged according to their tissue types and anterior-posterior order. Bold indicates genes that have been previously analyzed for patterns of gene expression (Table S1). Red indicates genes that are expressed mainly in the body wall muscle cells. Underlining indicates cell lineage-specific genes that are discussed in the main text. Data for single-cell gene expression can be found in Tables S2–S4. b.w.m., body wall muscle; blast, blast cells; epi., epithelial cells; hyp., hypodermal cells; int., intestine cells; neu., neurons; ph. m., pharyngeal muscle.

highly correlated (correlation efficient R > 0.99), indicating that the technical reproducibility of our procedure is very high. Third, we examined the biological variability of mCherry gene expression between individual worms from the same strain. For most strains, we found that different individual worms had correlation coefficients for mCherry expression of R > 0.80 (Figure S1A), indicating both that the annotation of cell nuclei is reliable and that the mCherry expression is reproducible. Finally, to test whether the site of integration has a large effect on expression, we generated different transgenic lines using the same mCherry reporter construct. We generated multiple lines for 12 mCherry reporter constructs and found that the level of expression could be different between different transgenic lines but that the correlation in mCherry expression was largely similar whether the worms were derived from the same strain or from different strains expressing the same construct (Figure S1B). This result indicates that the site of integration of the mCherry reporter in different transgenic lines affects the level but does not dramatically affect the pattern of mCherry expression.

The expression patterns for 53 of the 93 genes in our database have been described previously (Table S1). For 47 of these, our results match previous results. Overall, the automated single-cell lineage expression data shows a close match to previous expression data, but has much higher resolution and accuracy than was previously possible by subjectively viewing each image one at a time. The expression database also includes data for 40 genes whose expression had not been previously analyzed at the L1 stage.

Correlation of Gene Expression with Cell Fate and Cell Lineage

We analyzed the pattern of expression of every gene to determine the relative effect of cell fate and cell lineage. Cell fate has a strong influence on gene expression as highly differentiated cells must express specific genes to carry out terminal differentiation functions. Cell lineage could play a strong role in gene expression for a number of reasons, including stable segregation of lineage factors or stable transmission of chromatin structure. We compared the influence of cell lineage and cell fate on the expression pattern for each of the 93 reporter genes in this study. Specifically, for each gene, we examined whether it was expressed in cells that had the same fate (i.e., expressed in all of the body wall muscle cells) or in cells that were related by lineage (i.e., progeny of the blastomere AB.a).

For the majority of cases, gene expression correlated with cell fate rather than cell lineage (Figure 2). For example, ten genes are expressed mainly in the 81 body wall muscle cell nuclei, which are derived from four blastomere cells: AB, MS, C, and D. In addition, we observed tissue-specific expression for genes expressed in the hypodermis, neurons, pharyngeal muscle, blast cells, and the intestine (Figure 2). Each of these tissues is derived from multiple points in the cell lineage, except for the intestine, which is derived entirely from the E blastomere.

We found examples in which gene expression followed cell lineage more than cell fate. Body wall muscle cells are derived from the AB (one cell), MS (28 cells), C (32 cells), and D (20 cells) lineages. The muscle cells derived from MS and D are interspersed with each other in body muscle bundles and are thought

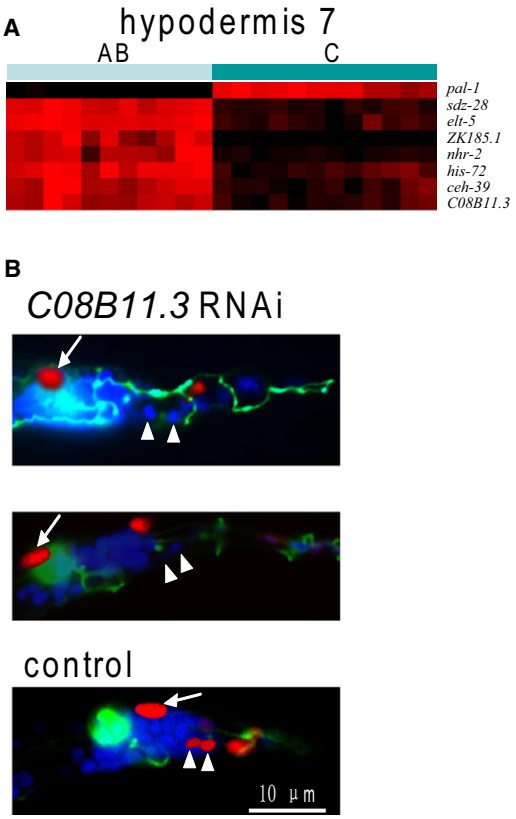

Figure 3. Cell Lineage-Dependent Gene Expression among Nuclei in the hyp7 Syncytium

(A) Genes are differentially expressed between AB-derived and C-derived nuclei ($p < 10^{-5}$, t test). Color indicates level of expression. Gene expression levels were normalized for each gene so that the minimal and maximal expression values are 0 and 1 for each gene. Expression levels in these nuclei and p values for all genes are in Table S5, part A.

(B) Different transcriptional control of AB- versus C-derived nuclei in the hyp7 syncytium. Shown are the tail areas of L1 stage worms expressing a *col-93:mCherry* reporter (expressed in hypodermal cells), *ajm-1*:*GFP*, (a hypodermal cell boundary marker), and stained with DAPI. The top and middle images show *C08B11.3(RNAi)* animals in which the two AB-derived nuclei do not express *col-93:mCherry*. In most cases, the AB-derived nuclei have not fused with the syncytium (top), but sometimes they fuse (middle). The bottom image shows a control with normal expression of *col-93:mCherry* in the hyp7 nuclei and cell fusion. Arrow head, AB-derived hyp7 nuclei; arrow, C-derived hyp7 nucleus.

to be physiologically indistinct. We found 18 genes that show different expression in body wall muscle cells depending on the cell lineage (Figure S2A). Fifteen of these encode transcription factors, many of which are known to be important for muscle cell fate. For *pal-1*, previous experiments have shown that this gene is important for generating body wall muscle cells derived from the C lineage but not from the MS lineage (Edgar et al., 2001).

We observed a surprising pattern of differential gene expression for different nuclei within the same cell syncytium (Figure 3A). Specifically, hypodermal 7 is a syncytium containing 23 nuclei that comprises a major section of the skin. Twelve hyp7 nuclei are derived from the C lineage, and eleven are

derived from the AB lineage. The molecular signature for nuclei derived from the C lineage is significantly different from that of nuclei derived from the AB lineage. *sdz-28*, *elt-5*, *ZK185.1*, *nhr-2*, *his-72*, *ceh-39*, and *C08B11.3* are expressed in hyp7 nuclei derived from AB, whereas *pal-1* is expressed in hyp7 nuclei derived from C. Since the hyp7 syncytium is formed by cell fusion, one possibility is that these genes might only be differentially expressed before cell fusion and might be evenly expressed once the cells have fused, such that mCherry reporter protein levels may be differentially localized immediately after cell fusion but would equalize rapidly within the syncytium after fusion. We ruled out this possibility for *C08B11.3* by showing that differential expression of *C08B11.3:mCherry* was stable for at least 8 hr, until the end of the L1 larval stage and that new expression appears after photobleaching (Figure S3). Thus, nuclei in the same syncytial cell can show large differences in gene expression pattern, indicating that there can be different transcriptional control in different nuclei and also that mRNAs expressed from one nucleus give rise to proteins that stay localized to the same nucleus.

We next performed a genetic experiment to show differential transcriptional control of AB- versus C-derived nuclei in the hyp7 syncytium. hyp7 cell nuclei fuse together to form one syncytium late in embryogenesis and then begin to express collagen genes such as *col-93*. We used RNA interference (RNAi) to reduce activity of the transcription factor gene *C08B11.3*, which is expressed in nuclei from AB- but not C-derived blastomeres, and then looked at the fates of the AB- versus C-derived nuclei in hyp7. We scored two AB-derived and two C-derived nuclei in hyp7 and found that *C08B11.3(RNAi)* affected the fates of the AB- but not C-derived nuclei. Specifically, the AB-derived nuclei did not express the *col-93* collagen reporter in seven of 51 cases examined (14%). In some cases, the AB-derived nuclei fused with the hypodermal syncytium as in the wild-type, but in most cases these hypodermal nuclei did not fuse with the rest of the syncytium. The C-derived nuclei appeared normal in all *C08B11.3(RNAi)* animals (Figure 3B). Together with our information about cell lineage-restricted expression, these observations suggest that different transcriptional networks can be used to produce cells with the same fate.

Molecular Signatures for Cell Fates

The combined expression profiles of the 93 reporter genes in each cell is a molecular signature for that cell, and can be used as a quantitative measure to determine whether cells have different, related or identical cell fates. We first clustered the cells into groups in a two-dimensional scatter plot according to their correlation in gene expression (Figure 4). In this scatter plot, the distance between two cells indicates similarity in molecular signatures. Cells that are placed close to each other express the 93 reporter genes at similar levels and cells that are far from each other have different molecular signatures. We find that cell clusters are consistent with known fates—intestinal nuclei cluster with other intestinal nuclei, as do nuclei for muscles, neurons, the hypodermis etc.

The map of molecular signatures shows an example of a spatial domain in gene expression for the pharynx. The pharynx is isolated from the rest of the worm anatomy by a layer of basal

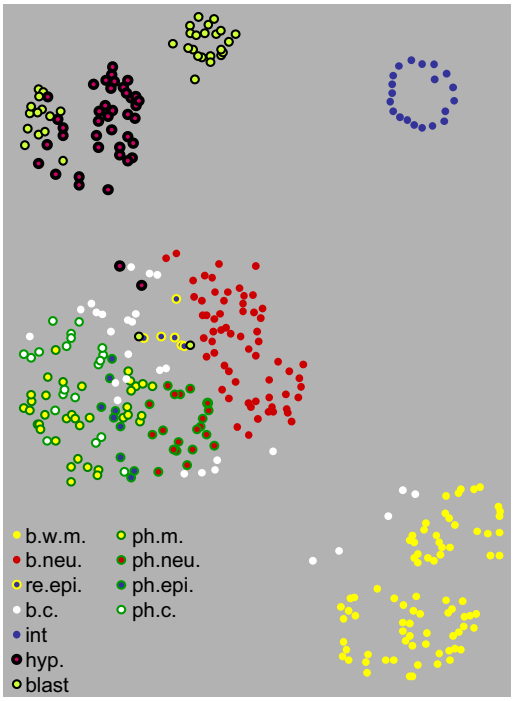

Figure 4. Clustering Diagram of 363 Cells According to the Similarity of Their Gene Expression Profiles

The terrain map of nuclei was generated by Genesis (Sturn et al., 2002). Colors indicate different tissue types. Distance between cells in the x-y plane indicate levels of molecular similarity, such that cells with similar gene expression patterns are placed close to each other and cells with different patterns are placed far apart. b. c., other body cells; b. neur., body neurons; b.w.m., body wall muscle; blast, blast cells; hyp, hypodermal cells; int., intestine cells; ph. c., other pharyngeal cells; ph.m., pharyngeal muscle; ph. neur., pharyngeal neurons; ph. rec., pharyngeal epithelial cells; re. epi., rectal epithelial cells.

lamina, and includes many distinct cell types, such as muscle, neural, and epithelial cells. The molecular signature map shows that pharyngeal muscle cells are clustered more closely to pharyngeal neural or epithelial cells than they are to body wall muscle cells. Similarly, pharyngeal epithelial and neuronal cells are clustered more tightly with other pharyngeal cells than to other epithelial or neuronal cells, respectively. These results indicate an underlying similarity in expression within the pharyngeal spatial domain.

The map of molecular signatures shows which tissues are relatively homogenous and which have diverse types of cells within that tissue. Cells from homogeneous tissues have much more similar correlations in gene expression to each other than do cell nuclei from heterogeneous tissues. For example, all 20 intestinal nuclei are clustered tightly on the molecular signature map, indicating that these cells have very similar gene expression signatures and are nearly homogeneous (Figure 5). Neuronal cell nuclei are not tightly clustered on the two-dimensional map of cell signatures, indicating diverse cellular functions within this tissue type. Body wall muscle and blast cells also show high levels of diversity in molecular signatures. Thus, molecular signatures obtained from the high-resolution expression database not only cluster cells according to tissue type,

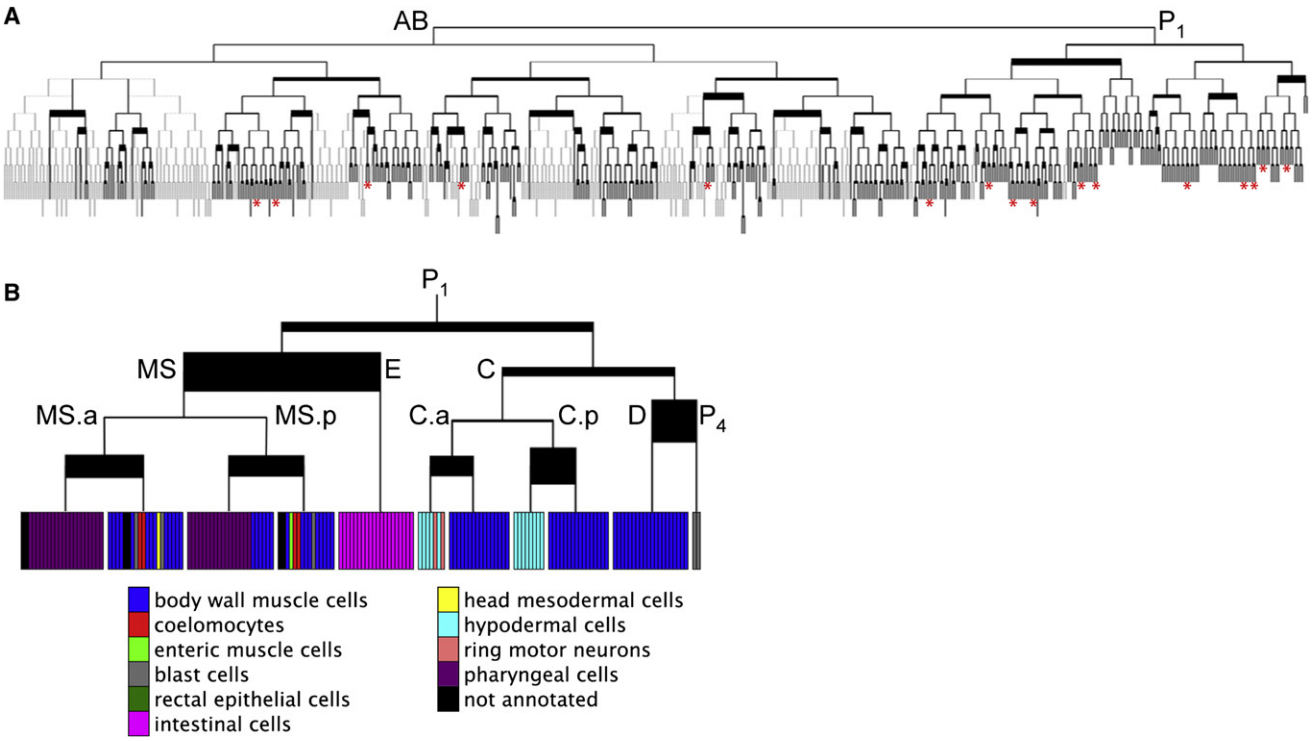

Figure 5. Molecular Differentiation Map for the *C. elegans* Cell Lineage

(A) Each cell in each bifurcation in the tree (corresponding to one or more cell divisions) is compared to its sister cell to determine whether they have similar or different gene expression states. Line thickness indicates degree of dissimilarity between these cells. The modified cell lineage is displayed. Dotted lines show the portions of the complete cell lineage that were unscored. The solid lines represent the modified lineage used for the analysis (Supplemental Experimental Procedures). Asterisks denote 16 new asymmetric terminal cell divisions.

(B) Expanded view of the P1 lineage, showing the fates of cells. Specific cells discussed in the text are shown. Colored bars represent tissue types of terminal cell nodes in the cell lineage tree.

but can also distinguish homogeneous from heterogeneous tissues.

In some cases, we found interesting trends that could explain some of the differences in gene expression between different cells in the same tissue, such as differences in expression between different body wall muscle cells. The anterior body wall muscle cells are larger and form different neuronal connections than posterior body wall muscle cells (Bird and Bird, 1991; White et al., 1986). We found that there is an anterior-posterior gradient of gene expression in these cells. Among 68 genes that are significantly expressed in the body wall muscle, 13 are expressed at higher levels in anterior body wall muscle cells and five are expressed at higher levels in posterior cells (Figure S2B).

A Map for Molecular Differentiation during Embryonic Development

We have created a molecular differentiation map based solely on molecular signatures, in which we identify regions of the cell lineage where developmental fates begin to diverge. Newly hatched worms have 558 cells resulting from 670 cell divisions from the one-celled zygote (Sulston et al., 1983). For each gene, we used the worm lineage and the observed expression levels at the 558-celled stage to predict when that gene became committed to be expressed in the embryonic lineage. We then searched for embryonic cell divisions in which daughter cells become committed to express a different battery of genes, thereby identifying cell divisions that are asymmetric and revealing when developmental potentials begin to diverge in the embryonic lineage.

We approached the problem of predicting gene commitment by adapting the parsimony algorithm used in molecular evolution, which determines ancestral sequences along a known phylogeny tree. Our algorithm assigns expression values to embryonic cells that minimize the changes in commitment needed to explain the expression pattern observed in the L1 worm from the known cell lineage. To do this, the gene commitment algorithm builds a graph based on the known cell lineage, where nodes signify cells that are connected by directed edges to their daughter cells. The terminal nodes are the 363 cells with observed expression values for 93 genes in the L1 worm. Our goal is to assign commitment values to every embryonic cell indicating how committed the cell is to expression of each gene. The algorithm assigns expression values to embryonic cells that minimize the changes in commitment to gene expression required to produce the observed expression profile in the L1 worm (see the Experimental Procedures).

The embryonic expression pattern is known in detail for nine of the genes from this study (Figure S4). We compared the known embryonic expression to predictions from the gene commitment

algorithm, and found a close match for seven genes. For *cnd-1*, there is transient expression in some embryonic lineages that was missed by the gene commitment algorithm (Figure S4D). For *lin-39*, the algorithm predicted commitment before protein expression was directly observed (Figure S4G). This time delay could be caused by a lag involving setting up the regulatory interactions that turn on expression, transcription of the gene, translation of the message, and accumulation of protein. For each of the remaining 84 reporter genes, we generated models predicting commitment to express a particular gene in the cell lineage (Figure S5).

For each cell, we combined the results from all 93 genes to generate a molecular signature of that cell (Experimental Procedures). We used this molecular signature as a quantitative measure to compare two cells to each other and to determine similarities and differences in their fates. We first used this approach to generate a molecular differentiation map, which shows points in the cell lineage when cell divisions generate daughter cells that are different.

For the 143 terminal cell divisions that we observed, we directly compared gene expression patterns of the 93 reporter genes in the daughter cells. Daughter cells that have different molecular signatures indicate cell divisions that are asymmetric. To find a cutoff that can distinguish symmetric from asymmetric cell divisions, we permuted the data such that every cell division is symmetric. Using a false discovery rate of 1%, we found 54 asymmetric cell divisions. Of these cell divisions, 38 were previously known to be asymmetric, and 16 asymmetric divisions were previously unknown (Figure 5A and Table S7).

For cell divisions that occur earlier in the embryo, we used the parsimony algorithm to predict whether sister cells (or cells separated by a common ancestor) are committed to express a similar set of genes. The amount of developmental change at each cell division is shown by the thickness of the line in Figure 5. Thick lines indicate cell divisions that generate daughters that are different from each other, whereas thin lines indicate symmetric cell divisions. We can thus overlay developmental activity onto the cell lineage and mark key points for cell differentiation during development, either due to cell-cell signaling or to asymmetric cell division.

One example of a highly asymmetric cell division is the division of EMS to generate E (which produces only intestinal cells) and MS (which produces pharyngeal and body wall muscle cells) daughters (Sulston et al., 1983) (Figure 5B). The E blastomere becomes different from the MS blastomere due to a Wnt signal from the P2 cell, which determines gut cell fate by inducing the sequential activation of the *end-1*, *end-3*, *elt-2* and *elt-7* GATA transcription factors (Maduro, 2006). By parsimony, 54 genes are predicted to be committed differently in the E versus MS daughter cells.

The division of MS.a and MS.p are also asymmetric, producing one daughter that generates pharyngeal cells (MS.aa and MS.pa) and another that produces body wall muscle cells (MS.ap and MS.pp), due to interaction with the AB.a cell (Schnabel, 1994). The molecular differentiation map shows that this cell division is highly asymmetric, as 38 and 35 genes are predicted to be differentially committed in the daughter cells of MS.a and MS.p, respectively.

C.a and C.p undergo an asymmetric cell division, as one daughter generates muscle cells (C.ap and C.pp), whereas the other daughter makes mostly hypodermal cell nuclei (C.aa and C.pa,). In the molecular differentiation map, the daughter cells of C.a and C.p differ in their developmental commitment for 32 and 51 genes, respectively.

In summary, the molecular differentiation map correctly annotates cell divisions that were previously known to be asymmetric but also predicts many new cases of asymmetric cell divisions that were previously unknown.

Developmental Clones and Sublineages

In order to systematically search for repeating use of developmental patterns in the cell lineage, we generated a heat map comparing the molecular signatures of each of the 363 cells to each other (Figure 6A). In this heat map, the cells are aligned according to their lineage along the x and y axes. We searched the heat map for two types of patterns: developmental clones and sublineages.

A developmental clone is a progenitor cell whose progeny have nearly identical cell fates. In the heat map, developmental clones appear as a discrete box along the diagonal, in which the molecular signature of every cell within the box is similar to each other. The clearest example of a developmental clone is the E cell, which is known to generate 20 intestinal cells. In the cell fate heat map, the 20 intestinal cells form a box along the diagonal showing that each cell in the E cell clone has a very similar molecular signature (Figure 6A). In addition to the E cell, other examples of developmental clones include C.pa (generates eight hypodermal cells), C.ap/C.pp (each generates 16 body wall muscle cells), and D (generates 20 body wall muscle cells).

A sublineage is a set of cells that undergoes the same pattern of cell divisions. In the cell fate heat map, sublineages appear as diagonal lines that are offset from the main diagonal, such as the diagonals generated by AB.pl and AB.pr. The length of the diagonal line includes all of the progeny of AB.pl and AB.pr, indicating that each homologous cell in the AB.pl and AB.pr lineage is equivalent to each other (Figure 6B). MS.a and MS.p also share a common sublineage.

C.a and C.p are a combination of a sublineage and a developmental clone, forming an off-center diagonal, indicating that each undergoes a similar sublineage (Figure 6C). C.ap and C.pp are developmental clones as each generates 16 body wall muscle cells. C.pa and C.aaa are developmental clones generating eight and four hypodermal nuclei, respectively.

In summary, the developmental clones and sublineages shown in Figure 6 extend earlier classic work that originally defined these lineage patterns using observation by Nomarski microscopy (Sulston and Horvitz, 1977; Sulston et al., 1983). With our approach, similarities and differences in cell lineages are revealed by quantitative comparisons of molecular signatures of cells.

DISCUSSION

We developed an automated method to quantify expression of 93 fluorescent reporter proteins in 363 specific cells from individual worms. For each gene, we analyzed its expression pattern to assess the relative effects of cell fate (defined as the

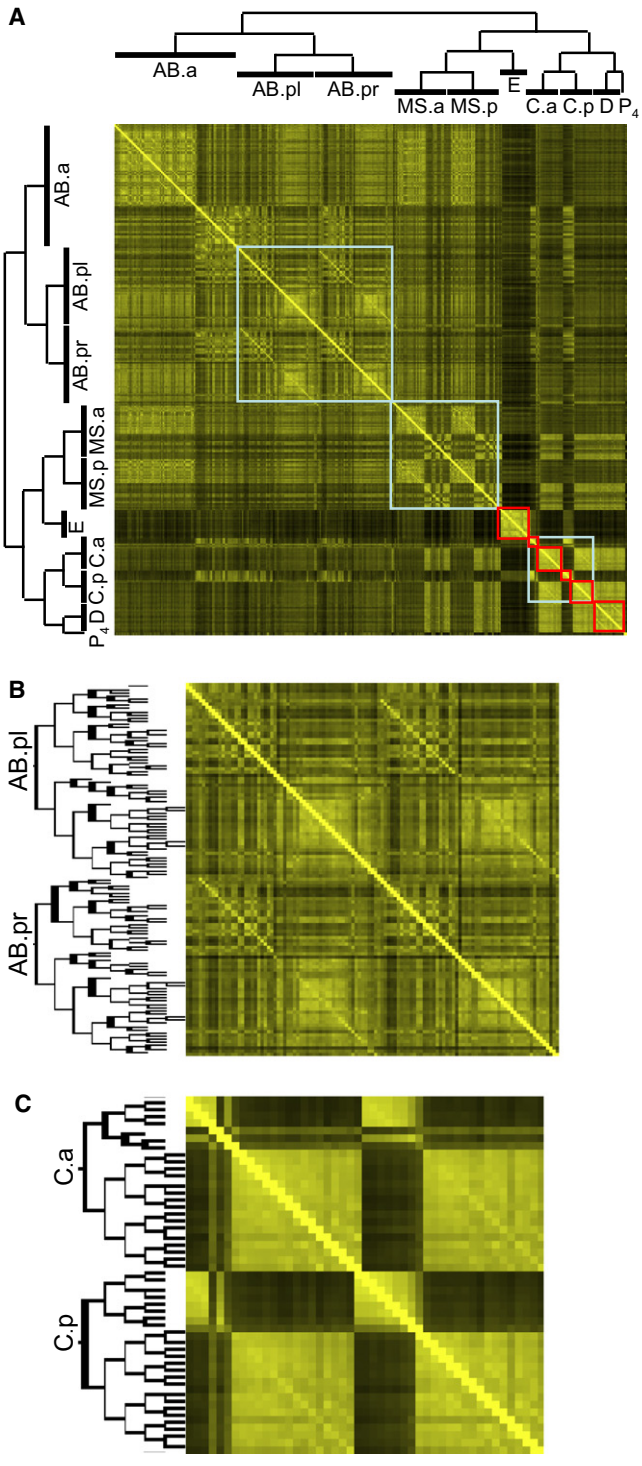

Figure 6. Developmental Clones and Sublineages in the *C. elegans* Lineage

(A) Cells are aligned according to their lineage along the x and y axes. The similarity of a pair of cells is a function of the activity score using the expression profile of the 93 reporter genes (s_{ij} *defined* in the Experimental Procedures). The lineage of the first several cell divisions is shown. Red boxes represent clones of cells of similar cell fate. Blue boxes represent developmental sublineages. Omitted cells are excluded from the map.

(B) AB.pl and AB.pr share a common sublineage.

(C) Developmental clones and sublineages from C.a and C.p.

physiological and functional capabilities of a cell, such as muscle, neuronal, or skin cells) and cell lineage (defined by the pattern of cell divisions that generated the cell, such as AB.pla). In *C. elegans*, many cell fates are derived from cells with similar lineages, such as the intestinal cells that are all derived from the E blastomere. For genes expressed in these cells, it is not possible to separate the effects of cell fate and cell lineage on gene expression. However, some types of cells, such as neurons and epithelial cells, are derived from distinct points in the cell lineage. As expected, expression of most genes was strongly linked to cell fate, as they were expressed in similar types of cells that were generated in distinct parts of the lineage.

In several cases, we found genes whose expression was linked to cell lineage but not cell fate, which is surprising because cells previously thought to have identical cell fates express different sets of genes depending on their cell lineage history. Previous work has also described some genes whose expression is linked to cell lineage (Baugh et al., 2005; Bowerman et al., 1997; Broitman-Maduro et al., 2006; Edgar et al., 2001; Good et al., 2004; Hunter and Kenyon, 1996). This connection between expression and cell lineage could only be found in model organisms such as *C. elegans*, where knowledge of the complete cell lineage allows one to keep track of the origin of cells that share the same cell fate.

Molecular Signatures of Cell Fate

For each of the 363 cells scored in this study, we used the expression levels of the 93 reporter genes to generate a molecular signature of that cell's fate. We used the molecular signatures in two ways—to find parts of the cell lineage when cell fates begin to diverge and to find repeating cell fate modules expressed from different parts of the cell lineage.

The molecular signatures were used to create a molecular differentiation map that shows when cell divisions generate daughter cells that have different commitments to express the 93 reporter genes. The molecular differentiation map closely matches results obtained by classical development studies, except that it is based on molecular signatures that graphically depict when and where major changes in developmental commitment occur. These changes in developmental commitment generate cell asymmetry in the *C. elegans* lineage and arise by one of two general mechanisms: asymmetric segregation of determinants during a cell division (an intrinsic mechanism) or extracellular signaling cues that affect one daughter differently than the other (an extrinsic mechanism) (Horvitz and Herskowitz, 1992).

Classical studies looking at the generation of cell fates from the *C. elegans* cell lineage defined several types of lineage patterns (Sulston and Horvitz, 1977; Sulston et al., 1983). One type is a developmental clone of cells, in which all of the progeny of a single progenitor cell adopt a single cell fate. Another pattern is a sublineage, in which cells distantly related by lineage adopt similar cell fates. We systematically compared the molecular signature of each cell to all other cells in order to see how cells with similar molecular signatures were generated in the worm cell lineage. We found clear examples of developmental clones

(such as the E blastomere) and of sublineages (such as AB.pl and AB.pr). By defining clones and sublineages, the generation of 363 individual cells can be broken down to simpler repeating patterns, likely representing developmental modules that are reused to generate the same cell fate in multiple instances. For example, a single gene regulatory network may be used repeatedly to generate all cells in a developmental clone or homologous cells in a sublineage.

Quantitative Analysis of Gene Expression Images

Previously, a simple method of extracting quantitative expression data has been developed by passing worms through a fluorescence-activated cell sorter and measuring fluorescence intensity along their anterior-posterior axis (Dupuy et al., 2007). This method is fast and allows one to measure expression from a large number of worms, but it contains very little information about GFP expression in specific tissues.

Recently, a computational method has been developed to analyze movies of GFP-expressing worms during embryonic development, from the one-celled zygote to the ~350-celled stage embryo (Murray et al., 2008). Similar to our automatic cell lineage analyzer, this approach generates quantitative expression data at the level of single cells. A fundamental difference in the two approaches is that nuclei in the embryo are annotated on the basis of their cell division pattern from continuous observation, whereas nuclei in the L1 larvae are named on the basis of their appearance and position relative to each other in a single confocal image. Another difference is that terminal fates such as neuronal, muscle, and intestinal fates are only established at after the majority of the embryonic cell divisions are complete, not at the 350-celled stage. Thus, tissue- and cell type-specific patterns of gene expression can be studied in the newly hatched worm but not in the early embryo. Beyond *C. elegans*, computational methods have been developed for analysis of expression from *Drosophila melanogaster* and *Danio rerio* (Fowlkes et al., 2008; Keller et al., 2008).

Gene Expression Database at Single-Cell Resolution

The complete cell lineage for *C. elegans* is known, providing a unique opportunity to analyze patterns of gene expression at the level of specific cells. By developing an automatic cell lineage analyzer, we extracted quantitative data from expression images showing expression of 93 genes in 363 specific cells. The data can be viewed with wormDB (http://www.computationalbio.com/Stanford/KimLab/wormDB/) and downloaded from Tables S2–S4.

This expression database could be greatly expanded in the future, which would greatly increase the resolution of this digital representation of worm development. Improvements in the automatic cell lineage analyzer will enable one to identify a larger number of cells and to analyze genes at a much faster rate. The automatic cell lineage analyzer can be modified so that it works on the other three larval stages as well as the adult, and confocal images of reporter genes can be generated from mutant worms or from worms grown under diverse growth conditions. Obtaining high-resolution expression data for each cell throughout development would provide a unique molecular framework for understanding gene regulation circuitry and cell fate patterning.

EXPERIMENTAL PROCEDURES

Strain Construction

pJIM20 was used for the promoter::reporter fusions, which contain a *his-24::mCherry* reporter and *unc-119* selection marker (Murray et al., 2008). Gene expression is driven by the promoter from the gene of interest. The upstream regions were inserted into pJIM20 either by Gateway recombination with promoter constructs from the promoterome (Dupuy et al., 2004) or from DNA fragments generated by PCR from genomic DNA.

DNA constructs were introduced into *unc-119(e3)* worms by microparticle bombardment (Praitis et al., 2001) and were then crossed with a strain containing a P*myo-3*::GFP reporter (Fire et al., 1998). Detailed information on promoters and strains is in Table S1.

Imaging Protocol

To obtain worms early in the L1 stage, we isolated eggs, and those that hatched within a 3 hr time window were used for image analysis. Although many genes show stable expression during this 3 hr time window, some genes (such as *sod-3*) show dynamic expression at this time in which case variability in expression could be caused by differences in developmental stage. L1 larvae were fixed and DAPI stained as described in Ruvkun and Giusto (1989) and then mounted in 60% glycerol.

Images were obtained with a Leica SP2 AOBS confocal microscope. The pixel size was 0.116 μm in the x-y plane and 0.122 μm in the z direction.

Automatic Cell Lineage Analysis

The three-dimensional image stacks of worms were straightened computationally along the anterior-posterior axis (Peng et al., 2008). The cell nuclei in each image stack was segmented automatically (Long et al., 2009) and then manually edited with the VANO interactive interface (Peng et al., 2009). The identities of 357 nuclei were automatically identified using the 82 GFP-labeled nuclei as landmarks with 86% accuracy.

Manual Editing of Nuclei

The named nuclei were manually corrected with VANO, and an additional six nuclei were manually annotated according to Sulston and Horvitz (1977) and http://www.wormatlas.org/. Several neighboring nuclei in hyp 7 (ABpraapppp, ABarppaapa, and ABarpaappp) have variable locations relative to each other and could not be reliably identified. Furthermore, some pairs of nuclei in the midline have ambiguous cell lineage identities. For example, for hyp3 nuclei, one cell nucleus (AB.plaapaaaa) migrates into the midline from the left, and another cell nucleus (AB.praapaaaa) migrates into the midline from the right. For convenience, we represent the anterior nucleus of a pair by (lr) and the posterior one by (rl). We did this similarly for other pairs of nuclei with ambiguous lineage identities. For the pharyngeal muscles, (ap) denotes the anterior nucleus and (pa) denotes the posterior one.

Many mCherry reporters are expressed in a small fraction of nuclei. For 72 images, we only identified those nuclei that show mCherry expression to expedite the annotation process. However, for each reporter, at least one image is fully annotated for all 363 nuclei.

Gene Expression Measurement

For every cell nucleus, the automatic cell lineage annotator measures the total volume of the nucleus, the total mCherry intensity summed over every voxel within the nucleus, and the total DAPI intensity summed over every voxel in the nucleus. The raw mCherry values were adjusted to account for background fluorescence and for loss of intensity due to distance of the focal plane from the objective.

For measurement of background mCherry fluorescence, ten pseudo-nuclei of equal size were drawn in the digestive tract of each worm. The background mCherry was measured in each false nucleus, and then the average fluorescence in the mCherry channel of all ten false nuclei was calculated. The density of the background mCherry is the average background fluorescence in the mCherry channel divided by the average size of the pseudo-nuclei. To find the amount of background fluorescence for each nucleus, the background mCherry density is multiplied by the size of the cell nucleus. To find the adjusted level of mCherry for each nucleus, the background mCherry level

was subtracted from the raw mCherry level. A similar approach was used to calculate the adjusted DAPI levels.

To account for effects on mCherry fluorescence caused by different depths in the confocal image stack, we used DAPI fluorescence as a normalization control because all nuclei in the newly hatched worm have the same DNA content. We calculated a normalized DAPI fluorescence level for each nucleus by dividing the adjusted DAPI fluorescence level of each nucleus by the median DAPI fluorescence level of all nuclei in the worm. We then calculated the normalized mCherry level for each nucleus by dividing its adjusted mCherry level by its normalized DAPI level. The normalized mCherry level is the level of fluorescence in a nucleus after background fluorescence has been subtracted and after correcting for variable distances on the z axis. If the normalized level was negative, we used 1 instead.

Multiple worms were imaged for each mCherry reporter, and 12 reporters were used to generate multiple transgenic lines. To show the average level of gene expression in each nucleus, we used the median normalized mCherry expression value from all images.

RNAi

Double-stranded RNA of *C08B11.3* was induced in *E. coli* with 100 μl of 0.1 M IPTG. Worms at the L4 larvae stage were added to the plates and incubated 2 days, and L1 progeny larvae were scored. *ajm-1*::GFP is described in Mohler et al. (1998).

Commitment Algorithm

The cell lineage was used to construct a graph of nodes and directed edges, where nodes represent cells that are connected by directed edges from parent, u, to daughter cell, v. The cell lineage was modified so that the tree consisted of only annotated cells and their common ancestral cells (Supplemental Experimental Procedures).

Using this fixed graph, we built a set of equations for a linear program to assign expression values to every remaining node. For every edge pair consisting of the source parent node u, and the target child node, v, we define the constraint

$$x_u + i_{uv} - d_{uv} = x_v$$

where x_u and x_v represent the expression value at parent u and child v, respectively, i_{uv} is the increase in expression between parent and child, and d_{uv} the decrease.

To allow flexibility in scoring penalties for increases and decreases in gene expression, we create constants C_i and C_d, respectively, which are able to affect the penalties for each type of change. We wish to minimize the amount of change between parent and daughter cells by minimizing the expression

$$C_i \sum_{(u,v)} i_{uv} + C_d \sum_{(u,v)} d_{uv}.$$

Since these constraints can yield multiple optimal solutions, we use the L2 norm to find the unique solution, which is the sample mean of all solutions. Therefore, we use a quadratic program solver and solve the minimization problem

$$C_i \sum_{(u,v)} \left(i_{uv} + \xi i_{uv}^2\right) + C_d \sum_{(u,v)} \left(d_{uv} + \xi d_{uv}^2\right),$$

where ξ is a small constant we set to 10^{-14}. This constant, ξ, is set to be arbitrarily small to ensure the scoring function still primarily minimizes the linear sum of changes. The small constant is included to add an additional, minimal, penalty to find the sample mean. We also chose to set the constants C_i and C_d to 1. To determine whether our results would be affected if we used another scoring function, we analyzed the data using sum of squares and obtained the same general results, indicating that the analysis is robust to type of scoring system used (Supplemental Experimental Procedures). To determine whether expression in the unannotated cells have a strong effect, we analyzed the data by assigning either a minimum or a maximum expression level to each un-annotated cell. We rederived the molecular differentiation map and found relatively little effect. As expected, regions of the lineage with few unannotated cells (e.g., P1 descendants) showed essentially no effect, and regions with a greater number of unannotated cells (AB descendants) showed a larger effect (Supplemental Experimental Procedures).

Molecular Differentiation Map

The results from the commitment algorithm were summed to determine the total amount of asymmetry at each cell division. For every nonleaf node included in the above described graph, p, that has two daughter cells u and v, then the asymmetry, a_{pg}, is defined as

$$a_{pg} = |x_u - x_v|$$

for a given gene, g. Therefore, the total asymmetry, a_p, is then set to

$$a_p = \sum_g \frac{a_{pg}}{\mu_{pg} + 500}$$

for the cell division at a given node, p, where μ_{pg} is the average of the commitment (calculated as described above) for the two daughters. The factor of 500 represents fluorescence noise.

Molecular Signature Heat Map

Pairwise asymmetry is computed between every pair of cells in the L1. For every cell pair (i, j), where $i \neq j$, the absolute difference between the observed expression values for those cells is computed

$$a_{ijg} = |o_{ig} - o_{ig}|,$$

where o_{ig} is the observed expression value for gene g in cell i. As with the molecular differentiation map, we divide this value by the average observed expression value of this pair of cells, μ_{ij}, and the baseline noise established at 500 units. By summing over all genes, we get the resulting asymmetry for this pair of L1 cells

$$a_{ij} = \sum_g \frac{a_{ijg}}{\mu_{ij} + 500}.$$

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, six figures, and seven tables and can be found with this article online at http://www.cell.com/supplemental/S0092-8674(09)01118-0.

ACKNOWLEDGMENTS

We thank Denis Dupuy and Marc Vidal for promoter DNA, Andy Fire for the *myo-3*:GFP strain, and Erik Jorgensen for the *dpy-30*:mCherry strain. We thank Robert Waterston for mCherry strains. We thank members of the Kim and Batzoglou labs for discussion. This work was funded by the National Institutes of Aging and the Howard Hughes Medical Foundation. X.L. was funded by the Larry L. Hillblom Foundation. S.J.A. is supported by a William R. Hewlett Stanford Graduate Fellowship and a National Science Foundation Fellowship.

Received: December 11, 2008
Revised: May 21, 2009
Accepted: August 18, 2009
Published: October 29, 2009

REFERENCES

Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development *132*, 1843–1854.

Bird, A.F., and Bird, J. (1991). The Structure of Nematodes (San Diego, CA: Academic Press).

Bowerman, B., Ingram, M.K., and Hunter, C.P. (1997). The maternal par genes and the segregation of cell fate specification activities in early Caenorhabditis elegans embryos. Development *124*, 3815–3826.

Broitman-Maduro, G., Lin, K.T., Hung, W.W., and Maduro, M.F. (2006). Specification of the C. elegans MS blastomere by the T-box factor TBX-35. Development *133*, 3097–3106.

Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., et al. (2004). A first version of the Caenorhabditis elegans promoterome. Genome Res. *14*, 2169–2175.

Dupuy, D., Bertin, N., Hidalgo, C.A., Venkatesan, K., Tu, D., Lee, D., Rosenberg, J., Svrzikapa, N., Blanc, A., Carnec, A., et al. (2007). Genome-scale analysis of in vivo spatiotemporal promoter activity in Caenorhabditis elegans. Nat. Biotechnol. *25*, 663–668.

Edgar, L.G., Carr, S., Wang, H., and Wood, W.B. (2001). Zygotic expression of the caudal homolog pal-1 is required for posterior patterning in Caenorhabditis elegans embryogenesis. Dev. Biol. *229*, 71–88.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Fowlkes, C.C., Hendriks, C.L., Keranen, S.V., Weber, G.H., Rubel, O., Huang, M.Y., Chatoor, S., DePace, A.H., Simirenko, L., Henriquez, C., et al. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell *133*, 364–374.

Good, K., Ciosk, R., Nance, J., Neves, A., Hill, R.J., and Priess, J.R. (2004). The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in C. elegans embryos. Development *131*, 1967–1978.

Horvitz, H.R., and Herskowitz, I. (1992). Mechanisms of asymmetric cell division: two Bs or not two Bs, that is the question. Cell *68*, 237–255.

Hunter, C.P., and Kenyon, C. (1996). Spatial and temporal controls target pal-1 blastomere-specification activity to a single blastomere lineage in C. elegans embryos. Cell *87*, 217–226.

Keller, P.J., Schmidt, A.D., Wittbrodt, J., and Stelzer, E.H. (2008). Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. Science *322*, 1065–1069.

Kimble, J., and Hirsh, D. (1979). The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. Dev. Biol. *70*, 396–417.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935.

Lecuyer, E., and Tomancak, P. (2008). Mapping the gene expression universe. Curr. Opin. Genet. Dev. *18*, 506–512.

Long, F., Peng, H., Liu, X., Kim, S.K., and Myers, E.W. (2009). A 3D Digital Atlas of C. elegans and Its Application to Single-Cell Analyses. Nat. Methods *6*, 667–672.

Maduro, M.F. (2006). Endomesoderm specification in Caenorhabditis elegans and other nematodes. Bioessays *28*, 1010–1022.

Mohler, W.A., Simske, J.S., Williams-Masson, E.M., Hardin, J.D., and White, J.G. (1998). Dynamics and ultrastructure of developmental cell fusions in the Caenorhabditis elegans hypodermis. Curr. Biol. *8*, 1087–1090.

Murray, J.I., Bao, Z., Boyle, T.J., Boeck, M.E., Mericle, B.L., Nicholas, T.J., Zhao, Z., Sandel, M.J., and Waterston, R.H. (2008). Automated analysis of embryonic gene expression with cellular resolution in C. elegans. Nat. Methods *5*, 703–709.

Peng, H., Long, F., Liu, X., Kim, S.K., and Myers, E.W. (2008). Straightening Caenorhabditis elegans images. Bioinformatics *24*, 234–242.

Peng, H., Long, F., and Myers, E.W. (2009). VANO: a volume-object image annotation system. Bioinformatics *25*, 695–697.

Potti, A., and Nevins, J.R. (2008). Utilization of genomic signatures to direct use of primary chemotherapy. Curr. Opin. Genet. Dev. *18*, 62–67.

Praitis, V., Casey, E., Collar, D., and Austin, J. (2001). Creation of low-copy integrated transgenic lines in Caenorhabditis elegans. Genetics *157*, 1217–1226.

Rodwell, G.E., Sonu, R., Zahn, J.M., Lund, J., Wilhelmy, J., Wang, L., Xiao, W., Mindrinos, M., Crane, E., Segal, E., et al. (2004). A transcriptional profile of aging in the human kidney. PLoS Biol. *2*, e427.

Ruvkun, G., and Giusto, J. (1989). The Caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. Nature *338*, 313–319.

Schnabel, R. (1994). Autonomy and nonautonomy in cell fate specification of muscle in the Caenorhabditis elegans embryo: a reciprocal induction. Science *263*, 1449–1452.

Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. Bioinformatics *18*, 207–208.

Sulston, J.E., and Horvitz, H.R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev. Biol. *56*, 110–156.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The Structure of the Nervous System of the Nematode Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B Biol. Sci. *314*, 1–340.