UNIVERSITY OF CALIFORNIA
Los Angeles

Graph Based Metrics of Lineages Used to Characterize Wild-Type *C. Elegans* Cell Cycles during Development

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Gunalan Kolaram Natesan

2022

ABSTRACT OF THE THESIS

Graph Based Metrics of Lineages Used to Characterize Wild-Type *C. Elegans* Cell Cycles during Development

By

Gunalan Kolaram Natesan
Master of Science in Bioinformatics
University of California, Los Angeles, 2022
Professor Eric Deeds, Chair

Comparing tree topologies is an invaluable biological tool, though it is mathematically difficult to systematize. Tree comparisons are simpler to compute on ordered/labeled graphs, which are produced through the eutelic lineages of *Caenorhabditis elegans* development. We make use of this property to create novel approaches to compare the values and topologies of weighted graphs. This involves generalizing the euclidean norm to the weighted edges of tree structures to create a "branch distance", incorporating the tree edit distance as a measure of topological distance, and normalizing the tree edit distance into the Jaccard Distance. These approaches are benchmarked by measuring cell cycle timing in embryonic cell lineages using this framework to categorize patterns of development, and corroborating these findings with known developmental properties of this model organism.

Alexander Hoffmann

Pavak Shah

Eric Deeds, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures and Equations

Acknowledgements

I would like to thank my parents, as well as my mentors, colleagues, and friends at the University of California, Los Angeles.

Introduction:

Finding a systematic way to characterize lineages is an active area of interest in various fields of biology [1], including development [2] and epidemiology [3]. The ability to compare the tree topologies and associated values of these lineages is useful, as the tree structure can be used to represent descent from a single point of origin, which is a recurring biological motif. The recursive nature of this also makes practical computation of tree alignments difficult [4], necessitating a further look into making these graph metrics applicable.

C. elegans is a commonly used animal model, due to its small size, transparency, ease in cultivation and maintenance, and low cost [5]. Its property of eutely, the lack of variation in somatic cell count in an adult multicellular organism, [6] yields insight into the processes of differentiation and tissue formation during development. The various cells in *C. elegans* embryos can be uniquely identified and thus named [7] , allowing comparisons of its developmental trees without computationally expensive graph alignment. Using *C. elegans* as a model organism, new tree comparison measures and metrics can be developed and benchmarked against known traits.

Regarding its use as a model organism, *C. elegans* has been extensively studied to understand a variety of developmental and biological processes, such as RNA interference [8] , addiction [9], and aging [10]. As the individual descendants of each progenitor cell are accurately known [11], the regulation of cell fate in *C. elegans* has been studied extensively. Since complete cell lineages are known and RNAi pathways are characterized, *C. elegans* is well suited to gene knockdown perturbation analysis [12]. Indeed, a large-scale screening of developmental genes and the resulting perturbed fate changes in specific lineages has been conducted [13]. Fate changes, in this context, are defined through changes in the expression patterns of specific marker genes. In sampling aggregate levels of specific markers for each cell, timing data that corresponds to the life cycle of each cell has also been produced. This timing data has mapped cell division down to the individual minute, delineated by cell position, lineage, and type, allowing for accurate measurements.

Since timing data exists for every individual cell it can be arranged in a way that mirrors the process of proliferation from a zygote, namely, a binary developmental tree. These trees can be compared across individuals and lineages using topological measures while avoiding the computationally expensive alignment problem that often plagues graph theoretic approaches.

In this project, we attempt to characterize the pattern of cell cycle times across individual lineages using novel and adapted graph theoretic approaches. This can be used to benchmark the sensitivity of our proposed metrics, as well as investigate the temporal development of wild-type *C. elegans* embryos on a lineage-by-lineage basis.

Results:

Comparing Measures of Embryonic Development Time:

We first looked at previous methods used to characterize cell birth times. Pearson's correlation and its square measure linear correlations between two sets of measurements. It has been used to measure correlation in developmental events in C. elegans embryos

[14], comparing absolute cell birth times. These comparisons yielded r^2 values from 0.995 to 0.997, which is indicative of tightly regulated data.

When comparing the absolute cell birth times between two embryos, there were no r^2 values smaller than 0.99, confirming methods found in [14]. Since the birth time for each cell is the sum of cycle times of its parents, we note that birth time might be altered by effects of random variable summation. To test this, the correlation coefficient was calculated on the embryonic cell cycle times that were scrambled (r^2 less than 0.005), before it was summed to simulate birth timing comparisons. The new r^2 were calculated as being in between 0.7 and 0.85 from initially completely uncorrelated data, suggesting that the assumed tight correlation between the two comparisons may be a result of statistical artifacts induced by the central limit theorem when individual cell cycle durations are summed to yield the absolute cell birth time as a measure of developmental timing [15].

Comparing the linear correlation between wild type embryos calculated on individual cell cycle durations yields a more heterogeneous view of normal embryogenesis, motivating further characterization. Given that the structure of the data was that of an ordered binary tree, we attempted to create a metric that would function on the topology of ordered binary trees.

Constructing Graph Metrics

We generalized the Euclidean notion of distance to the graph space to create the branch distance, as a metric for comparing cell cycle durations across embryos with shared topology. In the branch distance, the differences between cycle durations of aligned cells in each lineage tree under comparison are squared and summed, before the square root is taken. The eutelic nature of C. elegans embryogenesis produces ordered lineages where each cell can be uniquely named, allowing for 1-to-1 alignment between lineages. branch distances were calculated to form a distance matrix of 435 unique distance measurements between 30 WT embryos from the published dataset. The embryos were hierarchically clustered with single, average, and maximal linkage methods which consistently found two large groups with a branch distance of ~150 seconds between them. The larger of these 2 groups had a 21 embryo cluster with diameter ~ 100 seconds while the smaller of the 2 had 9 embryos and a diameter of ~ 100 seconds. WT

embryos were imaged from January 2011 to May 2011 in multiple locations. WT embryos in the smaller cluster are imaged from April 27 2011 to April 29 2011 suggesting some form of batch effect thatmay be present in the published data. Using the first eigenvector of the singular value decomposition of a matrix comprised of the aligned pairs of cell cycle durations between each pair of embryos from the two clusters as a non-parametric linear regression yielded slopes larger than ~1.1, suggesting that most of the difference between these clusters was produced by a global scaling of cell cycle durations.

Characterizing *C. elegans* Sublineages:

The branch distance, in theory, should also compare the timings of the progeny of specific cells, due to the recursive binary tree topology of each lineage. Because of the eutelic property of the WT embryos, each sublineage follows the exact same structure across embryos and can thus be standardized. However, when comparing different sublineages, the difference in topology must be accounted for before comparing branch distances. In this case we will use a conventional graph metric known as the tree edit distance to measure differences in the number and location of cells in each lineage under comparison. We define the tree edit distance, in this case, as the number of elements that need to be added and subtracted from one graph to turn it into another graph. This can be normalized by dividing the

tree edit distance by the number of unique elements in the combined set of both trees, which is analogous to the Jaccard distance.

The Jaccard distance was computed between each pair of sub-lineages originating from the 22 founder cells described in [14]. Each generation of the AB lineage was found to be topologically identical to every other AB lineage that was in the same generation. The descendants of the P1 lineage are far more varied, in that the Jaccard distance between the pairwise comparisons of each of the descendants of P1 is never below 0.2.

The pairwise Branch Distances between all of the 22 sub lineages was also computed between the 21 embryos belonging to the larger of the wild-type clusters identified above. In this case, it is noted that the branch distances between the sublineages follow a structure that resembles the jaccard distance calculation. AB sub lineages are closest in branch distance to other AB sub lineages in the same generation. The descendants of P1 tend to be more sporadic in branch distance comparisons. The most noticeable difference is the large distance found between the E and EMS lineages and all other sub-lineages (Distances from E & EMS are from ~110 to ~150 seconds while MS has branch distances from ~40 to ~90). This is corroborated when the E lineage has an average scaling parameter ~5 times larger than the MS lineage's scaling parameters. This is further supported as the E lineage has a jaccard distance of 0 from the AB lineage, with an average branch distance of ~80 to the AB lineage. This suggests that the E lineage has significant delays to its cell cycle.

Given our metric, can different sub-lineages be uniquely identified on the basis of cell cycle duration alone? I compared the distributions of each sub-lineage branch distance to additional examples of itself from other wild-type lineages to the distribution of branch distances measured between each sub-lineage and each other sub-lineage in other wild-type embryos using a permutation test to determine significance. We determined that each self-comparing branch distance distribution was significantly different from the branch distances between subtrees, with the exception that the distributions of branch distances between subtrees ABprp and ABplp were not distinguishable.

Methods:

The data that is used [13] consists of 30 WT embryos imaged to 350 Cells, with the concentration of one of three marker genes (pha-4, cnd-1, nhr-25) imaged every ~1.25 minutes. Data also contains 1322 RNAi embryos of 204 Essential Conserved Developmental regulators s.t. knockdown results in a lethal phenotype. Timing and marker expression data are also included in the same format as the WT image data. In addition, times where each RNAi embryo imaging process was stopped for each lineage are also included, as are a list of Homeotic and Undefined Fate transformations for the first 4 divisions of the lineage. Data was analyzed using numpy [16] and scipy [17] modules, and graphed with matplotlib on Python 3.9 [18] Jupyter Notebooks [19].

Measures of Relative Comparison:

Each of the WT and RNAi embryos can be represented as vectors, with dimensions analogous to the presence of certain cells and each cell’s corresponding timing data representing feature value. With this, vector operations can work on the predefined tree structure. The Coefficient of Determination (R^2) between an independent and dependent variable is indicative of the amount of variation in the dependent variable that is predictable from the independent variable. It is frequently used as a measure of correlation between two vectors of equal size. In a Similar vein, To show differences in the global clock between two embryos, the cells of each embryo are plotted against each other and linearly regressed, from which the slope is taken to provide a parameter which describes the time scaling between the various embryos. This version of the scaling coefficient is a parametric measure which implicitly assumes a causative relationship between the two distributions. Principal component analysis (PCA) is instead used as a nonparametric measure of finding primary and secondary orthogonal basis vectors, where the ratio between the two is taken to find the scaling coefficient. It is worth noting that the value produced by computing f(x,y) is not necessarily equal to f(y,x), rather, ln(f(x,y)) = -ln(f(y,x)). Previous studies have shown that the developmental clock of c elegans embryogenesis scales linearly with temperature.

The Plotted Birth time of each cell in an embryo against the Birth time of corresponding cell in another embryo is benchmarked by computing linear regression and r^2 values. We converted embryo birth times to cycle times by subtracting each cell's birth time from its child's birth time. Plotted cycle times of each cell in an embryo against cycle time of corresponding cell in another embryo, computed linear regression and r^2 value. We assigned each cell in a WT embryo to a random cell’s cycle time in the same embryo. Use these new assignments to compute cell birth time through adding parental

assignments to compute cell birth time through adding parental cycle times. Use this resampling as a method of bootstrapping linear regression of birth times.

Metrics on Graph Spaces:

Each embryo can be represented as a directed binary tree of cells, where each cell is a node connected to its parent, with up to 2 children. We adopted the standard naming convention of C. Elegans embryonic development [7].

The Eutelic property of WT C. Elegans embryos means that each developmental lineage is identical, such that the various developmental subtrees of each embryo also have identical topologies as well. In C. elegans embryos, the daughter of a specific cell tends to be named after its mother, with a suffix representing the axis and direction of it’s mothers division. Exceptions to this convention are found in the initial divisions of the zygote, where designated lineage founders are named distinctly. The divisions from the WT zygote are always arranged in the same way, where the zygote divides into the AB and P1 cells. The AB cell undergoes an anterior/posterior division, after which a left/right division takes place with the daughters, and another anterior/posterior division takes place to form the 8 sublineage progenitors. The P1 cell divides into EMS & P2, which divide into E and MS as well as C & P3 respectively. After P3 divides

into P4 and D, the E, MS, C, & D founding cells continue to divide into their own lineages. Each of these lineage founders, as well as the parents, are designated as founding cells, and thus have distinct phenotypes that can be categorized and compared.

Each embryo can be thought of as a tree of cells, where each cell is positionally defined by its parent. The Tree edit distance is the number of specific cells that need to be added or subtracted from one tree to convert it into another. In terms of direct operations, Tree edit Distance between embryo A & B is Calculated by the magnitude of the XOR set of cells between embryo A & B. It is a rigorous distance metric, as it is a generalization of the hamming distances on binary dimensional spaces.

The Jaccard Distance can be thought of as the intersection of the two embryo sets divided by the union, which is subtracted from one. It can be thought of as the Tree Edit Distance normalized by the union of the embryo sets. It follows the distance metric definition rigorously, though its range is normalized to values between 0 and 1.

Branch Edit Distance is defined as the square root of the sum of the squares of the differences between the times assigned to each cell. It is the euclidean norm on a dimensional space where tree structures function as vectors such that each dimension is represented by a cell in the embryos. As such, it is a rigorous distance metric that follows the properties set by the euclidean norm in high

dimensions. The TED and JD were run on the WT embryos to ensure total correspondence, since all WT C. elegans embryos have an invariant developmental lineage.

The Branch Distances were calculated between each pair of the 30 WT embryos to characterize natural heterogeneity. These were then hierarchically clustered using single, average, and maximum linkage methods to find distinct groups that are robust under various grouping schemes. The larger of these groups, which also has a proportionally smaller variance, is deemed an inlier cluster while the other is deemed an outlier cluster. The PCA scaling parameters are also calculated between all pairs of WT embryos to see if variances between clusters are due to scaling. The experimental conditions of the two clusters are then compared to see if residual effects are caused by obvious batch effects.

Subtree Analysis:

P0, P1, and AB lineages are not consistently present in imaged embryos and will thus be excluded from subtree analysis. Each subtree has one of three dividing orientations (anterior/posterior, dorsal/ventral, left/right). When comparing subtrees with varying topologies, dorsal and left oriented divisions are equated to anterior divisions and ventral and right oriented divisions are equated to posterior divisions.

For the WT data, each of the 21 subtrees in the embryos was compared against the others using Jaccard and Branch Distances. This results in a structure where each of the 21 subtrees was compared against the 20 other subtrees (to produce 441 distance distributions for each comparison) and against itself once (producing 231 unique distributions consisting of 21 self comparisons and 220 cross-sublineage comparison distributions). These comparisons were also computed with pairwise PCA based scaling coefficients. This scaling parameter matrix was sanitized (removed errata of values above 54.6, below 0.01, and NaN values) and log-normalized.

Significance tests between the Distribution of a Self Subtree BD and the Distribution of distinct subtree BD's were conducted. This serves to see if a subtree is distinguishable from another reference subtree given only branch distances to the reference subtree as information. For each lineage, its branch distance against a reference lineage was computed, and compared to the reference lineage's Branch Distance against itself. The comparisons took place using a standard t test, the brunner munzel test, and a permutation test with 10000 iterations completed, using a p value of 0.05. When Bonferonni Corrected using this p value of 0.05 with a total number of 400 total hypotheses tested, we get a new cutoff of 0.000125, which each of these statistical tests also used as a cutoff.

Discussion:

C. elegans has been used as a model organism for over 50 years [20] , due to the roundworm's small size, ease of cultivation and short lifespan [21]. In addition to its genetic homology with humans, *C. elegans* organisms are eutelic [6]. Every adult *C. elegans* has the same number of cells, with no variance under normal conditions. Working backwards from this property, the specific divisions that take place in development can be mapped. This can be used to compare specific cells between individuals. Eutely in *C. elegans* has been the basis of many assumptions regarding a high degree of homogeneity in the mechanism of the developmental clock [22] [23]. Indeed, previous studies on C. elegans developmental timing [14] state that "95% of the cell divisions of an embryo deviate less than 2% from its general clock," suggesting a high degree of consistency of cell cycle timing. However, bootstrapping the linear regression technique reveals that lineages with no correlation can produce tight correlations by summing constituent variables. This suggests that smoothing by way of variable summation [15] may affect the interpretation of developmental clock heterogeneity, which necessitates further investigation. It is likely that the high degree of previously noted consistency is due, in part, to the significant statistical impact of the artifact introduced by the central limit theorem.

In addition, the slope produced by the linear regression of cycle/birth times has been stated to be indicative of scaling differences caused by temperature, parental health, or other extraneous factors [14]. However, linearly regressing the timings of each embryo implicitly assumes a causative relationship between the variable on the x axis and the variable on the y axis [24]. In addition, linear regressions have no commutative properties regarding variable order as regressing the x variable on the y variable does not have the same degree of precision as regressing the y variable on the x variable. A nonparametric method of linear correlation is needed to compute slope. Principal Component Analysis does so [25] by identifying a set of orthogonal basis vectors which minimize aggregate variance normal to the basis. The first principal component is used to compute slope, which after log-normalization [26] , is anticommutative.

The introduced novel branch distance can find consistent differences in the branches of trees more effectively than either PCA or linear regression, as the branch distance is a metric and thus follows the triangle inequality. This means that there is an upper and lower bound on the distances between two structures (embryos in this case) based on the distances to other structures in that space [27]. Neither the coefficient of regression [28] nor any slope-based measurement have this property, and are thus less useful in the context of clustering and comparison [29] [30]. Due to this property, the branch distance metric is very effective at the hierarchical clustering of lineage trees [31]. We were able to use the branch distance to characterize batch effects in WT embryonic cell lineages, while the slope was able to identify, but not compare the extent of batch effects, and the coefficient of regression was able to do neither. In benchmarking this metric, we have revealed greater heterogeneity in this dataset than previously noticed, while showcasing the sensitivity of the branch distance. Indeed, there are many more applications to the branch distance, as it is a nonparametric comparative measure that can work on any ordered graph structure.

The generalization of the branch distance to compare non-identical graph structures is non-trivial. The tree edit distance is an established rigorous metric to quantify the difference in topology of non-identical graphs by counting the necessary addition/operations to transform one tree into another [32]. When input graphs are unlabeled/unordered, the time complexity of the computation increases drastically, with the Zhang-Sasha Tree Edit

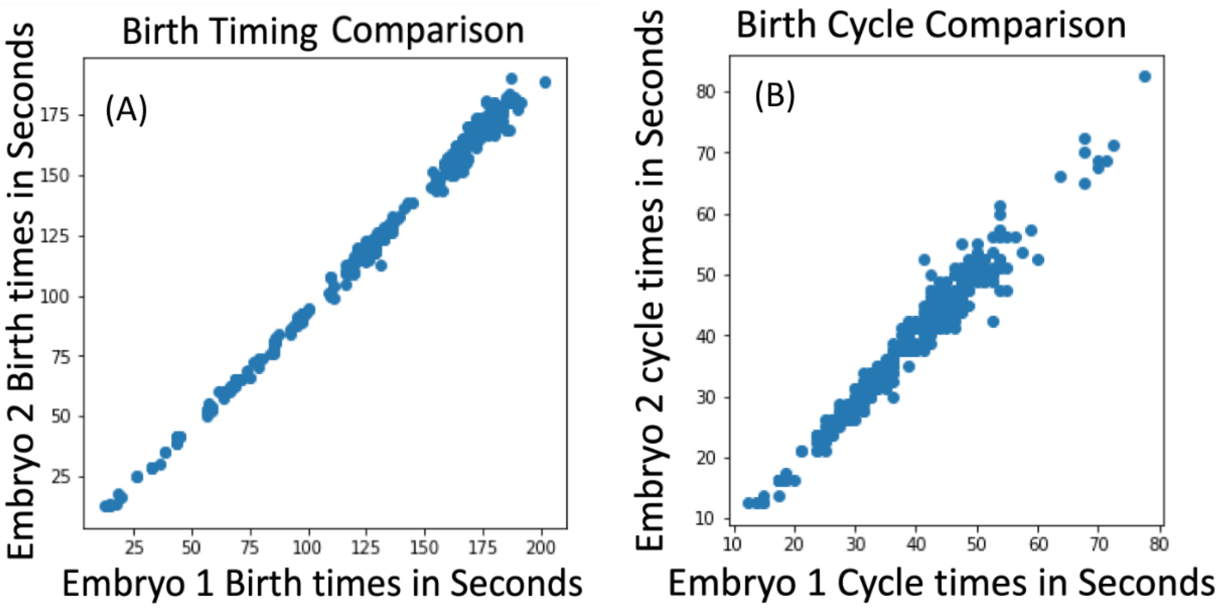
Distance Algorithm [33] having a worst case of O(n^4) [34]. This is indicative of the broader NP-complete problem of graph alignment [35] , which the branch distance metric necessarily operates on. Tree alignment is a relevant problem with various approaches to solve, such as introducing unique metrics [36] and semimetrics [37].

The eutely property and subsequent naming convention of *C. elegans* [38] sidesteps this problem. Since each cell can be labeled

by virtue of its parents with no discrepancies, cross-organism comparisons can directly take place on a cell by cell basis. This changes the tree edit distance calculation into an exclusive disjunction (XOR) operator on the cell labels, which can be finished much more quickly. The jaccard distance [39] can function as a normalized tree edit distance when embryos are converted to sets of labels. The tree edit distance and jaccard distance can be used as measurements that show topological differences between non-identical trees as a reference. If the branch distance is computed between non-identical trees, the tree edit distance and jaccard distances can be used to provide a baseline for the tree topological difference to contextualize the branch distance measurement.

This concept is shown when taking subtree comparisons, where the branch distance, jaccard distance (normalized tree edit distance) and scaling parameters are calculated for each subtree of each embryo. When looking at branch distance matrices, it is worth noting that large distances between lineages may be a result of large topological changes on the lineage instead of large changes in the edge lengths of the tree itself. This is best exemplified with the E lineage, which has significant scaling delays when compared to other lineages and thus a large branch distance. However, the E lineage, as imaged in this context, has no topological changes from the fourth generation of the AB lineage, suggesting that the E lineage is comparatively significantly slowed. The *C. elegans* E cell is the origin of all intestinal cells [40] , which undergo extended cell cycles [41] due to the gap (g2) phase that is exclusive to the E descendants [42]. To examine our metrics further, there are variations within single sublineage blocks that correspond to batch affected embryos. This may be a sign that the timings of certain sub lineages are more sensitive to these batch effects or developmental delays [43] , which warrants further investigation.

Regarding the comparison of the distinct sublineage branch distributions against self comparison branch distributions, a nonparametric permutation test not assuming independence or identically distributed points is used [44]. Since non-self-compared pairwise branch distance distributions are distinguishable from the distributions of self comparison, it can be inferred that distinct lineages have distinct patterns of timing. The sole exception is that the distribution of the branch distances between ABplp and ABprp and the distribution of ABprp branch distance self comparisons are not distinguishable. It is worth noting that the branch distances between ABplp and ABprp are distinguishable from the ABprp branch distance self comparison distribution, implying that ABprp is more tightly regulated than ABplp, and that ABprp and ABplp have similar developmental clocks. Indeed, *C. elegans* fates have been completely mapped [6] , such that all descendants and cell types of each progenitor cell can be identified [45]. The ABplp and ABprp lineages have similarities in lineage development [46] in that both lineages produce neurons with each cell’s anterior sister producing hypodermal cells [47]. The two lineages also exhibit similar migratory behavior [48] and show similar behavior when perturbed [49].


Birth Timing Comparison
(A)
Embryo 2 Birth times in Seconds
Embryo 1 Birth times in Seconds
25
50
75
100
125
150
175
200
Birth Cycle Comparison
(B)
Embryo 2 cycle times in Seconds
Embryo 1 Cycle times in Seconds
10
20
30
40
50
60
70
80

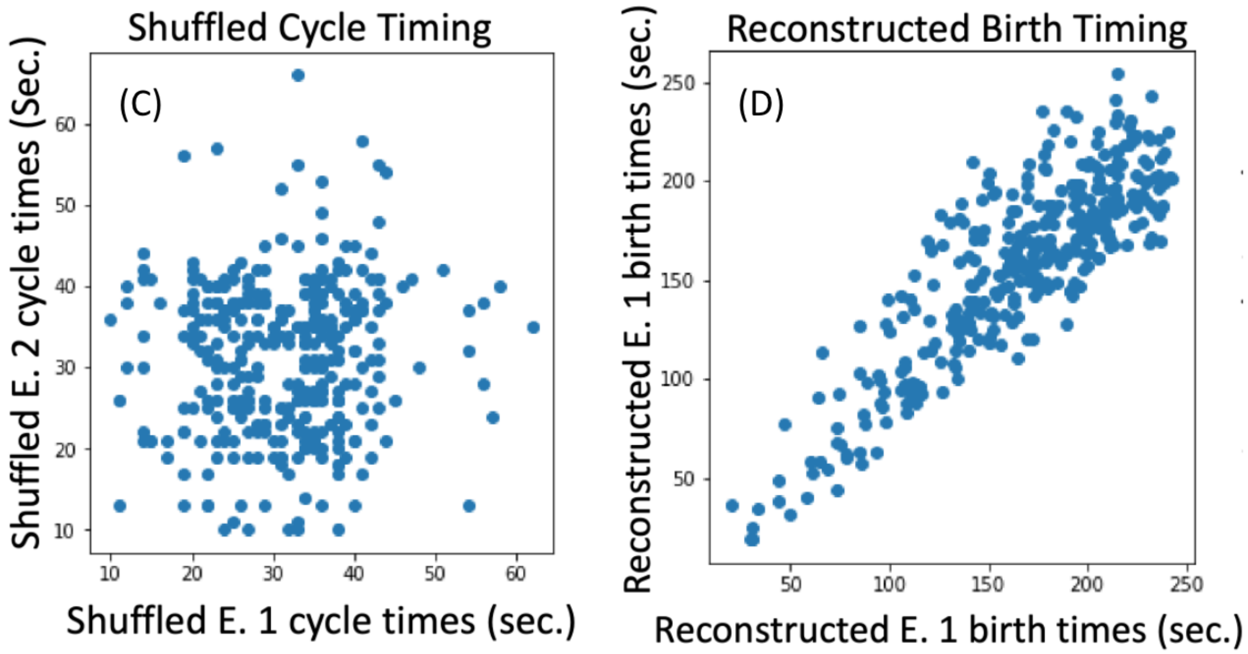


Figure 1: **Benchmarking Comparisons of Absolute Cell birth time and Cell Cycle Time in Wild-Type *C. elegans* embryos.**

(A) The absolute birth times of each cell in some WT Embryo 1 are

plotted against the absolute birth times of each corresponding cell in some WT Embryo 2, which has a coefficient of correlation r^2 = 0.99. **(B)** The cell cycle times for each cell are computed by subtracting the birth time of each cell and its parent. These cell cycle times for each cell are computed for WT Embryo 1 and are plotted against the cell cycle times of each corresponding cell in WT Embryo 2. The calculated coefficient of correlation for these cell cycle times is r^2 = 0.89.

(C) The cell cycle times for each cell in embryo 1 were bootstrapped by assigning a random cell cycle time to each cell. The shuffled embryo 1 cell cycle times are plotted against each the shuffled embryo 2 cell cycle times, with a coefficient of correlation r^2 = 0.005.

(D) The scrambled cell cycle times of each embryo are summed to all of its parent cells to produce reconstructed birth times for each cell. Each cell’s reconstructed birth times in embryo 1 are plotted against each cell’s reconstructed birth times in embryo 2, with a coefficient of correlation r^2 = 0.81.

$$BD(A,B) = \sqrt{\sum_{\epsilon \in E(A) \cap E(B)} (\epsilon_A - \epsilon_B)^2}$$

Equation 1: **Branch Distance Formula for Labeled and Ordered Binary Trees.**

A and B are ordered binary tree representations of lineage trees with uniquely labeled nodes representing specific cells and weighted edges representing cycle times of division events. For each edge that is present in both A and B, the difference in weights between the edges are squared and summed to produce a single value. The square root of this value is then computed to produce the branch distance.

(A) **(B)**

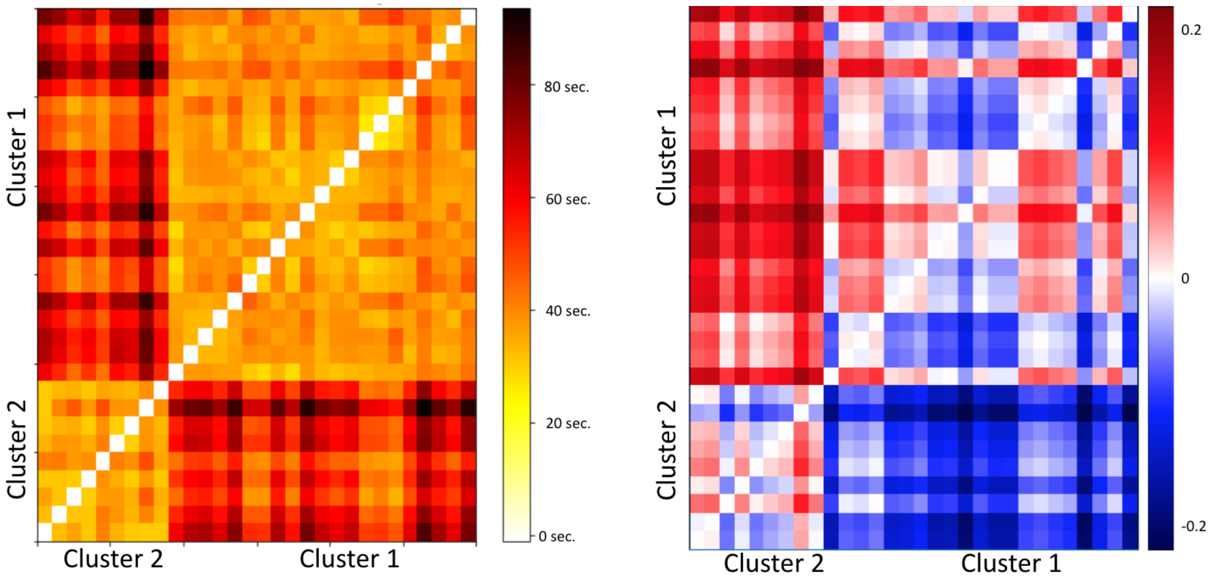


Figure 2: **Branch Distances and Scaling Coefficients between Cell Cycle times of Wild-Type *C. elegans* Embryos, Hierarchically Clustered and Sorted.**

(A) The cell cycles for each cell of thirty Wild-Type embryos were computed before the branch distance matrix between each pair of embryos present was calculated, hierarchically clustered using single, average, and maximal linkage methods, and sorted on the hierarchical clustering.

(B) PCA is used to find the primary and secondary principal components between the regression of the cell cycle times of each embryo. The ratio of these principal components yields the slope, which is used as pairwise scaling coefficient which is calculated between each Wild-Type embryo before being plotted on the heatmap shown with a log scale. The matrix of these scaling coefficients is sorted into two clusters using the hierarchical clustering labels from the branch distance matrix.

$$TED(A,B) = |V(A) \cup V(B) - V(A) \cap V(B)| = |XOR(V(A), V(B))|$$

Equation 2: **Tree Edit Distance Formula for Labeled and Ordered Binary Trees.**

Assuming A and B are ordered, labeled, binary lineage trees, the Tree Edit Distance is computed between ordered and labeled trees by computing the magnitude of the exclusive disjunction between the sets of vertices/nodes.

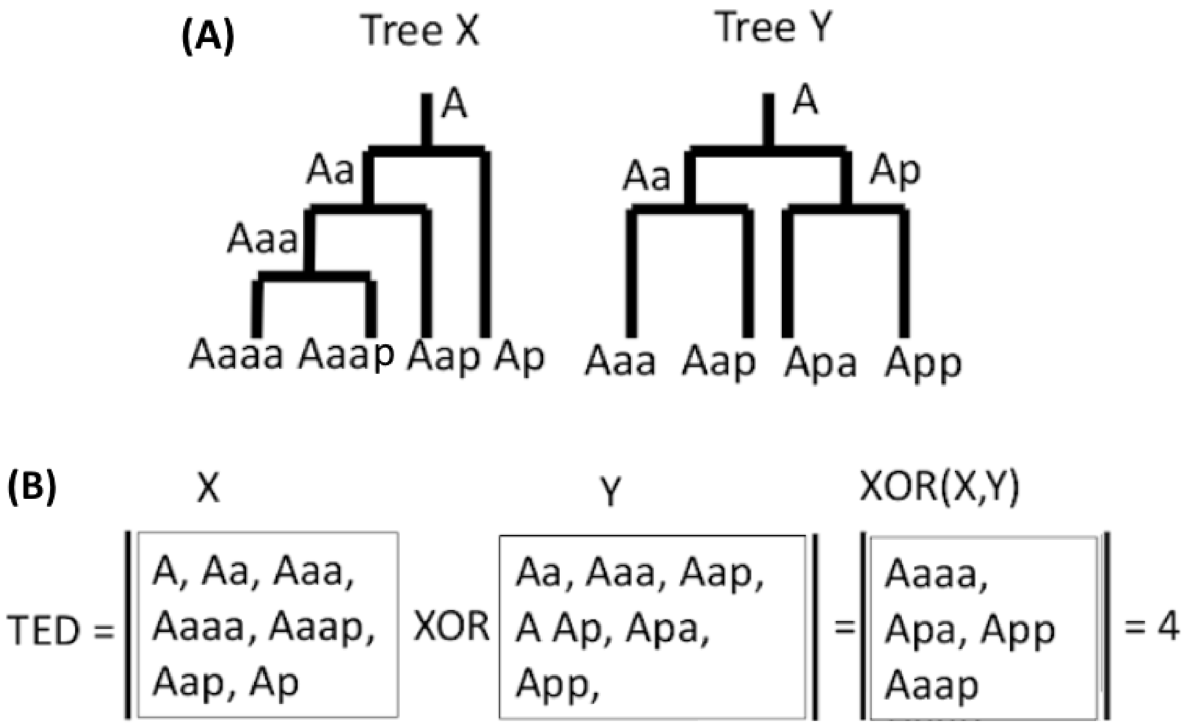


Figure 3: **Tree Edit Distance between Ordered and Labeled Trees.**

(A) Ordered binary trees X and Y are shown, each following C. elegans naming convention with a root cell A and subsequent anterior/posterior divisions, producing a unique topology.

(B) The Tree Edit Distances between example Trees X and Y is computed and illustrated.

$$JD(A,B) = \frac{|V(A) \cup V(B) - V(A) \cap V(B)|}{|V(A) \cup V(B)|} = \frac{|XOR(V(A), V(B))|}{|V(A) \cup V(B)|} = \frac{TED(A,B)}{|V(A) \cup V(B)|}$$

Equation 3: **Jaccard Distance Formula for Labeled and Ordered Binary Trees.**

Assuming A and B are ordered, labeled, binary lineage trees, the Jaccard Distance is computed between ordered and labeled trees by computing the magnitude of the exclusive disjunction between the sets of vertices/nodes and dividing by the magnitude of the union of sets of vertices/nodes.

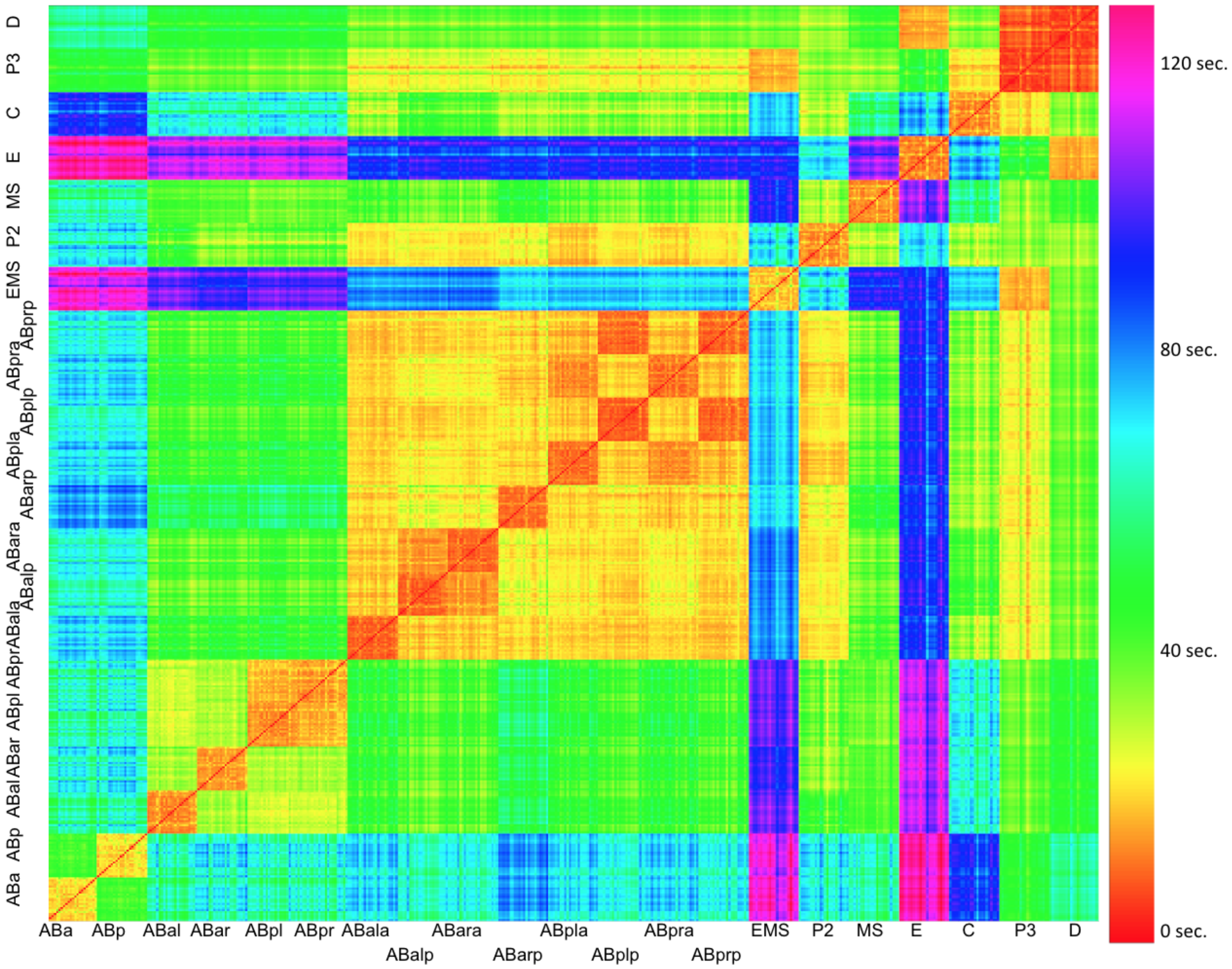


Figure 4: **Pairwise Wild-Type Embryo Sublineage Branch Distances.** For each of the WT embryos present, 21 distinct sublineage founding cells that are produced in the second, third, and fourth generation are cataloged. The descendants of these 21 distinct sublineage founding cells are recorded into 21 distinct sublineage trees. The Pairwise branch distances between each of the 21 subtrees of the 30 WT embryos is computed.

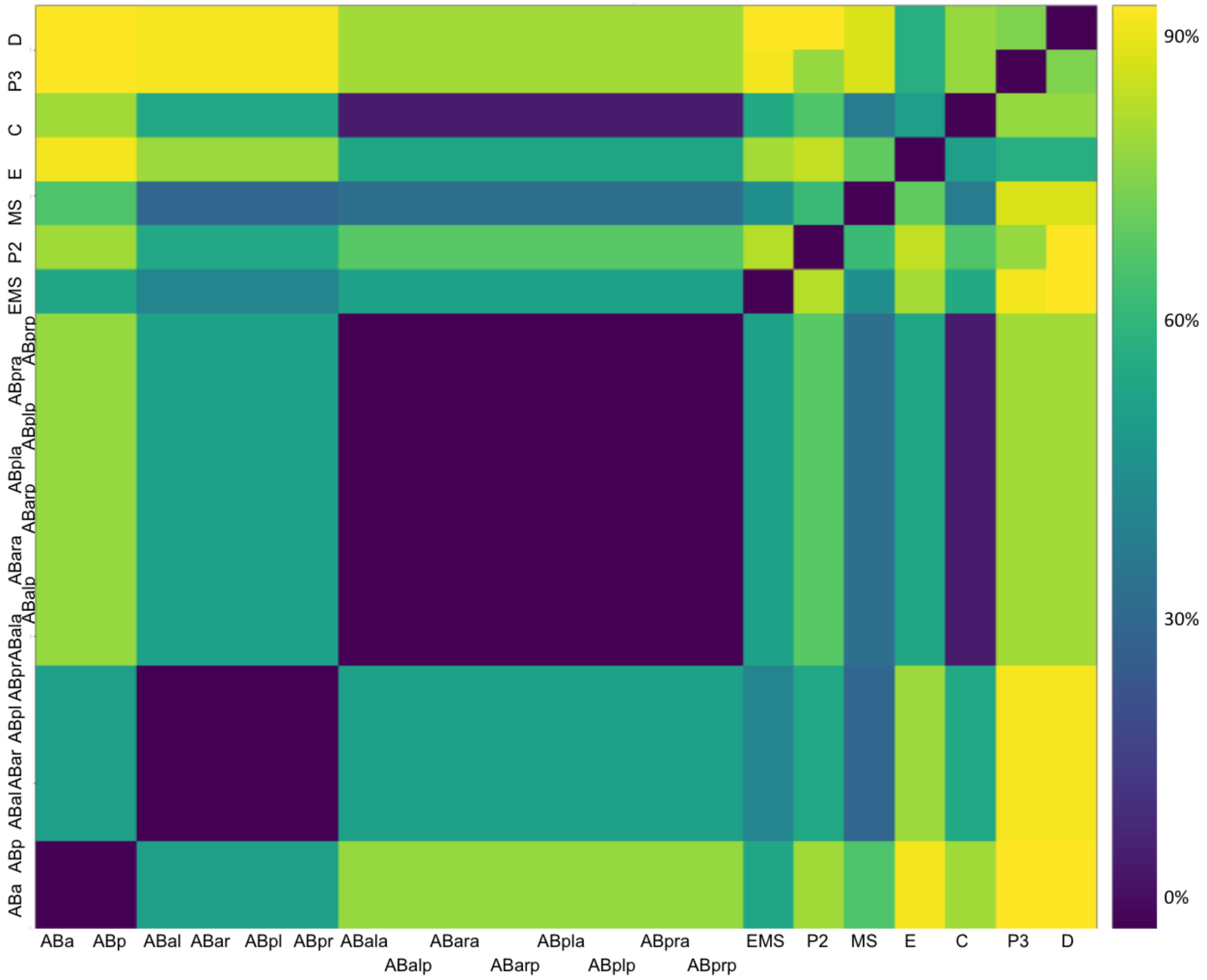


Figure 5: **Pairwise Wild-Type Embryo Sublineage Jaccard Distances.** For each of the WT embryos present, 21 distinct sublineage founding cells that are produced in the second, third, and fourth generation are cataloged. The descendants of these 21 distinct sublineage founding cells are recorded into 21 distinct sublineage trees. The Pairwise Jaccard distances between each subtree of the 30 WT embryos is computed, using units of percent similarity and difference in regards to cells present in the embryo set.

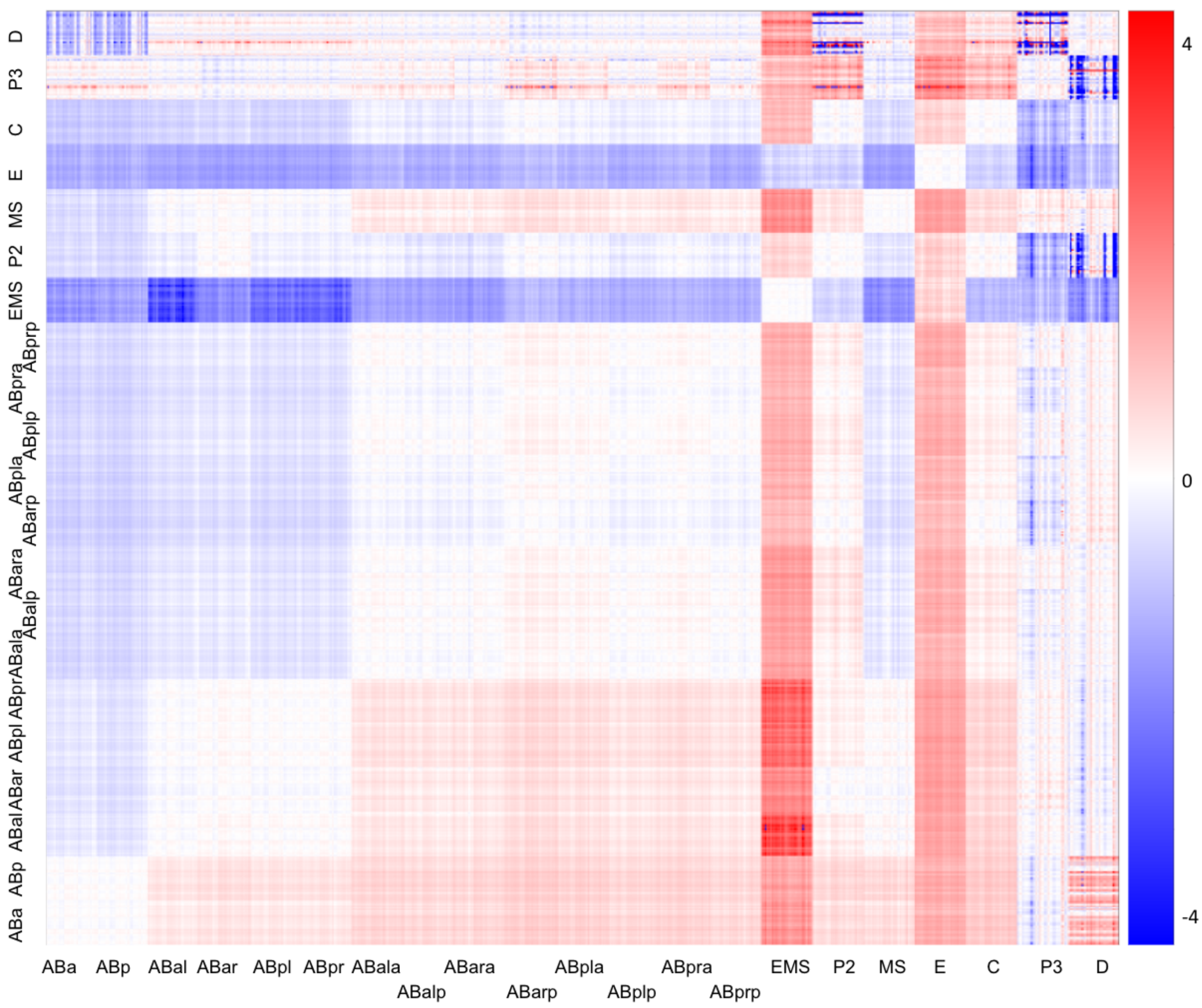


Figure 6: **Pairwise Wild-Type Embryo Sublineage Scaling Coefficients.**

For each of the WT embryos present, 21 distinct sublineage founding cells that are produced in the second, third, and fourth generation are cataloged. The descendants of these 21 distinct sublineage founding cells are recorded into 21 distinct sublineage trees. PCA is used to find the primary and secondary principal components between the regression of the cell cycle times of each WT embryo subtree. The ratio between the first and second principal components are computed as the scaling coefficient before errata are removed (values above 54.6 and below 0.01), before all values are plotted on the heatmap on a log scale. As such, negative values correspond to the logarithms of slopes below 1, while positive values correspond to the logarithms of above 1.

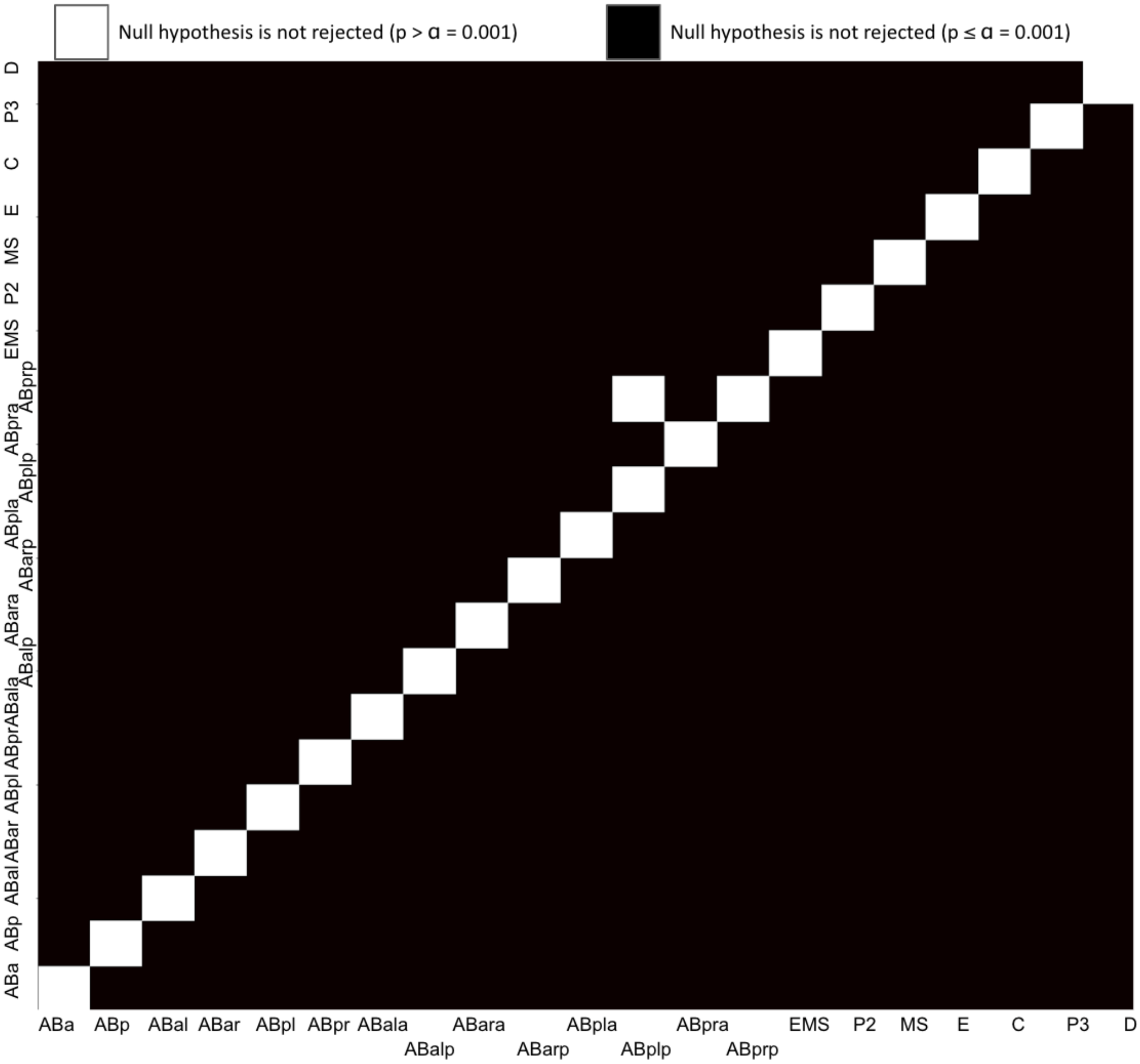


Figure 7: **Distinguishing Distributions of Branch Distances in Between Wild-Type Subtrees using Nonparametric Hypothesis Testing.**

Distributions of Pairwise branch distances between 21 distinct subtrees produces 120 distributions of branch distances between different sublineages and 21 sublineage self comparisons.

Distributions of branch distances between two different subtrees are hypotheses tested against each subtree’s self comparison distribution. Using an initial p value of 0.05, Bonferroni corrected to $\alpha = 0.001$, a permutation test of 10,000 iterations is calculated between each set of distributions.

Bibliography:

[1] Choi, K., & Gomez, S. M. (2009). Comparison of phylogenetic trees through alignment of embedded evolutionary distances. In
BMC Bioinformatics (Vol. 10, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/1471-2105-10-423

[2] De Ugarte, D. A., Morizono, K., Elbarbary, A., Alfonso, Z., Zuk, P. A., Zhu, M., Dragoo, J. L., Ashjian, P., Thomas, B., Benhaim, P., Chen, I., Fraser, J., & Hedrick, M. H. (2003). Comparison of Multi-Lineage Cells from Human Adipose Tissue and Bone Marrow. In Cells Tissues Organs (Vol. 174, Issue 3, pp. 101–109). S. Karger AG. https://doi.org/10.1159/000071150

[3] Schroeder, S., Mache, C., Kleine-Weber, H., Corman, V. M., Muth, D., Richter, A., Fatykhova, D., Memish, Z. A., Stanifer, M. L., Boulant, S., Gultom, M., Dijkman, R., Eggeling, S., Hocke, A., Hippenstiel, S., Thiel, V., Pöhlmann, S., Wolff, T., Müller, M. A., & Drosten, C. (2021). Functional comparison of MERS-coronavirus lineages reveals increased replicative fitness of the recombinant lineage 5. In Nature Communications (Vol. 12, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-021-25519-1

[4] Ong, C.-H. L. (n.d.). On Model-Checking Trees Generated by Higher-Order Recursion Schemes. In 21st Annual IEEE Symposium on Logic in Computer Science (LICS’06). 21st Annual IEEE Symposium on Logic in Computer Science (LICS’06). IEEE. https://doi.org/10.1109/lics.2006.38

[5] Meneely, P. M., Dahlberg, C. L., & Rose, J. K. (2019). Working with Worms:Caenorhabditis elegansas a Model Organism. In Current Protocols Essential Laboratory Techniques (Vol. 19, Issue 1). Wiley. https://doi.org/10.1002/cpet.35

[6] Maduro, M. F. (2010). Cell fate specification in the C. elegans embryo. In Developmental Dynamics (p. NA-NA). Wiley. https://doi.org/10.1002/dvdy.22233

[7] Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. In Developmental Biology (Vol. 100, Issue 1, pp. 64–119). Elsevier BV. https://doi.org/10.1016/0012-1606(83)90201-4

[8] Winston, W. M., Sutherlin, M., Wright, A. J., Feinberg, E. H., & Hunter, C. P. (2007). Caenorhabditis elegans SID-2 is required for environmental RNA interference. In Proceedings of the National Academy of Sciences (Vol. 104, Issue 25, pp. 10565–10570). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0611282104

[9] Feng, Z., Li, W., Ward, A., Piggott, B. J., Larkspur, E. R., Sternberg, P. W., & Xu, X. Z. S. (2006). A C. elegans Model of Nicotine-Dependent Behavior: Regulation by TRP-Family Channels. In Cell (Vol. 127, Issue 3, pp. 621–633). Elsevier BV. https://doi.org/10.1016/j.cell.2006.09.035

[10] Ewald, C. Y., Landis, J. N., Abate, J. P., Murphy, C. T., & Blackwell, T. K. (2014). Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity. In Nature (Vol. 519, Issue 7541, pp. 97–101). Springer Science and Business Media LLC. https://doi.org/10.1038/nature14021

[11] Gönczy, P. (2005). Asymmetric cell division and axis formation in the embryo. In WormBook (pp. 1–20). WormBook. https://doi.org/10.1895/wormbook.1.30.1

[12] Du, Z., Santella, A., He, F., Tiongson, M., & Bao, Z. (2014). De Novo Inference of Systems-Level Mechanistic Models of Development from Live-Imaging-Based Phenotype Analysis. In Cell (Vol. 156, Issues 1–2, pp. 359–372). Elsevier BV. https://doi.org/10.1016/j.cell.2013.11.046

[13] Du, Z., Santella, A., He, F., Shah, P. K., Kamikawa, Y., & Bao, Z. (2015). The Regulatory Landscape of Lineage Differentiation in a Metazoan Embryo. In Developmental Cell (Vol. 34, Issue 5, pp. 592–607). Elsevier BV. https://doi.org/10.1016/j.devcel.2015.07.014

[14] Bao, Z., Zhao, Z., Boyle, T. J., Murray, J. I., & Waterston, R. H.
(2008). Control of cell cycle timing during C. elegans embryogenesis.
In Developmental Biology (Vol. 318, Issue 1, pp. 65–72). Elsevier BV.
https://doi.org/10.1016/j.ydbio.2008.02.054

[15] Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. In Korean Journal of Anesthesiology (Vol. 70, Issue 2, p. 144). The Korean Society of Anesthesiologists. https://doi.org/10.4097/kjae.2017.70.2.144

[16] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. *Array programming with NumPy*. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).

[17] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. In Nature (Vol. 585, Issue 7825, pp. 357–362). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-020-2649-2

[18] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. In Computing in Science &amp; Engineering (Vol. 9, Issue 3, pp. 90–95). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mcse.2007.55

[19] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Team, J. D. (2016, May 31). *Jupyter notebooks – a publishing format for reproducible computational workflows*. IOS Press Ebooks. https://doi.org/10.3233/978-1-61499-649-1-87

[20] Goldstein, B. (2016). Sydney Brenner on the Genetics of Caenorhabditis elegans. In Genetics (Vol. 204, Issue 1, pp. 1–2). Oxford University Press (OUP). https://doi.org/10.1534/genetics.116.194084

[21] Zhang, S., Li, F., Zhou, T., Wang, G., & Li, Z. (2020). Caenorhabditis elegans as a Useful Model for Studying Aging Mutations. In Frontiers in Endocrinology (Vol. 11). Frontiers Media SA. https://doi.org/10.3389/fendo.2020.554994

[22] Fay, D. (2005). The cell cycle and development: Lessons from C. elegans. In Seminars in Cell &amp; Developmental Biology (Vol. 16, Issue 3, pp. 397–406). Elsevier BV. https://doi.org/10.1016/j.semcdb.2005.02.002

[23] van den Heuvel, S., & Kipreos, E. T. (2012). C. elegans Cell Cycle Analysis. In Methods in Cell Biology (pp. 265–294). Elsevier. https://doi.org/10.1016/b978-0-12-394620-1.00009-6

[24] Luck, S. (2022). A parametric framework for multidimensional linear measurement error regression. In S. Saha (Ed.), PLOS ONE (Vol. 17, Issue 1, p. e0262148). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0262148

[25] Alhusain, L., & Hafez, A. M. (2018). Nonparametric approaches for population structure analysis. In Human Genomics (Vol. 12, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s40246-018-0156-4

[26] West, R. M. (2021). Best practice in statistics: The use of log transformation. In Annals of Clinical Biochemistry: International Journal of Laboratory Medicine (Vol. 59, Issue 3, pp. 162–165). SAGE Publications. https://doi.org/10.1177/00045632211050531

[27] Moore, A. (2013). The Anchors Hierachy: Using the triangle inequality to survive high dimensional data (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1301.3877

[28] Jiaxing, C., Ng, Y. K., Lin, L., Jiang, Y., & Li, S. (2019). On triangular inequalities of correlation-based distances for gene expression profiles. Cold Spring Harbor Laboratory. https://doi.org/10.1101/582106

[29] Baraty, S., Simovici, D. A., & Zara, C. (2011). The Impact of Triangular Inequality Violations on Medoid-Based Clustering. In Lecture Notes in Computer Science (pp. 280–289). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21916-0_31

[30] Chang, C.-S., Liao, W., Chen, Y.-S., & Liou, L.-H. (2016). A Mathematical Theory for Clustering in Metric Spaces. In IEEE Transactions on Network Science and Engineering (Vol. 3, Issue 1, pp. 2–16). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tnse.2016.2516339

[31] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. In H. A. Kestler (Ed.), PLOS ONE (Vol. 14, Issue 1, p. e0210236). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0210236

[32] Schwarz, S., Pawlik, M., & Augsten, N. (2017). A New Perspective on the Tree Edit Distance. In Similarity Search and Applications (pp. 156–170). Springer International Publishing. https://doi.org/10.1007/978-3-319-68474-1_11

[33] Zhang, K., & Shasha, D. (1989). Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. In SIAM Journal on Computing (Vol. 18, Issue 6, pp. 1245–1262). Society for Industrial & Applied Mathematics (SIAM). https://doi.org/10.1137/0218082

[34] Dulucq, S., & Touzet, H. (2005). Decomposition algorithms for the tree edit distance problem. In Journal of Discrete Algorithms (Vol. 3, Issues 2–4, pp. 448–471). Elsevier BV. https://doi.org/10.1016/j.jda.2004.08.018

[35] Hall, G., & Massoulié, L. (2020). Partial Recovery in the Graph Alignment Problem (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2007.00533

[36] Kim, J., Rosenberg, N. A., & Palacios, J. A. (2020). Distance metrics for ranked evolutionary trees. In Proceedings of the National Academy of Sciences (Vol. 117, Issue 46, pp. 28876–28886). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1922851117

[37] Wang, Y. (2021). Two Metrics on Rooted Unordered Trees with Labels (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2103.11553

[38] Bao, Z., Murray, J. I., Boyle, T., Ooi, S. L., Sandel, M. J., & Waterston, R. H. (2006). Automated cell lineage tracing in Caenorhabditis elegans. In Proceedings of the National Academy of Sciences (Vol. 103, Issue 8, pp. 2707–2712). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0511111103

[39] Levandowsky, M., & Winter, D. (1971). Distance between Sets. In Nature (Vol. 234, Issue 5323, pp. 34–35). Springer Science and Business Media LLC. https://doi.org/10.1038/234034a0

[40] Goldstein, B. (1993). Establishment of gut fate in the E lineage of C. elegans: the roles of lineage-dependent mechanisms and cell interactions. In Development (Vol. 118, Issue 4, pp. 1267–1277). The Company of Biologists. https://doi.org/10.1242/dev.118.4.1267

[41] Robertson, S. M., Medina, J., & Lin, R. (2014). Uncoupling Different Characteristics of the C. elegans E Lineage from Differentiation of Intestinal Markers. In B. Goldstein (Ed.), PLoS ONE (Vol. 9, Issue 9, p. e106309). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0106309

[42] Arata, Y., Takagi, H., Sako, Y., & Sawa, H. (2015). Power law relationship between cell cycle duration and cell volume in the early embryonic development of Caenorhabditis elegans. In Frontiers in Physiology (Vol. 5). Frontiers Media SA.

https://doi.org/10.3389/fphys.2014.00529

[43] Singhal, A., & Shaham, S. (2017). Infrared laser-induced gene expression for tracking development and function of single C. elegans embryonic neurons. In Nature Communications (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/ncomms14100

[44] Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. In Statistics &amp; Probability Letters (Vol. 36, Issue 1, pp. 9–21). Elsevier BV. https://doi.org/10.1016/s0167-7152(97)00043-6

[45] Krause, M. (1999). Cell Fate Determination in Caenorhabditis elegans. In Development (pp. 251–267). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-59828-9_16

[46] Ho, V. W. S., Wong, M., An, X., Guan, D., Shao, J., Ng, H. C. K., Ren, X., He, K., Liao, J., Ang, Y., Chen, L., Huang, X., Yan, B., Xia, Y., Chan, L. L. H., Chow, K. L., Yan, H., & Zhao, Z. (2015). Systems‐level quantification of division timing reveals a common genetic architecture controlling asynchrony and fate asymmetry. In Molecular Systems Biology (Vol. 11, Issue 6, p. 814). EMBO. https://doi.org/10.15252/msb.20145857

[47] Riddle DL, Blumenthal T, Meyer BJ, et al., editors. C. elegans II. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. Section II, Specification of Cell Fates in the AB Lineage. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20169/

[48] Schnabel, R., Bischoff, M., Hintze, A., Schulz, A.-K., Hejnol, A., Meinhardt, H., & Hutter, H. (2006). Global cell sorting in the C. elegans embryo defines a new mechanism for pattern formation. In Developmental Biology (Vol. 294, Issue 2, pp. 418–431). Elsevier BV. https://doi.org/10.1016/j.ydbio.2006.03.004

[49] Hutter, H., & Schnabel, R. (1995). Specification of anterior-posterior differences within the AB lineage in the C. elegans embryo: a polarising induction. In Development (Vol. 121, Issue 5, pp. 1559–1568). The Company of Biologists. https://doi.org/10.1242/dev.121.5.1559