C. elegans exhibits coordinated oscillation in gene expression during development

Luke A. D. Hutchison[a], Bonnie Berger[*,a], Isaac Kohane[†,b]

[a]*Massachusetts Institute of Technology*
[b]*Harvard Medical School*

Abstract

The advent of automated cell lineaging and the ability to track gene expression at single-cell resolution *in vivo* in *C. elegans* has yielded a dramatically more complete view into developmental processes. We present a novel meta-analysis of the EPIC single-cell-resolution gene expression dataset of Bao, Murray, Waterston *et al.*, and show that a linear combination of the expression levels of the developmental genes in that dataset is strongly correlated with the wall-clock developmental age of the organism during early development, irrespective of the cell division rate of different cell lineages. We uncover a pattern of collective sinusoidal oscillation in gene expression, with multiple dominant sinusoidal frequencies, pointing to the existence of a global coordinated mechanism governing the timing of gene expression during development. We furthermore present a new method derived from Fisher's Discriminant Analysis that can be used to produce sinusoidal oscillations of any frequency and phase using just a linear weighting of the expression patterns of the genes in this dataset, strengthening the view that oscillatory timing mechanisms play an important role in development. The Fisher's Discriminant Analysis method also constitutes a generally useful tool for identifying the differential gene expression patterns that most strongly separate two distinct phenotypic or developmental traits.

Keywords: *C. elegans*, body plan, development, developmental clock, developmental age, transcription factor, transcriptional regulation

Introduction

Despite evidence that a global biological clock may govern the fate of cells and the timing of development, mechanisms regulating the timing of development that have been discovered so far appear to be localized, and a comprehensive control mechanism for global developmental timing has yet to be determined [1–10]. Recent work on single-cell gene expression has illuminated interesting regulatory processes at work during development [11–16], but thus far, there is comparatively little research into patterns of expression across multiple genes, in all cells of an organism, across the entire developmental timeline.

Single-cell-resolution 4D microscopy is enabling unprecedented study of how the developmental program unfolds [17–24]. A recent 4D confocal microscopy technique by Bao, Murray, Waterston *et al.*[21, 23–25] employs histone-mCherry reporters under the control of upstream promoters to detect expression levels of genes of interest, while employing fluorescent cell labeling to enable automated cell lineaging. This process yields a complete three-dimensional map of gene expression at single-cell resolution across the entire developmental timeline while simulta-

[*]bab@csail.mit.edu
[†]kohane@hms.harvard.edu

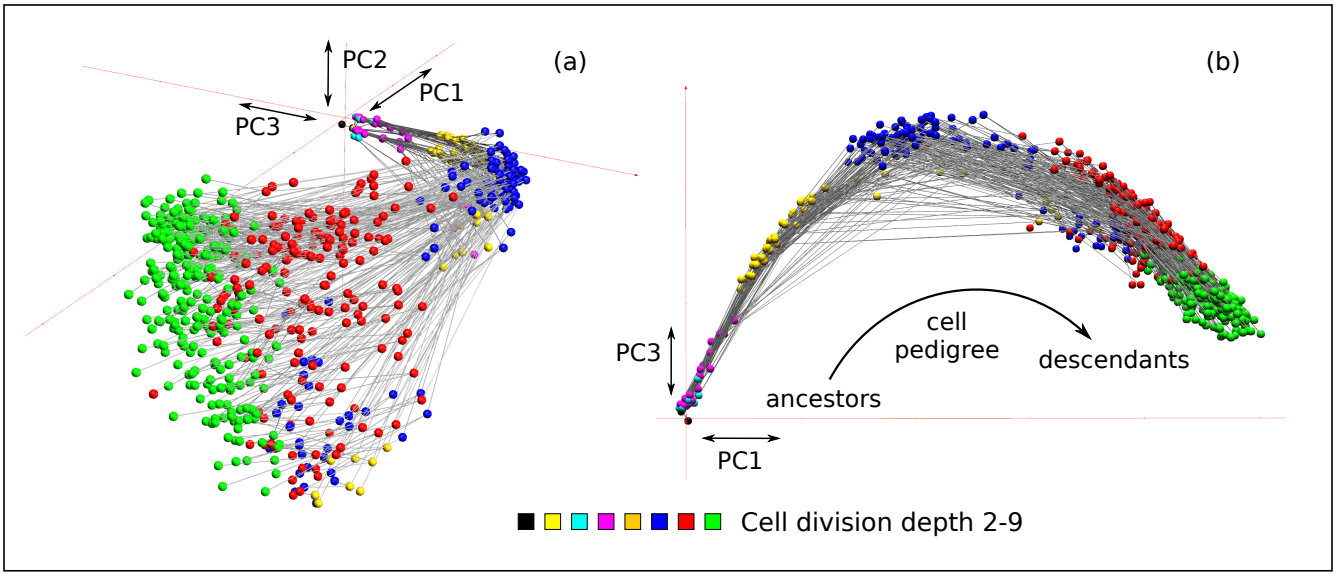

Figure 1: Projection of binarized gene expression profiles onto the first three principal component axes. Each node represents a cell, and the edges between the nodes connect a cell with each of its two daughter cells. The color of each node indicates the division depth in the cell pedigree. (a) A perspective view of the projection of the data onto the first three principal components PC1-PC3, showing that the cell pedigree is embedded in a curved manifold that sweeps through space in a direction primarily aligned with PC1 as development proceeds. (b) A top-down view of PC3 vs. PC1, showing the curved path of the manifold relative to the first principal component axis.

neously tracing the cell division pedigree, resulting in a 4D dataset, "the EPIC dataset", published at http://epic.gs.washington.edu/ .

We present a meta-analysis of a dense subset of the EPIC dataset. This subset consists of the expression levels of 102 developmentally-relevant genes across the first 686 cell identities in the *C. elegans* cell pedigree (i.e. all ancestral cell identities up to approximately the 350-cell stage) – in particular, we did not include genes with low coverage over these core cell identities, or cells beyond this point for which there was little gene expression data. Significant work was required to prepare the EPIC data for meta-analysis (see Methods), but producing the complete gene expression matrix for these genes and cells allowed us to apply novel analysis techniques to the data.

After applying principal component analysis to the gene expression data, we observed gene expression following a sweeping manifold shape through expression eigen-space as the developmental timeline proceeded (Figure 1). We discovered a strong linear correlation ($R^2 = 0.94$) between the first principal component of the gene expression profiles in the EPIC data over a span of the developmental timeline of the worm during cell proliferation (Figure 2). Over the entire available timeline, we observed multiple apparent sinusoidal oscillations in gene expression (Figure 3), with different frequencies of oscillation manifest in different principal components, indicating that the period of linear monotonic correlation between gene expression and wall-clock time may also have been an observation of the nearly linear part of a sinusoidal oscillation around the zero-crossing of the sinusoid graph.

We also present a technique based on Fisher's Discriminant Analysis for uncovering the relative contributions of genes to an attribute of interest. This technique does not appear to yet be in wide usage in developmental biology, however it is a powerful and simple mechanism to determine which genes are most strongly differentially expressed between cells of two different phenotypes or lineages, yielding a set of gene weights that maximizes inter-class variance while simultaneously minimizing intra-class variance (Figure 4). We used this technique to establish that

simple linear weightings of the expression levels of genes of interest were able to produce sinusoidal oscillations of any desired phase or frequency (Figure 5), suggesting that oscillatory mechanisms may be used extensively by developmental processes.

The presence of whole-organism sinusoidal and/or linear trends in gene expression would suggest the existence of a global, coordinated mechanism regulating developmental timing, in the form of a monotonic clock, and/or one or more global oscillatory mechanisms. This global timing mechanism may either involve the genes directly observed in this dataset, or may be due to an unobserved mechanism acting on these genes.

The results presented here suggest numerous opportunities for further research into the specific mechanism or mechanisms controlling the observed gene expression time-correlations and oscillations.

Results

Most variance in gene expression is characterized by the first 3-10 principal components

To examine the contribution of the principal axes towards overall dataset variance, we produced a *scree plot* (Figure S1) from the binarized gene expression matrix (Table S1), showing the eigenvalues of gene expression sorted into decreasing order. Most of the variance in the expression patterns of the 102 genes is embodied in the first 10 principal components, and in particular by the first three.

The cell pedigree monotonically sweeps a curved manifold through gene expression space

To understand how patterns in gene expression changed in relation to cell division, we produced a novel three-dimensional visualization of the cell pedigree directly overlaid on the first three principal components of gene expression (Figure 1). In this plot, nodes represent the 686 unique cells in the dataset (specifically, the specific identities of cells between cell division events), and edges indicate the relationship between a cell and its two daughter cells. The color of a cell in this plot indicates the cell division depth. The position of a cell in the three-dimensional space is the projection of that cell's gene expression profile (the binarized expression levels of the 102 genes for the cell) onto the first three principal components, i.e. the cell's position in this 3D plot is a simple linear combination of the activity levels of the cell's genes. By definition, the first principal component (PC1) is aligned with the axis of greatest variance in gene expression, the second principal component (PC2) is aligned with the axis of second greatest variance in gene expression orthogonal to PC1, and the third principal component (PC3) is aligned with the axis of third greatest variance in gene expression orthogonal to both PC1 and PC2.

Because the edges that connect each non-leaf cell to its two daughter cells clearly show the cell pedigree, trends in gene expression can be clearly seen as development proceeds. Remarkably, the cell pedigree sweeps across a curved manifold surface embedded in the three-dimensional "eigengene" expression space. The sweep direction of the cell pedigree across the manifold is monotonic, in the sense that pedigree edges between cells and their daughter cells all follow the same approximate sweep direction; there are no pedigree edges directed opposite to this general sweep direction after the first two or three cell divisions.

The strongest vector component of this sweeping manifold path is aligned with the first principal component of gene expression (PC1), indicating that movement along the manifold in the direction of the pedigree is correlated with the most significant orthogonal direction of variance in gene expression.

The strong, monotonically smooth manifold shape collectively swept out by the expression patterns of these genes suggests a strong and tightly coordinated global gene expression control mechanism. Since a large number of genes are collectively involved in sweeping this broad path through gene expression space (as seen in

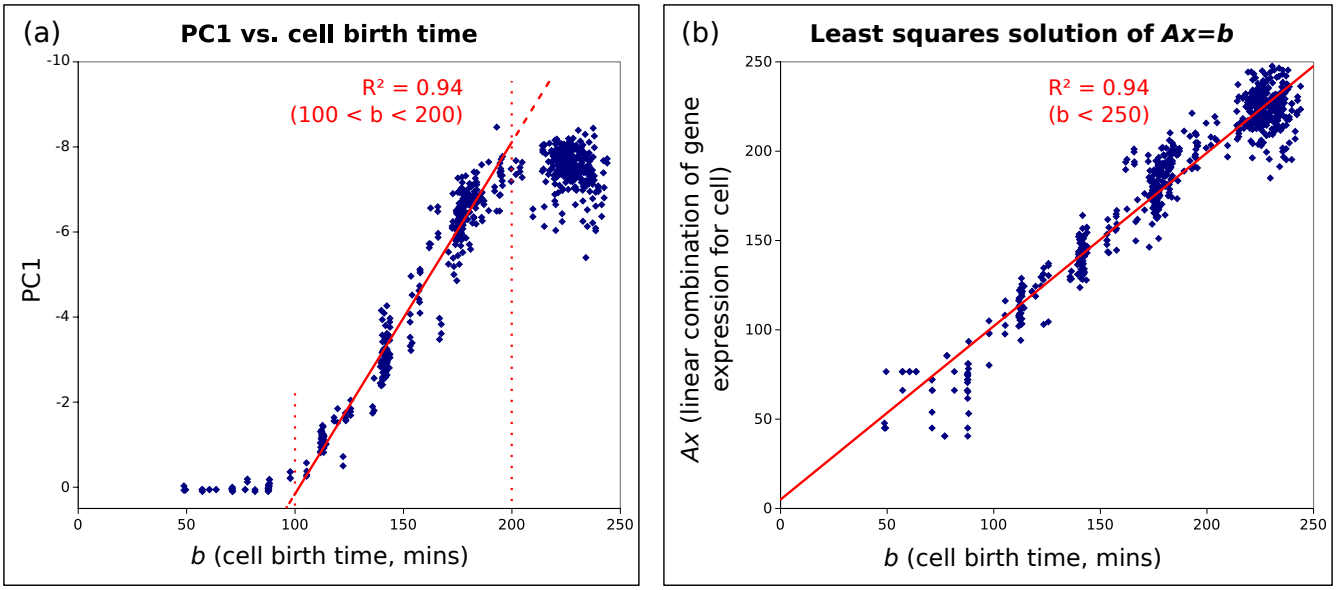

Figure 2: (a) the projection of gene expression onto the first principal component, PC1, vs. b, the cell birth time (i.e. the cell onset time) in minutes. The plot is strongly linear from 100 to 200 minutes. (b) After solving the linear equation $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is the binarized gene expression matrix and $\boldsymbol{b}$ is the vector of cell birth times, this plot shows the components of $\boldsymbol{Ax}$ (the best-fit linear estimators of each component of $\boldsymbol{b}$) against the corresponding component of $\boldsymbol{b}$. The model is linearized across the entire developmental timeline, using a fixed linear combination of gene weights.

the principal component weightings, Figure S4), this could imply that the developmental program is redundantly encoded across many genes for robustness.

As cells divide during development, and as gene expression trends in the direction of PC1, the cloud of cells at a given cell division depth (indicated by a given node color) expands in width along PC2 relative to the previous cell generation. This behavior indicates a general diversification in gene expression patterns as development proceeds, consistent with lineage-specific differentiation. However, the total spread in PC2 is less than half the distance swept through PC1 as development proceeds, suggesting that variation in expression levels of genes in this dataset is more strongly associated with the progress of development than the details of tissue-specific differentiation, since PC1 is most strongly aligned with the developmental timeline.

In gene expression research, in order to gain insights into the functioning of genes, cells are often clustered according to similarity in gene expression profile (e.g. [20]). However, Figure 1 illustrates a possible caveat to understanding these clustering results: cells with the most similar gene expression profiles may be more strongly *temporally* related than they are *functionally* related. In other words, the greatest sources of differential gene expression over the developmental timeline may be due to large-scale developmental coordination, rather than due to tissue-specific expression arising from differentiation. This was at least the case for the this set of *C. elegans* genes we studied, where the first principal component of gene expression was strongly temporally-correlated, and where non-temporally-correlated diversification in gene expression was primarily observed only in the third and subsequent principal components. Note, however, that in private correspondence, Murray [23] pointed out that the genes selected for analysis in the EPIC dataset may have been biased for temporal patterns, as opposed to spatial patterns. Therefore, a larger study, involving a

wider array of genes, would be needed to determine whether or not the observed effect was due to the selection of genes.

Gene expression is correlated with developmental age of the organism

Remarkably, since each cell pedigree edge from a cell to its two daughter cells follows the arrow of time, and since the cell pedigree sweeps a "monotonic" manifold shape through gene expression space as development proceeds, the expression patterns of the selected genes must be related to the developmental age of the organism, or at least to the cell division depth. Figure 2(a) shows PC1, the first principal component of gene expression, plotted against b, the birth time of each cell in minutes since fertilization. For much of the recorded development time, a striking linear correlation is evident ($R^2 = 0.94$ from 100-200 minutes). This correlation is notable, because projection of the data onto the PC1 axis represents a simple weighted linear combination of the binarized gene expression levels, which indicates there is a weighting of the gene expression levels that is directly predictive of the wall clock developmental age in minutes. This linear weighting can be obtained directly from the eigenvector loadings for the first principal component (i.e. the eigenvectors multiplied by the square root of the corresponding eigenvalue – Figure S4; Table S2). This correlation appears independent of the cell division depth of each of the cells present at a given developmental age, in agreement with the findings in Nair *et al.* [10] that expression and proliferation are independently entrained to separate clock-like processes.

Before 100 minutes and after 200 minutes, however, PC1 diverges from being linearly correlated with the cell birth age b (Figure 2(a)). It is possible gene expression is linearizable across the entire developmental timeline, but that PCA does not recover the optimal set of gene weights to expose the direct linear correlation, and for this purpose we consider the means to fit a linear model to the data below. Another possible explanation for the observed distribution is that the underlying gene expression is fundamentally sinusoidal, not linear – this possibility is discussed in the next section. It is also possible that entirely different developmental programs or processes are active before 100 minutes, between 100 and 200 minutes, and after 200 minutes: one plausible explanation for the difference in gene activity before and after 100 minutes is the maternal-to-zygotic transition (MZT) [26].

To find the optimal function mapping gene expression to cell birth time, we could try using supervised learning with randomized k-fold cross validation to learn a regression function. However, this would tend to overfit the data and underestimate the validation error, because so many cells share similar gene expression profiles (meaning that there would be "sample pollution" between the training set and the test set). Instead, we expressed the relationship between gene expression and cell birth time as a linear system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is the binarized gene expression matrix (i.e. Table S1, with cells in the rows, and genes in the columns) and $\boldsymbol{b}$ is the vector of birth times for each cell. We then solved the system for $\boldsymbol{x}$, a set of gene weights that map the expression matrix onto cell birth times. (A small amount of random noise was added to the binarized gene expression levels for regularization.) The resulting vector of gene expression weights $\boldsymbol{x}$ (Table S5) gives us the linear weighting of genes that maps the gene expression profile of a cell onto the developmental age of the cell with the minimum squared error.

Given this set of weights, gene activity was close to linearly correlated with cell birth time for all cells in the pedigree (not just from 100 to 200 minutes), retaining approximately the same linear correlation strength of $R^2 = 0.94$, but across the entire developmental timeline (Figure 2(b)). If PC1 is evidence of a linear correlation with developmental age, rather than a sinusoidal correlation, then the genes with high-magnitude positive or negative weights in this table would be indicated as important in the timing of development. The gene with the largest negative weight in Table S5, *med-2*, is necessary for endoderm specification [27]. Other genes with high negative weights include *sdz-*

28, an *SKN1*-dependent zygotic transcript only active in early development, triggered by the *SKN1* maternally-deposited transcription factor, and *glp-1*, which encodes a transmembrane protein essential for mitotic proliferation of germ cells and maintenance of germline stem cells, and important in many differentiation decisions in somatic tissues. Genes with strong positive weights include *egl-5*, a Hox gene [28], as well as a number of genes that are expressed nearly ubiquitously across the entire developmental timeline.

Table S6 lists, for each cell, the cell birth time, the PC1 projection of gene expression for the cell, and the linearized gene expression for the cell, and compares these with the Sulston onset time of the cell.

Sinusoidal oscillation observed in the principal components of gene expression

Figure 1 shows that gene expression in the third principal component, PC3, traces half a sinusoidal cycle with respect to PC1. The Waterston technique [21, 23, 24] is not able to reliably track development once the organism begins to move, so the end of the current dataset is not the end of the developmental timeline. It is unclear whether a complete sinusoid would be traced in PC3 if more of the developmental timeline were captured – in other words, it is not clear whether the semi-sinusoidal oscillation in PC3 relative to PC1 is in fact half of a sinusoidal oscillation. Figure 3(a) shows all pairings of principal components between PC1 and PC10 as a mechanism of visualizing the first ten principal component dimensions in 2D, and interestingly, PC6 traces a complete sinusoid with respect to PC1 across the same time period that PC3 traces half a sinusoid. If PC6 is paired with PC3, the half sinusoid paired with the complete sinusoid causes the developmental path to trace a spiraling alpha shape (α), as shown enlarged in Figure 3(b). The presence of a complete and clear sinusoidal oscillation in PC6 relative to PC1 lends strength to the hypothesis that the semisinusoidal curve of PC3 relative to PC1 is in fact half of a sinusoidal oscillation.

The semisinusoidal oscillation observed in PC3 and the sinusoidal oscillation observed in PC6 may also explain the nonlinearity observed in plotting PC1 against developmental age (Figure 2(a)) before 100 minutes and after 200 minutes – the plot overall exhibits a sigmoidal shape. This suggests that in fact PC1 may be oscillating sinusoidally over the developmental timeline, and that the observed linearizable portion of PC1 between 100 and 200 minutes is in fact only linear because a sinusoidal wave is also nearly linear around its zero-crossing point. Given more data about later stages of development, sinusoidal oscillation may be exhibited for all three of PC1, PC3 and PC6, with periods of roughly 500 minutes for both PC1 and PC3, and 250 minutes for PC6. If these oscillations are in fact sinusoidal, then given that the phases of the apparent oscillations in PC1 and PC3 are offset by $\pi/2$ from each other, and that they have similar periods, the plot of PC1 vs. PC3 (Figure 3(b)) should trace a roughly circular motion.

Circular oscillatory paths in gene expression have been previously observed with PCA dimension reduction on whole-organism RNA-Seq profiles in frog, mosquito, fly, and zebrafish by Anavy *et al.*, who applied the Traveling Salesman algorithm to a series of RNA-Seq profiles to arrange the samples into approximate order of developmental age, finding a minimum-distance simple path between all samples [29]. We show that the observed sinusoidal trend in gene expression occurs not just as an aggregate, whole-organism measure, but within each individual cell in the organism. This is significant, as it lends credence to the presence of a global clock mechanism coordinating development.

Discovering ubiquitous large-scale sinusoidal oscillations in gene expression would not be unexpected, as oscillations have been previously observed in RNA-Seq profiles of whole organisms [29], as well as in the expression levels of individual mRNAs over time [30, 26]. Both oscillatory and temporally graded activity has been observed in transcript levels [31]. Some key regulators of the timing of heterochronic miRNA expression have been discovered, including *lin-4*

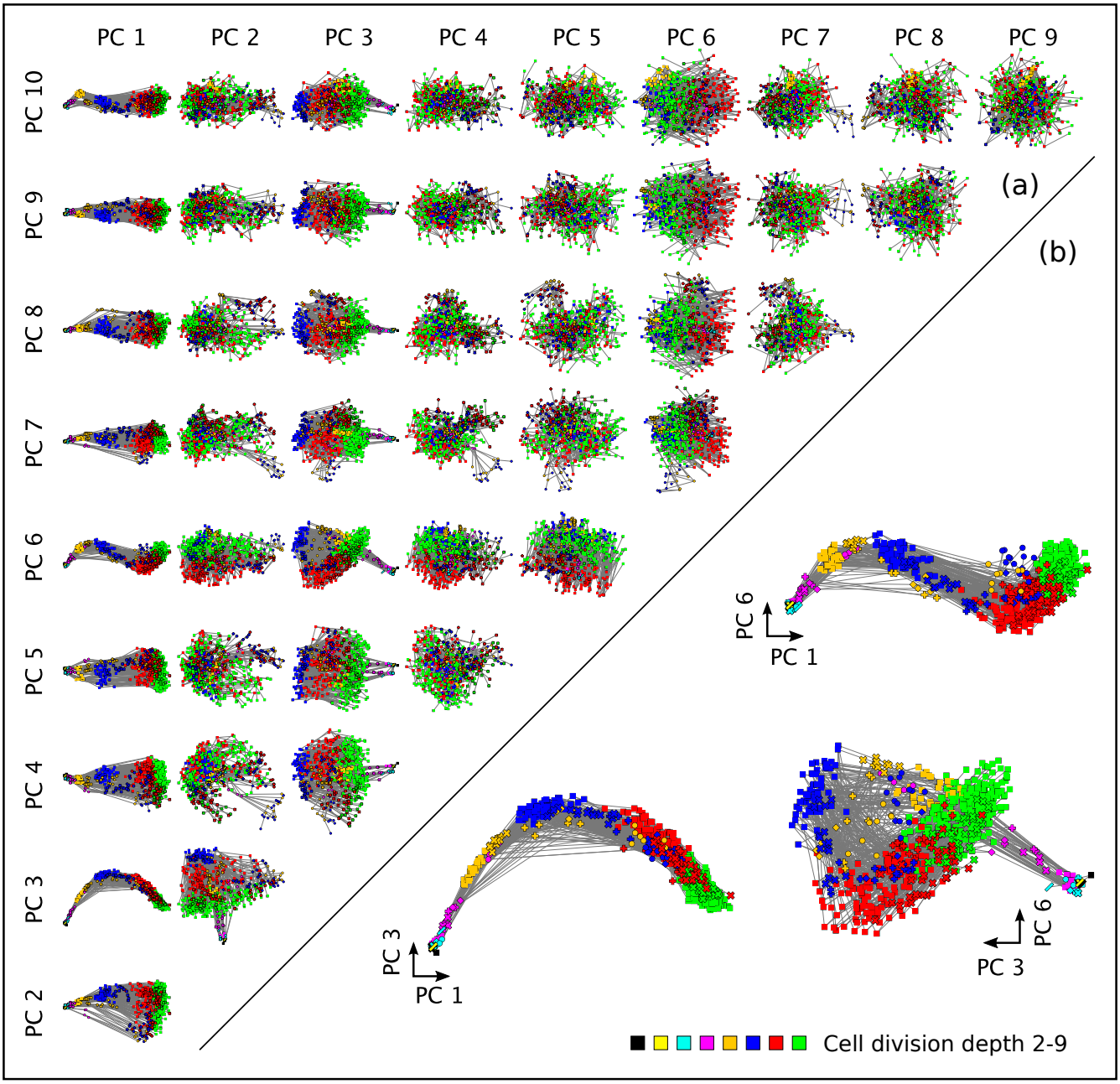

Figure 3: (a) The projection of gene expression onto the first 10 principal component axes, PC1-PC10, shown as 2-dimensional projections: the row and column labels indicate the pairing of two component axes that produces a given projection. (b) Expanded views of PC3 vs. PC1, PC6 vs. PC1, and PC6 vs. PC3. Gene expression appears to oscillate sinusoidally in PC3 and PC6, with approximately double the sinusoidal frequency in PC6 (although insufficient data was available to determine if either PC3 or PC6 would continue a sinusoidal path later in development). Plotting PC6 against PC3 causes the cell pedigree to trace an "alpha"-shaped path (α) through gene expression space as development proceeds.

and *let-7*-family miRNAs, under control of *lin-42* [32], however data on these miRNAs was not available in the EPIC dataset.

Robust oscillation in transcription at multiple temporal scales is consistent with the results of Hendricks [30] and Kim [31], both documenting ultraradian cycles involving approximately 1/6 of the transcriptome with changes in expression

of up to an order of magnitude during the cycles. The authors identify several periodic developmental phenomena, such as cuticle development and cuticle molting, and speculate that the timing of other developmental processes are similarly controlled by one or more of these transcriptional cycles.

What is unique about the findings presented here is that the pattern of oscillation in gene expression are not limited to specific cell types, or to an averaged pattern of expression across the entire organism, but rather we have shown that the apparent oscillations occur separately and simultaneously in each individual cell, as part of a globally synchronous oscillatory pattern.

Principal components other than PC1, PC3 and PC6 did not exhibit strong linear or sinusoidal structure, but may capture variation in gene expression due to tissue-specific variation. The widening of the point cloud in all principal components other than PC3 and PC6 when plotted against PC1 is consistent with cells differentiating as development proceeds. See for example the widening in PC2 vs. PC1 in Figure 1.

Fisher's Discriminant Analysis used to identify lineage-specific genes

It is possible to determine a linear combination of gene expression that maximizes the separation between a group of cells with any specific trait (class 0) and all other cells (class 1), using Linear Discriminant Analysis (LDA). Specifically, we used Fisher's Discriminant Analysis (FDA), which finds a hyperplane decision boundary that maximizes inter-class variance while minimizing intra-class variance in the one-dimensional projection of class 0 and class 1 onto $\boldsymbol{\omega}$, the normal vector to the hyperplane (see Methods). Using the FDA method, we obtained gene weightings that maximally separate each major lineage in the *C. elegans* pedigree (AB, MS, C, D and E) from cells in other lineages. In Figure 4, the FDA projection maximally separating the gene expression patterns of cells in one lineage against all other cells is used to select the y-axis value, and the birth time of each cell is used to select the x-axis value for each cell. The clear separability of each cells in each lineage from the remainder of the cells indicates that gene expression patterns in each lineage are distinct. Table S4 lists the gene weights required to produce maximum class separation in the projection onto $\boldsymbol{\omega}$ for each lineage. Genes with weights that are more strongly negative are generally expressed in the opposite class from genes with weights that are more strongly positive, allowing for the identification of strongly lineage-correlated genes.

Note that the FDA method can be used to identify gene weightings for any binary trait, or any separation of two traits. Given a gene expression matrix consisting of positive values for expression and zero values for non-expression of genes, the strongly positively weighted genes produced by FDA analysis are more strongly correlated with the first of the two binary classes, and the strongly negatively weighted genes are more strongly correlated with the second of the two binary classes. The weights recovered by FDA not only yield maximal inter-class separation, but also maximal intra-class compactness. FDA analysis could be a powerful tool for understanding differential gene expression, but this method does not yet appear to be widely used for this purpose.

Of note, *elt-1*, which has been identified as a master regulator of epidermis specification [33], is strongly weighted in separating the AB epidermal lineage from the rest of the cells, but not as strongly weighted in separating the C epidermal lineage from the rest of the cells (Table S4). This is consistent with the observation that these two lineages rely on different developmental regulators [33]. Also, several factors are weighted more highly than *elt-1* in separating the AB lineage from the rest of the cells (*sdz-38*, *tlp-1*, *hlh-26*, *hlh-3*, *pax-3*), and several factors are weighted similarly highly in separating the C lineage from the rest of the cells (*nob-1*, *cwn-1*, *C25D7.10*, *tbx-9*, *rad-26*). Some of these genes are known to play a central role in epidermal development, including *nob-1* [34], *pax-3* [35], and *tbx-9* [36], confirming the utility of this FDA method for

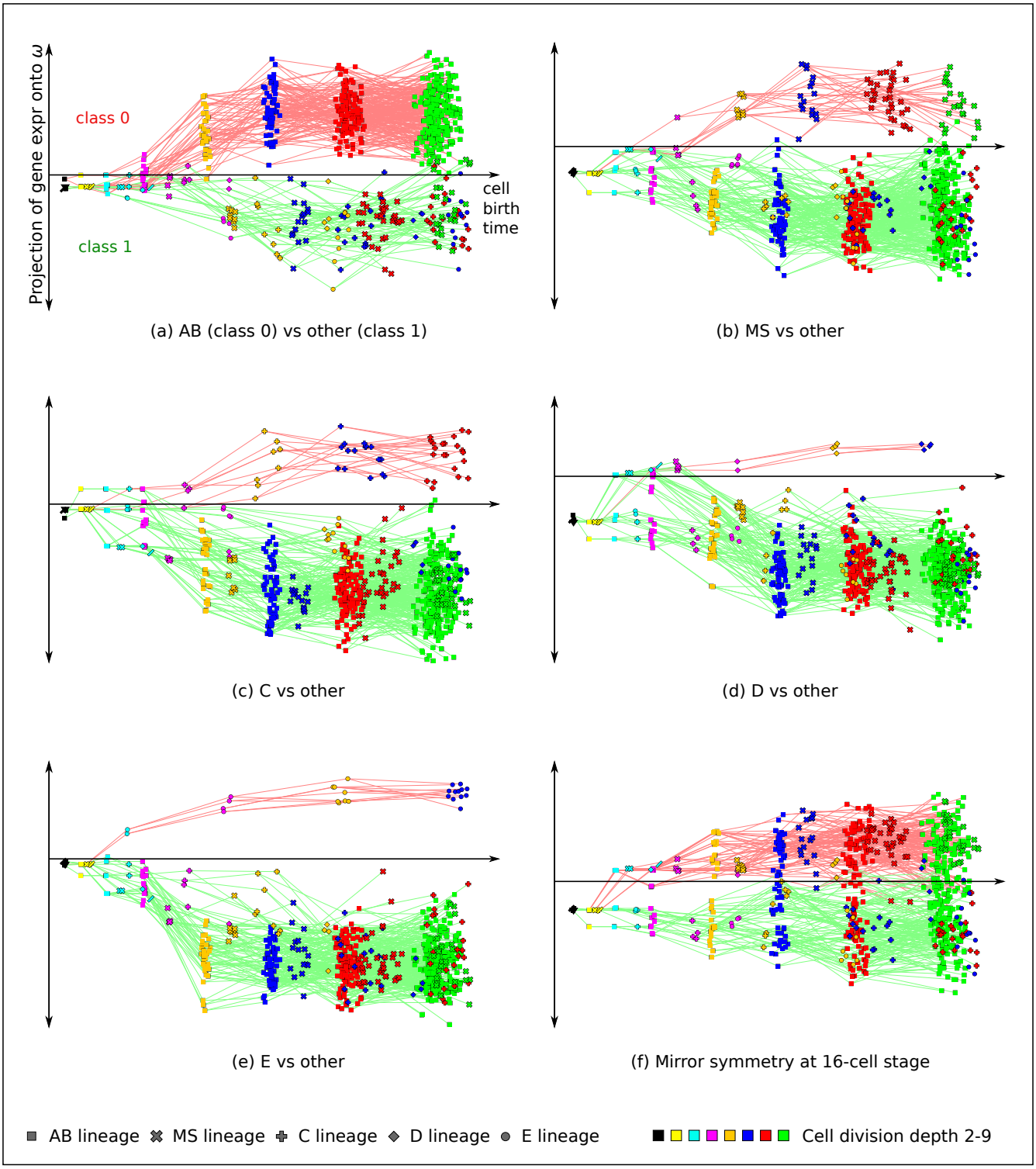

Figure 4: (a)-(e) Fisher's Discriminant Analysis (FDA) is used to identify the key genes that are maximally differentially expressed in each major cell lineage (AB, MS, C, D, E) relative to the rest of the cell lineages. Table S4 gives the weightings of genes that produce these maximal separations between cell lineages. In (f), the same method is used to find gene expression differences between the approximate bi-fold (mirror) symmetry in the structure of the cell pedigree that arises at the 16-cell stage.

identifying genes relevant to developmental processes. However, the other highly-weighted factors for the AB and C lineages do not appear to have yet been closely studied in relation to their broader role in epidermal development.

In Figure 4(f) and Table Table S4, we also found strong separability between the two halves of the cell pedigree that exhibit near-mirror symmetry at the 16 cell stage: (1) the descendants of P0.pppa (i.e. D), P0.pppp, Z.paap (MS.p), P0.paaa (MS.a), P0.aalp (AB.alp), P0.aala (AB.ala), P0.aara (AB.ara) and P0.aarp (AB.arp), where P0 is the zygote; and (2) the descendants of P0.papa (E.a), P0.papp (E.p), P0.ppap (C.p), P0.ppaa (C.a), P0.aplp, P0.apla, P0.aprp and P0.apra. The strong linear separability of these two mirror-symmetric lineage groups, and the weightings of the involved genes, may provide useful information about one genetic basis for left/right and anterior/posterior body plan layout.

Identification of gene weights producing arbitrary sinusoidal oscillations

To determine whether other sinusoidal gene activation patterns were evident in the data, we generated sinusoidal waves with a range of frequencies and phase offsets (Figure 5(a)), then labeled the cells in the pedigree as being in class 0 or 1 depending on whether the cell birth time fell within a trough or peak of the sinusoid (Figure 5(b)). Fisher's Discriminant Analysis was used to find the hyperplane that separates class 0 and class 1 with minimum intra-class variance and maximum inter-class variance. The projection of gene expression onto the normal vector of the hyperplane (w) was plotted against cell birth time (Figure 5(c)) to produce an approximate fit of gene expression to the original sinusoidal wave. (This is technically the best square-tooth approximation of the data to the original sinusoidal wave, because FDA simply tries to make the two classes as compact and as separated from each other as possible.) Across a range of different phase and frequency values, a reasonable approximation of a sinusoid was obtained (Figure 54(d)), indicating that oscillatory behavior of almost any required frequency or phase can be obtained relative to the timeline as a simple fixed linear combination of the expression levels of these genes. Figure 5(e) shows the FDA weights for each gene that are required to achieve maximum class separation at a given frequency and phase. The largest positive and negative FDA weights in each gene's heatmap correspond to the gene's largest contributions to separating class 0 and 1 at a given frequency and phase.

Whether or not this mechanism is directly used in the biology of the organism, it is quite remarkable that a simple fixed linear combination of gene expression levels can be used to produce a sinusoidal oscillatory signal of any frequency of phase as a function of the developmental age of the organism.

As remarked previously, the sorted PC1 component weights approximately correspond to a time ordering of gene activation during development. However, Figure 5(e) indicates that time ordering is not the only factor in play; otherwise, all FDA weightings would all look roughly similar, except for being phase-offset relative to each other (i.e. vertically offset, but wrapped).

Analysis of transpose PCA

For reference, Table S7 gives the PCA of the transpose of the gene expression matrix, which shows the result of dimension reduction in the cell dimensions rather than the gene dimensions. The distance between any pair of genes in this transpose-PCA space, calculated using the Euclidean norm on the top k principal components (e.g. for $k = 2$), gives a measure of the similarity of expression patterns for the pair of genes. For example, *sdc-2*, *hmg-11*, *F28C6.1*, *tbx-11*, *B0310.2* and *ceh-26* are close together in transpose-PCA space (their first two transpose PCA coordinates are very similar), and the gene expression patterns for those genes are similar, as shown in Figure S3, whereas *egl-5* is distant from all other genes in transpose-PCA space, and has the most unique gene expression profile, also shown in Figure S3.

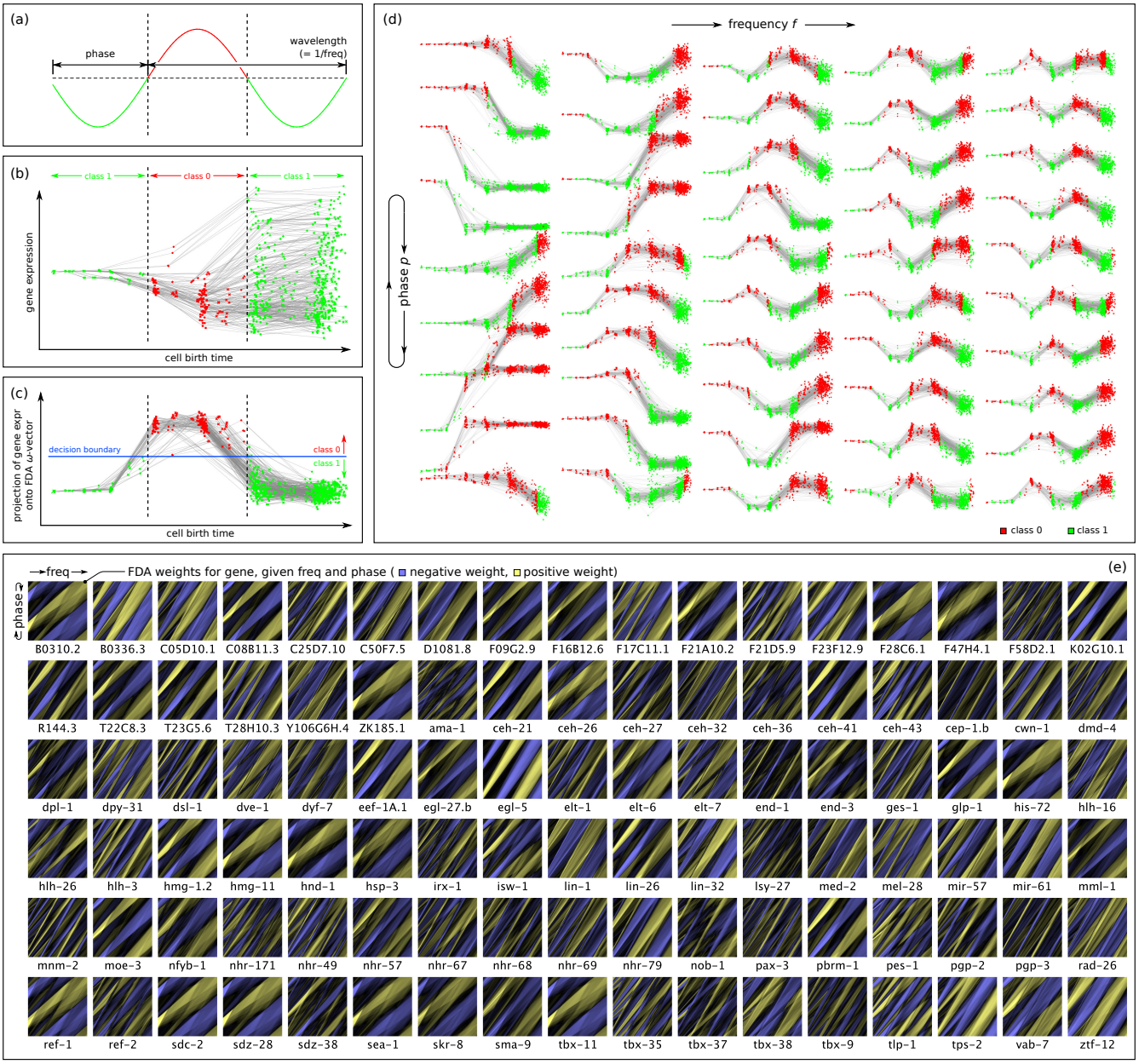

Figure 5: Identification of simple linear weightings of gene expression levels that can produce oscillations across a range of sinusoidal frequencies and phases. (a) A target sine wave is generated. (b) Cells are assigned to class 0, at developmental times when the sine wave is positive, or class 1, at times when the sine wave is negative. (c) Fisher's Discriminant Analysis (FDA) is used to maximally separate class 0 from class 1 in the vertical axis, minimizing intra-class variance and maximizing inter-class variance, producing a best-fit square wave approximation of the target sine wave. (d) The best-fit FDA results are plotted across a range of phases in the rows and frequencies in the columns, with phase wrapping vertically ($0 \to \pi \to 0$), and with frequency increasing across the columns. (e) A heatmap of FDA weight given phase and frequency for each gene, with the largest negative weight for the gene in blue, and the largest positive weight for the gene in yellow.

Discussion

Effects of gene selection and reporter mechanism on results

Gene activity data was collected by Murray *et al.* using the following criteria and methodology: "We identified a list of transcription factors and other regulatory proteins for which prior microarray or phenotype data suggested embryonic function and targeted these for expression anal-

ysis. For these, we constructed stable *C. elegans* strains expressing a histone-mCherry reporter under the control of the gene's upstream intergenic sequences. We analyzed expression of reporter strains whose expression begins before the last round of embryonic cleavage (the 350-cell stage) by crossing in a ubiquitous histone-GFP marker, collecting three-dimensional confocal time-lapse movies, and tracing the cell lineage as described previously." [23]

In private correspondence, Murray suggested the use of histone-mCherry reporters may impact analysis, because these reporters tend to persist and even increase in the descendants of expressing cells, even if the endogenous gene is degraded (the reporter mRNA has a "stable" *let-858* 3' UTR, and the histone-mCherry itself has a halflife that is likely to be longer than the length of embryogenesis). It is unclear what the effect of gene selection and reporter mechanism may be on our results, and further work is needed to determine whether other gene sets and/or different reporter methods for obtaining single-cell-resolution gene expression data exhibit the same properties.

Cylindrical projection of gene expression manifold

The manifold swept by the cell pedigree through the space spanned by the first three principal component axes was roughly semicylindrical. We flattened out this "principal manifold" of the data [37] using a cylindrical projection. This involved radially projecting cell positions in principal component space outwards from a central axis onto the surface of a cylinder, and then flattening out the cylinder (Figure S2). Table S6 gives the two dimensional coordinates (θ, y) of the cells in the cylindrical projection.

This cylindrical projection can be used to visualize the cell pedigree in two dimensions rather than three, while eliminating most problems with occlusion and perspective distortion. We plotted the binarized expression data for each gene using the cylindrical projection in Figure S3, so that large-scale patterns of gene expression can be examined across the developmental timeline, and across the surface of the principal manifold traced through the first three principal components.

Examination of principal component weights

The principal component weights (i.e. the gene weights that project the gene expression data onto the principal component axes) indicate which genes are the strongest sources of variance in a given principal component axis (Table S2(b); Figure S4). If variance in a principal component axis is only due to a small number of genes, the weights corresponding to those genes will be large in positive or negative magnitude, and the other genes will be close to zero. However, the distribution of PC1 weights in Figure S4 demonstrates that the contribution towards variance is close to zero for very few of the 102 genes in this dataset, indicating that most or all of the 102 genes under consideration are involved in establishing the linear correlation with developmental age. If the time-correlated nature of PC1 is indeed due to a global developmental clock mechanism, then the fact that many genes are involved in this mechanism could indicate a redundancy in the temporal functioning of these genes, affording the opportunity for adaptation in the function of developmental regulators without disrupting the global developmental clock. Redundancy adds resilience and flexibility, giving a system more degrees of freedom over which to adapt.

However, comparing the sorted weights in Figure S4 to the gene expression patterns in Figure S3, it can be seen that the strongest negative weight (*egl-5*) corresponds to gene expression in all cells excluding the last generation measured, whereas the strongest positive weights (*lin-26*, *K02G10.1*, *ceh-41*, etc.) correspond to high levels of gene expression commencing later during development. Consequently, the sorted PC1 weights roughly correspond to a time ordering of gene activation. It is possible then that the apparent time-correlatedness of PC1 is due to a sequential pattern of gene activation. However, cross-comparison with FigureS3 suggests

that the situation is not as simple as the genes being switched on in a specific ordered sequence.

Methods

Dataset preprocessing

We obtained the Waterston EPIC dataset from http://epic.gs.washington.edu/ . This dataset comprises image data (obtained using 4D confocal microscopy of developing *C. elegans* embryos) as well as gene expression data for 127 developmentally related genes at single-cell resolution up to the point at which the embryo gains motor control.

We parsed the available data, discarding genes and cell pedigree subtrees with significant numbers of missing values (i.e. where gene expression had not been recorded for significant numbers of cells). For genes that were run multiple times, we averaged the expression levels across the runs.

We took the maximum expression level of each gene across the lifetime of each cell, using the "global" intensity measurement method (out of "global", "local", "blot" and "cross", as described in the original paper), and binarized gene expression to 0 or 1 using a reporter intensity threshold of 2500. This threshold value was conservatively chosen based on Murray *et al.* [23], where it was stated that spurious gene activity was not observed below a measured reporter intensity value of 2000, and that strong expression signals were observed at values over 4500. We discarded a number of additional genes that were expressed in all (or nearly all) cells, as well as genes that were not expressed in any (or almost any) cells given this intensity threshold.

The resulting binarized gene expression matrix (Table S1) consists of the binarized gene expression values (0 or 1) for 102 genes, measured in 686 cells. The 686 cells can be broken down into 341 internal nodes in the cell pedigree (cells that divide within the measured developmental timeline) and 345 leaf cells (cells that are either terminally-differentiated, die through apoptosis, or divide later than the end of the recorded timeline).

Timescale correction

We extracted cell birth times from the dataset by finding the time point for each cell at which reporter intensity level data first became available (these times were used as the x-axis for the linear regression in Figure 2). Despite the claim in [23] that the EPIC data was sampled "with ~1-min temporal resolution", the raw data was actually sampled at different time scales for each gene, with time scale factors including at least 1.0, 1.35, 1.4, 1.5 and 2.0 minutes per 3D scan, and with the datasets listing only scan indices, not timestamps. There is no available data source on the EPIC website that indicates the time scale for a given run, and in private correspondence, the original authors were not able to easily retrieve the timescales used for each run. Therefore, to get all data on the same timescale, we had to perform some slightly tricky analysis to recover a best-fit time scale factor for each run. We used a custom multiple-alignment regression technique to warp the gene expression timepoints for each run to a consensus cell pedigree. In this process, we also discovered that not only a multiplicative offset was needed to scale the data indices to fit the timeline of the canonical Sulston pedigree, but there was also an additive offset averaging approximately 45 minutes between the Sulston time zero and the dataset index zero (i.e. the sample indices in the raw datafiles are zero-indexed, but the sampling started at a developmental age of approximately 45 minutes). Cell birth times, appropriately scaled and offset as described above, were used for the linear correlations in Figure 2.

Note that despite our best efforts to align these datasets, a small degree of time-spreading has probably been unavoidably been introduced into our cell birth time predictions, due to the fact that the raw data was not properly timestamped. Properly timestamping future scans using minutes since time of fertilization (rather than sample index) would increase the strength of linear correlation (R^2) between developmental age and gene expression.

The EPIC data includes multiple runs for some genes, sometimes with nontrivial variation

in gene expression between runs. After adjusting for the unspecified time dilation as described above, we combined data from these different experiment repetitions by averaging, across all runs for a gene, the maximum reporter level achieved by the gene during the lifetime of each cell.

Principal component analysis

Principal component analysis was run on the binarized gene expression matrix, producing Table S2(a) and Table S2(b), the PCA eigenvalues and eigenvectors respectively. We projected the binarized gene expression matrix onto the eigenvectors, producing Table S3.

The figures in this paper were produced using our own custom 2D and 3D visualization software. Columns 1-3 of Table S3 were plotted in a 3D view to yield Figure 1. All pairings of principal component axes for PC1-PC10 were plotted as 2D projections, yielding Figure 3.

Cylindrical projection

The cylindrical projection of gene expression in Figures S1 and S2 were produced by identifying a central axis in the 3D-embedded two-dimensional manifold in which lies the cell pedigree, as shown in Figure 1. This central axis was chosen such that cells in Figure 1 were all approximately equidistant from this axis. The rotation about this axis (θ) and position along the axis of the line segment orthogonal to the axis that passes through each cell (y) were used as 2D coordinates for the cylindrical projection plot.

Fisher's Discriminant Analysis for identifying genetic basis of differentiation

Figure S4 and Figure 5 were generated by implementing Fisher's Discriminant Analysis (FDA). This method employs Fisher's linear discriminant to provide a closed-form solution for maximizing inter-class variance while minimizing intra-class variance between two classes of interest. The data matrix $\boldsymbol{A}$ is separated into two matrices $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$, each containing the subset of rows (representing cells) for the corresponding class of interest. For example, the rows of $\boldsymbol{A}$ representing cells in the MS lineage can be placed in $\boldsymbol{A}_0$, and the other rows (representing cells in other lineages) can be placed in $\boldsymbol{A}_1$. The maximum class separation occurs when the data are projected onto the vector

$$\boldsymbol{\omega} = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\,, \qquad (1)$$

i.e. the inverse of the sum of the covariance matrices of the data matrices for each class, $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$, multiplied by the vector difference in class means. The data matrices can be projected onto $\boldsymbol{\omega}$ by simple matrix-vector multiplication ($\boldsymbol{A}_0\,\boldsymbol{\omega}$ and $\boldsymbol{A}_1\,\boldsymbol{\omega}$, or projected collectively as $\boldsymbol{A}\,\boldsymbol{\omega}$) to obtain a one-dimensional representation of the data points, maximally separated into the two classes.

In our use case, the matrices $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$ have cells of their respective class in the rows and genes in the columns, i.e. they are of dimensions $(n_0^{cell} \times n^{gene})$ and $(n_1^{cell} \times n^{gene})$ respectively. The column covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are both of dimension $(n^{gene} \times n^{gene})$. The mean gene expression vectors $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are both of dimension $(n^{gene} \times 1)$, derived from the column means of $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$ respectively (i.e. these vectors are the mean expression level for each gene within the class). The resulting FDA projection vector ω is of dimension $(n^{gene} \times 1)$.

The vector $\boldsymbol{\omega}$ gives the set of gene weightings that maximally separates two classes of cells when expressed as a linear combination ($\boldsymbol{A}_0\,\boldsymbol{\omega}$ and $\boldsymbol{A}_1\,\boldsymbol{\omega}$). Each component of $\boldsymbol{\omega}$ is the weight of a specific gene, and therefore the absolute magnitude of these weights is directly related to how differentially expressed a gene is between the two classes, relative to other genes. FDA is therefore a particularly useful technique for identifying genes involved in differentiation (or in other A/B testing scenarios, such as case vs. control). However, it appears that this type of discriminant analysis is not yet widely used for this purpose in the biological sciences.

Conclusion

We have presented a comprehensive meta-analysis of the *C. elegans* single cell resolution EPIC gene expression dataset of Waterston *et al.*. Our analyses show multiple temporal patterns in the expression data, including oscillatory and/or linear correlations vs. developmental age, hinting at a global regulatory mechanism or developmental clock. We show that a simple linear weighting of gene expression can be chosen to exhibit roughly sinusoidal oscillation of any desired phase or frequency, suggesting that sinusoidal oscillations may be pervasive in regulating development. These results warrant further study, in order to understand and characterize the mechanisms of global regulation of gene expression during the developmental process.

Supplementary Tables

Table S1: The binarized gene expression matrix, with cells in the rows, genes in the columns, and 0 or 1 in the cells indicating whether or not gene expression crossed the minimum expression threshold during the lifetime of the cell.

Table S2: (a) The eigenvalues of the binarized gene expression matrix, with eigendecomposition performed on the columns (i.e. the genes). (b) The corresponding eigenvectors for each eigenvalue, i.e. the principal axes of gene expression.

Table S3: The projection of the binarized gene expression matrix onto the principal component axes. Rows represent cells; columns represent principal component axes. The PC1, PC2 and PC3 columns are the source of the data plotted in Figure 1.

Table S4: The Fisher's Discriminant Analysis (FDA) weights for genes that project the gene expression profile of a cell onto the FDA ω-vector, which is the axis that maximizes inter-class variance and minimizes intra-class variance between two classes. Gene weights are given for each of the FDA results depicted in Figure 4.

Table S5: The linear weighting x of genes that maps the gene expression profile of a cell onto the developmental age of the cell with the minimum squared error. This is derived by solving the linear system $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is the binarized gene expression matrix and $\boldsymbol{b}$ is the vector of cell birth times. Large negative and positive weights here correspond with genes that require a stronger contribution to linearize the mapping from gene expression to cell birth time.

Table S6: Comparison for each cell of b (cell birth time), PC1 (the projection of gene expression for the cell onto the first principal component axis), Ax (the linear combination of gene expression that best maps gene the expression profile of the cell onto the cell birth time), the Sulston cell onset time (obtained from [25], not available for all cells), the average offset from the cell birth time to the first point at which a gene's expression levels cross the binarization threshold, the average TTL (time to live) for the cell (i.e. the time until cell division or apoptosis), and the θ- and y-coordinates for the cylindrical projection of the gene expression manifold.

Table S7: The projection of gene expression values onto the principal component axes for the PCA of the transpose of the binarized gene expression matrix (i.e. the PCA of the transpose of Table S1). Consequently, this represents dimension reduction in the cell dimensions rather than the gene dimensions. Whereas the first k columns of Table S3 represents a dimension-reduced gene expression profiles for each cell, the first k columns of Table S7 represents a dimension-reduced profile of gene expression patterns for each gene. Genes that have similar dimension-reduced profiles (i.e. genes that are close together in the PCA space) are expressed in many of the same cells.

Acknowledgments

L.H. and B.B. were partially supported by NIH GM081871 (to B.B.).

Disclosure Declaration

There is no known conflict of interest.

References

[1] J. Cooke, Control of somite number during morphogenesis of a vertebrate, *Xenopus laevis* (1975).

[2] K. J. Dale, O. Pourquie, A clock-work somite, Bioessays 22 (2000) 72–83.

[3] C. Lorthongpanich, T. P. Y. Doris, V. Limviphuvadh, B. B. Knowles, D. Solter, Developmental fate and lineage commitment of singled mouse blastomeres, Development 139 (2012) 3722–3731.

[4] A. R. Desai, S. K. McConnell, Progressive restriction in fate potential by neural progenitors during cerebral cortical development, Development 127 (2000) 2863–2872.

[5] I. Palmeirim, D. Henrique, D. Ish-Horowicz, O. Pourquié, Avian *hairy* gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis, Cell 91 (1997) 639–648.

[6] N. Satoh, Timing mechanisms in early embryonic development, Differentiation 22 (1982) 156–163.

[7] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, G. Ruvkun, The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*, nature 403 (2000) 901–906.

[8] Z. Zhao, T. J. Boyle, Z. Liu, J. I. Murray, W. B. Wood, R. H. Waterston, A negative regulatory loop between microRNA and *Hox* gene controls posterior identities in *Caenorhabditis elegans*, PLoS Genet 6 (2010) e1001089.

[9] J. Hench, J. Henriksson, A. M. Abou-Zied, M. Lüppert, J. Dethlefsen, K. Mukherjee, Y. G. Tong, L. Tang, U. Gangishetti, D. L. Baillie, et al., The homeobox genes of *Caenorhabditis elegans* and insights into their spatio-temporal expression dynamics during embryogenesis, PloS one 10 (2015) e0126947.

[10] G. Nair, T. Walton, J. I. Murray, A. Raj, Gene transcription is coordinated with, but not dependent on, cell divisions during *C. elegans* embryonic fate specification, Development 140 (2013) 3385–3394.

[11] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, S. R. Quake, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq, Nature 509 (2014) 371–375.

[12] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al., Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, Nature structural & molecular biology 20 (2013) 1131–1139.

[13] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nature biotechnology 32 (2014) 381–386.

[14] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, D. PeâĂŹer, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development, Cell 157 (2014) 714–725.

[15] K. L. Davis, S. C. Bendall, D. A. El-ad, E. F. Simonds, A. Jager, A. Trejo, D. Pe'er, G. P. Nolan, Single cell trajectory detection

orders hallmarks of early human B cell development, Blood 120 (2012) 1044–1044.

[16] C. L. Araya, T. Kawli, A. Kundaje, L. Jiang, B. Wu, D. Vafeados, R. Terrell, P. Weissdepp, L. Gevirtzman, D. Mace, et al., Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution, Nature 512 (2014) 400–405.

[17] J. L. Moore, Z. Du, Z. Bao, Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis, Development 140 (2013) 3266–3274.

[18] J. L. Richards, A. L. Zacharias, T. Walton, J. T. Burdick, J. I. Murray, A quantitative model of normal *Caenorhabditis elegans* embryogenesis and its disruption after stress, Developmental biology 374 (2013) 12–23.

[19] C. A. Giurumescu, S. Kang, T. A. Planchon, E. Betzig, J. Bloomekatz, D. Yelon, P. Cosman, A. D. Chisholm, Quantitative semi-automated analysis of morphogenesis with single-cell resolution in complex embryos, Development 139 (2012) 4271–4279.

[20] X. Liu, F. Long, H. Peng, S. J. Aerni, M. Jiang, A. Sánchez-Blanco, J. I. Murray, E. Preston, B. Mericle, S. Batzoglou, et al., Analysis of cell fate from single-cell gene expression profiles in *C. elegans*, Cell 139 (2009) 623–633.

[21] Z. Bao, J. I. Murray, T. Boyle, S. L. Ooi, M. J. Sandel, R. H. Waterston, Automated cell lineage tracing in *Caenorhabditis elegans*, Proceedings of the National Academy of Sciences of the United States of America 103 (2006) 2707–2712.

[22] T. J. Boyle, Z. Bao, J. I. Murray, C. L. Araya, R. H. Waterston, AceTree: a tool for visual analysis of *Caenorhabditis elegans* embryogenesis, BMC bioinformatics 7 (2006) 275.

[23] J. I. Murray, Z. Bao, T. J. Boyle, R. H. Waterston, The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree, Nature protocols 1 (2006) 1468–1476.

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

[25] J. I. Murray, T. J. Boyle, E. Preston, D. Vafeados, B. Mericle, P. Weisdepp, Z. Zhao, Z. Bao, M. Boeck, R. H. Waterston, Multidimensional regulation of gene expression in the *C. elegans* embryo, Genome research 22 (2012) 1282–1294.

[26] M. T. Lee, A. R. Bonneau, A. J. Giraldez, Zygotic genome activation during the maternal-to-zygotic transition, Annual review of cell and developmental biology 30 (2014) 581–613.

[27] B. Goszczynski, J. D. McGhee, Reevaluation of the role of the *med-1* and *med-2* genes in specifying the *Caenorhabditis elegans* endoderm, Genetics 171 (2005) 545–555.

[28] H. R. Nicholas, J. Hodgkin, The *C. elegans Hox* gene *egl-5* is required for correct development of the hermaphrodite hindgut and for the response to rectal infection by Microbacterium nematophilum, Developmental biology 329 (2009) 16–24.

[29] L. Anavy, M. Levin, S. Khair, N. Nakanishi, S. L. Fernandez-Valverde, B. M. Degnan, I. Yanai, BLIND ordering of large-scale transcriptomic developmental timecourses, Development 141 (2014) 1161–1166.

[30] G.-J. Hendriks, D. Gaidatzis, F. Aeschimann, H. Großhans, Extensive oscillatory gene expression during *C. elegans* larval development, Molecular cell 53 (2014) 380–392.

[31] D. h. Kim, D. Grün, A. van Oudenaarden, Dampening of expression oscillations by synchronous regulation of a microRNA and its target, Nature genetics 45 (2013) 1337–1344.

[32] K. A. McCulloch, A. E. Rougvie, *Caenorhabditis elegans* period homolog *lin-42* regulates the timing of heterochronic miRNA expression, Proceedings of the National Academy of Sciences 111 (2014) 15450–15455.

[33] J. Shao, K. He, H. Wang, W. S. Ho, X. Ren, X. An, M. K. Wong, B. Yan, D. Xie, J. Stamatoyannopoulos, et al., Collaborative regulation of development but independent control of metabolism by two epidermis-specific transcription factors in *Caenorhabditis elegans*, Journal of Biological Chemistry 288 (2013) 33411–33426.

[34] Z. Chen, D. J. Eastburn, M. Han, The *Caenorhabditis elegans* nuclear receptor gene *nhr-25* regulates epidermal cell development, Molecular and cellular biology 24 (2004) 7345–7358.

[35] K. W. Thompson, P. Joshi, J. S. Dymond, L. Gorrepati, H. E. Smith, M. W. Krause, D. M. Eisenmann, The paired-box protein PAX-3 regulates the choice between lateral and ventral epidermal cell fates in *C. elegans*, Developmental biology 412 (2016) 191–207.

[36] Y. Andachi, *Caenorhabditis elegans* T-box genes *tbx-9* and *tbx-8* are required for formation of hypodermis and body-wall muscle in embryogenesis, Genes to Cells 9 (2004) 331–344.

[37] A. N. Gorban, A. Y. Zinovyev, Principal graphs and manifolds, in: E. S. Olivas (Ed.), Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI Global, 2009, pp. 28–59.

Supplementary Figures

Figure S1: The scree plot from Principal Components Analysis of gene expression (i.e. the eigenvalues of gene expression space, sorted in decreasing order of magnitude). Note the vertical log scale. This plot shows that a significant amount of variance in gene expression across cells is captured by the first 10 components of the PCA.

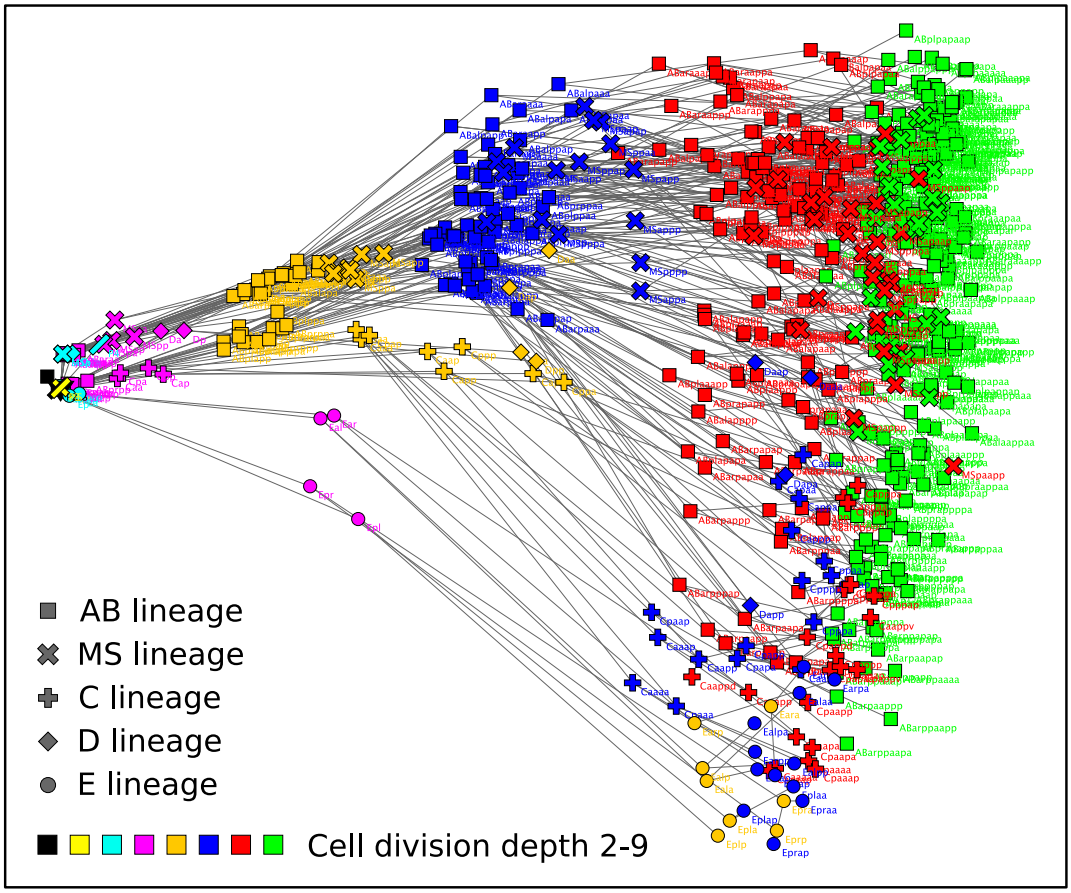

Figure S2: The cylindrical projection of the gene expression manifold seen in Figure 1 onto a flat 2D surface.The horizontal axis is θ, the angle of rotation about the cylindrical axis, and the vertical axis, y, represents the distance of the cell along the cylindrical axis. The cell identities are labeled for illustration purposes (though are not intended to be highly legible, due to label overlap).

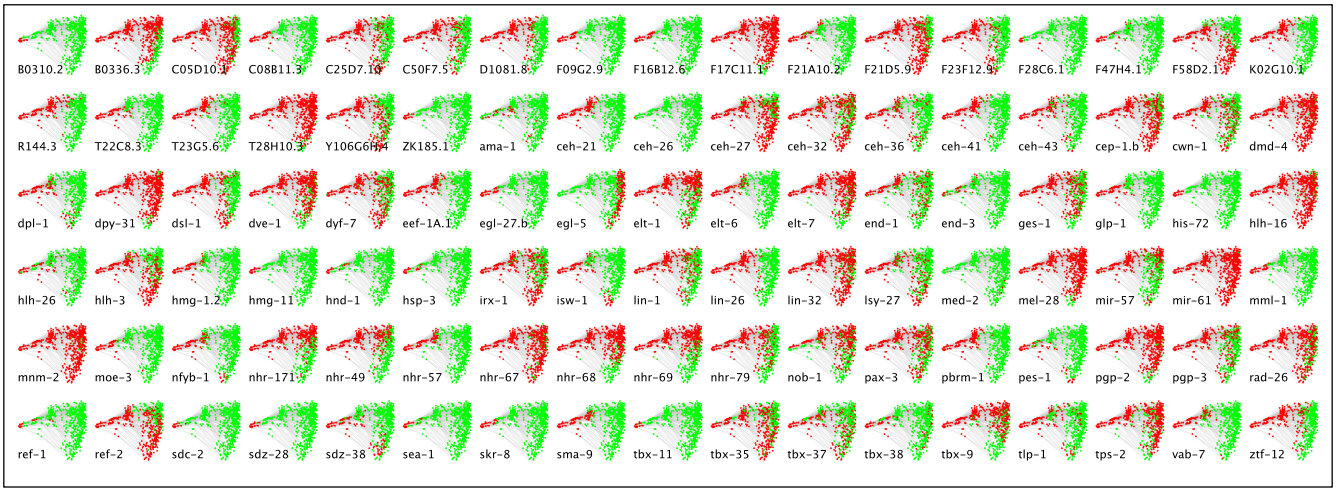

Figure S3: The cylindrical projection (as in Figure S2) of the gene expression manifold for each of the 102 genes, with red and green indicating that the gene was not expressed or expressed in a given cell respectively.

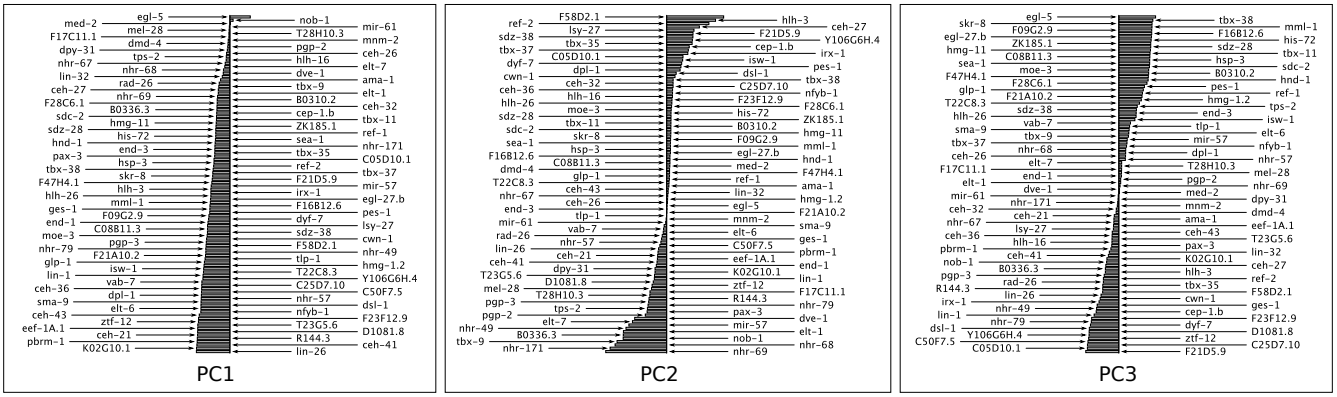

Figure S4: Bar chart of eigenvector components for PC1-PC3, with the components sorted vertically in order of magnitude (positive to the right of the midline, negative to the left), showing the relative magnitude of contribution of each gene to the position of cells in each of the three PC axes. Multiplying these weights by the square root of the corresponding eigenvalue gives the *principle component loadings*. Larger positive or negative weights indicate more relative contribution of the gene to the position of a cell in the corresponding PC axis. For PC2, only a minority of genes have large negative or positive weight, indicating this subset of genes is involved in differentiation. For PC1, the comparative lack of low-weighted genes indicates that the strong correlation between PC1 and the developmental age of the organism is not a simple relationship governed by a small subset of genes, but that most of the genes are involved in this correlation.