Bioinformatics, 36(13), 2020, 4047–4057
doi: 10.1093/bioinformatics/btz864
Advance Access Publication Date: 20 December 2019
Original Paper

Systems biology

Meta-analysis of *Caenorhabditis elegans* single-cell developmental data reveals multi-frequency oscillation in gene activation

Luke A.D. Hutchison[1], **Bonnie Berger**[1,*] **and Isaac S. Kohane**[2,*]

[1]MIT Computer Science and AI Lab, Cambridge, MA 02139, USA and [2]Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 23, 2019; revised on September 23, 2019; editorial decision on November 14, 2019; accepted on December 18, 2019

Abstract

Motivation: The advent of *in vivo* automated techniques for single-cell lineaging, sequencing and analysis of gene expression has begun to dramatically increase our understanding of organismal development. We applied novel meta-analysis and visualization techniques to the EPIC single-cell-resolution developmental gene expression dataset for *Caenorhabditis elegans* from Bao, Murray, Waterston *et al.* to gain insights into regulatory mechanisms governing the timing of development.
Results: Our meta-analysis of the EPIC dataset revealed that a simple linear combination of the expression levels of the developmental genes is strongly correlated with the developmental age of the organism, irrespective of the cell division rate of different cell lineages. We uncovered a pattern of collective sinusoidal oscillation in gene activation, in multiple dominant frequencies and in multiple orthogonal axes of gene expression, pointing to the existence of a coordinated, multi-frequency global timing mechanism. We developed a novel method based on Fisher's Discriminant Analysis to identify gene expression weightings that maximally separate traits of interest, and found that remarkably, simple linear gene expression weightings are capable of producing sinusoidal oscillations of any frequency and phase, adding to the growing body of evidence that oscillatory mechanisms likely play an important role in the timing of development. We cross-linked EPIC with gene ontology and anatomy ontology terms, employing Fisher's Discriminant Analysis methods to identify previously unknown positive and negative genetic contributions to developmental processes and cell phenotypes. This meta-analysis demonstrates new evidence for direct linear and/or sinusoidal mechanisms regulating the timing of development. We uncovered a number of previously unknown positive and negative correlations between developmental genes and developmental processes or cell phenotypes. Our results highlight both the continued relevance of the EPIC technique, and the value of meta-analysis of previously published results. The presented analysis and visualization techniques are broadly applicable across developmental and systems biology.
Availability and implementation: Analysis software available upon request.
Contact: bab@mit.edu or isaac_kohane@harvard.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Four-dimensional microscopy (imaging three dimensions over time) and a range of single-cell sequencing and profiling techniques are enabling unprecedented study of how the developmental program unfolds *in vivo* (Bao *et al.*, 2006; Boyle *et al.*, 2006; Cao *et al.*, 2017; Giurumescu *et al.*, 2012; Liu *et al.*, 2009; Moore *et al.*, 2013; Murray *et al.*, 2006; 2008; Richards *et al.*, 2013; Wolf *et al.*, 2018). However, as the amount of available single-cell data increases, e.g. through the advent of single-cell RNA sequencing (scRNA-seq) (Chen *et al.*, 2018), we need better tools for analyzing and visualizing these datasets [e.g. Cho *et al.* (2018) and Hie *et al.* (2019)], in order to make the data more comprehensible and useful.

One of the early groundbreaking whole-organism single-cell gene expression analysis techniques, by Bao *et al.* (2006) (Murray *et al.*, 2006, 2008, 2012) employed histone-mCherry reporters under the control of upstream promoters to detect expression levels of genes of interest at single-cell resolution, while employing fluorescent cell labeling and 4D confocal microscopy to enable automated cell lineaging. This process yielded a complete 3D map of gene expression at single-cell resolution across the entire developmental timeline while simultaneously tracing the cell division pedigree, resulting in a 4D gene expression dataset superimposed over the cell division tree. This dataset was published as Expression Patterns in Caenorhabditis (EPIC) (Waterston *et al.*, 2006). After the initial impact of this publication, the technique and the results were not extensively scrutinized, meta-analyzed or reproduced in subsequent published work, and scRNA-seq has subsequently drawn the attention of the research community [e.g. Ramsköld *et al.* (2012) and Cao *et al.* (2017)].

However, even today, the EPIC technique remains a goldmine for understanding development, and retains several major advantages over RNA-Seq, including that in EPIC, gene expression is tracked continuously *in vivo* for each cell, throughout the lifetime of each cell and over the entire developmental timeline of the organism (until rapid movement of the organism prevents further tracking), without sacrificing the organism and without requiring the mixing or combination of data from different cells or different organisms to produce sufficient data to generate a fine-grained timeline of development at different time points. Optical scanning in EPIC provides continuous gene expression profiling at single-cell resolution, at least for optically translucent organisms, such as *Caenorhabditis elegans*, and does not suffer from either the cell type mixing issues of RNA-Seq or the single-cell extraction and isolation complexity issues of scRNA-seq. The EPIC technique is also capable of not only tracking the expression levels of genes of interest within each cell, but also simultaneously tracing the cell division lineage of the organism, which would be a particularly powerful technique for studying the effect of specific mutations on development, as well as organisms whose cell lineages vary more widely than that of *C.elegans*. Due to each of these advantages over the dominant technique of RNA-Seq, the EPIC technique (and the published EPIC dataset) deserves significantly greater exposure and further study than it has thus far received. This paper aims to provide a deeper look into the published EPIC dataset, to determine whether interesting signals could be found within the dataset that were missed in the original authors' analyses and visualizations. Our meta-analysis of the EPIC dataset focused on understanding the timing of development as a function of gene expression.

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 (Cooke, 1975; Dale and Pourquié, 2000; Desai and McConnell, 2000; Hench *et al.*, 2015; Lorthongpanich *et al.*, 2012; Nair *et al.*, 2013; Palmeirim *et al.*, 1997; Reinhart *et al.*, 2000; Satoh, 1982; Zhao *et al.*, 2010). Recent work on single-cell gene expression has illuminated interesting regulatory processes at work during development (Araya *et al.*, 2014; Bendall *et al.*, 2014; Davis *et al.*, 2012; Trapnell *et al.*, 2014; Treutlein *et al.*, 2014; Yan *et al.*, 2013). New methods are needed for determining the mechanisms behind the regulation and timing of development, utilizing the data produced by single-cell techniques (Briggs *et al.*, 2018; Cao *et al.*, 2017; Farrell *et al.*, 2018; Wagner *et al.*, 2018). In our meta-analysis, we attempted to identify any time correlation present in gene activity in the EPIC dataset.

2 Materials and methods

2.1 Dataset preprocessing

We obtained the EPIC dataset (http://epic.gs.washington.edu/), comprising 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). We selected the largest possible dense subset of the data, i.e. the largest subset for which there were no missing values in the table consisting of genes in the columns and cells in the rows. We truncated the timeline after 686 unique cell identities had appeared in the cell pedigree, because the data sparsity increased sharply after that point, as a result of tracking difficulty once the nematode begins to move within its egg sac.

We found 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, in order to prevent outliers (genes with very high activation levels) from skewing the results (since we were only interested in whether or not a gene was active at a given point in development, not in the magnitude of activity, beyond a minimal threshold). We used the reporter intensity threshold of 2500 as the boundary between active and inactive, which was conservatively chosen based on Murray *et al.* (2006), 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. Our chosen intensity threshold of 2500 errs more toward false positives than false negatives in classifying gene activity, but we found that this threshold provided an adequate balance between the gene expression matrix tending toward being filled with zeroes (for a threshold value set too high) and being filled with all ones (for a threshold value set too low). We discarded a number of genes that were expressed in all (or nearly all) cells at this threshold level, as well as genes that were not expressed in any (or almost any) cells at this threshold level. As can be seen in Supplementary Figure S3, the number of positive (green) and negative (red) gene activations is roughly balanced across all cells for the remaining genes.

The resulting binarized gene expression matrix (Supplementary Table S3) consists of the binarized gene expression values (0 or 1) for 102 genes, measured across the lifespan of 686 cell identities. 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).

2.2 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. Despite the claim in Murray *et al.* (2006) that the EPIC data was sampled 'with ~1-min temporal resolution', the raw data was quite clearly sampled at a wide range of different time intervals for each gene, with time scale factors including at least 1.0, 1.35, 1.4, 1.5 and 2.0 min 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 correspondence, the original authors stated they 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 careful time series analysis to recover a best-fit time scale factor for each run. We used a custom multiple-alignment regression technique to linearly stretch the timeline such that the birth time and division or death time of each cell best fit a consensus cell pedigree [the canonical *Sulston pedigree* (Sulston *et al.*, 1983)]. The sample interval was known for some genes, which was used as a cross check of the timeline fitting method. In this process, we also discovered that not only was a multiplicative offset needed to scale the data samples to fit the timeline of the canonical Sulston pedigree, but a varying additive offset was also required to normalize the 'zero point', averaging ~45 min 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 ~45 min).

The EPIC data includes multiple runs for some genes, sometimes with non-trivial 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. For a discussion of sources of noise and variability between runs for the same gene, we refer the reader back to the original publication.

2.3 Principal component analysis

PCA was run on the binarized gene expression matrix, by mean-centering the gene expression matrix (adjusting the expression levels to have a mean expression level of zero for each gene), calculating the covariance matrix, finding the eigenvalues and eigenvectors of the covariance matrix, and then sorting the eigenvectors into decreasing order of corresponding eigenvalue. This produced Supplementary Table S4a and b, the PCA eigenvalues and eigenvectors, respectively. We then projected the gene expression matrix onto the first 10 principal component axes to produce the principal components PC1–PC10, shown in Supplementary Table S5, and depicted in many of the figures. [Note that it is the raw (non-mean-centered) data that was projected onto the eigenvectors, not the mean-centered gene expression matrix. This choice results in only a translation of the data in all axes, it does not change the shape of the projected data.] Columns 1–3 of Supplementary Table S5, plotted in 3D, yielded Figure 1. All 2D pairings of principal component axes PC1–PC10 yielded Figure 2. (The figures in this paper were all rendered using custom 2D and 3D visualization software.)

To examine the contribution of the principal axes toward overall dataset variance, we produced a *scree plot* (Supplementary Fig. S1) from the binarized gene expression matrix (Supplementary Table S3), 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 principal components.

2.4 FDA for identifying genetic basis of differentiation

Figure 3 and Supplementary Figure S4 were generated by implementing 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 A is separated into two matrices A_0 and A_1, each containing the subset of rows (representing cells) for the corresponding class of interest. For example, the rows of A representing cells in the MS lineage can be placed in A_0, and the other rows (representing cells in other lineages) can be placed in A_1. The maximum class separation occurs when the data are projected onto the vector

$$\boldsymbol{\omega} = (\Sigma_0 + \Sigma_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \tag{1}$$

i.e. the inverse of the sum of the covariance matrices of the data matrices for each class, A_0 and A_1, multiplied by the vector difference in class means. The data matrices can be projected onto $\boldsymbol{\omega}$ by simple matrix-vector multiplication ($A_0\,\boldsymbol{\omega}$ and $A_1\,\boldsymbol{\omega}$, or projected collectively as $A\boldsymbol{\omega}$) to obtain a 1D representation of the data points, maximally separated into the two classes.

In our use case, the matrices A_0 and 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 Σ_0 and Σ_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 A_0 and 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 ($A_0\boldsymbol{\omega}$ and $A_1\boldsymbol{\omega}$). 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 does not yield an innately optimal class separation boundary along its projection vector, $\boldsymbol{\omega}$. For Figure 4 and the figures in Supplementary Table S6, we plotted as the decision boundary the line which equalized the percentage of cells in Class 1 above the line and the percentage of cells in Class 0 below the line.

2.5 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 semi-cylindrical. We flattened out this 'principal manifold' of the data (Gorban and Zinovyev, 2009) 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 (Supplementary Fig. S2). Supplementary Table S6 gives the 2D 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 Supplementary 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.

3 Results

We present a meta-analysis and visualization of a subset of the EPIC dataset, which suggests a global mechanism governing the timing of

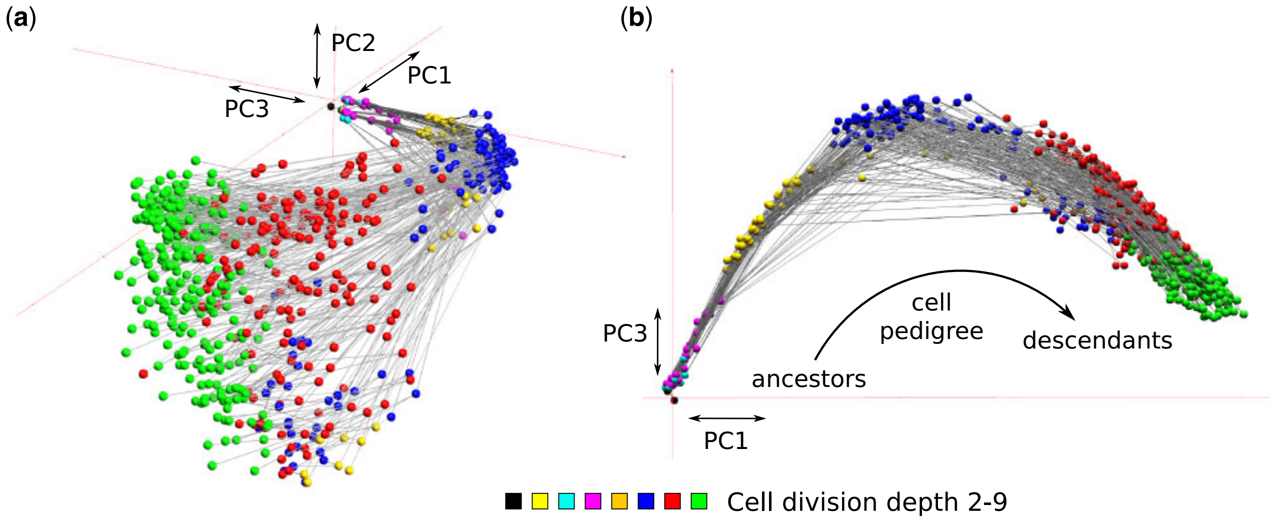

Fig. 1. Projection of binarized gene expression profiles onto the first three principal component axes. Each node represents a cell, and the edges between the nodes connect a cell with each of its two daughter cells. The color of each node indicates the division depth in the cell pedigree. (**a**) A perspective view of the first three principal components. (**b**) A top-down view of PC3 versus PC1, showing the curved path of the manifold relative to the first principal component axis

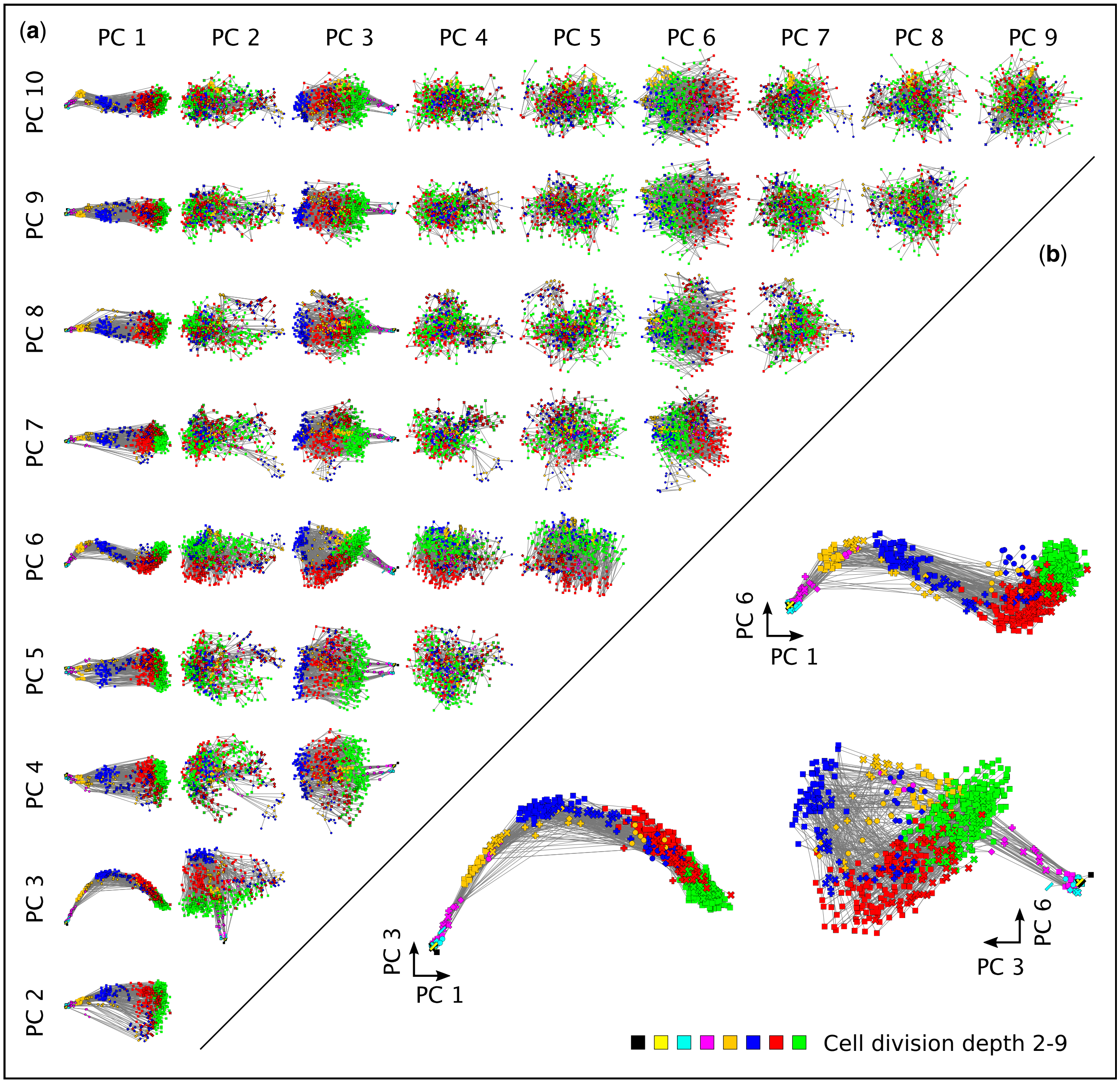

Fig. 2. (**a**) The projection of gene expression onto the first 10 principal component axes, PC1–PC10, shown as 2D projections. (**b**) Larger views of PC3 versus PC1, PC6 versus PC1 and PC6 versus PC3, showing what appears to be sinusoidal oscillation of two different frequencies. Plotting PC6 against PC3 causes the cell pedigree to trace an α-shaped path as development proceeds

development in *C.elegans*. This subset of EPIC consists of the expression levels of 102 developmentally relevant genes (out of 127 genes total) 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). We did not include genes or cells with incomplete data. Significant work was required to prepare the EPIC data for meta-analysis (see Section 2). Producing the complete gene expression matrix for these genes and cells has allowed us to apply analysis techniques to the dataset as a whole, including principal component analysis (PCA), which allowed us to discover that gene expression follows a sweeping manifold shape through expression eigenspace as the developmental timeline proceeds.

We discovered a strong linear correlation ($R^2 = 0.94$) between the first principal component of gene expression and wall-clock time during cell proliferation. Over the entire available timeline, we observed multiple apparent sinusoidal oscillations in gene expression, with different frequencies of oscillation manifest in different principal components, indicating that our observation of linear monotonic correlation between gene expression and wall-clock time have been an observation of the nearly linear part of a sinusoidal graph around the zero-crossing point (it is difficult to know without the availability of data from later in development).

We devised a novel technique from Fisher's Discriminant Analysis (FDA) for uncovering the relative contributions of genes to an attribute of interest, and used this method to study lineage-specific gene expression patterns, and the separability of lineages based on gene expression profiles. By applying this technique, remarkably we were able to find simple linear weightings of gene activation that were able to produce sinusoidal oscillations of any desired frequency and phase in the weighted sum of gene expression. This result suggests that oscillatory mechanisms may be used extensively to regulate the timing of development.

To determine the functionality of developmental genes of interest, we also applied our FDA technique to the cross-linked EPIC dataset with the gene ontology (GO) and anatomy ontology from Wormbase (Stein *et al.*, 2001) to identify weightings for each

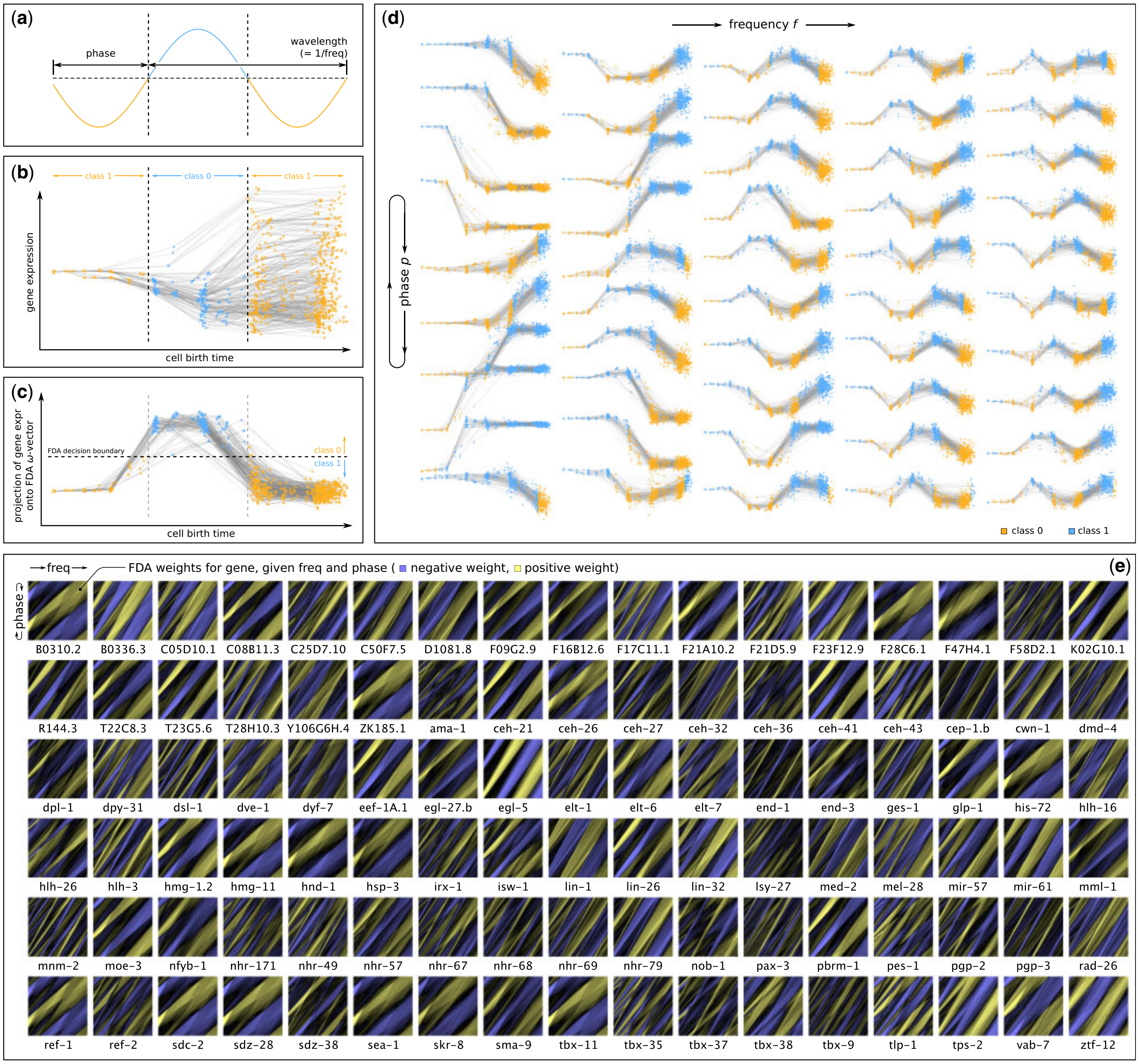

Fig. 3. Identification of simple linear weightings of gene expression levels that can produce oscillations across a range of sinusoidal frequencies and phases. (**a**) A target sine wave is generated. (**b**) Cells are assigned to Class 0, at developmental times when the sine wave is positive, or Class 1, at times when the sine wave is negative. (**c**) 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 ($0 \rightarrow \pi \rightarrow 0$), 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

gene that specify how strongly a gene appears to be correlated with the presence or absence of a given phenotypic trait or developmental process in each cell. These gene weightings provided a large number of previously unknown implications about the functioning of various genes across a wide range of developmental processes.

Individual specific results are discussed inline as they are presented in the relevant Sections 2 and 4.

The methods described in this paper are simple to apply, but have the potential to be broadly useful in understanding single-cell experimental data. The specific results we produced from our meta-analysis of the EPIC *C.elegans* data, cross-linked with Wormbase ontologies, presents strong evidence for global control of developmental timing, suggesting numerous opportunities for further research into the time-correlated and oscillatory mechanisms of developmental regulation.

4 Discussion

4.1 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 3D plot of the cell pedigree directly overlaid on the first three principal components of gene expression (Fig. 1). In this plot, nodes represent the 686 unique cells in the dataset (specifically, the unique 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 3D 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.

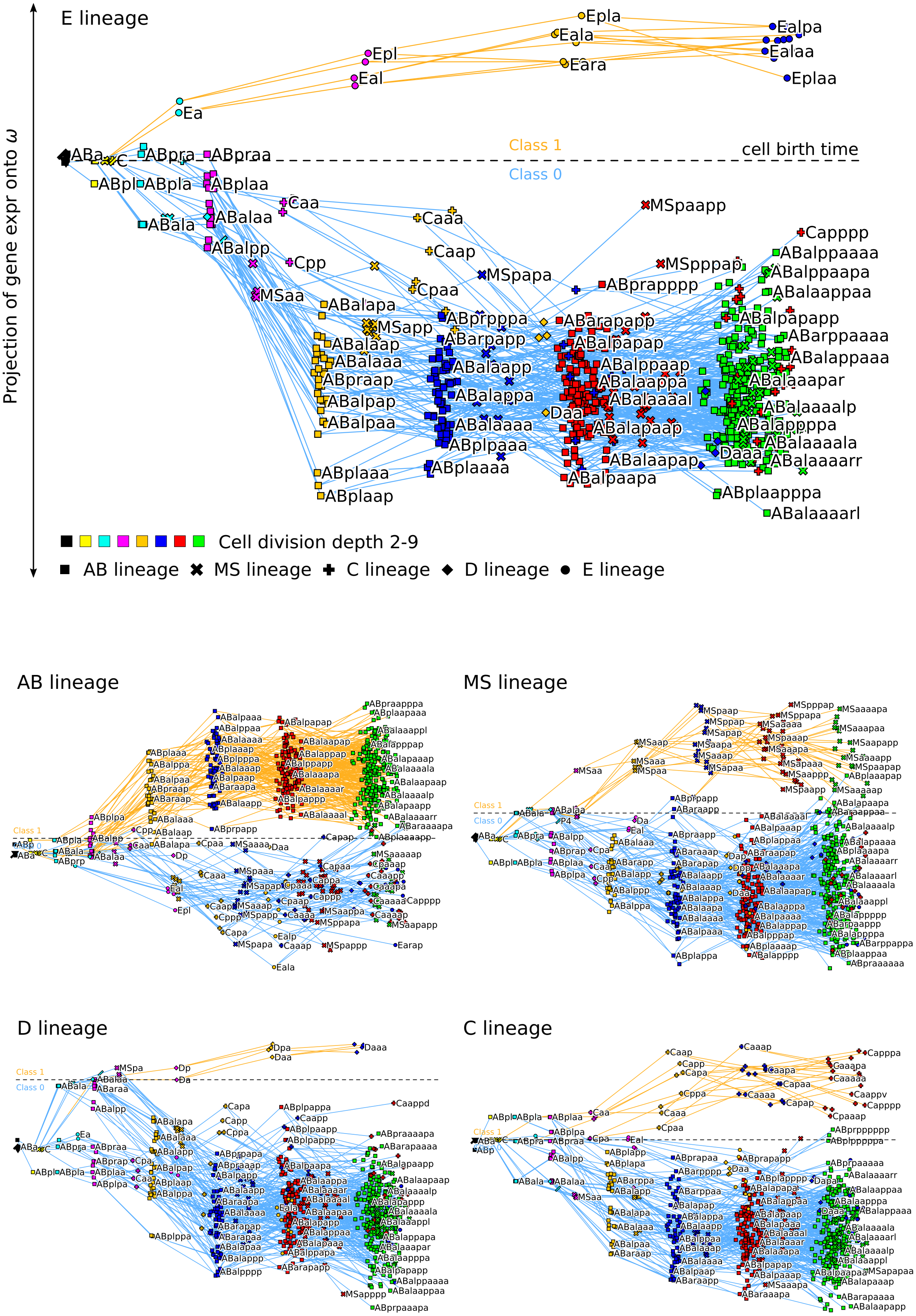

Fig. 4. Separation of the E cell lineage (Class 1) from other cells in the cell pedigree (Class 0) using FDA. The horizontal axis indicates developmental time, and the vertical axis represents the linear projection onto the 1D vector ω that gives maximal inter-class variance and minimal intra-class variance. (Only a subset of cell labels is shown for legibility.) The resulting gene weights are strongly positive for genes primarily expressed in the E lineage, and strongly negative for genes primarily expressed in cells other than those in the E lineage

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 3D '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. It has been observed previously that gene expression can follow a trajectory over time through expression space (Qiu *et al.*, 2017), a concept that has been referred to as 'pseudotime' (Trapnell *et al.*, 2014), but in Figure 1, we rather show that the observed cell lineages all collectively trace not a single path, but a set of paths lying approximately on the surface of a manifold, suggesting the existence of a strong and globally coordinated developmental control mechanism.

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 monotonically correlated with the most significant orthogonal direction of variance in gene expression. The implication of this is that the largest variation in gene expression across all cells is time-correlated, and this time correlation is monotonic – in other words, variations in gene activation is collectively coordinated and *sequenced* such that development moves in a specific direction.

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, most plausibly corresponding to a general diversification in gene expression, consistent with lineage-specific differentiation. 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 with tissue-specific differentiation.

4.2 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. Figure 5a 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 min). 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 for the first principal component (i.e. the eigenvector multiplied by the square root of the corresponding eigenvalue—Supplementary Fig. S4 and Table S4).

This strong correlation between PC1 and developmental age appears to be independent of the cell division rate, in particular implying that the gene expression profiles of cells are all similarly correlated with developmental age regardless of cell division depth within the cell pedigree. This finding can be seen by the skew in cell division depth in Figure 1, evidenced by the increased mixing of cell colors, representing cell division depth, as development proceeds. Each of the major cell lineages MS, C, D and E have successively slower cell division rates relative to the AB lineage (the difference in cell division rates between the lineages can be seen in Figure 4, but gene expression is not correlated with cell division depth as strongly as with developmental age). The finding that wide-scale gene expression patterns are decoupled from cell division rate agrees with the findings in Nair *et al.* (2013), who found that expression and proliferation are independently linked to separate clock-like processes. Whereas Nair *et al.* found that the *relative timing* of cell division was not directly correlated with large-scale transcriptional regulation, in our analyses we observed that the *depth* of cell division was not directly correlated with large-scale transcriptional regulation, since lineages with dramatically different cell division intervals exhibit similar expression trends.

Before 100 min and after 200 min, however, PC1 diverges from being linearly correlated with the cell birth age b (Fig. 5a). 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 below a linear regression method for fitting a linear model to the data. Another possible explanation for the observed distribution is that the underlying gene expression is fundamentally sinusoidal, not linear, and that we merely observed the section of the sinusoid that is most linear, around the zero-crossing point (discussed below). It is also possible that entirely different developmental programs or processes are active before 100 min, between 100 and 200 min, and after 200 min: one plausible explanation for the difference in gene activity before and after 100 min is the maternal-

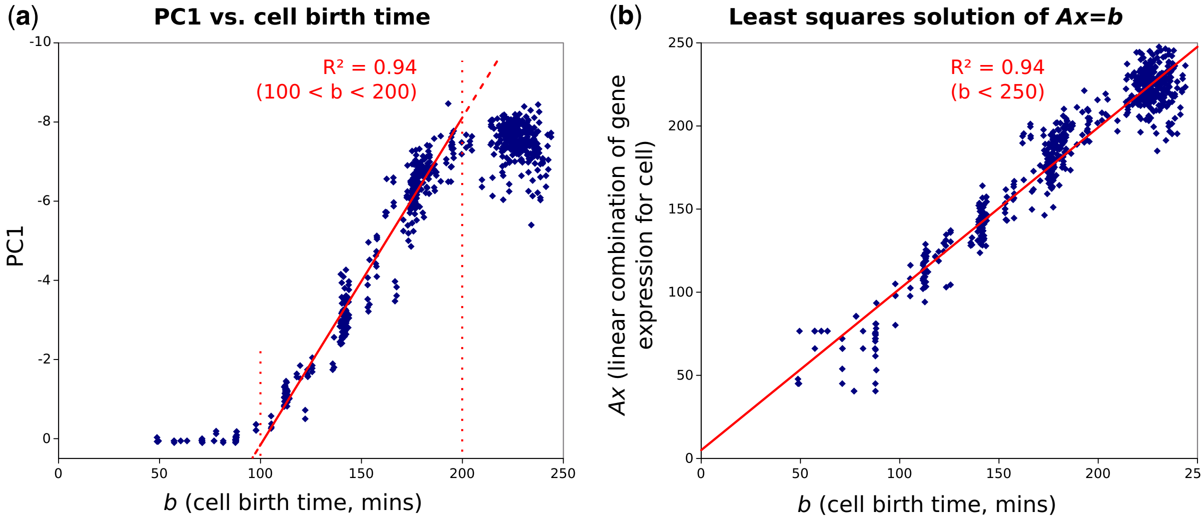

Fig. 5. (**a**) The projection of gene expression onto the first principal component, PC1, versus b, the cell birth time (i.e. the cell onset time) in minutes. The plot is strongly linear from 100 to 200 min. (**b**) After solving the linear equation $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A}$ is the binarized gene expression matrix and $\boldsymbol{b}$ is the vector of cell birth times, this plot shows the best-fit linear estimator mapping birth time to gene expression

to-zygotic transition (Lee *et al.*, 2014). Whether what we observed is linear or sinusoidal in nature, the correlation we observed between the first principal component of gene expression and wall-clock time is remarkable, since a simple linear projection of gene activation patterns can accurately predict the age of the organism across much of the studied developmental timeline.

To find the extent to which gene expression can be linearly correlated with time, we reformulated 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. Supplementary Table S3, 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 this linear 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 $\boldsymbol{x}$ (Supplementary Table S8) 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 min), retaining approximately the same linear correlation strength of $R^2 = 0.94$, but across the entire developmental timeline (Fig. 5b). 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 Supplementary Table S5, *med-2*, is necessary for endoderm specification (Goszczynski and McGhee, 2005). Other genes with high negative weights include *sdz-28*, an SKN1-dependent zygotic transcript active only 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 (Nicholas and Hodgkin, 2009), as well as a number of genes that are expressed nearly ubiquitously across the entire developmental timeline.

Supplementary Table S9 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.

4.3 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 EPIC technique 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 2a shows all pairings of principal components between PC1 and PC10, as a way of visualizing the first ten principal component dimensions in two dimensions. Interestingly, PC6 traces a complete sinusoid with respect to PC1 across the same time period that PC3 traces half a sinusoid. The presence of a complete and clear sinusoidal oscillation in PC6 relative to PC1 lends strength to the hypothesis that the apparent curve of PC3 relative to PC1 is in fact the first half of a sinusoidal oscillation with half the frequency compared to the oscillation in PC6. However, without data for the entire developmental timeline of the organism, this cannot be confirmed.

These oscillations of different frequency are concurrent and mutually orthogonal (hence the reference to 'multi-frequency oscillation' in the title of this paper). If PC6 is paired with PC3, the half sinusoid paired with the higher-frequency complete sinusoid causes the developmental path to trace a spiraling alpha shape (α), as shown enlarged in Figure 2b.

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 versus PC1 in Figure 1.

We have not identified a plausible biological mechanism behind the apparent sinusoidal or semi-sinusoidal oscillations observed in PC3 and PC6 (and perhaps also in PC10) relative to PC1, other than that they may be related to the coordination of global developmental timing, rather than some specific developmental process itself. We believe this question deserves further study.

4.4 Connection to previous observations of oscillatory patterns during development

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 applying 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 (Anavy *et al.*, 2014). This is significant, as it lends credence to the presence of a global clock mechanism coordinating development. The expression levels of individual mRNAs were also observed to oscillate over time (Hendriks *et al.*, 2014; Lee *et al.*, 2014). Both oscillatory and temporally gradated activity has been observed in transcript levels (Kim *et al.*, 2013). Some key regulators of the timing of heterochronic miRNA expression have been discovered, including *lin-4* and *let-7*-family miRNAs, under control of *lin-42* (McCulloch and Rougvie, 2014), however data on these miRNAs was not available in the EPIC dataset for comparison.

A range of new techniques have been proposed to track gene expression trajectories over time (Domany, 2014). Robust oscillation in transcription has been observed previously at multiple temporal scales (Hendriks *et al.*, 2014; Kim *et al.*, 2013), involving ultraradian cycles affecting ~1/6th of the transcriptome, with changes in expression of up to an order of magnitude during the cycles. The authors identified several periodic developmental phenomena, such as cuticle development and cuticle molting, and speculated that the timing of other developmental processes are similarly controlled by one or more of these transcriptional cycles.

What is unique about our findings is that the patterns of oscillation in gene expression that we observed were not limited to specific cell types, or to specific cell lineages, and were not affected by the cell division rate of different lineages. Our observations were not made by taking samples of transcripts or other genetic activity, averaged across the whole organism at specific time points. Rather, we observed oscillations occurring *separately and simultaneously within each individual cell at single-cell resolution*, as part of a globally synchronous oscillatory pattern exhibited by all cells in existence at each point in the developmental timeline of the organism. Also notable is our observation of *multiple superposed oscillations of different frequency and phase.* This superposition of time-correlated oscillation was *contemporaneous with non-oscillatory patterns of gene expression involved with cell differentiation:* the observed patterns of gene activation simultaneously and collectively encoded multiple oscillatory mechanisms, in an almost 'holographic' sense, based on gene activations at individual cells—and yet each gene simultaneously and separately also served its own unique role in development, unrelated *per se* to the oscillation to which it contributed.

4.5 Generation of sinusoidal oscillation as a linear weighting of gene activations

To understand the extent to which linear combinations of gene activation could give rise to a sinusoidal oscillation, we temporally divided cells into two classes based on the cell birth time in the developmental timeline, corresponding to peaks (Class 0) or troughs (Class 1) in a target sinusoidal function of a given frequency and phase (Fig. 3a). FDA was used to find the hyperplane that separates Class 0 and Class 1 with minimum intra-class variance and maximum inter-class variance, which effectively created a square wave approximation of the target sine wave. The projection of gene expression onto the normal vector ω of the separating hyperplane was plotted against the cell birth time to produce a best fit of gene expression to the original sinusoidal wave (Fig. 3b and c).

Figure 3d shows that a simple linear combination of gene activations can produce an approximate sinusoidal wave of any desired frequency or phase. This result lends support to the idea that the patterns of sinusoidal oscillation that we found in our PCA, as a simple linear projection onto the principal component axes, were not an anomaly, but may have been due to an underlying process that relies upon the combined activation of positively weighted genes and the combined inactivation of negatively weighted genes to measure oscillation during development, with several important frequencies being tracked by the developmental processes of the organism. The fact that we could recreate this phenomenon for any desired frequency or phase may indicate a deeper pattern—that oscillatory timing of developmental processes as a simple function of gene activations may be a mechanism that is heavily relied upon for orchestration of development.

One possible explanation for the ability to generate oscillations of any phase and frequency lies in the fact that the different genes studied become active and inactive at different times during development (Supplementary Fig. S3). As long as the predominant timespans of the activity of different genes do not perfectly overlap, an appropriate weighting of genes could select a subset of genes that demonstrate activation during the peaks of a sine wave, and do not demonstrate activation during the troughs of a sine wave, as a function of time. This raises a ‘chicken or egg’ question about whether sinusoidal oscillations may control gene expression, or whether gene expression collectively gives rise to sinusoidal oscillation. Conceivably both processes occur, and are interrelated. This question deserves further research.

4.6 Genes differentially expressed between E lineage and other cells

Supplementary Table S1 lists the gene weights required to produce maximum class separation in the projection of the gene expression of cells onto the axis of maximum class separation, $\boldsymbol{\omega}$, for the E lineage versus all other cells. Gene weights that maximally separate each major cell lineage from the other cells are given in Supplementary Table S7. Note that *end-1*, the top-weighted gene in the FDA weightings for the E lineage, is well known as an important developmental gene for the E lineage (the intestine) (Robertson *et al.*, 2014). Heat shock-driven expression of END-1 has been found to cause a majority of embryonic cells to express intestine-specific genes and form intestinal structures (Zhu *et al.*, 1998). It has also been discovered than mutations in the second-highest ranked gene, *nob-1*, along with *pgp-3* (ranked in the top 25% of gene weights for the E lineage), both posterior-group *Hox* genes, results in gross posterior embryonic defects (Van Auken *et al.*, 2000). However, not all of the highly weighted genes have been previously linked to intestinal development in *C.elegans*. More importantly, the *negatively weighted* genes for the E lineage or most other developmental processes, body structures or phenotypic traits, have rarely been examined. This table raises the question as to whether strongly negatively weighted genes such as *isw-1* (chromatin-remodeling complex ATPase chain) are specifically not expressed in the E lineage—but similar questions could be asked about negative gene weights for each cell lineage, or each cell attribute, tested using FDA.

Of note, *elt-1*, which has been identified as a master regulator of epidermis specification, 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 (Supplementary Table S6). This is consistent with the observation that these two lineages rely on different developmental regulators (Shao *et al.*, 2013). 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* (Chen *et al.*, 2004), *pax-3* (Thompson *et al.*, 2016) and *tbx-9* (Andachi, 2004), confirming the utility of this FDA method for 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. The same is true of the highly weighted genes for each of the major lineages—some but not all of the highly positively weighted genes have documented roles in the development of these lineages, and, importantly, most of the strongly negatively weighted genes have not previously been identified as specifically being inactive only in a given lineage.

4.7 Insights into roles of genes in development

We next sought to identify overall patterns in the weightings of each gene in the FDA results, given known functions of each gene. To this end, we cross-linked the FDA results with the Wormbase (Stein *et al.*, 2001) GO terms for each of the 102 genes under study. Given a set of gene weights, we looked up the GO terms for each gene, and contributed the weight of the corresponding gene into an accumulator for the GO term. The sum of the gene weights for all genes associated with each GO term, for the E lineage FDA weightings, are presented in Supplementary Table S2. These aggregate weightings give an idea of which biological processes are more characteristically active during intestinal development compared to the development of the rest of the organism. There are RNA polymerase II terms at both the high positive and low negative ends of the scale, which could be related to inconsistent application of GO terms to genes in the GO database, or the inconsistent application of multiple related and similar but differently-coded terms. Either way, there are numerous strong developmental signals at the positive and negative ends of the scale, as would be expected, and specifically, ‘endodermal cell fate specification’ is predictably highly weighted for gut development (the E lineage). Other terms stand out as interesting, such as ‘nematode male tip morphogenesis’—indicating that development of the male tail tip is coordinated with development of the gut, or that development is regulated by some of the same genes in both cases.

Note that Figure 4 and Supplementary Tables S1 and S2, constitute the FDA results for just one attribute, where Class 1 is comprised of cells in the E cell lineage and Class 0 is comprised of all other cells in the cell pedigree. For comparison, we also applied FDA analysis to each of the other major lineages in *C.elegans* (AB, MS, C and D) versus the other cells in the pedigree (also visible in Fig. 4), and found that all the major lineages were cleanly separable from cells not in that lineage, indicating that each lineage had a distinctive gene expression profile. We also tested a couple of hundred other cell attributes, derived from Wormbase anatomy ontology terms, including tissue type and cell function, as well as division depth, and other attributes. The full set of FDA result can be seen in Supplementary Table S6.

We also applied the FDA technique to a number of cell features, including anatomy ontology terms from Wormbase, as well as developmental stage, tissue types, etc. For each, we generated a figure indicating the separation of cells that were labeled with the trait from cells that were not labeled with the trait, as well as FDA gene weightings and GO term weightings for each FDA result (Supplementary Table S6). We also produced variants of each case, for cells that were parent cells or daughter cells of cells that possess each given trait. In each case, we measured the ratio of inter-class to intra-class variance, as a measure of separability of the two classes. This ratio is given in the filename of the FDA results.

4.8 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 (Supplementary Table S4b and Fig. 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 value, and the other genes will be close to zero. However, the distribution of PC1 weights in Supplementary Figure S4 demonstrates that the contribution toward 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 Supplementary Figure S4 to the gene expression patterns in Supplementary 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 Supplementary Figure S3 suggests that the situation is not as simple as the genes being switched on in a specific ordered sequence.

4.9 Possible conflation of similarity in cell function with co-temporality

In a wide array of gene expression research, in order to gain insights into the functioning of genes, cells have been clustered according to similarity in their gene expression profile [e.g. Liu *et al.* (2009)]. However, Figure 1 illustrates a possible important 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. This is true at least of the dataset we studied: the greatest source of variance in gene expression that we observed (PC1 in our PCA results) was *time-correlated*. This implies that, at least for the genes selected for this dataset, *differential gene expression is more strongly affected by developmental age than it is by tissue-specific differences in expression arising from differentiation processes*.

Note however, as a caveat, that in correspondence, Murray *et al.* (2006) 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.

If however the above observations are generally true of larger sets of genes, across longer spans of the developmental timeline, then many previous research conclusions about cell clustering and cell similarity may need to be revisited in the light of how gene expression is globally coordinated across all cells at each stage of development, e.g. by the oscillatory processes we observed in PC3, PC6 and PC10, and in particular, by the highest-variance component, the monotonic progress of the gene expression manifold through gene expression space that we observed in PC1. The magnitude of the correlated variance observed in these principal components cannot be discounted in the analysis of cell similarity based on gene expression similarity alone.

4.10 Effects of gene selection and reporter mechanism on results

Gene activity data were 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 analysis. 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' (Murray *et al.*, 2006).

In 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.

5 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 versus 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 employed pervasively by developmental processes. We presented a number of novel techniques, and novel applications of existing statistical techniques to whole-organism developmental data, yielding novel insight into regulatory genes and regulatory control mechanisms. These techniques are broadly applicable to similar single-cell or whole-organism datasets. Our results highlight the value of the EPIC technique for continuously tracking gene expression over the developmental timeline, and the value of the published EPIC dataset, and of meta-analysis of previously published results in general.

Acknowledgements

We thank Sumaiya Nazeen and Hoon Cho for helpful comments and suggestions.

Funding

L.H. and B.B. were partially supported by NIH R01GM081871 [to B.B.].

Conflict of Interest: none declared.

References

Anavy,L. *et al.* (2014) BLIND ordering of large-scale transcriptomic developmental timecourses. *Development*, **141**, 1161–1166.

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

Araya,C.L. *et al.* (2014) Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature*, **512**, 400–405.

Bao,Z. *et al.* (2006) Automated cell lineage tracing in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA*, **103**, 2707–2712.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, **157**, 714–725.

Boyle,T.J. *et al.* (2006) AceTree: a tool for visual analysis of *Caenorhabditis elegans* embryogenesis. *BMC Bioinformatics*, **7**, 275.

Briggs,J.A. *et al.* (2018) The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, **360**, eaar5780.

Cao,J. *et al.* (2017) Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *Science*, **357**, 661–667.

Chen,X. *et al.* (2018) From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Ann. Rev. Biomed. Data Sci.*, **1**, 29–51.

Chen,Z. *et al.* (2004) The *Caenorhabditis elegans* nuclear receptor gene *nhr-25* regulates epidermal cell development. *Mol. Cell. Biol.*, **24**, 7345–7358.

Cho,H. *et al.* (2018) Neural data visualization for scalable and generalizable single cell analysis. *Cell Syst.*, **7**, 185.

Cooke,J. (1975) Control of somite number during morphogenesis of a vertebrate, *Xenopus laevis*. *Nature*, **254**, 196–199.

Dale,K.J. and Pourquié,O. (2000) A clock-work somite. *Bioessays*, **22**, 72–83.

Davis,K.L. *et al.* (2012) Single cell trajectory detection orders hallmarks of early human B cell development. *Blood*, **120**, 1044.

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

Domany,E. (2014) Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer Res.*, **74**, 4612–4621.

Farrell,J.A. *et al.* (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, **360**, eaar3131.

Giurumescu,C.A. *et al.* (2012) Quantitative semi-automated analysis of morphogenesis with single-cell resolution in complex embryos. *Development*, **139**, 4271–4279.

Gorban,A.N. and Zinovyev,A.Y. (2009) Principal graphs and manifolds. In: Olivas, E.S. (ed.) *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, Pennsylvania, USA, pp. 28–59.

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

Hench,J. *et al.* (2015) The homeobox genes of *Caenorhabditis elegans* and insights into their spatio-temporal expression dynamics during embryogenesis. *PLoS One*, **10**, e0126947.

Hendriks,G.-J. *et al.* (2014) Extensive oscillatory gene expression during *C. elegans* larval development. *Mol. Cell*, **53**, 380–392.

Hie,B. *et al.* (2019) Panoramic stitching of heterogeneous single-cell transcriptomic data. *Nat. Biotechnol.*, **37**, 685.

Kim,Dh. *et al.* (2013) Dampening of expression oscillations by synchronous regulation of a microRNA and its target. *Nat. Genet.*, **45**, 1337–1344.

Lee,M.T. *et al.* (2014) Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.*, **30**, 581–613.

Liu,X. *et al.* (2009) Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell*, **139**, 623–633.

Lorthongpanich,C. *et al.* (2012) Developmental fate and lineage commitment of singled mouse blastomeres. *Development*, **139**, 3722–3731.

McCulloch,K.A. and Rougvie,A.E. (2014) *Caenorhabditis elegans* period homolog *lin-42* regulates the timing of heterochronic miRNA expression. *Proc. Natl. Acad. Sci. USA*, **111**, 15450–15455.

Moore,J.L. *et al.* (2013) Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. *Development*, **140**, 3266–3274.

Murray,J.I. *et al.* (2006) The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree. *Nat. Protoc.*, **1**, 1468–1476.

Murray,J.I. *et al.* (2008) Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods*, **5**, 703–709.

Murray,J.I. *et al.* (2012) Multidimensional regulation of gene expression in the *C. elegans* embryo. *Genome Res.*, **22**, 1282–1294.

Nair,G. *et al.* (2013) Gene transcription is coordinated with, but not dependent on, cell divisions during *C. elegans* embryonic fate specification. *Development*, **140**, 3385–3394.

Nicholas,H.R. and Hodgkin,J. (2009) 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. *Dev. Biol.*, **329**, 16–24.

Palmeirim,I. *et al.* (1997) Avian *hairy* gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell*, **91**, 639–648.

Qiu,X. *et al.* (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14**, 979–982.

Ramsköld,D. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777.

Reinhart,B.J. *et al.* (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.

Richards,J.L. *et al.* (2013) A quantitative model of normal *Caenorhabditis elegans* embryogenesis and its disruption after stress. *Dev. Biol.*, **374**, 12–23.

Robertson,S. *et al.* (2014) Uncoupling different characteristics of the *C. elegans* E lineage from differentiation of intestinal markers. *PLoS One*, **9**, e106309.

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

Shao,J. *et al.* (2013) Collaborative regulation of development but independent control of metabolism by two epidermis-specific transcription factors in *Caenorhabditis elegans*. *J. Biol. Chem.*, **288**, 33411–33426.

Stein,L. *et al.* (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **29**, 82–86.

Sulston,J.E. *et al.* (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.*, **100**, 64–119.

Thompson,K.W. *et al.* (2016) The paired-box protein PAX-3 regulates the choice between lateral and ventral epidermal cell fates in *C. elegans*. *Dev. Biol.*, **412**, 191–207.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

Treutlein,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Van Auken,K. *et al.* (2000) *Caenorhabditis elegans* embryonic axial patterning requires two recently discovered posterior-group *Hox* genes. *Proc. Natl. Acad. Sci. USA*, **97**, 4499–4503.

Wagner,D.E. *et al.* (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, **360**, 981–987.

Waterston R.H. *et al.* (2006) *Expression Patterns in Caenorhabditis (EPIC)*. http://epic.gs.washington.edu/ (28 January 2018, date last accessed).

Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.

Yan,L. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.

Zhao,Z. *et al.* (2010) A negative regulatory loop between microRNA and *Hox* gene controls posterior identities in *Caenorhabditis elegans*. *PLoS Genet.*, **6**, e1001089.

Zhu,J. *et al.* (1998) Reprogramming of early embryonic blastomeres into endodermal progenitors by a *Caenorhabditis elegans GATA* factor. *Genes Dev.*, **12**, 3809–3814.