# Novel metrics reveal new structure and unappreciated heterogeneity in C. elegans development

Gunalan Natesan[1], Timothy Hamilton[2], Eric J. Deeds[3,4,*], Pavak K. Shah[1,4,*]

1. Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA
2. Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA
3. Department of Integrative Biology and Physiology, University of California, Los Angeles, CA
4. Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, CA

***** Correspondence: deeds@ucla.edu (E.J.D.), pavak@ucla.edu (P.K.S.)

Abstract:

High throughput experimental approaches are increasingly allowing for the quantitative description of cellular and organismal phenotypes. Distilling these large volumes of complex data into meaningful measures that can drive biological insight remains a central challenge. In the quantitative study of development, for instance, one can resolve phenotypic measures for single cells onto their lineage history, enabling joint consideration of heritable signals and cell fate decisions. Most attempts to analyze this type of data, however, discard much of the information content contained within lineage trees. In this work we introduce a generalized metric, which we term the branch edit distance, that allows us to compare any two embryos based on phenotypic measurements in individual cells. This approach aligns those phenotypic measurements to the underlying lineage tree, providing a flexible and intuitive framework for quantitative comparisons between, for instance, Wild-Type (WT) and mutant developmental programs. We apply this novel metric to data on cell-cycle timing from over 1300 WT and RNAi-treated *Caenorhabditis elegans* embryos. Our new metric revealed surprising heterogeneity within this data set, including subtle batch effects in WT embryos and dramatic variability in RNAi-induced developmental phenotypes, all of which had been missed in previous analyses. Further investigation of these results suggests a novel, quantitative link between pathways that govern cell fate decisions and pathways that pattern cell cycle timing in the early embryo. Our work demonstrates that the branch edit distance we propose, and similar metrics like it, have the potential to revolutionize our quantitative understanding of organismal phenotype.

## Introduction:

The differentiation of cell types in the developing embryo depends on both cell autonomous processes and signaling from neighbors, diffusible cues, and mechanical forces. In metazoa, lineal history plays an important role in patterning many of these factors and thus in establishing the basic animal body plan[1]. The study of cell lineages in eutelic organisms, which possess a fixed number of somatic cells and thus exhibit stereotypical cell lineages, has been a powerful driving force in our understanding of fundamental developmental and biological processes[2,3]. Cell lineages in these organisms represent valuable scientific resources, providing a spatial and temporal index of the animal onto which multimodal measurements can be aligned[4]. Aligning measurements such as gene expression[5,6], chromatin accessibility[7], cell size and shape[8], and the effects of genetic perturbations[9,10] onto the *C. elegans* lineage has contributed to an increasingly holistic view of development. Advances in light microscopy and computer vision have dramatically expanded the reach of these approaches, and datasets are now available containing

measurements aligned to thousands of embryonic cell lineages[10,11]. The scale of these data poses interesting challenges for data exploration and analysis.

Cell lineages map intuitively to mathematical graphs, and the alignment of cell divisions along body axes in *C. elegans* allows for unambiguous names to be assigned to each cell produced from a division[3]. This allows *C. elegans* cell lineages, and those of any other eutelic species, to be considered as *ordered binary trees*, a type of graph that allows straightforward one-to-one alignments to be made between any pair of lineages within or between individual embryos. This property dramatically simplifies the application of metrics computed on lineage trees, since a single unique value can be calculated for each comparison. While the inference of lineage relationships is common practice in the study of evolutionary relationships[12,13], the distinct problem in comparing phenotypic measurements aligned to lineages has been less extensively explored. Comparisons of the topology of cell lineages has been previously performed using the Robinson-Foulds distance and triplet distance, which each rely on the generation and comparison of sub-trees, accumulating a count of shared sub-trees between lineages normalized against the total number of possible sub-trees to arrive at a metric[14].

In this work, we developed several metrics that operate on the topology of lineages and applied them to the analysis of the *C. elegans* embryonic cell lineage. The first of these is the tree edit distance[15], which has previously been applied to the comparison of neuron morphology[16] and RNA secondary structure[17]. The tree edit distance allows one to quantify the topological changes in lineage between different embryos, different sublineages, or different experimental conditions. For ordered binary trees, the tree edit distance can be computed more efficiently and is more directly interpretable than previously used sub-tree-based metrics[15]. The second metric that we developed is the "branch distance," which measures the similarity between lineages based on quantitative measurements of properties of either cells or branches within the lineage. In this case, our particular focus was on the timing of cell division events in the lineage.

Both metrics represent intuitive notions of distance, which greatly aids their use in downstream analyses such as unsupervised clustering and hypothesis testing. We benchmarked these metrics using a published database of wild type and RNAi-perturbed *C. elegans* embryonic cell lineages[10]. Our analysis describes previously uncharacterized heterogeneity in wild type lineages and in the phenotypic consequences of RNAi variability on developmental timing. We extended these measurements of lineage-specific timing to detect patterns of similarity in the anterior cell lineages of the embryo and identify a previously unappreciated role of Notch signaling in the control of developmental timing. Finally, we apply this approach to a systematic analysis of RNAi perturbations that result in cell fate transformations where we find that, while developmental timing appears to be highly sensitive to genetic perturbation, RNAi against genes in a subset of important developmental regulators generate transformations that preserve lineage-specific developmental clocks.

## Results:

### *Defining metrics on spaces of cell lineages*

Multicellular organisms develop from a single cell through a sequence of divisions. The stereotypical nature of *C. elegans* development makes it possible to uniquely identify every cell in its somatic lineage based on the orientation of the division of its predecessor relative to the embryo's body axes[3]. This feature of *C. elegans* development has been a major advantage of its use as a model system, enabling systematic and quantitative studies of developmental processes. Here, we take advantage of the structured nature of the *C. elegans* cell lineage to represent it digitally as an ordered binary tree, such that nodes and edges represent cells and division events respectively. The cells in the embryo and the corresponding nodes in the tree can be labeled using the convention based on the orientation of cell divisions along body axes[3] and can

be associated with quantitative measurements on a cell-by-cell basis. This natural representation of the lineage as a binary tree suggests several straightforward metrics for comparing lineages to quantify how, say, a gene knockout impacts development (Fig. 1A). The first is the *tree edit distance*, which is derived from the graph edit distance in graph theory and is based on counting the minimum number of operations (such as adding or removing a node or edge) that is needed to convert one tree into another. Since *C. elegans* development is stereotypical, there is a natural alignment between any two trees based on the naming convention described above. This makes computing the tree edit distance very straightforward, essentially reducing the calculation to determining the number of nodes that are different between the two trees (Fig. 1B). This metric captures how perturbations like gene knockouts influence the *topology* of the lineage.

The second is a class of metrics that can be computed on any measurement associated with individual cells within the tree, such as gene expression or time. Again, here we use the stereotyped nature of *C. elegans* development to directly align any two lineage trees on a cell-by-cell basis, meaning that a single and unique distance can be calculated for any comparison between lineages. Given this alignment, any numerical property of cells during development can be unambiguously converted into a vector representation (Fig. 1C). This allows us to use any metric on such vectors to compare the trees. The most straightforward metric that might be used is simply the Euclidean distance (i.e. the $L^2$ norm). Here we focus on the application of the $L^2$ norm to comparisons of cellular division timing between embryos, which has been shown to vary under genetic perturbation[18], and is a measurement produced by every method for lineage tracing by cell tracking. We call this metric the "*branch distance*" since these division timings represent the length of the branches in the tree (Fig. 1C). Of course, a gene knockout or mutation might impact *both* the topology of the tree and the timing of cell divisions. To account for this, we need to have a way for dealing with cases where a cell/node exists in one tree, but not in the other (Fig. 1C). Here, we define two different types of branch distance to account for this problem. The first is the *intersection branch distance*, where the vector of division timings is constructed only on the basis of cells that are shared between two lineages (i.e., they are in the *intersection* of the list of cells in the two) (Fig. 1D). The second is the *union branch distance*, where we simply set the cell cycle timing of any cell that is missing from any embryo to 0 and calculate the distance in the normal way (Fig. 1D). The union branch distance thus simultaneously includes information about differences in both timing and topology. As described below, these two metrics capture different aspects of variation between trees.

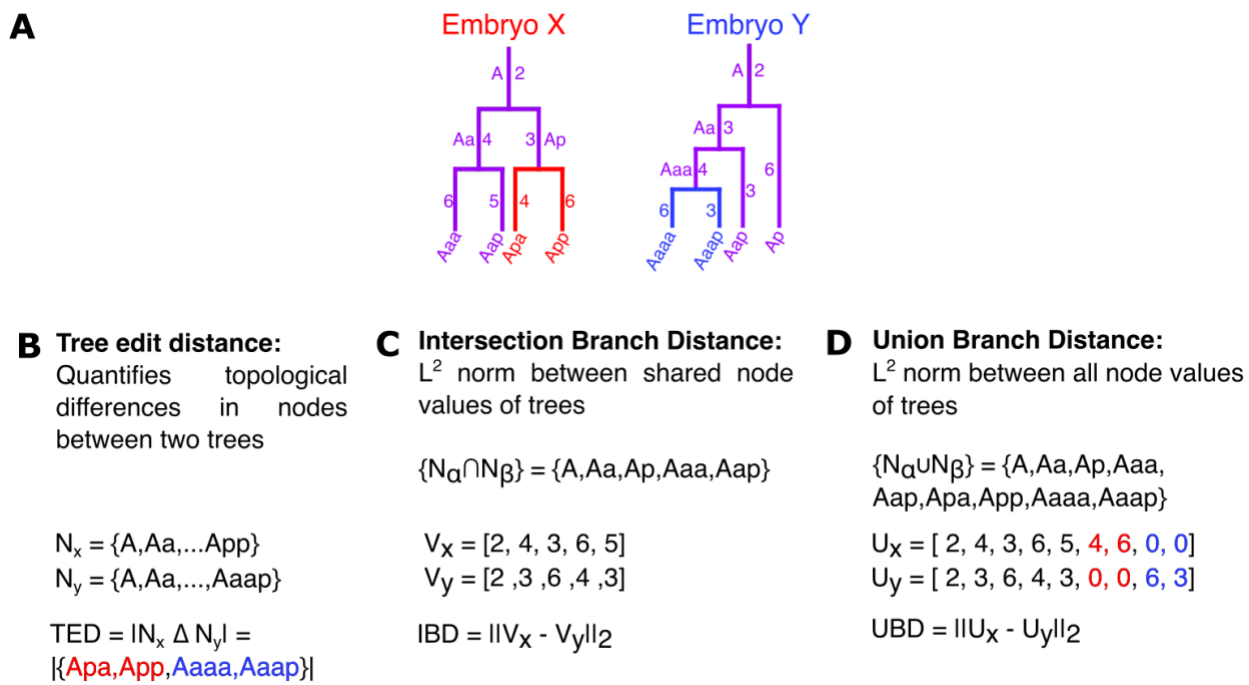



**Fig 1. Defining Distance Metrics on Lineage Based Tree Structures: (A)** Cell lineages can be expressed as binary trees, with parent, sibling, and child cells relationship reflected in node topology. *C. elegans* has a naming convention that allows for direct comparisons between cells in distinct lineages. Here we show schematics of two lineages with different topologies, corresponding to embryos "X" and "Y." The canonical names of the cells are shown either next to or at the end of the corresponding edges. Lineage tracing data also provides information about how long each cell persists between when it is "born" through division and when it divides itself. The numerical values next to each edge indicate these cell cycle times in this schematic. **(B)** The tree edit distance describes the topological differences between trees by counting the number of additive/subtractive operations required to transform one tree into another. In the case of *C. elegans* lineage trees, this corresponds to the size of the symmetric difference between the set of nodes present in one embryo vs. another. **(C)** The intersection branch distance is the Euclidean or $L^2$ norm between measurements associated with shared nodes or edges of trees, disregarding topological differences between trees by only considering nodes/edges present in both trees. **(D)** The union branch distance is the Euclidean or $L^2$ norm between values on the union set of nodes or edges between trees. Nodes or edges that are absent from one tree in any comparison are given a 0 value.

## *The branch distance reveals unexpected batch effects in WT embryos*

The most straightforward method for cell lineage tracing is via direct observation and cell tracking. Even in the absence of visible reporters, this approach inherently generates both spatial (3D cell positions and cell trajectories) and temporal (the timing of cell divisions) measurements of the embryo. While the spatial organization of the embryo has been extensively characterized[19–21], the distribution of cell cycle times within lineages have not been as deeply explored, making it an attractive target for analysis via our branch metrics. During lineage tracing experiments, every cell in developing embryos is tracked, and each cell division can be mapped to a particular time $t$, with $t = 0$ corresponding to, say, the first division of the zygote. There are thus two ways of thinking about the "branch length" value for each cell in the tree (Fig. 1C). In one scenario, we could label each cell with its "birth time," which is just the time $t$ at which the cell was generated through a division event. Another alternative is to consider the "cycle time" for that cell, which is just the length of time between when the cell is born through a division event until it divides itself.

Prior work has focused primarily on comparisons of cell birth times. In particular, Bao et al. showed that birth times for cognate cells from different embryos are highly correlated, with $R^2$ values that range between 0.995 and 0.997 (Fig. 2A).[18] While comparing birth times between

embryos seems natural, there is a potential issue with that approach. In particular, the birth time of a given cell is the sum of the cell cycle times of all the previous division events (Fig. 1C). Since there is some randomness in these cycle times, we can think of those times as

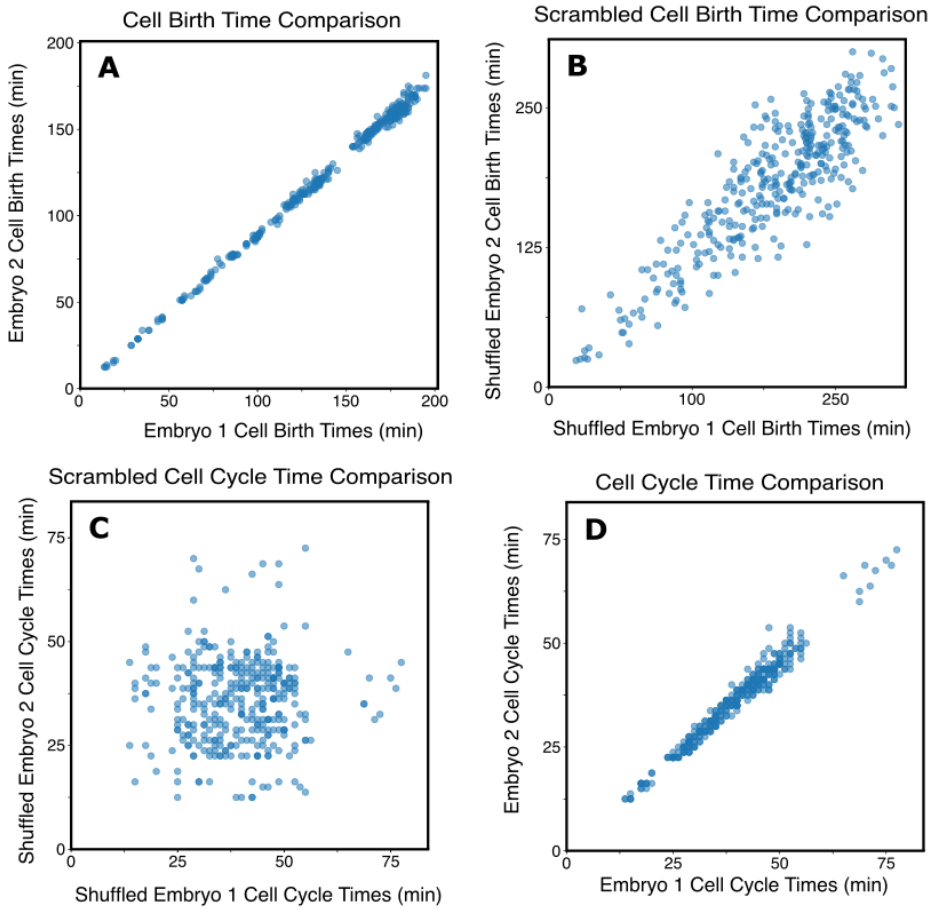



**Fig 2. Summation of Cycle Times into Birth Times Suppresses Variation. (A)** A comparison between the birth time of each cell (calculated as the sum of cell cycle times of each cell's ancestors) in two wild type *C. elegans* embryos. **(B)** Comparison between birth times calculated from two randomly shuffled wild type embryos, where each cell is assigned another random cell's birth time from within the lineage of the same embryo. Note that a significant correlation in birth times exists even in this shuffled data. **(C)** Comparison between shuffled cell cycle times rather than birth times. In this case, there is no correlation, as would be expected. . **(D)** Comparison between the cell cycle times of each cell in two wild type embryos. Note that the same two embryos were used for all comparisons in panels A-D.

random variables, noting that summing over random variables always reduces variation[22]. In other words, the "birth time" is essentially equivalent to averaging the previous cycle times, and averaging generally suppresses variation (i.e., the standard error of the mean is generally less than the standard deviation). In addition, because the birth times are a sum of previous cycle times, birth times for cells born later in development will always be larger than birth times for cells born earlier. Both effects can spuriously increase the correlation in cell birth times between embryos. To demonstrate this, we completely randomized the cell cycle times in the embryo, intentionally destroying any correlation in the length of cell cycles for the same cells across each randomized embryo (see Methods). After randomizing these cycle times, we found birth time correlations with $R^2$ values between 0.65 and 0.85 (Fig. 2B), despite a complete absence of correlation in the individual cell cycle times (Fig. 2C). Thus, while the cycle times are still highly correlated between WT embryos ($R^2$ between 0.97 and 0.99, Fig. 2D), the correlation is less than we observe with birth times. Since using the cycle time avoids spurious correlations and reveals

more variation and structure in the data (Fig. 2D), we focused on using the cycle time to calculate branch distances in this work.

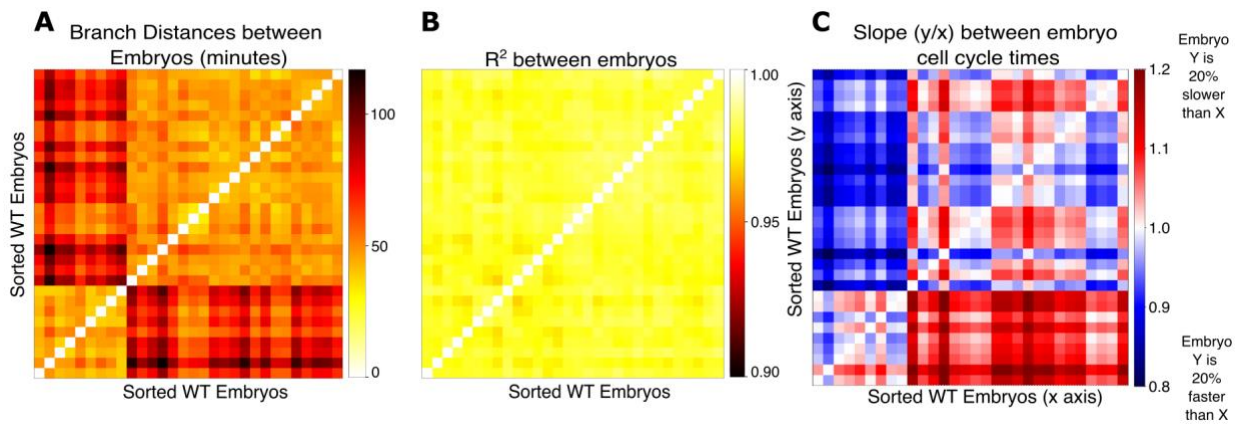



**Fig 3. The Branch Distance Reveals previously undetected Batch Effects in WT Embryo Cell Cycle Timing (A)** Heatmap showing the union branch distance calculated between each pair of wild type embryos in the dataset. The ordering of embryos was sorted based on their assignment to two clusters computed using hierarchical clustering. **(B)** Heatmap showing the $R^2$ in cell cycle times between all pairs of WT embryos, sorted as in (A). **(C)** The slope calculated between cell cycle times between all pairs of WT embryos, sorted as in (A).

We used the cycle times to calculate the branch distance between each pair of 30 WT embryos with cycle times taken from lineage tracing data from Du *et al.*[10] We then hierarchically clustered the embryos based on these distances. Surprisingly, clustering on the branch distance revealed two distinct populations of WT embryos in the published dataset (Fig. 3A). If we calculate the $R^2$ value between each pair of embryos, the two different clusters vanish (Fig. 3B). To understand how these embryos could be highly correlated and still cluster into two groups based on the branch distance, we considered not just the correlation of the birth time relationship, but also its *slope* $m$ (see Methods). Intuitively, this slope quantifies the systematic variation in developmental timing between two embryos and can be thought of as the slope of the best-fit line to the data in Figure 2D. When this slope $m = 1$, it means that on average, each cell has similar cycle times between the two embryos; if the slope $m < 1$, that means that on average, cells in the embryo on the x-axis have cycle times that are systematically *longer* than cells in the embryo on the y-axis (Fig. 3C). We can interpret changes in this slope as representing changes in the relative "global clock" that times cell divisions in the two embryos. We calculated this slope for each pair of WT embryos in the data (Fig. 3C). Note that here the embryo on the x-axis of the heat map is also used for the x-axis of the slope calculation. In this case, we used Principal Component Analysis (PCA) rather than linear regression to estimate the slope, since in any given comparison between embryos the choice of dependent vs. independent variable would be arbitrary (see Methods). These slope calculations, along with the high correlations in Fig. 3B, indicates that the primary difference between these two groups of embryos is indeed the global rate of development. In particular, "Cluster 1" embryos develop systematically slower than "Cluster 2" embryos. This analysis exemplifies how the branch distance can reveal systematic differences in the data, such as this batch effect, that a focus on correlations cannot identify (Fig. 3).

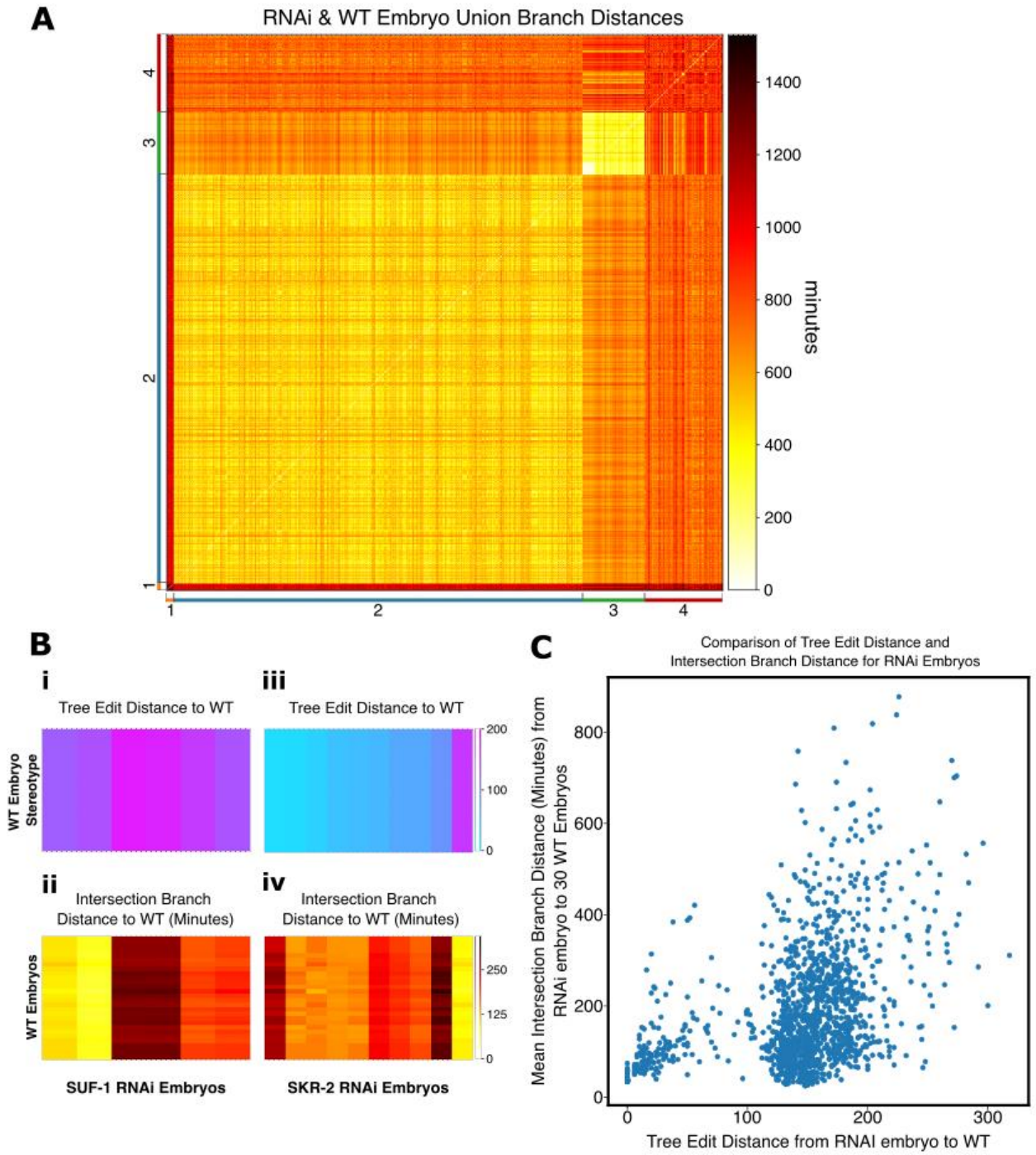



**Fig 4. The union branch distance reveals Heterogeneity in RNAi Cell Cycle Timing Coordination (A)** Heatmap showing the union branch distance between all 30 WT embryos and 1322 RNAi Embryos in the dataset. Embryos were hierarchically clustered and sorted into 4 clusters shown along the axes of the heatmap. **(B) i.** Distribution of the tree edit distance between 6 SUF-1 embryos and 30 WT embryos. **ii.** Distribution of the intersection branch distance between 6 SUF-1 embryos and 30 WT embryos. **iii.** Distribution of the tree edit distance between 10 SKR-2 embryos and 30 WT embryos. **iv.** Distribution of the intersection branch distance between 10 SKR-2 embryos and 30 WT embryos. **(C)** Comparison between the tree edit dIstance and intersection branch distance for each of the 1322 RNAi embryos relative to a single WT reference embryo.

*The branch distance reveals heterogeneity between RNAi replicates*

We then computed the tree edit and branch distances between the 1352 embryos treated with RNAi against 204 genes described by Du *et al*[10]. We hierarchically clustered these embryos based on the union branch distance into 4 major groups (Figure 4A), where the number of partitions was decided by analysis of the union branch distance dendrogram (Supplemental Figure 1). Of these, 2 clusters shown in the upper right corner and lower left corner of Figure 4A likely represent many outliers, as these embryos are approximately as different from one another as they are from the other 2 groups. Even among the remaining 2 clusters, we observe a significant degree of heterogeneity (Supplemental Figure 2). This heterogeneity exists not just between embryos treated with RNAi against different gene targets, but also between embryos treated with RNAi against the same gene (Figure 4B). The examples in Figure 4B highlight just two patterns that we observed. In the case of embryos treated with RNAi against *suf-1*, three pairs of embryos exhibit distinct levels of divergence from wild type lineage topologies (as indicated by the tree edit distance, Figure 4Bi) as well as from wild type patterns of cell cycle timing (as indicated by the branch distance, Figure 4Bii). RNAi against *skr-2*, on the other hand, induces minor defects in lineage topology (Figure 4Biii) but a broad spectrum of defects in the distribution of cell cycle times (Figure 4Biv). Surprisingly, this variability isn't a simple manifestation of variable phenotypic severity, as these embryos often differ from one another as much as they differ from wild type (Supplemental datasets 1 and 2, Supplemental Figure 2). Prior work has demonstrated that some mutant phenotypes manifest variable penetrance due to underlying variation in endogenous gene expression[23], which may contribute to the variability we observe in combination with embryo-to-embryo variation in RNAi penetrance.

Our findings in Figures 4A and B show that perturbations through RNAi can impact both lineage topology and the timing of cell cycle events. To separate these effects, we chose a single representative WT embryo from Cluster 1 (Figure 3A) and used this embryo to calculate both the tree edit distance between each RNAi embryo and WT and the intersection branch distance between each RNAi embryo and WT. We chose to focus here on the intersection branch distance because it focuses on just the duration of the cell cycle events among cells that are present in both the WT and RNAi-treated embryos; the union branch distance reflects both changes in timing and topology (Figure 1). In Figure 4C, we plot the tree edit distance to a WT reference embryo for each RNAi-treated embryo on the x-axis, and the intersection branch distance to WT on the y-axis. We use a single WT reference embryo, selected from Cluster 2 among the WT embryos since the inter-embryo branch distance between WT embryos is much smaller than the distances between any WT embryo and any RNAi embryo, and all WT embryos possess an identical topology such that the Tree Edit Distance between them is always 0 in this dataset. It is immediately clear that there is a bimodal distribution of tree edit distances, with a smaller subset of RNAi embryos having WT-like lineage topologies (with tree edit distances near 0) and most RNAi perturbations having a large impact on the structure of the lineage. Interestingly, we see that there is a general lack of correlation between tree edit distance and intersection branch distance, indicating that some RNAi perturbations have a large impact on topology, but the duration of the cell cycle is similar to WT amongst lineages with preserved topologies, while other perturbations leave the topology of the lineage almost intact but have a relatively large impact on cycle duration (Figure 4C).

We then examined whether RNAi against genes with related functions generated phenotypes of similar severity based on our graph metrics. We grouped RNAi embryos together based on their functions as annotated by Du *et al.*[10] and observed a weak correlation between tree edit distance and intersection branch distance relative to WT (Supplemental Figure 3) although for most groups of genes intra-class variability is greater than inter-class.

A*pplication of the branch distance to sublineages in WT and RNAi embryos*

In all the work above, we applied our metrics to the cell lineage of entire embryos. While informative, this approach ignores the fact that certain developmental processes are specific to certain sublineages and might be lost in a global analysis. For instance, previous work on developmental timing in *C. elegans* focused on cell-by-cell comparisons and found that while cell birth times were globally well correlated[18], the specific ordering of cell divisions within the AB lineage was variable[24]. Given these observations, we wondered whether any structure existed in the distribution of cell cycle durations within the sublineages of each individual embryo.

We computed the intersection branch distance, which does not reflect differences in lineage topology, between each pair of canonical founder lineages in the early *C. elegans* embryo (Figure 5A). Cells in the *C. elegans* embryo are named based on their lineage history. A few founding cells in the early embryo possess unique names, but all cells derived from these are incrementally named according to the body axis it was born along. Cells with names containing the same number of characters following the unique name of the originating cell in the early embryo are thus born in the same generation of cell divisions and cells whose name only differs in the last character are siblings. The distribution of the intersection branch distance between each of the major differentiated lineages of the embryo show consistent patterns across the wild type samples that match the intuitive prediction that the posterior mesodermal and endodermal lineages derived from P1 are quite different from each other and from the AB lineage. We also found the same patterns reflected in the union branch distance which is sensitive to differences in topology (Supplemental Figure 4). We were surprised to find consistent and statistically significant differences (Supplemental Figure 5) between every pair of lineages derived from AB except for ABplp and ABprp. ABa and ABp derived lineages show distinct patterns of similarity, with the two lineages rooted at ABpl and ABpr being closer to one another than ABal and ABar, reflected in the left/right symmetric pattern of similarity in the lineages rooted at ABpxx and the lack of any such symmetry in ABaxx lineages. We were thus interested in identifying the origin of these patterns in cell cycle timing among the AB-derived sub-lineages.

*Notch signaling modulates cell cycle duration in a lineage-specific manner.*

Since Notch is responsible for breaking fate symmetry between the ABa and ABp lineages in the 4-cell embryo[25,26] and the pattern of branch distance distributions among the wild type AB lineages align with known Notch signaling events in the early embryo[27] (Figure 5B,C), we were interested in whether this pattern might be generated by Notch signaling. In *glp-1* RNAi embryos, the structure visible in the intersection branch distance between AB sublineages in wild embryos is clearly lost (Figure 5D). At the $AB_4$ stage, the two left/right symmetric lineages produced by ABp, which received a Notch signal from P2, are closer to one another by the branch distance than the two lineages produced by ABa are to each other (Figure 5E). Lineages derived from cells that independently receive Notch induction (ABalp and ABara by MS) also have a smaller branch distance between each other than between their direct siblings (Figure 5E). Embryos treated with RNAi against *glp-1* lose these differences and *glp-1* RNAi produces a nearly uniform pattern of branch distances among the AB lineages. Both ABp derived lineages in the wild type and all AB derived lineages in *glp-1* RNAi embryos exhibit a distinct pattern where left/right homologous lineages are closer to each other based on the branch distance than sibling lineages (Figure 5E). Perhaps Wnt signaling, which has been shown to incrementally accumulate in the posterior child of each cell division[28], continues to act to break fate symmetries between sibling cells in *glp-1* depleted embryos. Does the decreased intersection branch distance between Notch-stimulated sublineages represent a consistent effect on the duration of the cell cycle or increased variability between cognate cells across sub-lineages? To answer this, we compared the overall clock speeds of the AB8 lineages, indicating that Notch affected lineages were faster compared to unstimulated lineages (Supplemental Figure 6) suggesting a consistent impact on cell cycle

durations within sub-lineages, even when the stimulated sublineages are derived from separate founder cells that independently receive Notch stimulation.

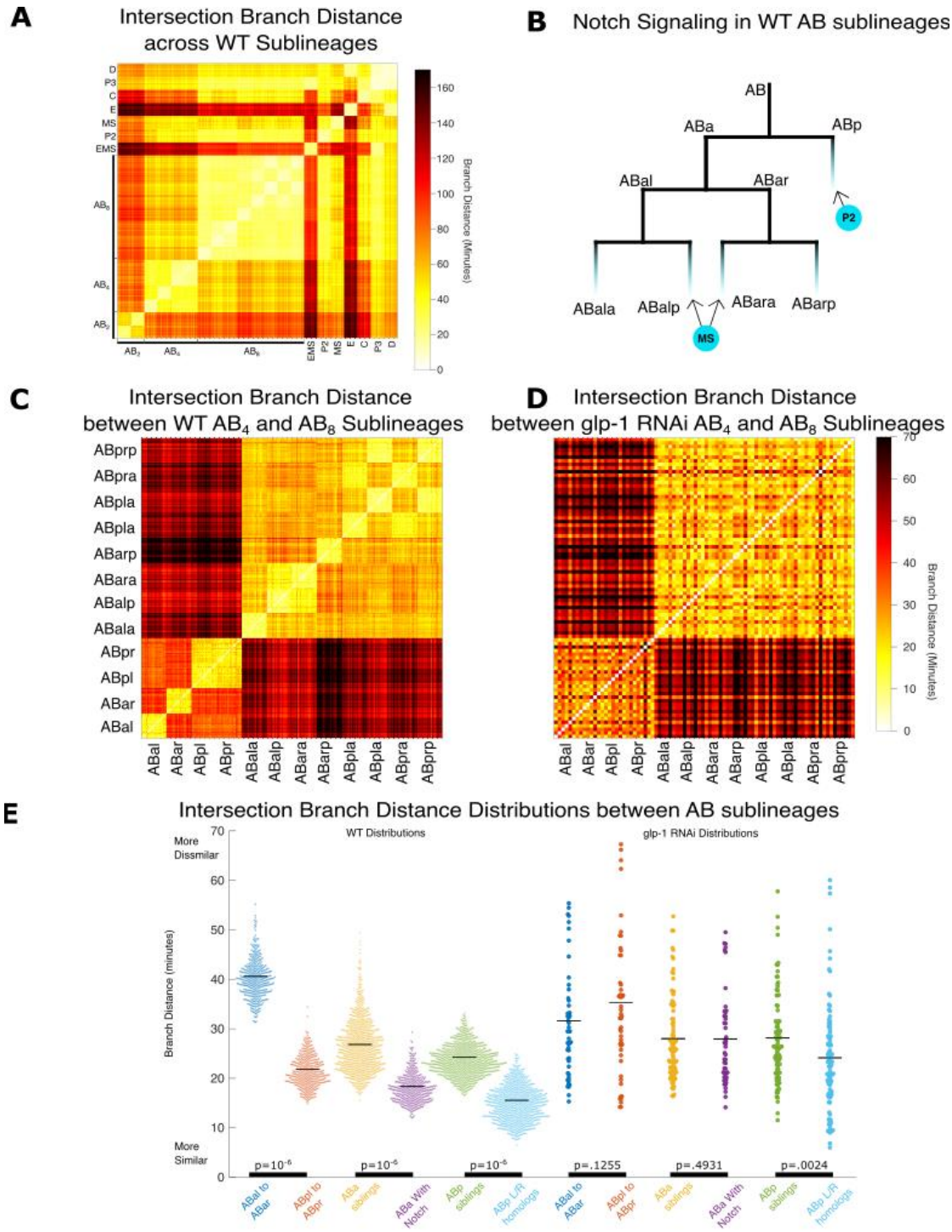



**Fig 5. Branch Distance reveals structure in the AB lineage. A)** Heatmap showing the intersection branch distance between every pair of sublineages in every pair of 22 wild type embryos. **B)** Illustration of the first two Notch signaling events in the early AB lineage. **C)** Heatmap showing a zoomed in view of the intersection branch distance between the 22 wild type embryos for each pair of AB-derived sublineages. Colormap is scaled from 0 to the max intersection

branch distances between same-generation AB sublineages. **D)** Heatmap showing a zoomed in view of the intersection branch distance between 6 embryos treated with RNAi against *glp-1* for each pair of AB-derived sublineages. Colormap is scaled from 0 to the max intersection branch distances between same-generation AB sublineages. **E)** Distributions of intersection branch distances between subsets of AB-derived sublineages in WT embryos and embryos treated with RNAi against glp-1. P-values calculated using $10^6$ iterations of a permutation test.

## *Preservation of cell cycle timing structure through lineage transformations*

A key process in the early embryo is the differentiation of cell lineages needed for the formation of different organs and body parts. The large-scale RNAi screen performed by Du *et al.* systematically explored these phenomena using genetic markers of tissue fate[10]. They characterized diverse genes whose depletion results in homeotic transformations, where some cell lineages adopt the pattern of tissue fates normally produced by another lineage, and genes whose loss results in patterns of tissue fates not normally seen in any wild type lineage. Most of the lineages in the wild type embryo have both qualitatively and, as we showed above in Figure 5, quantitatively distinct patterns of cell cycle times.

We wondered whether patterns in cell cycle timing are a product of the same differentiation processes that define the tissue fate of these lineages. In other words, in an embryo where one lineage adopts the fate of another, does the pattern of cell cycle lengths in the transformed lineage change to match the pattern of the newly acquired fate? We designed a heuristic based on the branch distance to search for cases where this is true. For each homeotically transformed lineage identified by Du *et al.* we refer to the transformed lineage as the *origin* and the acquired lineage fate as the *destination*. To account for natural variation in each of the wild type lineages, we first define a diameter $D$ equal to the maximum pairwise branch distance between wild type examples of the *destination* fate (Figure 6A). We then assigned a transformation score to each *origin* lineage based on how many of the wild type *destination* lineages lie within $D$ minutes of the *origin* sub-lineage in any particular RNAi embryo (Figure 6B). While most lineage transformations do not adopt the pattern of cell cycle times normally expressed by the acquired fate, we identified 99 cases where they do. Interestingly, cases where the transformed lineage falls within the neighborhood of all 22 wild type examples of the destination fate are more common than cases where the transformed lineage is proximal but not fully overlapping the neighborhood of the destination fate.

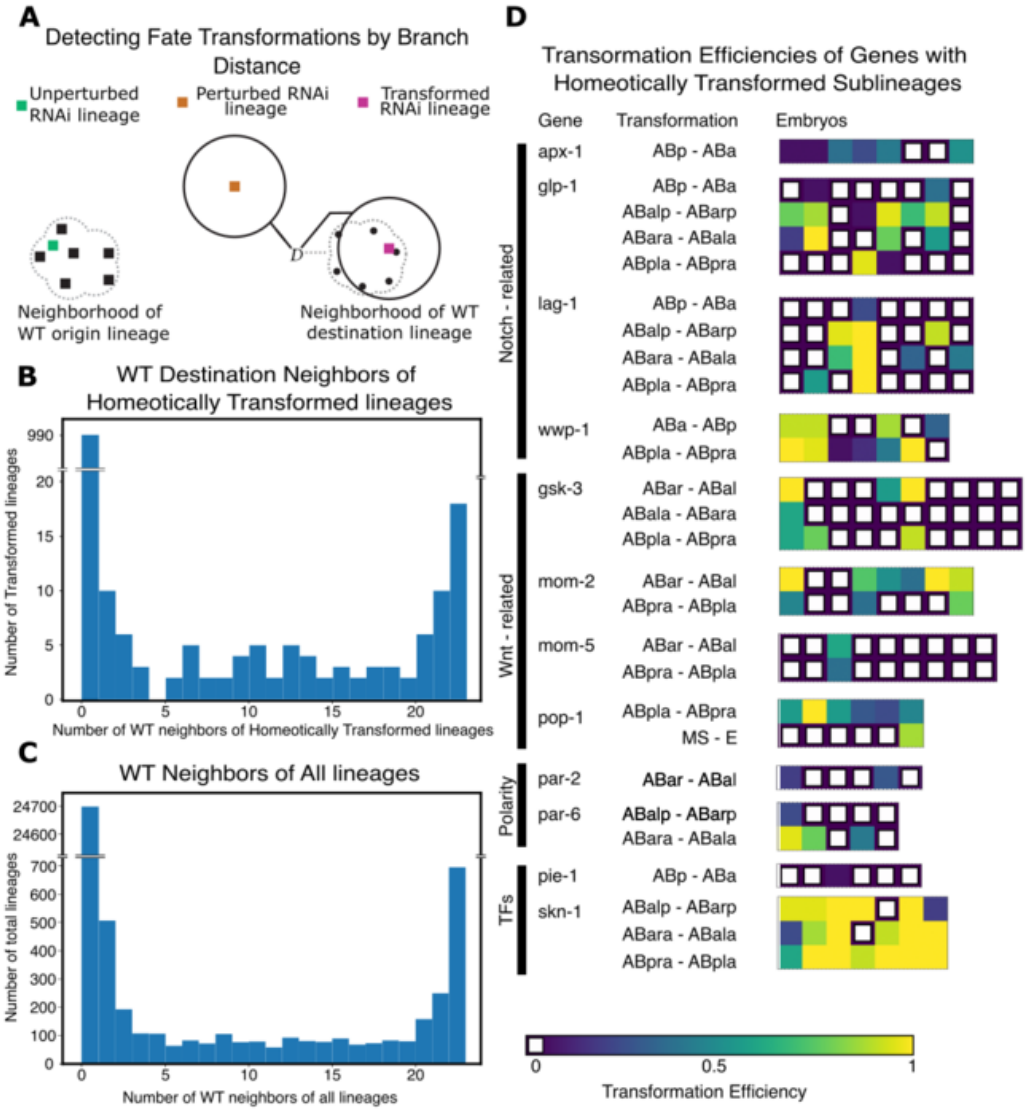



**Figure 6. Evidence for cell fate control over cell cycle timing. A)** Illustration of the transformation heuristic. For each WT destination lineage (black dots) a diameter *D* is calculated as the maximum intragroup intersection branch distance. The transformation efficiency is then defined as the fraction of WT destination lineages that fall within diameter *D* of each RNAi origin lineage (colored squares). In some cases, this value is 0 but overlaps the WT origin lineage neighborhood (green square) suggesting that the RNAi perturbed lineage maintained its original fate, at least in terms of cell cycle timing. In some cases, this value is 0 and the RNAi origin lineage has no WT neighbors (orange square) suggesting that the RNAi perturbed lineage has both lost its original fate and failed to acquire the pattern of cell cycle timing of the destination lineage. In a minority of cases the RNAi origin lineage is within *D* of 1 or more WT destination sublineages and a transformation efficiency is reported (magenta square). **B)** Histogram of the number of WT destination neighbors that homeotically transformed RNAi lineages have, using the heuristic defined in **A**. **C)** Histogram of the number of new WT neighbors that perturbed RNAi lineages have. **D)** Heatmaps representing the transformation heuristic in **A.** for homeotically transformed lineages with at least 1 WT destination neighbor. The genes that induce these transformations and functions are listed alongside the corresponding heatmap of transformation.

One advantage of using cell cycle timing and the branch distance as a phenotypic marker of lineage identity is that it can be assessed even in the absence of visible markers of cell fates, so we generalized our approach to measure the frequency of transformations between all possible pairs of lineages (Figure 6C). In this case, we find the nearest neighbor among possible

*destination* fates for each *origin* lineage and count the number of wild type examples of the *destination* that fall within $D$ minutes of the *origin* lineage in the RNAi treated embryo. For genes with or without homeotic transformations identified on the basis of marker gene expression by Du *et al.*, the majority of *origin* lineages fall outside the range of variation of all wild type *destination* lineages, suggesting that the patterns of cell cycle times in the wild type lineage are very sensitive to genetic perturbation. In both figures, the two most populated bins are the bin with 0 WT neighbors followed by the bin with 22 WT neighbors. (Figure 6B, 6C) noting a region of high density around WT neighborhoods (Supplemental Figure 7). Still, we identified 12 genes for which RNAi generates homeotic transformations based on both marker expression by Du *et al.* and our approach of using the branch distance. These genes belong to a small set of key pathways that have well-known roles in specifying cell fate in the early embryo including Notch, Wnt, PAR polarity genes, and the maternally derived transcription factors *pie-1* and *skn-1* (Figure 6D). These pathways operate to break symmetry in the early embryo. This set is likely an underestimate since we have not accounted for two common types of perturbations to cell cycle timing and lineage structure: changes to the "global clock" (since the branch distance is not scale invariant), and the partial transformation of lineages since we examined only lineages rooted at the major founder cells in the early embryo.

Interestingly, examining each RNAi lineage on an embryo-by-embryo basis reveals striking diversity in penetrance and phenotypic consistency in homeotic transformations detected by Du et al. (Figure 6D) and previously uninvestigated lineages (Supplemental Dataset 3). Several genes, such as *wwp-1*, *pop-1*, and *skn-1*, have sublineages that are within 1 diameter of their original and acquired fates, suggesting a degree of mixture in the neighborhoods of the third-generation descendants of the AB lineage. Furthermore, the variance in the number of embryos transformed and relative strength of each transformation suggests that RNAi penetrance and phenotypic severity are separable phenomena. For example, Notch pathway components *apx-1*, *glp-1*, and *lag-1* all have transformations from ABp to ABa, where the transformed lineage lies within 1 diameter of a similar number of wild type examples. The more consistent presence of ABp to ABa transformations in *apx-1* might imply more complete penetrance of *apx-1* RNAi than that of *glp-1*, and *lag-1*, or that the knockdown of *apx-1* more consistently induces temporal transformations, suggesting a more central role for *apx-1* in fate specification for first-generation AB lineages. In contrast, for the E3 ligase *wwp-1*, the penetrance of the ABa to ABp transformation is both fairly high and occurs with a higher degree of transformation (as indicated by the fraction of wild type ABp sublineages within 1 diameter of the transformed lineage), or the transformation from MS to E in *pop-1* RNAi which is only observed in 1 embryo but is a perfect match for the neighborhood of all wild type E lineages, suggesting an extremely precise transformation of identity.

## Discussion

Analyses of the structure of lineages in biology have focused principally on the construction of phylogenies[12,13] and measurements of inter-node distances within individual trees[29]. This approach has been shaped by the requirements of taxonomic work, where no ground truth topology exists, and multiple measures of distance might be employed. Cell lineages, on the other hand, have clearly defined structure, and recent work has explored strategies for measuring differences in tree topology[14]. Recently, techniques from spectral analysis were applied to phenotypic measures aligned to cell lineages, including in *C. elegans,* but with an emphasis on characterizing these phenotypes in the context of lineages with variable structure[30]. Automated cell lineage tracing is an increasingly mature technology, having been applied to *C. elegans*[31], *Drosophila*[32], zebrafish[33], and mouse[34] development as well as to the study of lineage relationships in stem cell[35] and immune cell[36] culture systems. A limited set of metrics have been

applied to the comparison of cell lineages[14,30], in part driven by the unordered nature of cell lineages reconstructed in most organisms. In the case of *C. elegans* lineages, for which there are now public repositories containing measured lineages from thousands of wild type and perturbed embryos[10,11,19], the ordered and stereotypical nature of its somatic lineage removes the need to align the lineages, thus dramatically simplifying the application of graph-based approaches to the problem of quantitatively comparing lineage trees. We thus applied the intuitive graph-theoretic notion of the tree edit distance[15], and its extension in the branch distance, to dissect the structure of *C. elegans* embryonic lineage. These metrics allowed us to uncover previously unknown heterogeneity between populations of wild type embryos, to quantify the variability of RNAi-induced lineage phenotypes, and to shed light on key mechanisms of patterning in early embryogenesis. Most prior analyses of developmental timing in the *C. elegans* embryo used the time of a cell's birth relative to an early reference point (ex. the first division of the zygote) and correlation between the birth time of the same cell across embryos as a measure of developmental similarity. We showed here that these two choices mask heterogeneity present in previously published records of wild type development. We wondered then whether these effects, combined with 1-to-1 comparisons of timing between the same cell across multiple embryos may have obscured patterns in developmental timing across lineages within the embryo.

Using the tremendous volume of existing lineage data available from the work of Du *et al*[10], we sought to benchmark our metrics on wild type and RNAi-perturbed embryos and explore whether a lineage-centric view of developmental timing may reveal previously unappreciated patterns. It is well known that RNAi, especially by feeding in *C. elegans,* induces phenotypes with variable penetrance[37]. It has been shown that variable penetrance in mutants may occur due to underlying heterogeneity in gene expression[23]. Our graph metrics show that phenotypic variability under RNAi can exhibit a wide range of patterns of severity. This includes patterns that correspond to linear gradients of severity, multimodal distributions of phenotypes, and apparently random variation between individual embryos. These measurements show that, for many genes, RNAi induces variability between individual embryos that is often on par with the phenotypic distance between wild type and individual RNAi embryos.

Taking advantage of the ordered nature of cell divisions in the *C. elegans* embryo to align arbitrary pairs lineages within the embryo, we sought to characterize the structure of cell cycle duration in the wild type lineage. We were especially surprised to find reproducible patterns within the lineages derived from AB, the larger of two cells born from the first asymmetric division of the zygote. In *C. elegans,* the major patterns of cell fate that are established by intercellular Notch signaling are well known, and the pattern of branch distances between the AB-derived lineages we observed aligns perfectly to the first two Notch signaling events in the early embryo. RNAi against Notch/*glp-1* abolishes this structure, demonstrating that this pattern of cell cycle timing in the AB lineage is a product of Notch signaling. Lineages that receive Notch signals also exhibit on average shorter cell cycle lengths than lineages that do not (Supplemental Figure 6).

Biophysical parameters such as cell volume affect cell cycle duration[38,39], but genetic regulation of subtle differences in cell cycle timing may occur via many potential mechanisms. Du *et al.*[9,10] demonstrated using transcriptional reporters of tissue fate that the loss of any one of many genes essential for development can induce homeotic transformations between the major founder lineages in the early embryo[9,10]. We set out to determine, using the branch distance, whether developmental timing in transformed lineages is independent of lineage fate, is transformed along with fate, or is lost upon fate perturbation. We devised a simple heuristic to assess the proximity of RNAi-treated origin lineages to the wild type destination lineage that expresses the closest pattern of cell fates as defined by Du *et al.*[10] by counting the number of wild type examples of the destination lineage that are less than the maximum inter-wild type branch

distance away from the RNAi origin lineage. Using this conservative approach, which would fail to detect transformations in cases where the global embryo clock is altered or where subsets of individual lineages are transformed, we find that only a handful of genes (12 genes out of 204 characterize) induce homeotic fate transformations where developmental timing in the transformed lineages also transforms to match that of the newly acquired fate. This set is composed of genes in the Notch and Wnt pathways, two PAR polarity genes, and the maternally derived transcription factors *skn-1* and *pie-1.* The fact that most perturbations produce homeotic transformations generate patterns of cell cycle duration that match neither that of the original wild type lineage or the newly acquired fate suggests that lineage-specific developmental timing is likely quite sensitive to genetic perturbation. When homeotic transformations in cell cycle timing do occur, a perfect match to wild type lineages outnumbers incomplete matches suggesting that, despite its sensitivity to perturbation, the wild type patterns of cell cycle timing may represent stable states.

Our analysis demonstrates that the genetic identity of cell lineages can reproducibly and finely tune the distribution of cell cycle duration within cell lineages. It is interesting that the pathways that preserve lineage-specific developmental timing across homeotic transformations are known to play a critical role in cell fate specifications upstream of most tissue-specific transcriptional programs. It is thus likely that either a specific subset of factors downstream of fate regulators or finely tuned expression levels of tissue-specific genes are required for the proper patterning of cell cycle duration within lineages. Whether this tuning is itself a functional element of the developmental program remains unclear. Perturbations to key cell cycle regulators generate dramatic changes in cell cycle duration as well as homeotic fate transformations in *C. elegans*[40–43], and changes in the duration of the cell cycle of stem cells in other systems are correlated with specific cell fates such as in the generation of bipolar cells during retinal development[44]. Our results demonstrate a precise relationship between cell fate and developmental timing that motivates revisiting gaps in our understanding of links between cell cycle regulation and cell fate control. More broadly, our findings highlight the ways in which quantitative analysis of phenotypic similarity can reveal unexpected structure in animal development. In particular, the use of pairwise distance metrics applied to lineage-resolve metrics allows for an intuitive extension of notions of cell state and identity. Reducing these multidimensional data types using such intuitive measures of distance simplifies the application of common data exploration and visualization strategies.

Acknowledgements

This work was supported by R21DC019485 from NIDCD to PKS. The authors would like to thank Dr. Zhuo Du and Dr. Anthony Santella for guidance in parsing their lineage data, and Dr. Roy Wollman and Dr. Alex Hoffman for their feedback and advice.

## Methods:

*Data Availability and Preprocessing*

Lineage data used in this study was retrieved from www.digital-development.org/download.html, the digital data repository provided by Du *et al.*[10]

Wild type and RNAi treated embryonic lineage data was retrieved as text files with each row corresponding to an individual cell in the lineage tree. Lineage relationships were reconstructed from the cell names, which are structured according to a common convention where a unique root cell ID indicates the identity of the founder cell of the lineage and each subsequent pair of cells is named according to the body axis along which its division was polarized. Cell cycle duration was extracted based on the number of columns associated with each cell from the data provided by Du *et al.* used to report tissue specific transgenic reporter signal intensity. Each column represents the intensity measure made for each timepoint of imaging where the corresponding cell existed, thus we calculate the duration of each cell's cell cycle as 1.25 minutes per timepoint, based on the imaging frequency reported. A total of 30 wild type and 1322 RNAi treated embryos were retrieved and time-resolved lineage trees were generated from these raw data.

While all wild type embryos covered a uniform set of cells, RNAi treated embryos were only partially curated by Du *et al.* to validate reporter expression. In order to address these discrepancies, we truncated each of these lineage trees based on the time cutoffs provided alongside the raw data and wild type embryos were pruned similarly for distance calculations. The implementation of our data import and pre-processing is available alongside a complete codebase implementing our distance metrics and analysis routines at https://github.com/shahlab-ucla/graph_distances

*Graph Based Distance Metrics Applied to C. elegans Lineages.*

In order to store the lineage data for each embryo as a binary tree, we take advantage of the naming convention for each cell using a standard hash table (or "dictionary" in python) data structure. A cell would be stored in the dictionary with name/reference X (i.e. the cell's name) and an element value representing its cell cycle time. The children of X would have the name/reference of X followed by suffix 'a', 'l', or 'd', representing anterior, left, and dorsal orientations of division respectively. This cell would have a corresponding sibling name/reference of X with a suffix 'p', 'r', or 'v' representing a posterior, right, or ventral division relative to its sibling. For instance, a cell in the data set might have the name "ABal." This represents a cell descended from the AB cell in the embryo, where it is the left daughter of the anterior cell of the first AB division. With all cells in the embryo following this convention in the dictionary, any cell and all of its ancestors can be referenced by looking at the cell name and truncating its suffix one letter at a time.

The tree edit distance is a metric defined by counting "the minimum number of node deletions, insertions, and replacements that are necessary to transform one tree into another" as a measure of topological distance between trees [Figure 1A]. This can be applied to the dictionaries that we use as proxies for graph structures. If a tree has a specific node not in another embryo, then a corresponding node must be inserted into the lacking embryo as a descendant of the appropriate shared node to produce topologically identical trees. Thus, a single operation has taken place to transform the structure of one tree into another. This approach can be generalized to describe any tree-based topological differences. Using the dictionary format allows us to take advantage of the naming convention. Any cell that is added contains information pertaining to the

connection to its parent nodes, allowing for trivial checks of hierarchical and topological relationships. Indeed, this can be expressed further by noting that addition/subtraction operations of nodes can be represented by the absence/presence of nodes in one embryo that is not in the other. Extending this concept allows us to calculate the number of transformation operations as the number of nodes that are in one, but not both, dictionary sets. This means that tree edit distance between two dictionaries with nodes under the naming convention of [sulston et al] is defined as the magnitude of the intersection set subtracted from the magnitude of the union set of the dictionaries (in other words, it is size of the symmetric difference between the sets of nodes). In terms of the python implementation, this is calculated as the length of the XOR set of cells between the two embryos. It is utilized in figure 4B (Looking at SUF-1 and SKR-2 RNAi embryos tree edit distance to WT stereotype) and in figure 4C (Tree edit distance from each RNA embryo to WT stereotype Plotted on x axis).

In order to compare the trees in terms of the division timings, we introduce the concept of the branch distance. We define the branch distance as the Euclidean or $L^2$ norm of a vectorized representation of each lineage under comparison. To generate the vector, cells within each lineage are aligned first on the basis of depth from the root cell of each lineage and then on the lineage name derived from division orientation. In other words, we determine the *components* of each vector in such a way that division times for one cell are always being compared to division times for that same cell in a different embryo. When we calculate the $L^2$ norm, the difference between the values ascribed to a cell in one embryo and the corresponding cell in another embryo is taken and squared. Summing up these values and then taking the square root allows for an extension of the euclidean norm to these weighted graphs.

To compensate for alternate topologies, we computed one of two variants of the branch distance. The Intersection branch distance only computes the distance on the intersection set of cells contained in both lineages [Figure 1B], treating values that are not shared as absent from the comparison. It is used in Figure 3A to look at distances between all 30 WT embryos, and hierarchically cluster them into 2 groups, with the larger group of 22 embryos representing the Wild Type in all future calculations (unless otherwise noted) . It is also used in figure 4B (Looking at SUF-1 and SKR-2 RNAi embryos intersection branch distance to all 30 WT embryos) and in figure 4C (Average branch distance from each RNAi embryos to 22 WT embryos Plotted on x axis).

Meanwhile, the Union branch distance treats any missing cells as having a cell cycle duration of 0 [Figure 1C]. Thus, the Union branch distance compensates for differences in topology by directly adding the squares of values of cells without counterparts to the distance value, increasing it depending on topological variance and the value of the missing node. It is used to calculate the distance matrix between all WT and RNAi embryos [Figure 4A].

In comparing any two trees with any of these metrics, we note that the metric should work on subtrees or trees with different root nodes. This necessitates a change to the naming convention in cases where we compared different sublineages. This is done by finding the root node of both subtrees and assigning them an arbitrary letter. In cases where descendants of a root node are to be compared but have different orientations of division, we treat ‘a’, ‘l’, and ‘d’ suffix letters as equivalent as well as ‘p’, ‘r’, and ‘v’. For example, the values of subtree [‘A’, ‘Aa’, ‘Ap’] and the values of subtree [‘B’,’Bl’,’Br’], if roots were normalized, would both have the naming convention [‘Q’,’Qa’,’Qp’]. Utilizing this convention allows us to apply the metrics described above to compute the distances between distinct sublineages. In Figure 5, this is used to compute the intersection branch distance between different sublineages in 22 selected embryos WT and all 7 *glp-1* knocked down embryos. In Figure 6, this is used to compute union branch distances between sublineages of 22 WT embryos and sublineages of RNAi embryos.

*Calculating Correlation of Timing Events between Embryos*

In previous work, some authors have used an alternative to the cell cycle time for comparing the timing of division events between embryos. Specifically, Bao *et al.*[18] compared embryos using the "cell birth time," defined as the time from the fertilization of the embryo to the birth time of a cell. It can be calculated as the sum of the cell cycle times of the ancestors of a cell. Previous authors have found extremely high correlations between different embryos using this birth time definition.

Since a cell's birth time is the sum of all previous division timings, comparing embryos using this parameter could suppress variation and introduce spurious correlations between embryos. A sum of random variables will often show less variation than the underlying variables themselves–this is the reason the "standard error of the mean" is generally less than the underlying standard deviation in the population. To test this, we shuffled all of the division times in the embryos in question. Specifically, we randomly assigned each cell to the cell timing parameter of a different cell, effectively removing any correlation between the division timing of cells in different embryos.

A simple method of comparing the differences in cell timing events [b] is by plotting the times for each cell of one embryo against the times for each corresponding cell of another embryo. We then calculated the linear correlation coefficient between the cell cycle times between the cells of two embryos [Figure 2D, Figure 3B] as well as the correlation coefficient between shuffled cell cycle times [Figure 2C]. Shuffled birth times are computed by calculating the sum of the Shuffled cycle times of all ancestors of a particular cell, and were also compared using the correlation coefficient [Figure 2B]. Our analysis clearly demonstrates a significant correlation in cell birth times even in the shuffled data. As such, our subsequent analyses focused on comparing embryos using the cell cycle times.

*Computing the time between WT embryos*

Our analysis in Figures 2 and 3A suggested that there are two distinct groups of WT embryos in the Du et al. data[10]. While the correlation between cell cycle times is lower than cell birth times (Figure 2), we nonetheless saw fairly high correlations between embryos of the two groups, despite their distinct branch distances (Figures 3A and B). We thus hypothesized that the difference between the two groups was due to a uniform rescaling of time–in other words, all of the division events in one group of embryos was likely slower than the events in the other group of embryos by a constant factor.

The plots in Figure 2 suggest a straightforward way to quantify this difference in timing: the slope of the timings in one embryo vs. another. If this slope is less than 1, this suggests that the embryo whose times are plotted on the x-axis develops slower than the one plotted on the y-axis; if the slope is greater than 1, that suggests the reverse. A natural way to estimate this slope would be to simply perform a linear regression between the two data sets. Doing so, however, involves selecting one set of timings as the "independent variable." Since both sets of timings in any comparison is subject to random variation, however, we chose a slightly different approach to calculating the slope.

To do this, we employed simple Principal Component Analysis (PCA) on each pair of embryos. The eigenvector corresponding to the largest eigenvalue corresponds to a line that best fits the principal axis of variation in the data. In all the embryo comparisons, this axis of variation corresponds naturally to the line that compares the cell cycle times between the two embryos (e.g. Figure 2D). We thus performed PCA on each pair of embryos with cell cycle times plotted against one another as in Figure 2D. The slope of this best fit line was then calculated by

comparing the resulting principal eigenvector to the standard basis (i.e. calculating the "rise over run" for the eigenvector in the plane of Figure 2D). This method is used in Figure 3C to find the cell cycle scaling by comparing the cell cycle times of all 30 WT embryos against each other and partitioning the embryos into the clusters indicated in Figure 3A. These findings confirmed our hypothesis, indicating that the "group 1" embryos develop about 20% faster than the group 2 embryos.

*Clustering Wild Type and Mutant Embryos.*

We generated a distance matrix consisting of all pairwise union branch distances between WT and mutant embryos. We then performed single linkage hierarchical clustering on this distance matrix to generate a dendrogram between the embryos Since the number of clusters must be selected before the clustering is performed in hierarchical clustering, we analyzed the dendrogram to find a point with a large distance between generations [see Supplement Figure 1 for further details]. In the case of Figure 3, this approach partitioned the WT embryos into two groups. In the case of Figure 4, this approach resulted in 4 distinct clusters. Note that the distance matrices in Figure 5 were not clustered in order to show the pattern of variation between sublineages.

*Nonparametric Permutation Significance Testing for Distributions of Distances.*

We found that the intersection branch distances between certain sublineages of WT embryos were generally smaller than the intersection branch distances between other sublineages. This difference seemed to be related to Notch signaling events during development (Figures 5A, B and C). We used a simple permutation test to evaluate the statistical significance of this observation. In this test, we had two sets of distances: for instance, we compared the distances between ABal and ABar to the distances between ABpl and ABpr. This data corresponded to an observed difference between the means of the distribution. We then pooled the datasets together and generated randomized datasets of the same sizes by sampling without replacement. We calculated the difference of means between these randomized datasets. The p-values reported in Figure 5 represent the number of random cases where the absolute value of the difference in the means in these randomized datasets was greater than or equal to the observed difference.

*Detecting Fate Transformations for Mutant Embryos.*

Homeotic transformations that were identified by Du *et al.* indicate cases where a sublineage of a RNAi embryo combinations of marker genes present more consistent with a *different* sublineage of the WT embryo. This transformation indicates that the RNAi treatment has transformed the cell-fate pattern from that typical of one sublineage (the "origin") into that of a different sublineage (the "destination"). A list of the transformations that take place is available at digital-development.org/download.html under the name "Excel spreadsheet of all homeotic transformation phenotypes". In our work, we considered whether these cell fate transformations also had an impact on cell cycle timing. To do this, we developed an approach to see whether one sublineage in an RNAi embryo was "close" to a different lineage in the WT embryos. Our approach represents the WT sublineage populations as point clouds in high dimensional space. Each such point cloud has associated with it the maximum distance between WT embryos in the population; we call this maximum distance the "diameter" of that sublineage (Figure 6A). In this case, we chose the larger group of 22 embryos with similar developmental timings that naturally cluster together in Figure 3A in order to avoid higher diameters that might arise from systematic differences in experimental conditions.

For any given RNAi embryo, we then compute the distance between each of its sublineages and each sublineage from each of the 22 WT embryos. Note that in this case we use the intersection branch distance. If the distance between an RNAi sublineage and a WT sublineage from *any* embryo is less than the diameter of the WT sublineage, we say that the RNAi sublineage is in the *neighborhood* of the WT sublineage from that embryo. If the RNAi sublineage is in the neighborhood of that same lineage in the WT embryos, then we say that the sublineage is "unperturbed." In other words, the origin and destination lineage for that sublineage is the same (in the sense of the neighborhood described above). If the RNAi lineage is in the neighborhood of a *different* lineage in the WT embryo, we say that that sublineage has been transformed (i.e. the origin and destination are different). If an RNAi sublineage is not in the neighborhood of any WT sublineage, then we say the sublineage is "perturbed" (Figure 6A). The degree of transformation/perturbation is quantified based on the number of origin/destination WT sublineages that are neighbors of each RNAi sublineage Figure [6D].

Using the transformation framework outlined in Figure 6A, we then used a bootstrapping method to determine the significance of the distributions of Figures 6B and 6C. In other words, we tried to determine whether lineages with randomly chosen cell cycle times lengths contain a number of WT neighbors comparable to homeotically Transformed and other RNAi lineages. In every homeotically Transformed Lineage, the length of every cell's cycle is shuffled among all cells of the same name across all embryos treated with RNAi against genes that produce homeotic transformations. For each of these shuffled lineages, the number of WT lineage neighbors (out of 22) is counted (Supplemental Figure 7B). Along with homeotic transformations, we also shuffled all other RNAi lineages, where the length of every cell's cycle is shuffled among all cells of the same name across all embryos treated with RNAi and counted the WT neighbors for each lineage (Supplemental Figure 7F). We then repeated this shuffling 10,000 times and counted the number of lineages with 22 WT neighbors (Supplemental Figures 7C,7G) along with the number of lineages with at least 1 WT neighbor (Supplemental Figures 7D,7H).

## References


1. Stent, G.S. (1985). The role of cell lineage in development. Philos Trans R Soc Lond B Biol Sci *312*, 3–19. 10.1098/rstb.1985.0174.

2. Conklin, E.G. (1905). The Organization and Cell-lineage of the Ascidian Egg (Academy of Natural Sciences).

3. Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol *100*, 64–119. 10.1016/0012-1606(83)90201-4.

4. Santella, A., Catena, R., Kovacevic, I., Shah, P., Yu, Z., Marquina-Solis, J., Kumar, A., Wu, Y., Schaff, J., Colón-Ramos, D., et al. (2015). WormGUIDES: an interactive single cell developmental atlas and tool for collaborative multidimensional data exploration. BMC Bioinformatics *16*, 189. 10.1186/s12859-015-0627-8.

5. Murray, J.I., Boyle, T.J., Preston, E., Vafeados, D., Mericle, B., Weisdepp, P., Zhao, Z., Bao, Z., Boeck, M., and Waterston, R.H. (2012). Multidimensional regulation of gene expression in the C. elegans embryo. Genome Res. *22*, 1282–1294. 10.1101/gr.131920.111.

6. Ma, X., Zhao, Z., Xiao, L., Xu, W., Kou, Y., Zhang, Y., Wu, G., Wang, Y., and Du, Z. (2021). A 4D single-cell protein atlas of transcription factors delineates spatiotemporal patterning during embryogenesis. Nat Methods *18*, 893–902. 10.1038/s41592-021-01216-1.

7. Zhao, Z., Fan, R., Xu, W., Kou, Y., Wang, Y., Ma, X., and Du, Z. (2021). Single-cell dynamics of chromatin activity during cell lineage differentiation in Caenorhabditis elegans embryos. Molecular Systems Biology *17*, e10075. 10.15252/msb.202010075.

8. Cao, J., Guan, G., Ho, V.W.S., Wong, M.-K., Chan, L.-Y., Tang, C., Zhao, Z., and Yan, H. (2020). Establishment of a morphological atlas of the Caenorhabditis elegans embryo using deep-learning-based 4D segmentation. Nat Commun *11*, 6254. 10.1038/s41467-020-19863-x.

9. Du, Z., Santella, A., He, F., Tiongson, M., and Bao, Z. (2014). De Novo Inference of Systems-Level Mechanistic Models of Development from Live-Imaging-Based Phenotype Analysis. Cell *156*, 359–372. 10.1016/j.cell.2013.11.046.

10. Du, Z., Santella, A., He, F., Shah, P.K., Kamikawa, Y., and Bao, Z. (2015). The Regulatory Landscape of Lineage Differentiation in a Metazoan Embryo. Developmental Cell *34*, 592–607. 10.1016/j.devcel.2015.07.014.

11. Xiao, L., Fan, D., Qi, H., Cong, Y., and Du, Z. (2022). Defect-buffering cellular plasticity increases robustness of metazoan embryogenesis. Cell Systems *13*, 615-630.e9. 10.1016/j.cels.2022.07.001.

12. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. Nat Rev Genet *21*, 428–444. 10.1038/s41576-020-0233-0.

13. Vaz, C., Nascimento, M., Carriço, J.A., Rocher, T., and Francisco, A.P. (2021). Distance-based phylogenetic inference from typing data: a unifying view. Brief Bioinform *22*, bbaa147. 10.1093/bib/bbaa147.

14. Gong, W., Granados, A.A., Hu, J., Jones, M.G., Raz, O., Salvador-Martínez, I., Zhang, H., Chow, K.-H.K., Kwak, I.-Y., Retkute, R., et al. (2021). Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of C. elegans and M. musculus developmental trees. Cell Systems *12*, 810-826.e4. 10.1016/j.cels.2021.05.008.

15. Bille, P. (2005). A survey on tree edit distance and related problems. Theoretical Computer Science *337*, 217–239. 10.1016/j.tcs.2004.12.030.

16. Heumann, H., and Wittum, G. (2009). The tree-edit-distance, a measure for quantifying neuronal morphology. Neuroinformatics *7*, 179–190. 10.1007/s12021-009-9051-4.

17. Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, Eliáš, R., and Hoksza, D. (2016). RNA Secondary Structure Visualization using Tree Edit Distance. IJBBB *6*, 9–17. 10.17706/ijbbb.2016.6.1.9-17.

18. Bao, Z., Zhao, Z., Boyle, T.J., Murray, J.I., and Waterston, R.H. (2008). Control of cell cycle timing during C. elegans embryogenesis. Developmental Biology *318*, 65–72. 10.1016/j.ydbio.2008.02.054.

19. Moore, J.L., Du, Z., and Bao, Z. (2013). Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. Development *140*, 3266–3274. 10.1242/dev.096040.

20. Hutter, H., and Schnabel, R. (1995). Establishment of left-right asymmetry in the Caenorhabditis elegans embryo: a multistep process involving a series of inductive events. Development *121*, 3417–3424. 10.1242/dev.121.10.3417.

21. Bischoff, M., and Schnabel, R. (2006). Global cell sorting is mediated by local cell–cell interactions in the C. elegans embryo. Developmental Biology *294*, 432–444. 10.1016/j.ydbio.2006.03.005.

22. Kwak, S.G., and Kim, J.H. (2017). Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol *70*, 144–156. 10.4097/kjae.2017.70.2.144.

23. Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913–918. 10.1038/nature08781.

24. Schnabel, R., Hutter, H., Moerman, D., and Schnabel, H. (1997). Assessing Normal Embryogenesis inCaenorhabditis eleganSUsing a 4D Microscope: Variability of Development and Regional Specification. Developmental Biology *184*, 234–265. 10.1006/dbio.1997.8509.

25. Hutter, H., and Schnabel, R. (1994). glp-1 and inductions establishing embryonic axes in C. elegans. Development *120*, 2051–2064. 10.1242/dev.120.7.2051.

26. Mello, C.C., Draper, B.W., and Prless, J.R. (1994). The maternal genes apx-1 and glp-1 and establishment of dorsal-ventral polarity in the early C. elegans embryo. Cell *77*, 95–106. 10.1016/0092-8674(94)90238-0.

27. Priess, J.R. (2005). Notch signaling in the C. elegans embryo. WormBook, 1–16. 10.1895/wormbook.1.4.1.

28. Zacharias, A.L., Walton, T., Preston, E., and Murray, J.I. (2015). Quantitative Differences in Nuclear β-catenin and TCF Pattern Embryonic Cells in C. elegans. PLOS Genetics *11*, e1005585. 10.1371/journal.pgen.1005585.

29. Gilbert, G.S., and Parker, I.M. (2022). Phylogenetic Distance Metrics for Studies of Focal Species in Communities: Quantiles and Cumulative Curves. Diversity *14*, 521. 10.3390/d14070521.

30. Hicks, D.G., Speed, T.P., Yassin, M., and Russell, S.M. (2019). Maps of variability in cell lineage trees. PLOS Computational Biology *15*, e1006745. 10.1371/journal.pcbi.1006745.

31. Bao, Z., Murray, J.I., Boyle, T., Ooi, S.L., Sandel, M.J., and Waterston, R.H. (2006). Automated cell lineage tracing in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U.S.A. *103*, 2707–2712. 10.1073/pnas.0511111103.

32. Amat, F., Lemon, W., Mossing, D.P., McDole, K., Wan, Y., Branson, K., Myers, E.W., and Keller, P.J. (2014). Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. Nat Methods *11*, 951–958. 10.1038/nmeth.3036.

33. Keller, P.J., Schmidt, A.D., Wittbrodt, J., and Stelzer, E.H.K. (2008). Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. Science *322*, 1065–1069. 10.1126/science.1162493.

34. Strnad, P., Gunther, S., Reichmann, J., Krzic, U., Balazs, B., de Medeiros, G., Norlin, N., Hiiragi, T., Hufnagel, L., and Ellenberg, J. (2016). Inverted light-sheet microscope for imaging mouse pre-implantation development. Nat Methods *13*, 139–142. 10.1038/nmeth.3690.

35. Barbaric, I., Biga, V., Gokhale, P.J., Jones, M., Stavish, D., Glen, A., Coca, D., and Andrews, P.W. (2014). Time-Lapse Analysis of Human Embryonic Stem Cells Reveals Multiple Bottlenecks Restricting Colony Formation and Their Relief upon Culture Adaptation. Stem Cell Reports *3*, 142–155. 10.1016/j.stemcr.2014.05.006.

36. Mitchell, S., Roy, K., Zangle, T.A., and Hoffmann, A. (2018). Nongenetic origins of cell-to-cell variability in B lymphocyte proliferation. Proceedings of the National Academy of Sciences *115*, E2888–E2897. 10.1073/pnas.1715639115.

37. Kamath, R.S., Martinez-Campos, M., Zipperlen, P., Fraser, A.G., and Ahringer, J. (2000). Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in Caenorhabditis elegans. Genome Biol *2*, research0002.1. 10.1186/gb-2000-2-1-research0002.

38. Arata, Y., Takagi, H., Sako, Y., and Sawa, H. (2015). Power law relationship between cell cycle duration and cell volume in the early embryonic development of Caenorhabditis elegans. Frontiers in Physiology *5*.

39. R. Jones, A., Forero-Vargas, M., Withers, S.P., Smith, R.S., Traas, J., Dewitte, W., and Murray, J.A.H. (2017). Cell-size dependent progression of the cell cycle creates homeostasis and flexibility of plant cell size. Nat Commun *8*, 15060. 10.1038/ncomms15060.

40. Clucas, C., Cabello, J., Büssing, I., Schnabel, R., and Johnstone, I.L. (2002). Oncogenic potential of a C.elegans cdc25 gene is demonstrated by a gain-of-function allele. The EMBO Journal *21*, 665–674. 10.1093/emboj/21.4.665.

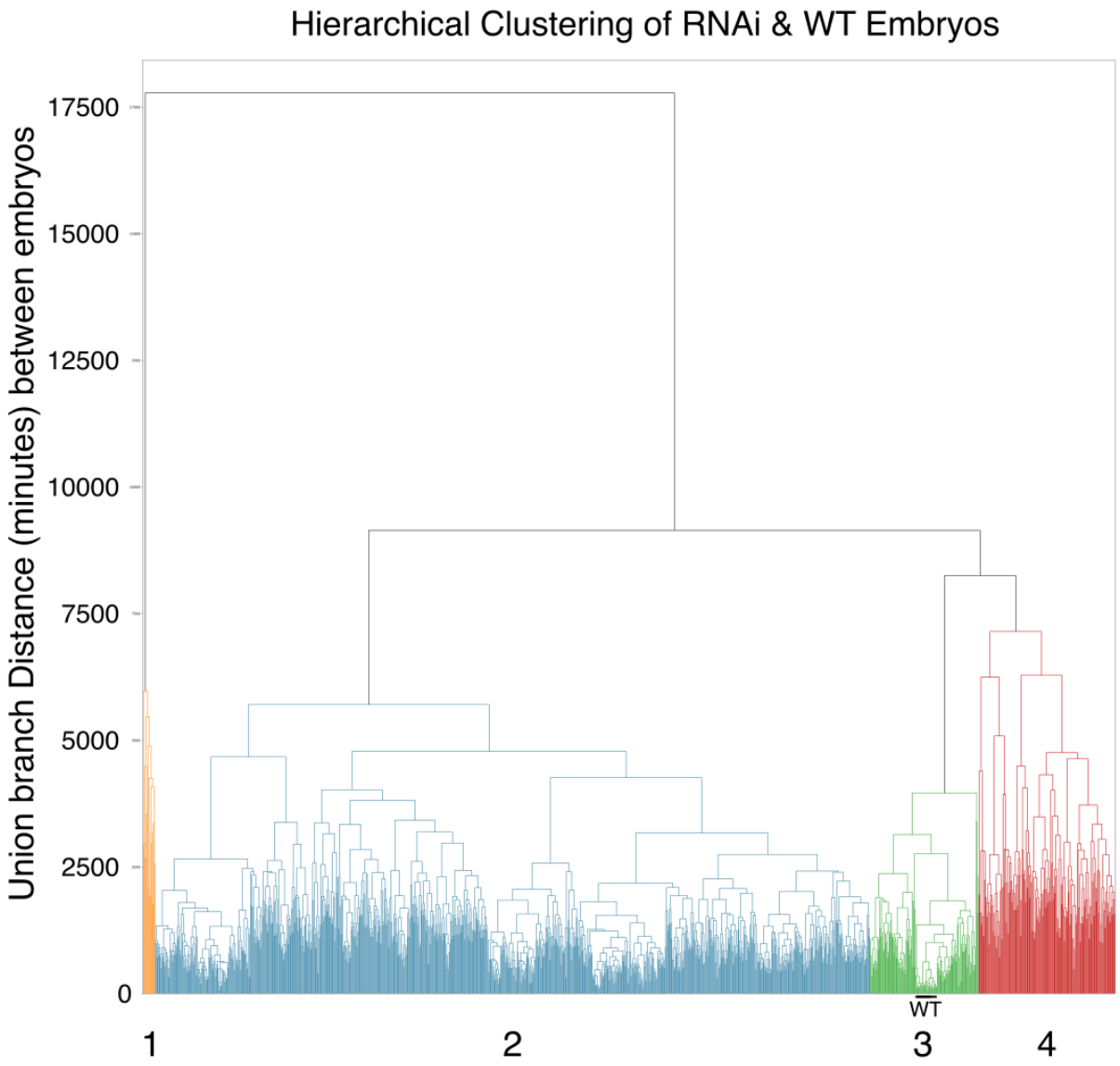
41. Roy, S.H., Clayton, J.E., Holmen, J., Beltz, E., and Saito, R.M. (2011). Control of Cdc14 activity coordinates cell cycle and development in Caenorhabditis elegans. Mechanisms of Development *128*, 317–326. 10.1016/j.mod.2011.06.001.

42. Budirahardja, Y., and Gönczy, P. (2009). Coupling the cell cycle to development. Development *136*, 2861–2872. 10.1242/dev.021931.

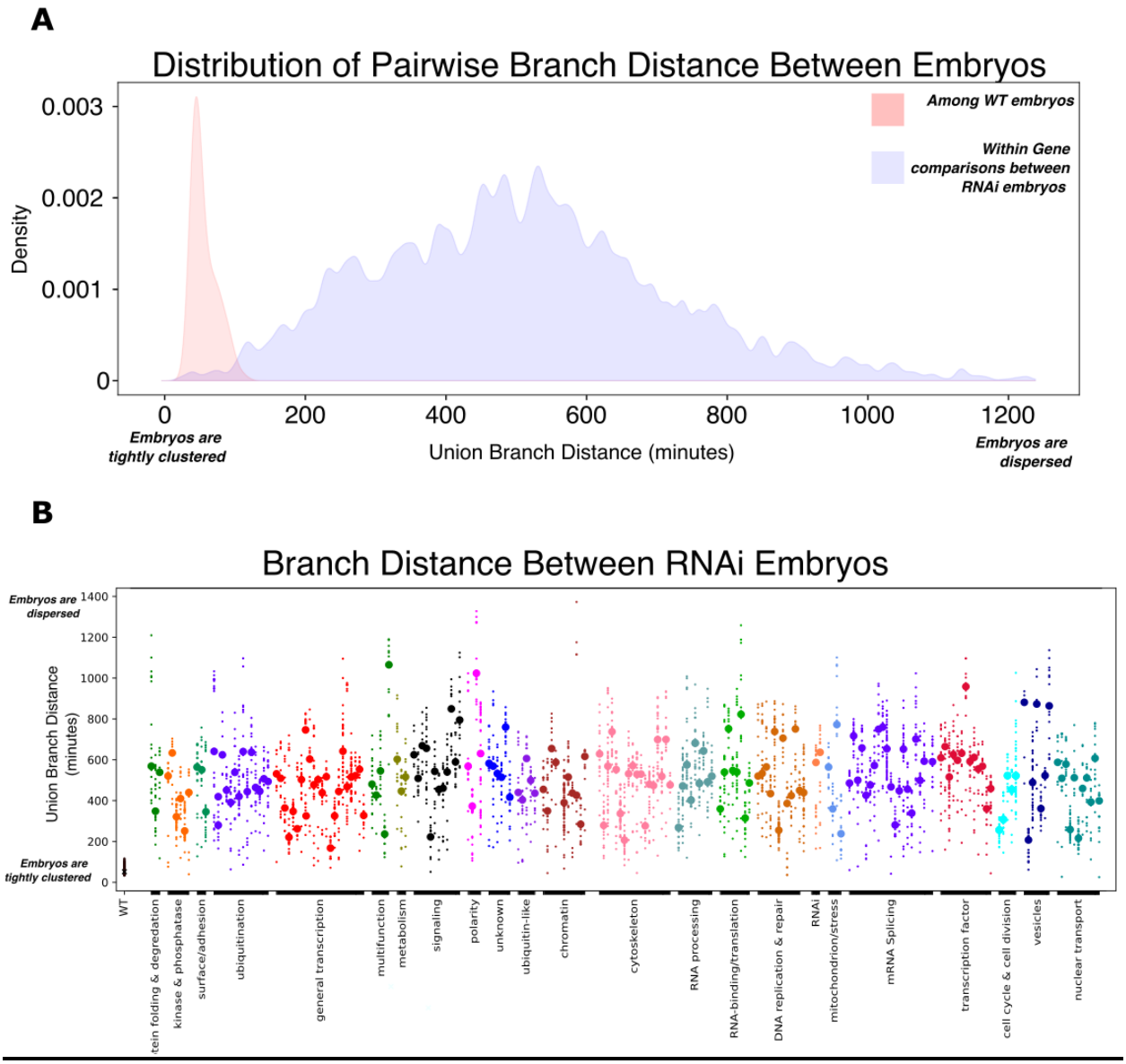
43. Fay, D.S. (2005). The cell cycle and development: Lessons from C. elegans. Seminars in Cell & Developmental Biology *16*, 397–406. 10.1016/j.semcdb.2005.02.002.

44. Cepko, C. (2014). Intrinsically different retinal progenitor cells produce specific types of progeny. Nat Rev Neurosci *15*, 615–627. 10.1038/nrn3767.

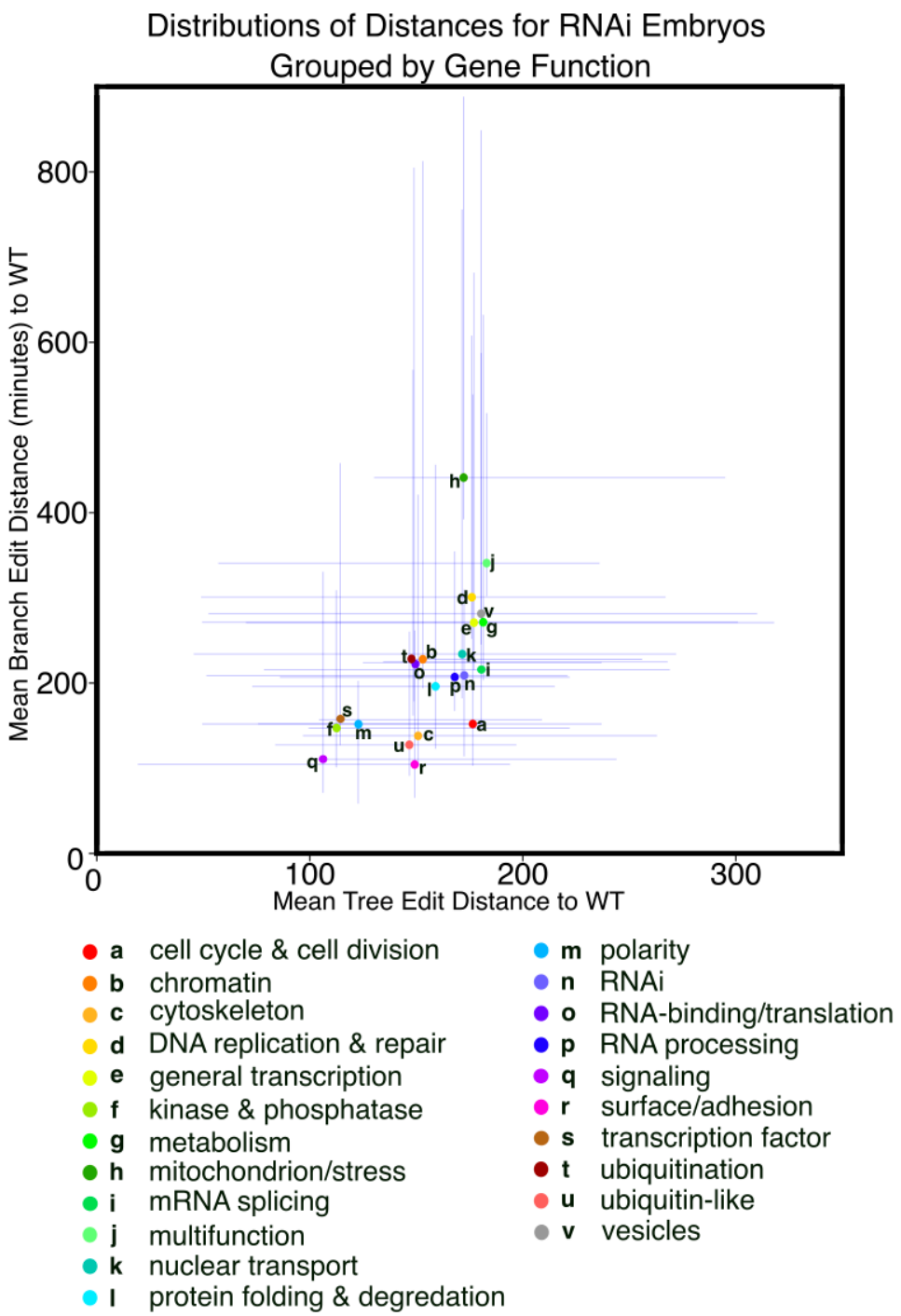
**Supplemental Figures**



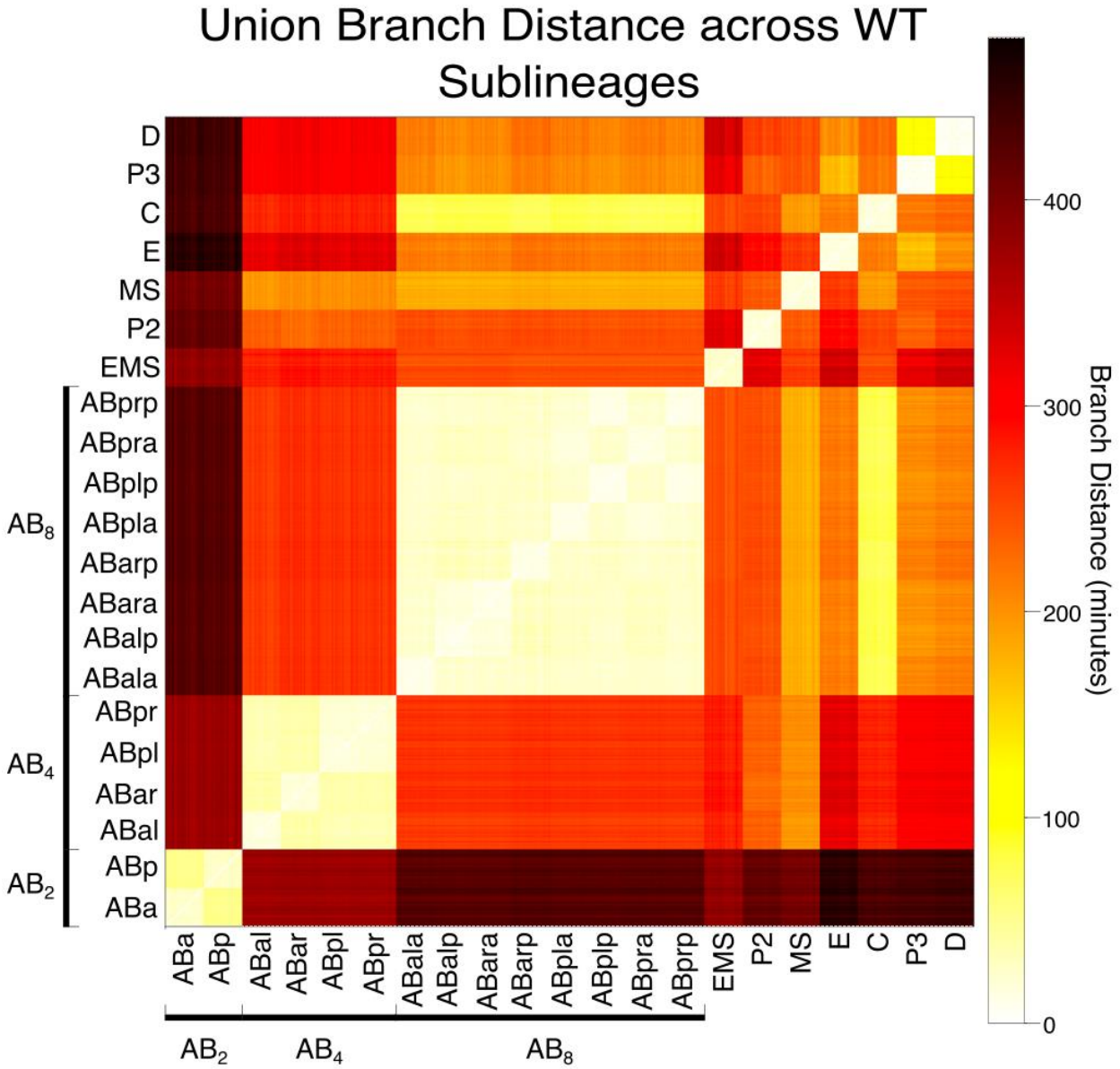

**Supplemental Fig 1. Structure in the distance between all WT and RNAi embryos.** Shown is a dendrogram constructed using the union branch distance measured between all WT and RNAi treated embryos along with the 4 classes we partition the dataset into. The WT embryos and WT-like RNAi-treated embryos are highlighted in cluster 3.



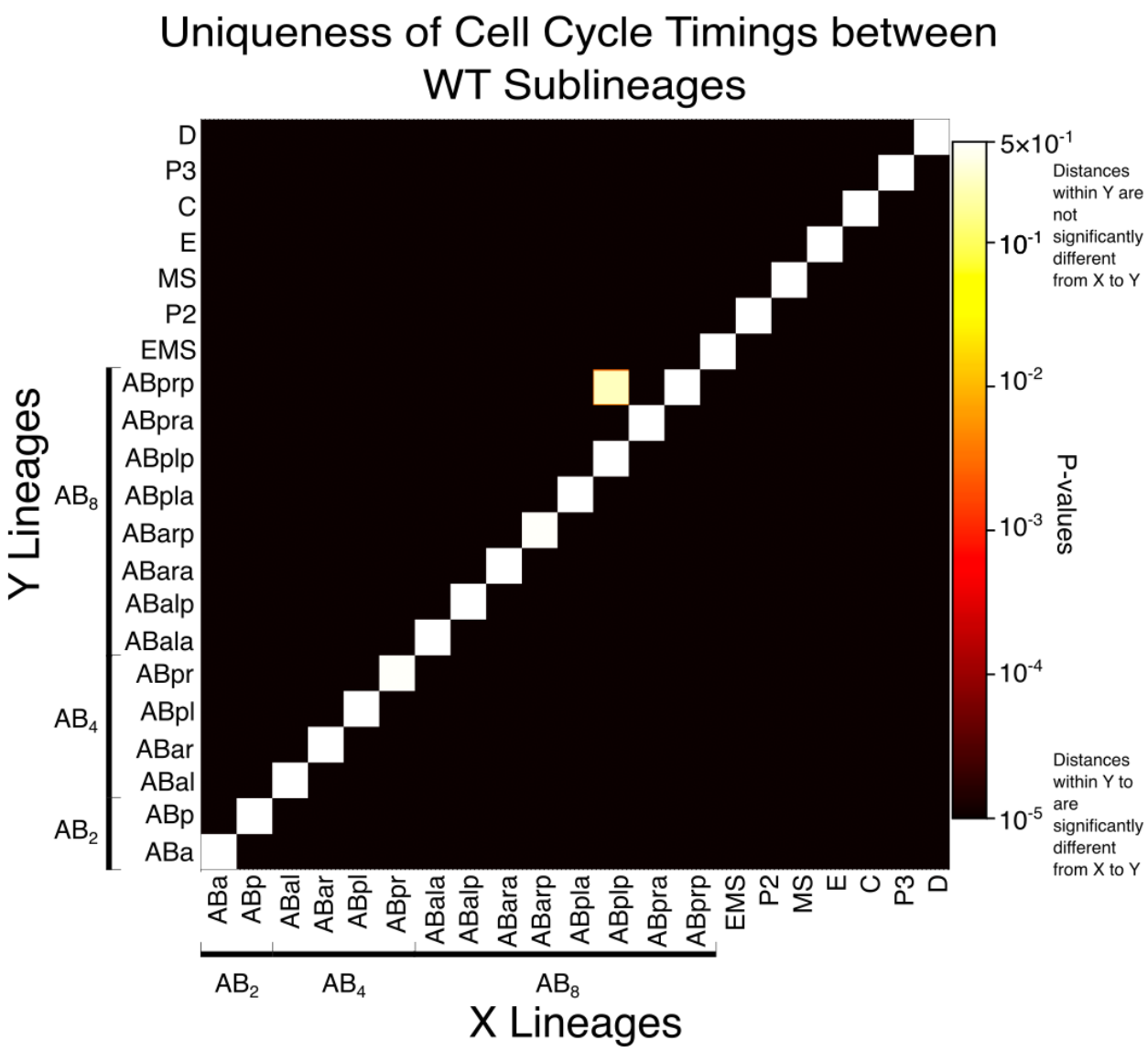

**Supplemental Figure 2. RNAi embryos are far more dispersed than WT embryos. (A)** The distribution of union branch distances between wild type embryos (red) and between embryos treated with RNAi against the same gene (blue). Densities were generated using a kernel density estimator. **(B)** Strip plot of all RNAi embryos showing the branch distances between embryos treated with RNAi against the same gene. Embryos treated with RNAi against genes with shared functions are grouped together. The median branch distance within each set of embryos is plotted as a large circle.



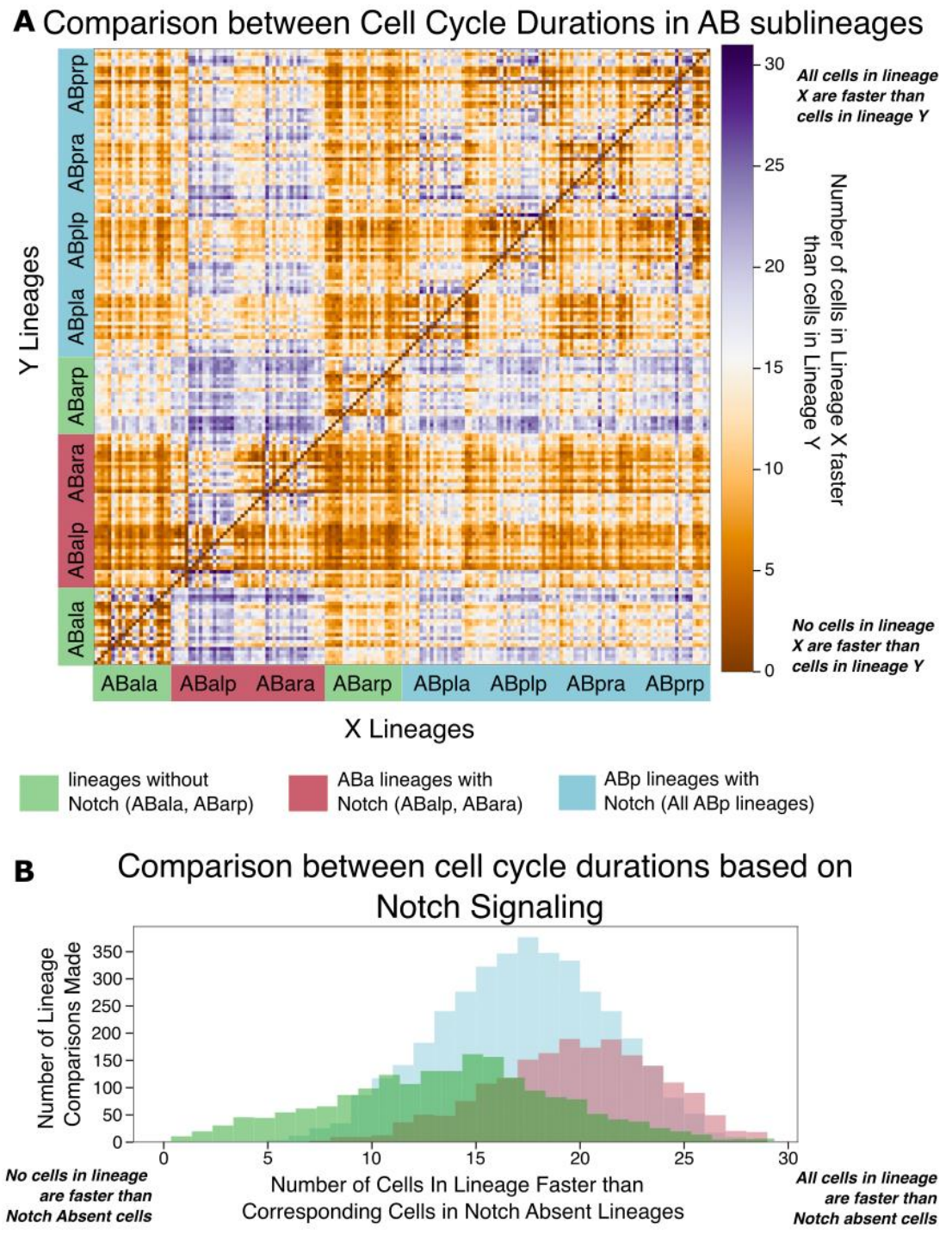

**Supplemental Figure 3. Correlation between the tree edit distance and branch distance between RNAi treated embryos and a single WT reference embryo.** The mean (circle) and standard deviation (blue lines) of the distance between embryos treated with RNAi against genes with common function and a single WT reference embryo are shown.



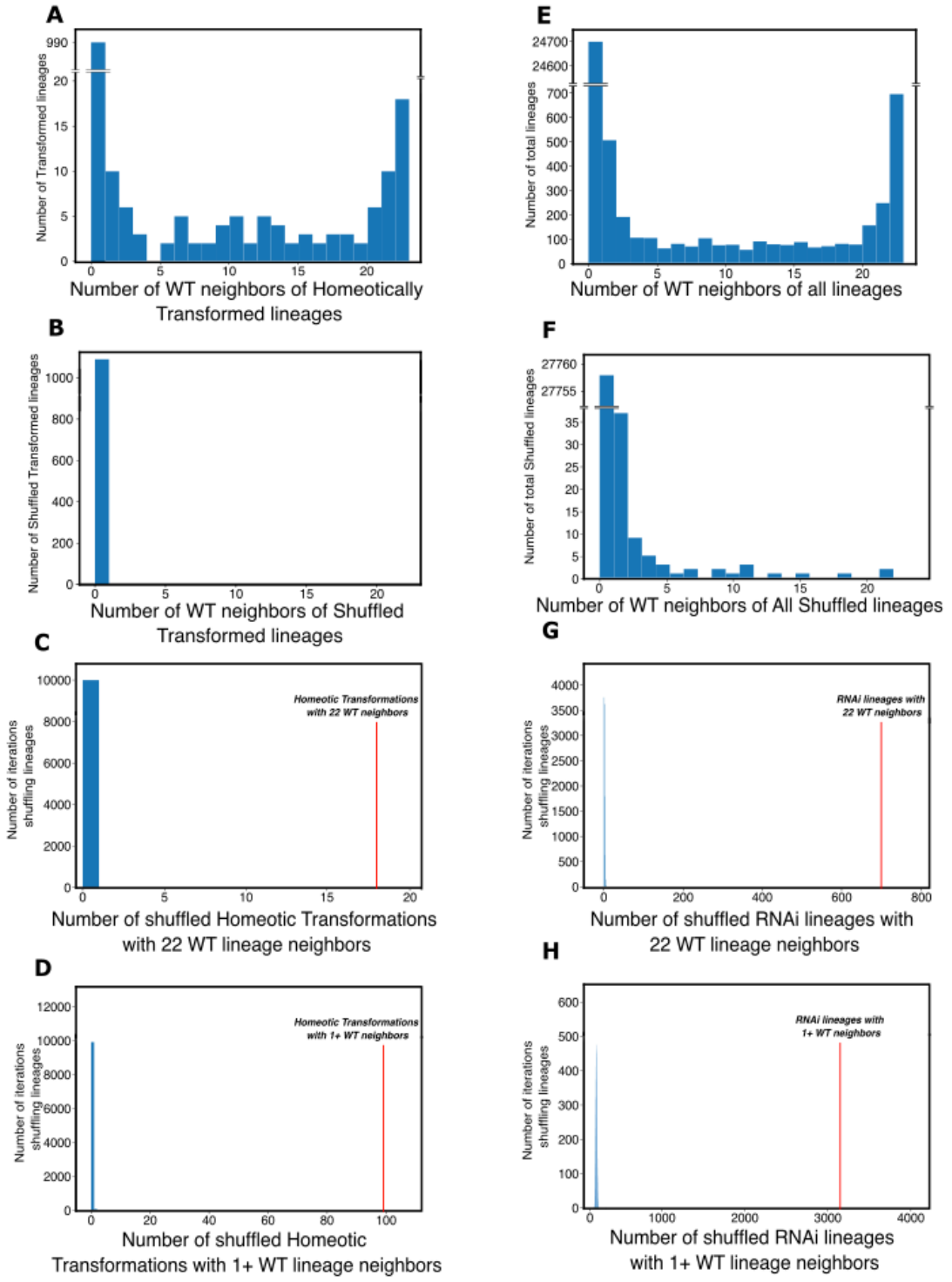

**Supplemental Fig. 4 Union Branch Distance between WT sublineages.** A heatmap showing the union branch distance between each pair of sub-lineages across 22 WT reference embryos.




**Supplemental Fig. 5 All but one WT sublineages possess unique patterns of cell cycle timing.** Heatmap showing the p-value calculated using 441 permutation tests (see methods) against the null hypothesis that sublineages share a common distribution of intersection branch distances. All but a single comparison (ABprp vs ABpla) are significant based on a Bonferroni corrected threshold of $2 * 10^{-5}$




**Supplemental Fig. 6 Notch induction systematically shortens cell cycles in affected sublineages. (A)** A heatmap showing the fraction of cells in the lineage listed along the Y axis that have a shorter cell cycle than the corresponding cell in the lineage along the X axis. Each sublineage contains 32 cells so a lineage in which >16 cells possess a shorter cell cycle is considered to be "faster" **(B)** The distribution of comparisons from the heatmap in panel A grouped based on lineages that do not have a history of Notch activation (green, ABala and ABarp), lineages derived from ABa in which Notch is activated (red, ABalp and ABara), and all lineages derived from ABp.




**Supplemental Fig. 7 Significantly more transformed lineages match the cell cycle timing of their ectopic fate than random. (A)** The number of WT destination lineages that fall within the transformed neighborhood of lineage annotated as homeotically transformed by Du *et al.*[10]. See Main Figure 6B. **(B)** The number of WT destination lineages that fall within the transformed neighborhood of homeotically transformed lineages where the length of every cell's cycle is shuffled among all cells of the same name across all embryos treated with RNAi against genes

that produce homeotic transformations. Note there are no WT neighbors to any of these shuffled lineages. **(C)** The number of cases where lineages from RNAi treated embryos that are shuffled as in (B) fall within the neighborhood of all 22 WT samples of the destination lineage. The red line shows the corresponding value for the unshuffled data. **(D)** The number of cases where lineages from RNAi treated embryos that are shuffled as in (B) fall within the neighborhood of 1 or more WT samples of the destination lineage. The red line shows the corresponding value for the unshuffled data. **(E)** The number of WT destination lineages that fall within the transformed neighborhood of all sublineages from all RNAi treated embryos. See Main Figure 6C. **(F)** The number of WT destination lineages that fall within the transformed neighborhood of all lineages from RNAi treated embryos where the length of every cell's cycle is shuffled among all cells of the same name. **(G)** The number of cases where lineages from RNAi treated embryos that are shuffled as in (F) fall within the neighborhood of all 22 WT samples of any lineage. The red line shows the corresponding value for the unshuffled data. **(H)** The number of cases where lineages from RNAi treated embryos that are shuffled as in (F) fall within the neighborhood of 1 or more sample of any lineage. The red line shows the corresponding value for the unshuffled data.