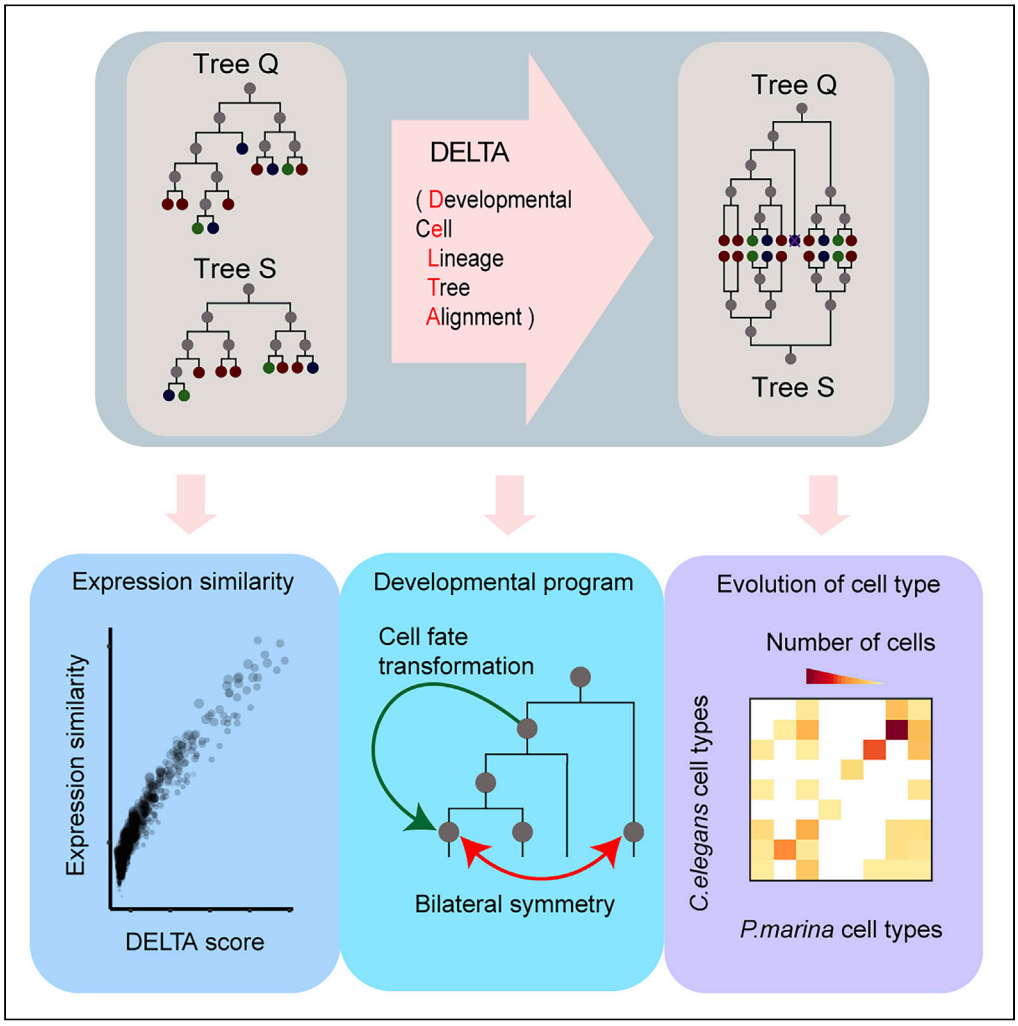
iScience

CellPress
OPEN ACCESS

Article

Alignment of Cell Lineage Trees Elucidates Genetic Programs for the Development and Evolution of Cell Types



Meng Yuan, Xujiang Yang, Jinghua Lin, ..., Xueqin Wang, Xiaoshu Chen, Jian-Rong Yang

chenxshu3@mail.sysu.edu.cn (X.C.)
yangjianrong@mail.sysu.edu.cn (J.-R.Y.)

HIGHLIGHTS

Align cell lineage trees (CLTs) to search/quantify their phenotypic similarities

Aligning worm CLTs captured known genetic/developmental programs

Similarities between knockdown CLTs revealed functional relationships between genes

CLT alignments between species gave insight on the evolution of cell types

Yuan et al., iScience 23, 101273
July 24, 2020
https://doi.org/10.1016/j.isci.2020.101273

iScience

CellPress
OPEN ACCESS

Article

Alignment of Cell Lineage Trees Elucidates Genetic Programs for the Development and Evolution of Cell Types

Meng Yuan,[1,7] Xujiang Yang,[1,7] Jinghua Lin,[2,7] Xiaolong Cao,[1,7] Feng Chen,[1] Xiaoyu Zhang,[1] Zizhang Li,[1] Guifeng Zheng,[2] Xueqin Wang,[3] Xiaoshu Chen,[4,*] and Jian-Rong Yang[1,5,6,8,*]

SUMMARY

A full understanding of the developmental process requires fine-scale characterization of cell divisions and cell types, which are naturally organized as the developmental cell lineage tree (CLT). Technological breakthroughs facilitated determination of more CLTs, but complete comprehension of the data remains difficult without quantitative comparison among CLTs. We hereby quantified phenotypic similarity between CLTs using a novel computational method that exhaustively searches for optimal correspondence between individual cells meanwhile retaining their topological relationships. The revealed CLT similarities allowed us to infer functional similarity at the transcriptome level, identify cell fate transformations, predict functional relationships between mutants, and find evolutionary correspondence between cell types of different species. By allowing quantitative comparison between CLTs, our work is expected to greatly enhance the interpretability of relevant data and help answer the myriad of questions surrounding the developmental process.

INTRODUCTION

The life of multicellular organisms typically starts from a zygote, which undergoes multiple rounds of cell divisions and simultaneous differentiation and eventually develops into an individual organism with multiple types of cells. The developmental *c*ell *l*ineage *t*ree (CLT) is a record of both the differentiation result of the cells appearing at a specific developmental time point (cell types the terminal nodes of the CLT) and the cell division events since the zygote that led to these cells (topology of the CLT) (Figure 1A). A more generalized CLT does not necessarily root at the zygote but may start from any dividable cell, in which case it is a subtree of the CLT rooted at the zygote or a sub-CLT (Figure 1A). As one of the most important traits of multicellular organisms, the CLT is the key to resolving many significant problems in the life sciences. For example, developmental CLTs record the process of development (Du et al., 2014, 2015; Junker et al., 2017; Kalhor et al., 2017; McKenna et al., 2016; Santella et al., 2016; Sulston et al., 1983) and help understanding the mechanism of developmental robustness (Reizel et al., 2012; Salipante et al., 2010; Yang et al., 2014). Other types of CLTs reveal the origin of relapsed or metastatic tumor cell populations (Frumkin et al., 2008; Shlush et al., 2012), the risk of carcinogenesis attributable to the number of cell divisions since the zygote (Tomasetti et al., 2017; Wasserstrom et al., 2008), and the origin and evolution of cell types and lineages (Arendt et al., 2016; Lescroart et al., 2015).

Since the resolution of the first complete cell lineage tree in *Caenorhabditis elegans* (Sulston et al., 1983), technological advancements ranging from 3D time-lapse imaging (Gritti et al., 2016) to genome editing in combination with single-cell high-throughput sequencing (Junker et al., 2017; Kalhor et al., 2017; McKenna et al., 2016; Raj et al., 2018a, 2018b) had fueled the accumulation of more CLT data. However, a general computational framework for quantitative comparison of CLTs has been lacking. Take the classical CLT of *Caenorhabditis elegans* for example, phenotypic comparison and functional inference were previously made on the predefined lists of developmental phenotypes (Gunsalus et al., 2005; Piano et al., 2002). This approach had not fully utilized the rich information embedded in the CLT and cannot reveal finer scale correspondence between individual cells. Quantitative comparison of CLTs should facilitate quality assessment of CLT data, relating new observations to the known, disentangling variation from the consensus, and evolutionary comparative studies.

[1]Department of Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China

[3]Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

[4]Department of Medical Genetics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

[5]RNA Biomedical Institute, Sun Yat-Sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China

[6]Key Laboratory of Tropical Disease Control, Ministry of Education, Sun Yat-sen University, Guangzhou 510080, China

[7]These authors contributed equally

[8]Lead Contact

*Correspondence: chenxshu3@mail.sysu.edu.cn (X.C.), yangjianrong@mail.sysu.edu.cn (J.-R.Y.)

https://doi.org/10.1016/j.isci.2020.101273

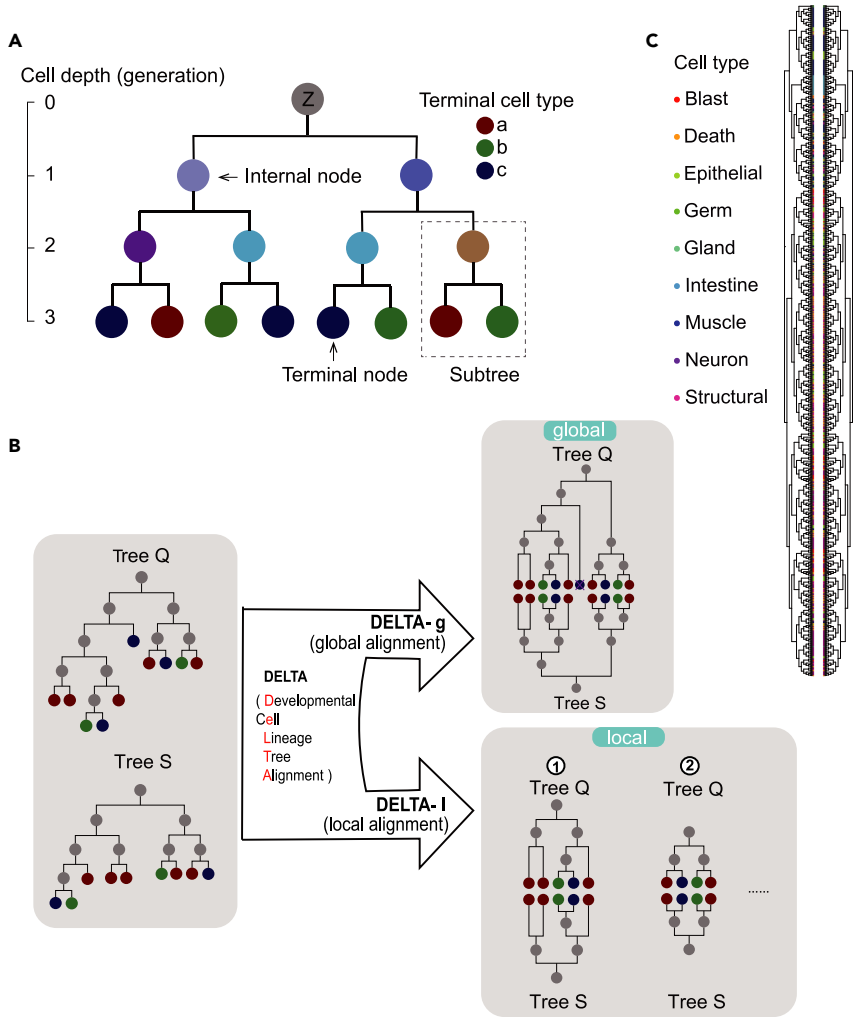


Figure 1. Overview of the DELTA Algorithm

(A) A very simple typical developmental CLT rooted at the zygote ("Z"). Cells undergoing further division are represented by internal nodes and others by terminal nodes. The cell types of the terminal nodes are indicated by the colors in the legend on the right, whereas the type of the internal cells, as inferred from the cell type of the two daughter cells, are also indicated by different node colors. The depth, or the number of divisions since the zygote, of a cell is indicated by the vertical axis. A subtree, or sub-CLT, is outlined by a dotted box.

(B) Two CLTs to be aligned, *Q* and *S*, are presented with their terminal cell types color coded. DELTA aligns them globally (DELTA-g), where all cells in respective CLTs are either pruned or aligned, or locally (DELTA-l), where only pairs of sub-CLTs with good enough alignments are reported (see Figure S1A for more details).

(C) The DELTA alignment of the *C. elegans* CLT of standard anatomical terminal cell type annotation, with an isomorphic version of itself, where 30% of randomly chosen sister sub-CLT pairs were swapped. The resulting CLT alignments were visualized by our newly developed R package, "ggVITA" (see also Figure S1C). See also Figure S1.

To address this critical demand, we designed *D*evelopmental *C*ell *L*ineage *T*ree *A*lignment (DELTA), an algorithm that aligns a pair of CLTs and quantifies the similarity between them by identifying homeomorphic sub-CLTs, with the assumption that similar genetic (or developmental) programs should give rise to similar sub-CLTs (Azevedo et al., 2005; Yang et al., 2014). A critical feature of DELTA is its compatibility with classical CLTs (such as those of nematodes) and the genome-editing-based lineage CLTs, as it only required the topology and the terminal cell types of the CLT to work. Using simulated CLTs (Lohaus et al., 2007) and real CLTs from *C. elegans* (Murray et al., 2012), we showed that homeomorphic sub-CLTs found by DELTA have highly similar expression profiles. Comparisons among CLTs of the wild-type and single-gene knockdown strains of *C. elegans* (Santella et al., 2016) revealed both known (Du et al., 2015) and novel homeotic transformations of cell fates in the knockdown strains and suggested for the knockdown genes functional

relationships compatible with evolutionary and experimental evidence. Finally, we compared the developmental CLTs of two nematodes and were able to pinpoint the evolutionary changes in fates between cells in these two CLTs. By maximizing the alignment score between real CLTs of the two species, we found biologically interpretable correspondence between their nonuniformly defined cell types, highlighting a conceptually new way of inferring the evolutionary relationship between cell types. Together, these results recapitulated known developmental patterns and demonstrated the usefulness of DELTA. In the way that sequence alignment algorithms fundamentally transformed genetics, CLT comparison/alignment enabled by DELTA will likely lead to new opportunities for a deeper understanding of the biology of multicellular organisms, such as assessing the repeatability of differentiation, linking sub-CLTs to developmental programs, and distinguishing autonomous and regulatory components involved in development.

RESULTS

Overview of the DELTA Algorithm

A typical developmental CLT, as analyzed here, is a binary tree (Figure 1A), where each node represents a single cell and each branch represents a descendant relationship from a mother cell to one of its daughter cells. The cells in the tree can be divided into internal or terminal cells/nodes based on whether they undergo further division as recorded by the CLT. A subtree rooted at any of the cells is a sub-CLT. The terminal cells of the CLT are all labeled by their cell types, which could be anatomically defined as, for example, muscle or neural cells, or defined by the expression state of one or more genes such as CD4+ cells. Note that, in contrast to the CLTs commonly discussed in nematodes such as *C. elegans*, we ignored the temporal duration of the cell cycle and the order of sister cells to ensure compatibility with CLTs determined by genome editing. In other words, the length of the branch contains no information about how long each cell exists and swapping any pair of sister sub-CLTs (i.e., two sub-CLTs whose roots are a pair of sister cells divided from the same mother) will not change the CLT.

We designed the DELTA algorithm with the purpose of identifying similarities in developmental programs using the phenotypic information represented by the CLT. Here the developmental program is a succession of cell fate choices made at every division event recorded by the CLT. We assumed that the developmental state of a cell is reflected by the states of its daughter cells, which are further defined by their own daughter cells until reaching the terminal cells with known cell types (Figure 1A, color of nodes). In other words, a pair of cells is similar if the two sub-CLTs rooted at them resemble each other in topology and lineal organization of terminal cell types. This assumption was deemed useful in demonstrating the simplicity (Azevedo et al., 2005) and robustness (Yang et al., 2014) of metazoan CLTs, as well as in identifying homeotic transformation of cell fates (Du et al., 2015). DELTA compares every sub-CLT from a query CLT with those from a subject CLT and exhaustively searches for their maximal resemblance in terms of topology and lineal organization of terminal cell types via a dynamic programming strategy (Figure S1A). As an analogy, sequence alignment algorithms align residues in biological sequences with the constraint of their sequential order, whereas DELTA aligns terminal and internal cells in CLTs with the constraint of their lineal organization. DELTA can align CLTs globally (DELTA-g), where all cells in respective CLTs are either pruned or aligned, or locally (DELTA-l), where only pairs of sub-CLTs with good enough alignments are reported, very similar to global and local sequence alignment, respectively (Figure 1B). DELTA also estimates the statistical significance of each CLT alignment relative to random pairs of CLTs with the same sizes and terminal cell type compositions as the aligned CLTs. More algorithmic details of DELTA are given in the Transparent Methods section and Supplemental Information.

As a basic validation of DELTA, we aligned a *C. elegans* CLT with an isomorphic version of itself, where 30% of randomly chosen sister sub-CLT pairs were swapped. DELTA-g successfully aligned the isomorphic CLT with the original CLT by matching all terminal nodes, yielding the same DELTA score as that of the alignment between two identical *C. elegans* CLTs (Figure 1C; see also Figure S1B). We also developed an accompanying R package named *gg*tree-based *vi*sualization of *t*ree *a*lignments (ggVITA) for the visualization of DELTA alignments (Figure 1C; see also Figure S1C).

CLT Simulations Suggest that DELTA Can Identify Developmental Similarities

To further demonstrate that DELTA alignment can indeed reveal developmental similarities, we simulated CLTs using a previously published model (Lohaus et al., 2007), in which the gene expression status (on/off) of each gene in each cell and at each discrete time point was calculated by a predefined regulatory network (Figures 2A and 2B; see Transparent Methods). DELTA-l was used to align the simulated CLT with itself. This

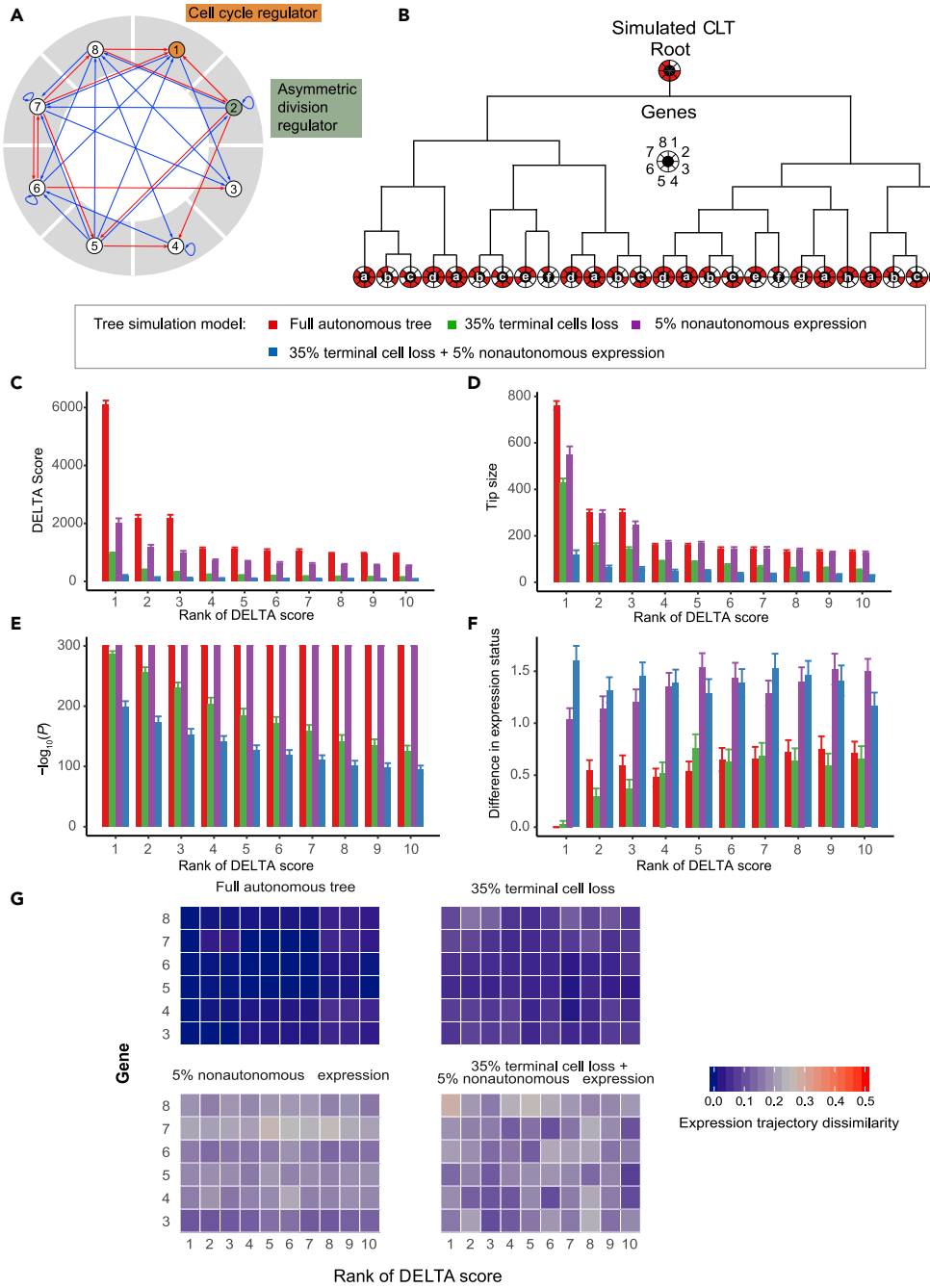


Figure 2. Validating DELTA by Simulated CLTs

(A) An example of a transcriptional regulatory network used to simulate CLTs. There are eight genes, each of which is regulated by an average of four other genes. Red and blue lines represent activations and repressions, respectively. Values in the matrix of regulatory interactions are detailed in Figure S2A. The sectors underneath the genes represent the sectors shown for each cell in (B).

(B) An example of a simulated CLT. Lowercase letters a to h represent eight terminal cell types based on the ON/OFF state of eight genes, which is shown as red (ON) or white (OFF) in the corresponding sector around the letter. The development of the lineage tree stops when one of the terminal cells reaches 12 rounds of divisions or the 50th discrete time point of the simulation. CLTs were also simulated under other parameter settings, as listed in Figure S2B and explained in the Transparent Methods.

(C) The score of the top 10 CLT alignments found by DELTA-l from self-comparison of simulated CLTs. In addition to the full simulated CLT (red bars), perturbations were added to mimic experimental/biological noises, such as the 35% loss of

Figure 2. *Continued*

terminal cells (green bars), non-autonomous cell fate (purple bars), or both (blue bars). Each bar shows the average score and its standard error assessed by 1,000 simulations with different regulatory networks and initial expression states.

(D–F) Similar to (C), except that the number of terminal cells in the aligned CLT (D), statistical significance of the alignment (E) (*P* values truncated at 10^{-300}), and difference in expression status between roots of aligned CLTs as measured by the Hamming distance (F) are plotted. Please see Figure S3 for results from simulations with cell loss rate of 5%, 10%, 20%, 50%, or 90%.

(G) For the same set of top 10 CLT alignments presented in (C) (*x* axis), the expression trajectory dissimilarity of every gene except genes 1 and 2 (*y* axis) is shown. The expression states of a specific gene at the last time point of every cell were compared for every pair of matched cells (terminal or internal) from the two aligned CLTs. The resulting Hamming distance is normalized by the number of matched cell pairs, giving rise to the expression trajectory dissimilarity, the value of which is scaled based on the color scale bar to the right. All expression trajectory dissimilarities are average values from the DELTA-I results of 1,000 simulations. The expression trajectory dissimilarities between the aligned cells are clearly much lower than the null expectation 0.5 for the normalized Hamming distance. See also Figures S2 and S3.

process was repeated with 1,000 different simulated CLTs, and the top ten local alignments from each simulation were examined to assess the performance of DELTA. Several results in support of the capability of DELTA were observed. First, we found that, for each simulated CLT, the self-alignment always had the highest DELTA score in the local alignment result (Figure 2C, left-most red bar). Second, DELTA tended to find alignments between large sub-CLTs, which contained more developmental information (Figure 2D, red bars). Third, the CLT alignments were statistically highly significant, indicating that DELTA scores were much higher than those between random CLTs of similar sizes and terminal cell type compositions (Figure 2E, red bars). Fourth, the differences in gene expression status, as measured by the Hamming distance (the number of genes with differences in expression status between two cells), between the roots of the aligned sub-CLTs was much smaller than that between two randomly chosen internal cells and tended to be lower for those with higher DELTA scores (Figure 2F, red bars). Fifth, by comparing the expression status of each gene for all aligned (terminal and internal) cells in a pair of (sub-)CLTs, we found that their expression trajectories were much more similar (Hamming distance scaled to [0,1], therefore having an expectation of 0.5) between aligned internal or terminal cells than expected (Figure 2G), suggesting that not only the initial state but also the subsequence changes in expression were highly similar between the aligned CLTs. These results demonstrated that DELTA can indeed pair internal cells with similar gene expression status.

Two practical considerations prompted us to further scrutinize the performance of DELTA using revised CLT simulation models. First, the aforementioned model used for simulated CLTs implicitly assumed that all cells differentiate autonomously, whereas the real differentiation process is believed to be highly regulated. To mimic such regulation by external signals from other cells or the environment, we introduced a probability (5%) of randomly flipping the gene expression status of a gene at every time point during the simulated development by negating its expression level (see Transparent Methods), excluding the cell cycle and asymmetric division regulator. Second, current experimental techniques are not perfect in capturing all cells, making the experimentally reconstructed CLTs incomplete. For example, the state-of-the-art lineage tree reconstruction method used a droplet-based single-cell sequencing system (Raj et al., 2018a) and the most commonly used commercial single-cell sequencing platform (10x Chromium) has a cell loss rate of ~35%. To reflect such technical limitations, we simulated two CLTs with identical initial parameters (regulatory network and expression of the root), randomly removed 35% of terminal cells from each CLT, reconstructed the simulated CLTs following the actual lineal relationship among the remaining terminal cells (see Transparent Methods), and aligned them by DELTA. As expected, these two perturbations reduced the DELTA score (Figure 2C, green and blue bars), CLT size (Figure 2D, green and blue bars), statistical significance (Figure 2E, green and blue bars), and gene expression similarity between the aligned sub-CLTs (green and blue bars in Figure 2F and the two panel on the right of Figure 2G). Nevertheless, DELTA is still capable of identifying statistically significant sub-CLTs with highly similar gene expression status (Figures 2C–2G). We found by additional simulations that statistically significant alignments could readily be found with 5%, 10%, 20%, or 50% cell losses, but not 90% (Figure S2). These results suggest that, despite the detection power reduction due to stochastic perturbations, associating CLT phenotypes with underlying gene expression status by DELTA remains feasible as long as $\geq$50% of the terminal cells of the real CLT were captured by the reconstructed CLT. Note, however, recent genome-editing-based CLT reconstruction efforts have not yielded CLTs with cell capture rate $\geq$50%; we therefore refrained from trying DELTA on genome-editing-based CLTs (see more details in Limitations of the Study).

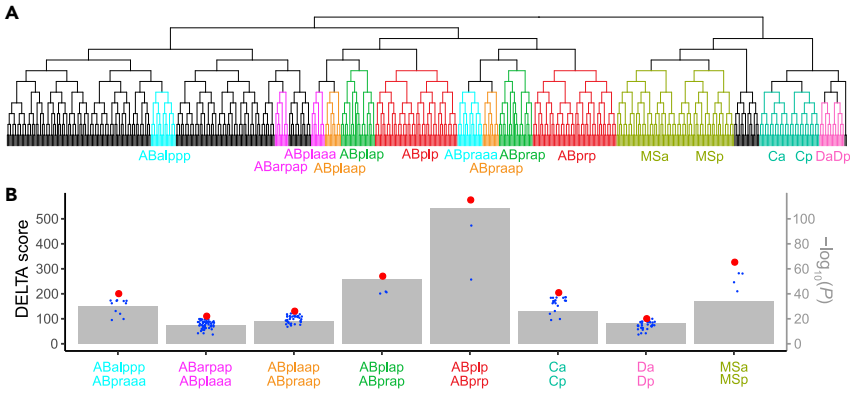


Figure 3. Bilaterally Symmetric sub-CLTs Yielded Highly Significant DELTA Alignments

(A) Bilaterally symmetric sub-CLTs with at least 10 terminal cells in the *C. elegans* CLT are highlighted by different colors, whereas the lineal names of their root are also marked below every sub-CLT.

(B) The alignment score (dots, scaled by the left *y* axis) and statistical significance (gray boxes, scaled by the right *y* axis, see Transparent Methods) found by DELTA-g alignment between the symmetric sub-CLTs are indicated by red dots. As controls, sub-CLTs with a number of terminal cells differing from the symmetric sub-CLTs by no more than 10% were also compared with one of the symmetric sub-CLTs, and the resulting DELTA scores are indicated by blue dots. Please see Table S1 for DELTA results between other bilaterally symmetric sub-CLTs with less than 10 terminal cells. See also Table S1.

Comparison of *C. elegans* CLTs by DELTA Reveals Cells with Highly Similar Developmental Programs

Next, we sought to test the performance of DELTA using real CLTs from *C. elegans*. The *C. elegans* embryonic CLT contains 31 pairs of bilaterally symmetrical sub-CLTs (Sulston et al., 1983), 8 of which have 10 or more terminal cells (Figure 3A). We aligned these symmetric pairs of sub-CLTs by DELTA-g and found that their DELTA scores were highly significant (Figure 3B, gray bars) and always higher than those from the alignments between one sub-CLT from the symmetric pairs and another sub-CLT with a similar number of terminal cells (Figure 3B, blue dots; see Transparent Methods). For the 17 symmetric pairs of sub-CLTs with three to nine terminal cells, all but one (ABarappa versus ABalpaaa) pair gave rise to statistically significant CLT alignments (Table S1). The remaining six symmetric pairs of sub-CLTs cannot be aligned by DELTA as they have <3 terminal cells (Table S1).

To further assess the capability of DELTA to find similarities in developmental programs, we take advantage of the Expression Patterns in *C. elegans* (EPIC) database, where the expression of 130 genes is tracked in each cell during the embryonic development of *C. elegans*, from the zygote to the last round of embryonic cleavage (the 350-cell stage) (Murray et al., 2012). We collected all statistically significant (nominal $P < 0.05$) entries from the top 2,000 sub-CLT alignments in the DELTA-l results of a *C. elegans* CLT versus itself and calculated the Pearson's correlation coefficient R between the aligned cells in the sub-CLTs (See Transparent Methods). Since the sizes of different alignments vary, the Pearson's R values are standardized by Fisher's r-to-z transformation before being compared. In support of the usefulness of DELTA, a higher z is observed in CLT alignments with higher DELTA scores (Figure 4A, Pearson's $R = 0.951$, $P < 10^{-300}$, Spearman's $\rho = 0.926$, $P < 10^{-300}$) and more significant alignment P values (Figure 4B, Pearson's $R = 0.505$, $P < 10^{-98}$, Spearman's $\rho = 0.177$, $P < 10^{-11}$). In combination with the results from the simulated CLTs, we demonstrated that DELTA can indeed identify CLTs with highly similar developmental programs.

Phenotypic Differences in Knockdown CLTs Quantified by DELTA Reveal Functional Relationships among Underlying Genes

Inspired by the capability of DELTA to identify similarities in developmental programs by homeomorphic (sub-)CLTs, we continued to test whether DELTA can associate CLT changes with their underlying genetic mechanisms. The Digital Development database (Santella et al., 2016), where CLTs are recorded for *C. elegans* strains with ~200 conserved genes individually knocked down (knockdown strains) provides a unique opportunity to compare phenotypic changes in CLTs with underlying genetic differences.

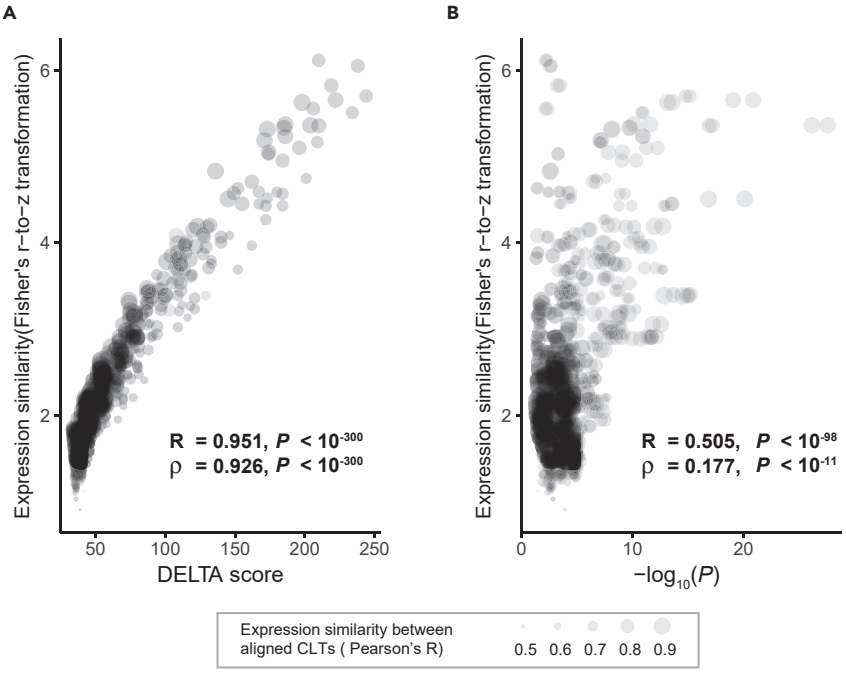


Figure 4. Expression Similarities between Aligned sub-CLTs in *C. elegans*

The top 2,000 local alignments between sub-CLTs of *C. elegans* were found by DELTA-l, and those with a nominal $P < 0.05$ were checked for expression similarities between aligned cells. For each sub-CLT alignment, the expression levels of 130 genes recorded in the EPIC database were compared for all aligned cells, giving rise to the expression similarity between the two aligned sub-CLTs in the form of Pearson's correlation coefficient (point size). The expression similarities were further processed by Fisher's *r*-to-*z* transformation to ensure comparability between sub-CLT alignments with different numbers of cells and then shown to be highly correlated with the DELTA score (A) and *p* value (B) of the sub-CLT alignment. A full list of all 1,526 results with nominal $P < 0.05$ within the top 2,000 local alignments was given as Table S2. Note that 30 alignments with DELTA score >250 were truncated from this figure to facilitate a better visualization of the trends for alignments with small DELTA scores. See also Table S2.

Specifically, homeotic transformations, where a cell *x* in a knockdown strain adopted the fate used by cell *y* in the wild-type strain (an *x*-to-*y* transformation), were previously observed using this dataset (Du et al., 2015). For a homeotic transformation of *x*-to-*y*, we extracted the sub-CLT rooted at *x* from the knockdown strain as well as the sub-CLT rooted at *y* from the wild-type strain. Using a scoring matrix defined by the number of markers with shared expression (Figure 5A) and a pruning cost of 1, we used DELTA-g to align all the extracted pairs of sub-CLTs from the homeotic transformations, i.e., the sub-CLT with an altered fate in the mutant strain and the sub-CLT from the wild-type strain representing the new fate resulting from transformation. We found that 87.3% of the pairs gave rise to statistically significant ($P < 0.05$) alignments (Figures 5B and 5C; see also Table S3), suggesting that DELTA can indeed identify homeotic cell fate transformations. Moreover, DELTA showed the correspondence between the terminal cells of these aligned CLTs (Figure 5D, alignments on the left), revealing the subtle differences between wild-type and transformed sub-CLTs.

We further examined the top 100 DELTA-l results between the wild-type and each mutant strain for homeotic transformation events. Some known homeotic transformations are among the top-ranking local alignments. For example, the cell fate transformation of the ABar lineage in the MOM-2 knockdown strain into ABal in the wild-type strain had the sixth highest DELTA score in the local alignment between the wild-type and MOM-2 knockdown CLTs, the detailed alignment of which between individual cells found by DELTA was visualized by ggVITA (Figure 5D, top left alignment). Similarly, the E cell in GLD-2 knockdown strains takes the cell fate of wild-type MS, which corresponds to the 32nd top alignment in the DELTA-l results between the wild-type and GLD-2 knockdown CLTs (Figure 5D, bottom left alignment). Furthermore, we found some alignments between sub-CLTs that likely correspond to additional homeotic cell fate transformations that have not been previously reported, such as the transformation of P1 into ABp when CAMT-1 is

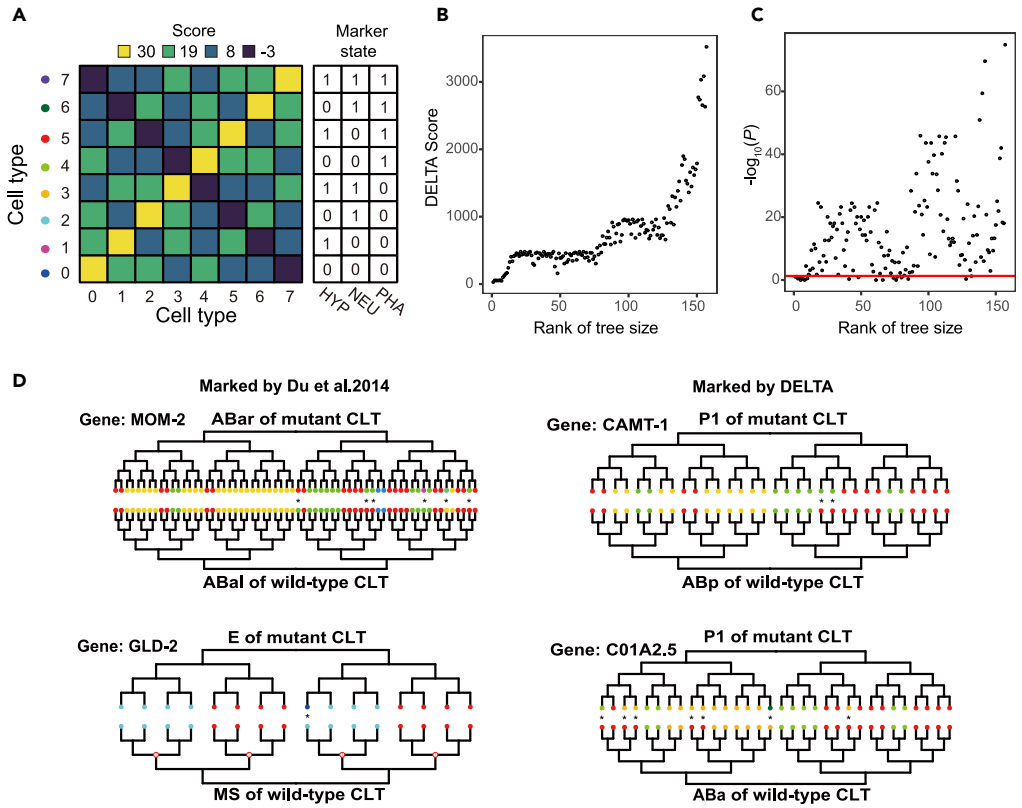


Figure 5. DELTA Reveals Homeotic Cell Fate Transformation in *C. elegans* Mutants

(A) The scoring matrix used in DELTA analyses of cell types annotated in the Digital Development database.

(B and C) For the 131 homeotic cell fate transformation events found by Du et al. between CLTs with at least five terminal cells, the DELTA score (B) and *p* value (C) between the sub-CLTs in mutant strains and those in a wild-type strain representing the adopted cell fate are shown.

(D) Detailed sub-CLT alignments visualized by our newly developed R package "ggVITA". The two alignments on the left were previously marked by Du et al., and the two on the right were newly found by DELTA. Terminal cell types are indicated by colors as in (A). Sub-CLT pruning was shown as red circles on the corresponding internal nodes, in which the number of pruned nodes was indicated. See also Table S3.

knocked down and P1 into ABa when C01A2.5 is knocked down (Figure 5D, right alignments). Note, however, that most local alignments found in DELTA-l between wild-type and mutant strains are between sub-CLTs unchanged by the gene knockdown or highly similar in the original CLT. Nevertheless, these results suggest that homeotic cell fate transformation can be readily found by DELTA.

Given the above results, we further hypothesized that the phenotypic impact of the gene knockdown, as approximated by the DELTA score between wild-type and knockdown CLTs, can reflect the functional importance of the underlying genes. To test this hypothesis, we compared the DELTA score, which was calculated from the global alignment between wild-type and knockdown CLTs, with the evolutionary rates of the genes being knocked down, since functionally more important genes generally evolve more slowly and thus are more conserved (Zhang and Yang, 2015). Here we used the dN/dS ratio to measure the protein evolutionary rate, where dN is the number of nonsynonymous nucleotide substitutions per nonsynonymous site and dS is the number of synonymous nucleotide substitutions per synonymous site (Nei and Kumar, 2000). We split the genes with knockdown CLTs into two groups with high or low DELTA scores with the wild-type CLT and compared the average evolutionary rates of the two groups. As we divided the two groups based on greater DELTA score differences, the deviation of the evolutionary rate of the two groups continued to increase up to ~ 8-fold when genes with DELTA scores >7400 and <2600 were compared (Figure 6A). Since genes with a more dramatic functional impact upon deletion are generally more constrained by natural selection (Zhang and Yang, 2015), this observation suggests that DELTA comparisons between

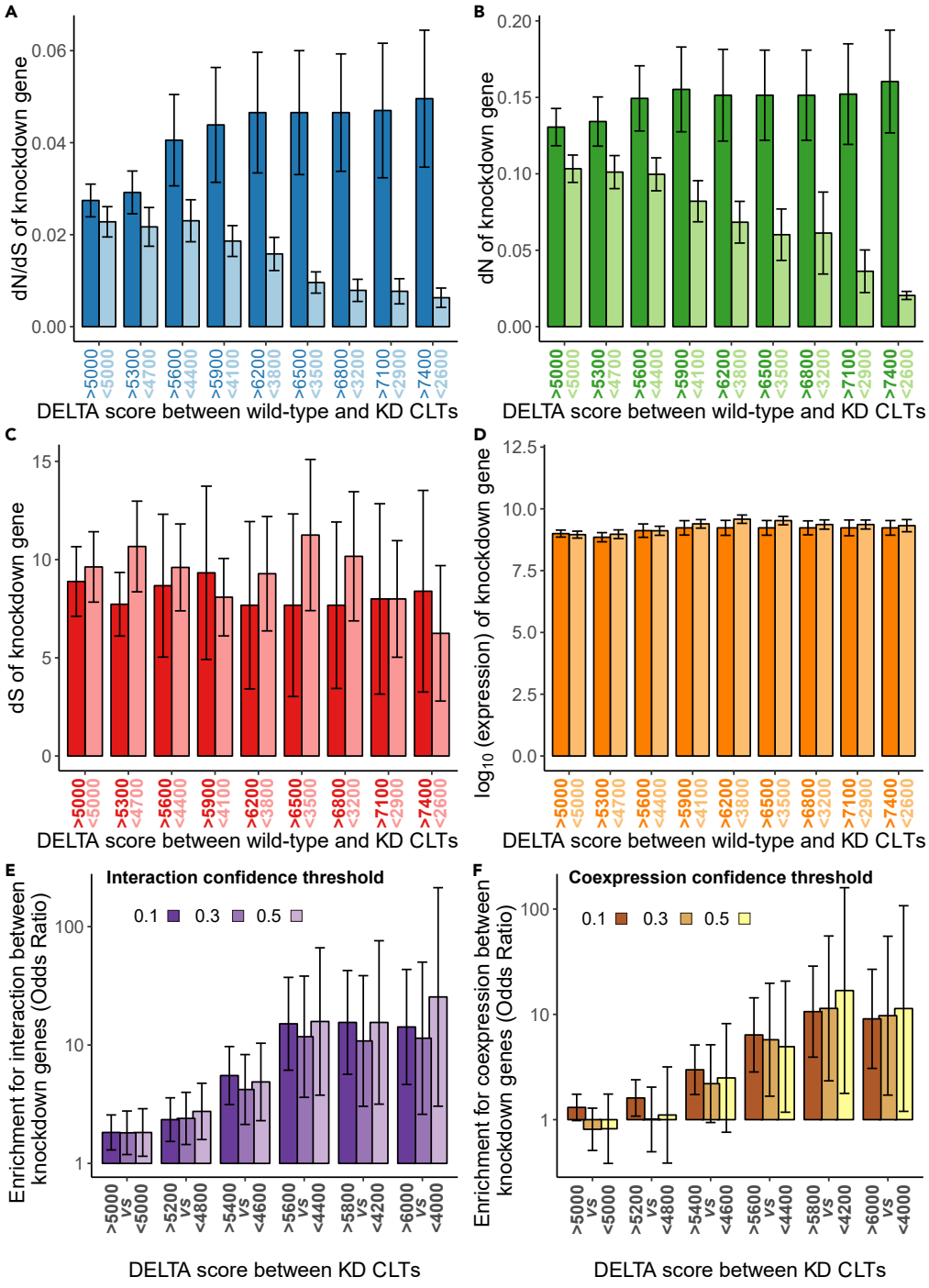


Figure 6. DELTA Relates Phenotypic Changes in CLTs to Underlying Genetic Mechanisms

(A–D) *C. elegans* single-gene knockdown (KD) CLTs were categorized into high (dark color) or low (light color) DELTA score groups by different thresholds (*x* axis) according to their DELTA score of global alignments with the wild-type CLT. The differences between the high and low DELTA score groups become more dramatic for dN/dS (A) and dN (B), but not the dS (B) or expression level (D) of the knockdown genes. The error bar indicates the standard deviation assessed by 1,000 bootstraps of the genes.

(E and F) Enrichment with experimentally determined interactions (E) and coexpression (F) in pairs of genes with knockdown CLTs having a high DELTA score relative to those with a low DELTA score was assessed by an odds ratio from the Mantel-Haenszel test, where the thresholds for high and low DELTA scores are indicated on the *x* axis. The

Figure 6. *Continued*
experimentally determined interactions and coexpression relationships were extracted from the STRING database with different confidence thresholds, as indicated by the color scale. The error bar indicates the 95% confidence interval of the odds ratio given by the Mantel-Haenszel test.

knockdown and wild-type CLTs can indeed quantify phenotypic changes in CLTs in relation to the functional importance of the knockdown gene.

To better understand this observation, we compared dN and dS separately with the DELTA score. We found that differences in DELTA scores were predictive of deviation in dN (Figure 6B) but not dS (Figure 6C). Since dS is primarily determined by the mutation rate, whereas dN is determined jointly by the mutation rate and natural selection (Nei and Kumar, 2000), these results suggest that the DELTA score indeed captured phenotypic changes in CLT that are subject to natural selection acting on the function of the gene, instead of a mutational bias in favor of less important genes. In addition, the gene expression level was found to be unrelated to the DELTA score (Figure 6D). Therefore, the correspondence between the evolutionary conservation of a gene and the impact of its knockdown quantified by DELTA is not confounded (Zhang and Yang, 2015) by the expression level of the gene.

We next asked whether comparisons between CLTs of two knockdown strains can reveal a functional relationship between the knockdown genes, as has been shown by CLT comparison using methods other than DELTA (Gunsalus et al., 2005; Lee et al., 2008; Piano et al., 2002). Since the DELTA score quantifies the phenotypic similarity between two CLTs, we hypothesized that, if the DELTA score between two knockdown CLTs is higher, the genes knocked down in these strains will more likely be functionally related. To test our hypothesis, we assessed the enrichment of experimentally determined interactions recorded by STRING (Szklarczyk et al., 2017) in pairs of genes with knockdown CLTs having a high DELTA score relative to those with a low DELTA score. To avoid interdependence between gene pairs due to involvement of the same gene, we constructed a 2x2 contingency table for each of the 204 genes separately based on (1) whether their DELTA scores with the 203 other knockdown CLTs were higher or lower than some thresholds (Figure 6E, *x* axis) and (ii) whether the confidence of experimentally determined interactions between the pair of knockdown genes in comparison surpassed a selected confidence threshold (Figure 6E, color scale). The 2x2 contingency tables for all genes were summarized by the Mantel-Haenszel procedure to calculate a combined odds ratio to reflect the enrichment of the interaction, such that a larger odds ratio indicates that the gene pairs underlying the knockdown CLTs are more likely to be coexpressed. A similar analysis was also performed for coexpression between knockdown genes (Figure 6F). For both interaction and coexpression, we found that the odds ratio increased as the DELTA score difference between the two groups of gene pairs became larger, regardless of the confidence threshold used. For example, the enrichment of experimentally determined interactions with a confidence value >0.5 yielded an odds ratio of 25.5 when gene pairs with a DELTA score >6,000 were compared with those with a DELTA score <4,000. These results suggest that genes whose knockdown yields similar phenotypic outcomes for CLTs tend to be functionally related. Our observations also provided novel CLT-based results that were consistent with previous observation derived from image-based developmental phenotypes (Gunsalus et al., 2005; Piano et al., 2002), yet offering an analytical method that is more versatile than a predefined list of image-based phenotypes. Altogether, our comparative analyses among knockdown and wild-type CLTs by DELTA successfully associate CLT phenotypes with the underlying genotypes.

CLT Comparison between Species by DELTA Hints at Evolutionary Correspondence between Cell Identities and Cell Types

The diversity of cell types is a significant feature of multicellular organisms, but how it evolves remains largely unexplored. Similar to the necessity of finding orthologous genes between species, the study of cell type evolution is impossible without mapping cell types or constructing the "cell type orthology" between different species. Traditionally, cell types from two species are considered "orthologous" according to structure- and/or function-based cell type definitions, such as neural or muscle cells. Recent technological development of single-cell RNA sequencing and other high-throughput approaches (Schwartzman and Tanay, 2015; Shapiro et al., 2013; Stegle et al., 2015) has promoted research efforts to revise cell type definitions by molecular similarities at the transcriptome or epigenome level, such as in the Human Cell Atlas Project (Regev et al., 2017). However, inferring cell type orthology by structural/functional/molecular similarities may not be reliable because such similarities could have emerged from phenotypic convergence

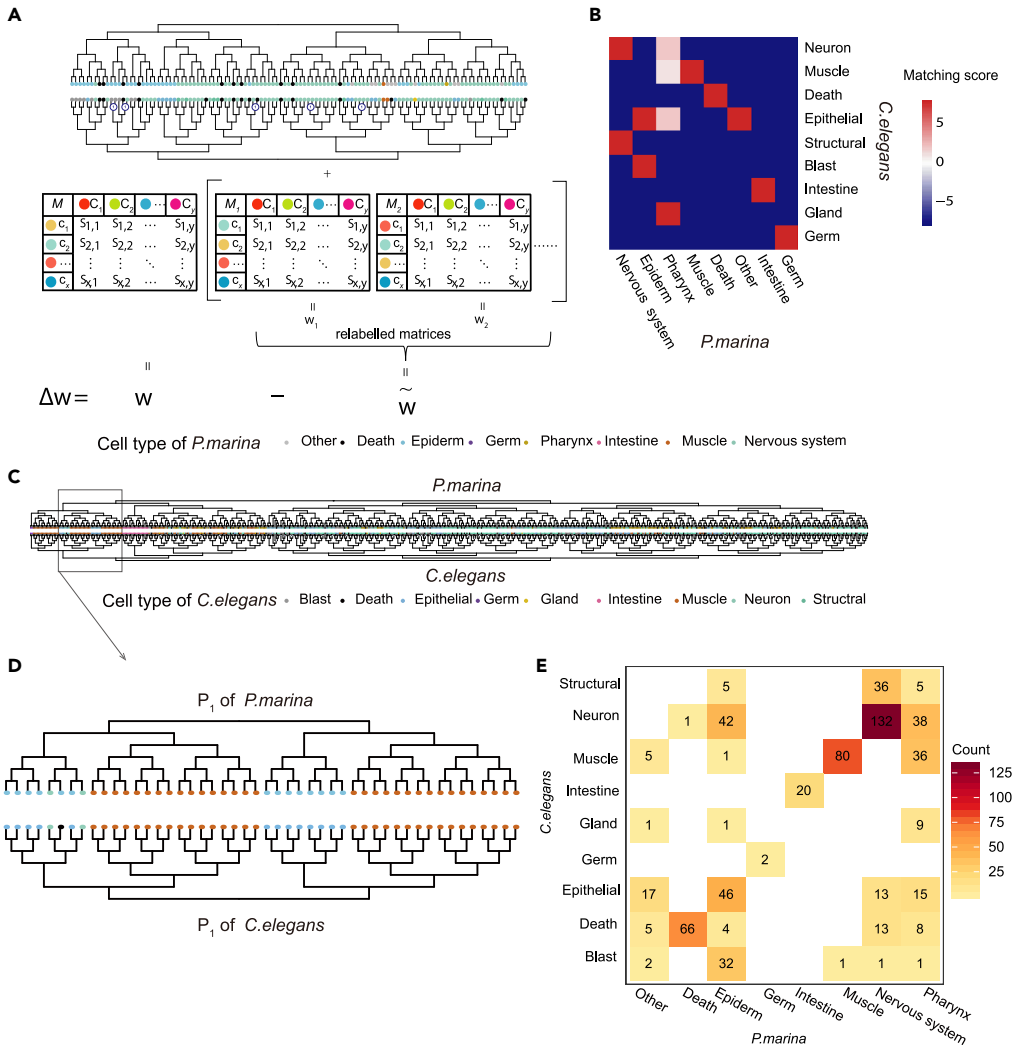


Figure 7. DELTA Comparison across Species Highlights the Evolutionary Correspondence of Cell Fates

(A) Optimization of the scoring matrix between cell types of two species. With the cell types from two species, we first defined a scoring matrix M, the row and column labels (cell types) of which were randomly permutated to generate control matrices M_1, M_2, …, and M_l. These 1 + l matrices were individually used by DELTA to align the CLTs from the two species, giving rise to 1 + l corresponding DELTA scores w, w_1, w_2, …, and w_l. The deviation of w from its random expectation $\tilde{w} = \frac{1}{l}\sum_{i=1}^{l} w_i$, i.e., $\Delta w = w - \tilde{w}$, is optimized by a greedy strategy. See Transparent Methods and Figure S4 for more details.

(B) The optimal scoring matrix for comparisons between *C. elegans* and *P. marina*.

(C) The alignment between *C. elegans* and *P. marina* found by DELTA using the scoring matrix shown in (B). Note that the small circled numbers on some internal branches indicate the size of the pruned sub-CLT.

(D) A sub-alignment from (C); note that the Epiderm "P1aabab" cell in *P. marina* is apoptotic in *C. elegans*.

(E) The number of aligned cell pairs in (E) that fall into each cell type pairs between *P. marina* (*x* axis) and *C. elegans* (*y* axis) is shown by the number and the color within the grid of a heatmap, with the color scale indicated by the scale bar to the right. See also Table S4 and Figure S4.

(Arendt et al., 2016), as found for the striated muscles of vertebrates and *Drosophila melanogaster* (Brunet et al., 2016). Alternatively, we hypothesized that the DELTA score between CLTs from two species with non-uniformly defined cell types would be maximized by a scoring matrix based on the actual evolutionary relationship between specific cell types.

To test our hypothesis, we compared the CLTs of *Pellioditis marina* and *C. elegans*, the cell type identities of which were previously defined by structure and function nonuniformly in the two species (Azevedo et al.,

2005). By a greedy strategy (see Transparent Methods, Figures 7A and S4), we optimized the scoring matrix such that the DELTA score from global alignments between the two CLTs was maximally higher than the scores from pairs of control CLTs created by relabeling all cells of a type as another random type. Intriguingly, the high matching scores in the optimized scoring matrix indeed hint at biologically reasonable correspondence between cell types from the two species (Figure 7B). For example, the optimized matrix suggests that the cells labeled as "Muscle," "Death," "Intestinal," and "Germ" have exact matches between the two species. On the other hand, cells labeled as "Nervous System" in *P. marina* correspond to the cells labeled as "Neurons" and "Structural" (neuronal structural) cells, and the cells labeled as "Pharynx" in *P. marina* correspond to "Gland," "Epithelial," "Muscle," and "Neuron" cells in *C. elegans*. Finally, the 30 cells in *P. marina* with the "Other" fate were suggested by the optimized matrix to be epithelial cells in *C. elegans*.

Furthermore, the DELTA-g alignment based on the optimized scoring matrix revealed the detailed correspondence between terminal cells of the two species (Figure 7C). Some evolutionary events of cell fate changes are clearly highlighted, such as an Epiderm cell in *P. marina* becoming apoptotic in *C. elegans* (Figure 7D. See Table S4 for a full list of cell pairs between *P. marina* and *C. elegans* aligned by DELTA-g). On a broader scale, the CLT alignment between *P. marina* and *C. elegans* showed results highly consistent with those found in previous reports (Houthoofd et al., 2003). For example, it was previously found that gut and somatic muscle of *P. marina* is highly conserved in comparison to *C. elegans* (Houthoofd et al., 2003). The same pattern was quite apparent if we count, for the CLT alignment by DELTA-g between *P. marina* and *C. elegans*, the number of aligned terminal cell pairs that fall into each combination of cell types between *P. marina* and *C. elegans* (Figure 7E). Indeed, all 20 intestine cells in *P. marina* were aligned to intestine cells in *C. elegans*, and 80 of 81 somatic muscle cells in *P. marina* were aligned to muscle cells in *C. elegans*. Given that DELTA only used the CLT (tree topology and terminal cell type), but not other information such as the cellular position, this observation therefore again supported the utility of DELTA.

DISCUSSION

A Computational Framework for Comparative Studies of CLTs

In this study, we designed and implemented DELTA, a computational framework for the alignment of developmental CLTs. Using simulated CLTs and real CLTs from *C. elegans*, we showed that DELTA can find sub-CLTs with highly similar developmental programs, such as bilaterally symmetric lineage pairs, and lineages with highly similar expression trajectories. Furthermore, DELTA alignments among knockdown and wild-type *C. elegans* strains identified homeotic cell fate transformations, showed more dramatic phenotypic changes for the knockdown of more important genes, and revealed higher CLT similarities between strains where functionally related genes are individually knocked down. Finally, we found that the scoring matrix, optimized for the DELTA score between CLTs from two species, shed light on the evolution of cell types and CLTs of the two species. Together, these results not only recapitulated previously known developmental patterns and therefore demonstrated the utility of DELTA but also pointed to some novel biological patterns that merit further investigation.

DELTA relies on two critical pieces of information to align CLTs, i.e., the topology and the terminal cell types of the CLT. This feature of DELTA ensures its compatibility with both classical CLTs (such as those of nematodes) and genome-editing-based CLTs. To the best of our knowledge, there is currently no other tool designed to quantify the similarity of CLTs other than DELTA. Algorithms for the quantitative comparison of other types of trees do exist. For example, similarities between phylogenetic trees can be measured by the edit distance between two trees with identical sets of unique leaves (species) (Nye et al., 2006). Additionally, an algorithm has been designed to find similarities between RNA structure trees, which are unrooted trees with nodes (internal or terminal) representing one of four nucleotides (A/U/G/C) (Milo et al., 2013). Both these methods are apparently not suitable for alignment and similarity quantification between CLTs, which are rooted trees with different non-uniquely labeled leaves. In conclusion, DELTA is expected to open new paths to the analysis of CLTs, which include both classical CLT (Houthoofd et al., 2003; Sulston et al., 1983) and novel genome-editing based CLT that will rapidly accumulate (Junker et al., 2017; Kalhor et al., 2017; McKenna et al., 2016; Raj et al., 2018a, 2018b).

Potential Applications of DELTA

As more CLTs are being determined, the application of DELTA to them shall provide critical insight into several important biological questions. First, the repeatability of development can be assessed by

comparing CLTs from different individuals or CLTs that root at different cells from an otherwise homogeneous cell population. This analysis is particularly relevant for the efficiency of induced pluripotent stem (iPS) cells, where a seemingly homogeneous population of cells are similarly treated but only a very small fraction are successfully transformed to the pluripotent state, with an even smaller fraction capable of growth into organoids. Comparisons between the CLTs rooted at failed or successful iPS cells may hint at the mechanisms underlying their differences.

Second, as demonstrated with simulated and wild-type *C. elegans* CLTs, DELTA is capable of associating CLT phenotypes with the genetic states of individual cells. Theoretically, the phenotypic consequence of genetic states, i.e., genotype-phenotype mapping (GPM), will become more complex when more cells are involved. As an intermediate phenotype to the phenotypes of single cells and of tissues/organisms, DELTA offers a novel path of bridging GPM at the unicellular and multicellular levels.

Third, we have shown in our study that DELTA can be used to find the evolutionary correspondence between cell types in two species. The advancement of single-cell transcriptome profiling experiments has allowed cell type classification at the finest scale with molecular signatures of gene expression. Experimental limitations aside, this approach for cell type identity determination has two biological difficulties. On the one hand, the similarity of the transcriptional profiles may arise from both cell type homology owing to inheritance from the common precursor and phenotypic convergence, which might lead to false combinations of different cell types into one. On the other hand, the stochastic nature of gene expression (Elowitz et al., 2002) may lead to erroneous separation of the homogeneous cell population of one type into two. This problem was recently reported (Arendt et al., 2016), where an evolutionary definition of cell types based on the "core regulatory complex" (CoRC) of transcription factors was proposed. As a complementary approach, DELTA utilizes the biological information in a CLT to find the evolutionary correspondence between cell types and simultaneously reveals how the CLT itself evolves.

Fourth, one critical unanswered question in development is the relative prevalence of autonomous and regulatory development. In nematodes such as *C. elegans*, development is autonomous except for a small number of sub-CLTs (Sulston and White, 1980). In most other animals, however, it is generally believed that regulatory development is the prevailing mode and autonomy is the exception. A direct quantitative answer to this question would emerge by examining the result from DELTA local alignments for the frequency of identical sub-CLTs, should a CLT or sub-CLT be available for those species.

Fifth, an ideal CLT comprises complete longitudinal and horizontal data. However, during experimental assessment of CLTs by single-cell transcriptome profiling and lineage barcoding, cell lysis and loss are inevitable even in the state-of-the-art method, which dictates that CLT is longitudinally (because the cells are killed at the time of experimentation) and horizontally incomplete. DELTA may provide a resolution to this problem by allowing assembly of temporally "sliced" incomplete CLTs, just as sequence alignment allows the assembly of the genome from short reads.

Limitations of the Study

There are several potential caveats in our study that are worth discussing. The DELTA result critically relies on the choice of its two parameters: the scoring matrix between cell types and the pruning cost. Although we have carefully chosen biologically informed parameters (see Transparent Methods), there is no objective estimate of how good or bad they are, which is likely impossible to obtain before more CLT data become available, similar to the refinement of the substitution matrix for sequence alignment when more sequences were determined (Dayhoff, 1972). Additionally, the value of the pruning cost relative to the matching score also affects the DELTA results. On the one hand, a lower pruning cost makes DELTA more sensitive because pruning of small sub-CLTs improves alignment compared with terminal cell type mismatches. On the other hand, a higher pruning cost makes DELTA more specific since terminal cell type mismatches are more likely to be retained than pruned. Nevertheless, a poor choice of parameters likely reduces biological signals. The significance of all patterns we have shown in this paper would thus be stronger if the parameters were further optimized, further enhancing the value of DELTA.

We have considered the qualitative cell type or gene expression status in our definition of CLTs. However, the dynamic programming scheme for comparison between CLTs is readily adaptable to the quantitative definition of cell types made by high-throughput experiments, such as single-cell transcriptomics. In this

case, instead of using the scoring matrix between qualitatively defined cell types, the alignment score between a pair of terminal cells is calculated by the similarities between their transcriptome profiles, using quantitative metrics such as correlation coefficients or negated Euclidean distances.

In the current study, CLTs were used without considering the temporal duration of the cell cycle for each cell. On the one hand, this is a necessary trade-off to ensure DELTA's compatibility with genome-editing-based CLTs, which contains minimal (if any) information on the temporal duration of the cell cycle. On the other hand, discarding the information of cell cycle duration did not prevent us from discovering the similarities in developmental programs by analyzing CLTs. There are two potential explanations for this observation. First, the molecular pathway that regulates cell cycles might be tightly coupled to the developmental program, such that cell cycle duration is an intrinsic property of the cell type. In other words, most cell types have their own specific cell cycle duration, such that the definition of cell types already contains the information concerning cell cycle duration. Second, the definition of cell types might have nothing to do with the cell cycle duration, in which case DELTA needs to be further improved by allowing cell cycle duration adjustment as an additional CLT revision (in addition to sub-CLT pruning), with properly defined costs to the DELTA score.

For both simulated CLTs and real CLTs from *C. elegans*, the developmental process is mostly autonomous, whereas in most complex organisms (except nematodes), it appears to be largely non-deterministic/regulatory. The critical difference between autonomous and regulatory development is whether the gene expression status of individual cells can be altered by external cues, such as environmental stress or signals from other cells. For example, when isolated from the 8-cell-stage mouse blastomere, one cell can grow into one individual mouse (Kelly, 1977) but not one-eighth of a mouse, as would be predicted by autonomous development. However, autonomous cell fate determination is certainly not absent in complex organisms, especially toward the end of the developmental process. Meiosis is one such example of re-occurring sub-CLT, where a primary oocyte divides twice (three division events) and creates one mature ovum and three polar bodies. As long as such autonomous sub-CLTs exist, DELTA alignment will be possible and informative, as demonstrated by our simulated CLTs with perturbed gene expression (Figure 2).

In the current study, we revisited previously identified homeotic cell fate transformation caused by gene knockdown. Although we found that >87% of previously found homeotic transformations gave rise to a statistically significant DELTA alignment, i.e., despite providing high sensitivity, the specificity of DELTA in identifying homeotic transformations was low because most local alignments found in DELTA-l between wild-type and mutant strains are between sub-CLTs that are unchanged by the gene knockdown or highly similar in the original CLT. We caution readers that additional systematic examination of these local alignments is required to exclude those false positives before one can identify candidate homeotic transformation warranting further investigation.

Last but not least, current single-cell high-throughput experiments suffer from the loss of cells; thus, the CLT reconstructed by lineage tracing of DNA barcodes is likely incomplete, with the majority of terminal cells missing. The robustness of DELTA to such data quality issues dictates the applicability of DELTA. As shown in our DELTA analysis by simulated CLTs with randomly dropped terminal cells (Figure S3), a cell loss rate of 5%, 10%, 20%, or 50%, but not 90%, might still give rise to statistically significant DELTA alignments with high gene expression similarity. In other words, the loss of terminal cells decreases the sensitivity of DELTA. Unfortunately, to the best of our knowledge, none of the CLTs reconstructed by genome editing and single-cell sequencing so far have achieved a cell loss rate $\leq$50%, which is why we refrained from analyzing genome-editing-based CLT in this current study. Nevertheless, we do believe this situation will be much improved in the near future, which is exactly why we decided to develop DELTA at this time. Moreover, if more CLT data were to become available, this limitation could potentially be alleviated by assembling small fragmented CLTs into big complete CLTs (DELTA or similar algorithms should be required in such assembling efforts), as short sequence fragments with sufficiently high coverage can be assembled into the full-length sequence.

Overall, DELTA establishes a computational foundation for the alignment of CLTs and potentiates systematic analyses of lineage trees. Albeit the limitations, DELTA will likely illuminate the connection between phenotypes represented by CLTs and their underlying genotypes, providing novel insights into many unresolved biological questions.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jian-Rong Yang (yangjianrong@mail.sysu.edu.cn).

Materials Availability

No new materials were generated in this study.

Data and Code Availability

This current study conducted computational analyses on publicly available datasets, whose source were all explicitly mentioned in main text. As for the source code, DELTA has been deposited in GitHub (https://github.com/yxj17173/DELTA), ggVITA has been deposited in GitHub (https://github.com/helloicyvodka/ggvita), scripts for various DELTA-based analyses presented in this study have been deposited in Github (https://github.com/helloicyvodka/DELTA_code).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101273.

ACKNOWLEDGMENTS

We are grateful to Xionglei He, Wenfeng Qian, Jianzhi Zhang, Xiao Liu, Zhuo Du, and three anonymous reviewers for their comments on the manuscript. This work was supported by the National Key R&D Program of China (grant number 2017YFA0103504 to Xiaoshu Chen, and grant number 2018ZX10301402 to J.-R.Y.), the National Natural Science Foundation of China (grant number 31671320, 31871320, and 81830103 to J.-R.Y., 31771406 to Xiaoshu Chen), the National Special Research Program of China for Important Infectious Diseases (grant number 2018ZX10302103 to Xiaoshu Chen), and the start-up grant from "100 Top Talents Program" of Sun Yat-sen University (grant number 50000-18821112 to Xiaoshu Chen and grant number 50000-18821117 to J.-R.Y.).

AUTHOR CONTRIBUTIONS

Xiaoshu Chen and J.-R.Y. conceptualized and supervised the study. J.L., Xiaolong Cao, G.Z., X.W., and J.-R.Y. implemented the algorithms. M.Y., X.Y., Xiaolong Cao, F.C., Z.L., and J.-R.Y. collected various datasets for the analyses. M.Y., X.Y., Xiaoshu Chen, and J.-R.Y. prepared the original draft. M.Y., Xiaoshu Chen, and J.-R.Y. revised the manuscript with input from all authors and anonymous reviewers.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 20, 2020
Revised: May 12, 2020
Accepted: June 10, 2020
Published: July 24, 2020

REFERENCES

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., et al. (2016). The origin and evolution of cell types. Nat. Rev. Genet. *17*, 744–757.

Azevedo, R.B., Lohaus, R., Braun, V., Gumbel, M., Umamaheshwar, M., Agapow, P.M., Houthoofd, W., Platzer, U., Borgonie, G., Meinzer, H.P., et al. (2005). The simplicity of metazoan cell lineages. Nature *433*, 152–156.

Brunet, T., Fischer, A.H., Steinmetz, P.R., Lauri, A., Bertucci, P., and Arendt, D. (2016). The evolutionary origin of bilaterian smooth and striated myocytes. Elife *5*, e19607.

Dayhoff, M.O. (1972). A model of evolutionary change in proteins. Atlas Protein Seq. Struct. *5*, 89–99.

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. Dev. Cell *34*, 592–607.

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.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G., and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. Cancer Res. *68*, 5924–5931.

Gritti, N., Kienle, S., Filina, O., and van Zon, J.S. (2016). Long-term time-lapse microscopy of *C. elegans* post-embryonic development. Nat. Commun. *7*, 12500.

Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.S., et al. (2005). Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. Nature *436*, 861–865.

Houthoofd, W., Jacobsen, K., Mertens, C., Vangestel, S., Coomans, A., and Borgonie, G. (2003). Embryonic cell lineage of the marine nematode *Pellioditis marina*. Dev. Biol. *258*, 57–69.

Junker, J.P., Spanjaard, B., Peterson-Maduro, J., Alemany, A., Hu, B., Florescu, M., and van Oudenaarden, A. (2017). Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. bioRxiv, 056499.

Kalhor, R., Mali, P., and Church, G.M. (2017). Rapidly evolving homing CRISPR barcodes. Nat. Methods *14*, 195–200.

Kelly, S.J. (1977). Studies of the developmental potential of 4- and 8-cell stage mouse blastomeres. J. Exp. Zool. *200*, 365–376.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. Nat. Genet. *40*, 181–188.

Lescroart, F., Hamou, W., Francou, A., Theveniau-Ruissy, M., Kelly, R.G., and Buckingham, M. (2015). Clonal analysis reveals a common origin between nonsomite-derived neck muscles and heart myocardium. Proc. Natl. Acad. Sci. U S A *112*, 1446–1451.

Lohaus, R., Geard, N.L., Wiles, J., and Azevedo, R.B. (2007). A generative bias towards average complexity in artificial cell lineages. Proc. Biol. Sci. *274*, 1741–1750.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

Milo, N., Zakov, S., Katzenelson, E., Bachmat, E., Dinitz, Y., and Ziv-Ukelson, M. (2013). Unrooted unordered homeomorphic subtree alignment of RNA trees. Algorithms Mol. Biol. *8*, 13.

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.

Nei, M., and Kumar, S. (2000). Molecular Evolution and Phylogenetics (Oxford university press).

Nye, T.M., Lio, P., and Gilks, W.R. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics *22*, 117–119.

Piano, F., Schetter, A.J., Morton, D.G., Gunsalus, K.C., Reinke, V., Kim, S.K., and Kemphues, K.J. (2002). Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. Curr. Biol. *12*, 1959–1964.

Raj, B., Gagnon, J.A., and Schier, A.F. (2018a). Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR-Cas9 barcodes by scGESTALT. Nat. Protoc. *13*, 2685–2713.

Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018b). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell Atlas. Elife *6*, e27041.

Reizel, Y., Itzkovitz, S., Adar, R., Elbaz, J., Jinich, A., Chapal-Ilani, N., Maruvka, Y.E., Nevo, N., Marx, Z., Horovitz, I., et al. (2012). Cell lineage analysis of the mammalian female germline. PLoS Genet. *8*, e1002477.

Salipante, S.J., Kas, A., McMonagle, E., and Horwitz, M.S. (2010). Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol. Dev. *12*, 84–94.

Santella, A., Kovacevic, I., Herndon, L.A., Hall, D.H., Du, Z., and Bao, Z. (2016). Digital development: a database of cell lineage differentiation in *C. elegans* with lineage phenotypes, cell-specific gene functions and a multiscale model. Nucleic Acids Res. *44*, D781–D785.

Schwartzman, O., and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. Nat. Rev. Genet. *16*, 716–726.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630.

Shlush, L.I., Chapal-Ilani, N., Adar, R., Pery, N., Maruvka, Y., Spiro, A., Shouval, R., Rowe, J.M., Tzukerman, M., Bercovich, D., et al. (2012). Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. Blood *120*, 603–612.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145.

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.

Sulston, J.E., and White, J.G. (1980). Regulation and cell autonomy during postembryonic development of *Caenorhabditis elegans*. Dev. Biol. *78*, 577–597.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*, 1330–1334.

Wasserstrom, A., Frumkin, D., Adar, R., Itzkovitz, S., Stern, T., Kaplan, S., Shefer, G., Shur, I., Zangi, L., Reizel, Y., et al. (2008). Estimating cell depth from somatic mutations. PLoS Comput. Biol. *4*, e1000058.

Yang, J.R., Ruan, S., and Zhang, J. (2014). Determinative developmental cell lineages are robust to cell deaths. PLoS Genet. *10*, e1004501.

Zhang, J., and Yang, J.R. (2015). Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. *16*, 409–420.

iScience, Volume *23*

Supplemental Information

Alignment of Cell Lineage Trees Elucidates Genetic Programs for the Development and Evolution of Cell Types

Meng Yuan, Xujiang Yang, Jinghua Lin, Xiaolong Cao, Feng Chen, Xiaoyu Zhang, Zizhang Li, Guifeng Zheng, Xueqin Wang, Xiaoshu Chen, and Jian-Rong Yang

Supplemental Figures

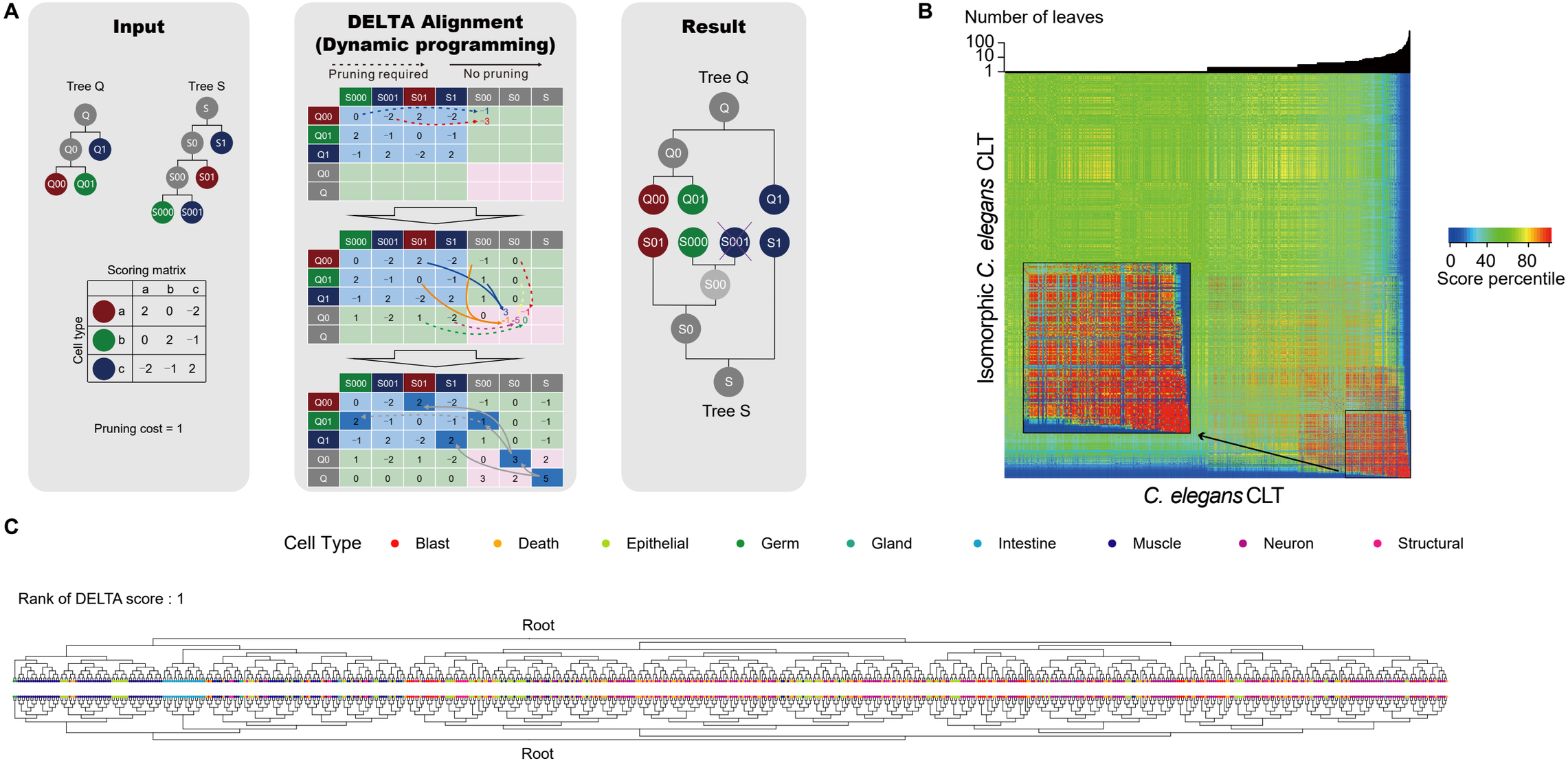
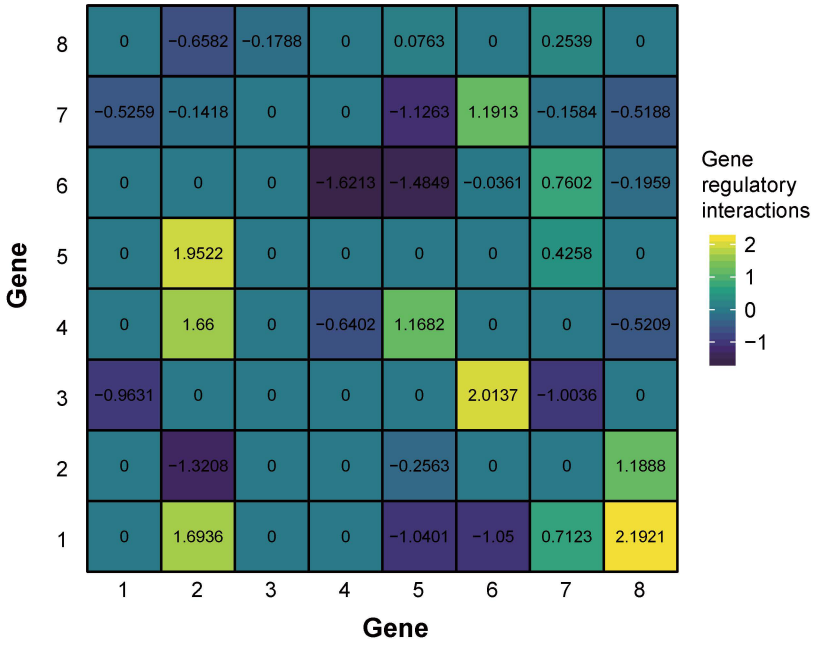

A
Input
Tree Q
Tree S
Scoring matrix
Cell type
a
b
c
Pruning cost = 1
DELTA Alignment
(Dynamic programming)
Pruning required
No pruning
Result
Tree Q
Tree S
B
Number of leaves
100
10
1
Isomorphic C. elegans CLT
C. elegans CLT
0 40 80
Score percentile
C
Cell Type
Blast
Death
Epithelial
Germ
Gland
Intestine
Muscle
Neuron
Structural
Rank of DELTA score : 1
Root
Rank of DELTA score : 2
ABa
ABp
Rank of DELTA score : 4
ABpa
ABpp
Rank of DELTA score : 3
ABp
ABa
Rank of DELTA score : 5
ABpp
ABpa

Figure S1

Figure S1. The DELTA algorithm, related to Fig. 1.

(**A**) A detailed example of DELTA alignment in Fig. 1B. Two CLTs, *Q* and *S*, with types of their terminal cells are colo-coded. Note the nontriviality of the problem when the correspondence between cells in *Q* and *S* is sought after, regardless of the extreme simplicity of both trees. The two CLTs, plus a scoring matrix describing how the correspondence between specific cell types should be awarded based on hypothetical biological similarities of each pair, and a pruning cost, are used as input. The DELTA algorithm then uses a dynamics programming (DP) strategy to search for the optimal alignment between the two CLTs. Some critical steps of the DP procedure are exemplified here. First, a DP matrix was constructed and filled by comparing the smallest sub-CLTs (those containing only one terminal cell) first, as shown by the light blue-shaded elements of the DP matrix. Each score between an internal node and a terminal node (green shaded elements) was then calculated by choosing the best combination of pruning and matching, with the source of best scored saved for each comparison (top panel). A similar procedure was carried out for scores between two internal nodes (pink shaded elements), allowing (mis-)matches for both daughter cells or necessary pruning (middle panel). In both the top and middle panels, all possible combinations of daughter cell alignments and the corresponding score are listed as arrows and scores with different colors, whereas the best score is always colored blue. Arrows with dashed lines indicate pruning is involved, or solid lines otherwise. The high-scored element in the DP matrix was used as a starting point for backtracing, where the optimal alignment between *Q* and *S* are represented by the backtraced route (gray arrows in bottom panel). The final result provides the optimal alignment between *Q* and *S*. Note that the aligned terminal cells are vertically matched and the aligned internal cells are indicated by double-headed arrows. (**B**) The actual DP matrix generated by a DELTA alignment of the *C. elegans* CLT of standard anatomical terminal cell type annotation, with an isomorphic version of itself, where 30% randomly chosen sister sub-CLT pairs were swapped. Each element of the DP matrix is colored from low (blue) to high (red) by the scored percentile, as indicated by the color scale bar on top. The bottom right corner inset is a magnification for finer details. The bar-plot on top shows the number of leaves for the sub-CLTs represented by the column of the DP matrix. (**C**) The top local alignments found by DELTA comparing *C. elegans* CLT with itself. The terminal cell types were color-coded as indicated by the legend at the bottom right corner. Note that except the first alignment of the full CLT root *vs* root, each pair of aligned sub-CLTs appears twice as both X *vs* Y and Y *vs* X, which is an expected behavior of the local alignment of a CLT *vs* itself.

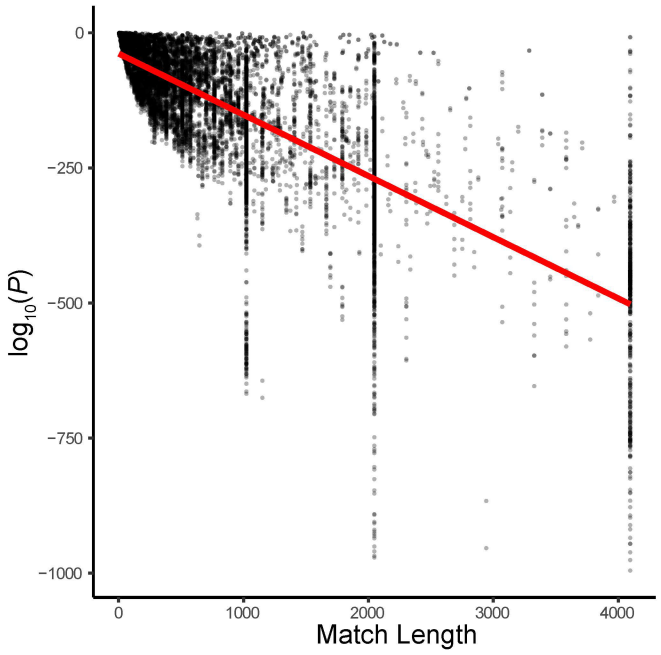
A



B

a	N	K	Dmax	tmax	prune	*P*<10^(-300)	*P*<10^(-250)	*P*<10^(-200)
100	8	4	12	50	-10	100%	100%	100%
100	16	2	12	50	-20	100%	100%	100%
100	16	4	12	50	-20	100%	100%	100%
100	16	8	12	50	-20	100%	100%	100%
10	16	4	12	50	-20	100%	100%	100%
1	16	4	12	50	-20	100%	100%	100%
100	32	4	12	50	-40	99.3%	99.7%	99.9%

C



D

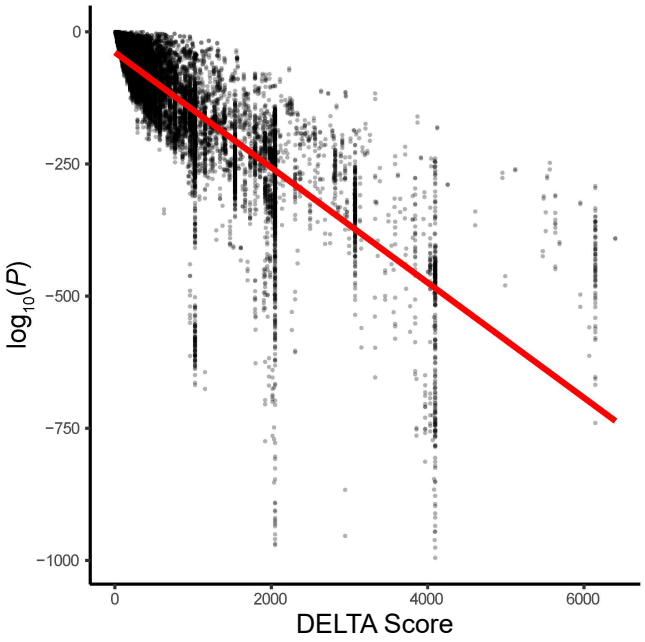


Figure S2

Figure S2. DELTA analyses with simulated CLT, related to Fig. 2.

(**A**) A matrix representing the regulatory network in Fig 2A, which is used to simulate CLTs presented in Fig. 2B. (**B**) Other parameters we tested to simulate CLTs. Each row represents a parameter combination (see Transparent Methods), and the last three columns indicate the fraction of self-alignments of the simulated CLTs with a *P* value below three different thresholds. (C and D) The 1,000 CLTs simulated with the parameter set (a = 100, N = 16, K = 2, t_{max} = 50 and d_{max} = 12) were self-aligned by DELTA-l. The top 100 CLT alignments were extracted from each DELTA-l run. The relationship between the statistical significance of the alignment and the number of matched cell (C) or DELTA score (D) is shown. Linear models fitted with the 100,000 points are represented by red lines in both (C) and (D).

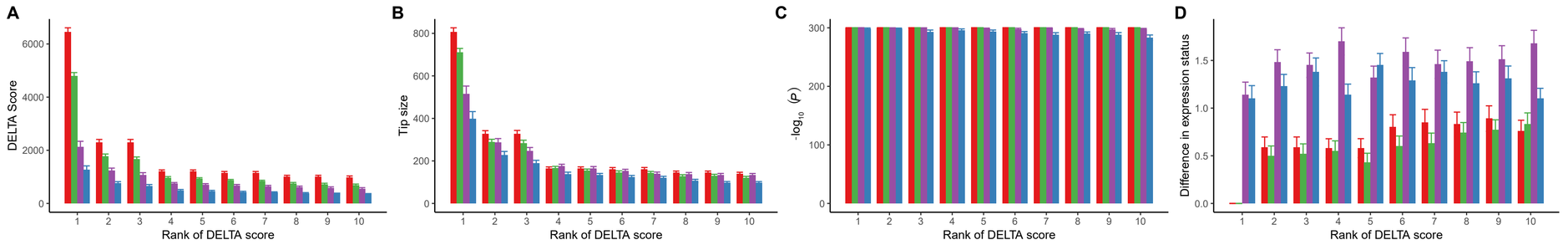
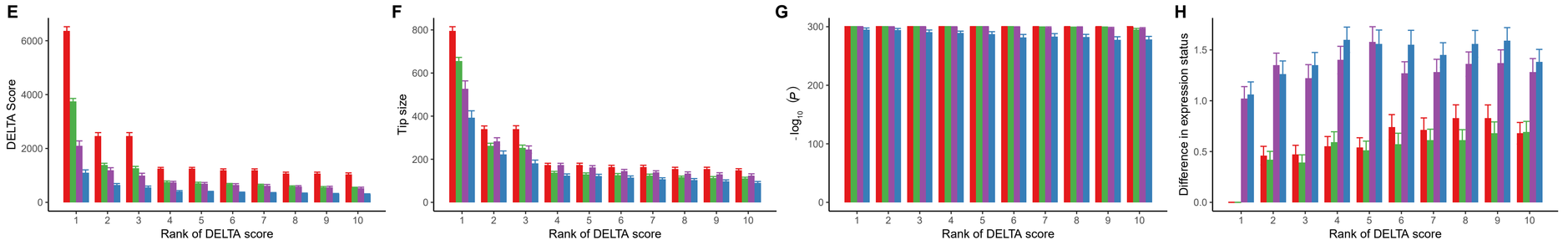
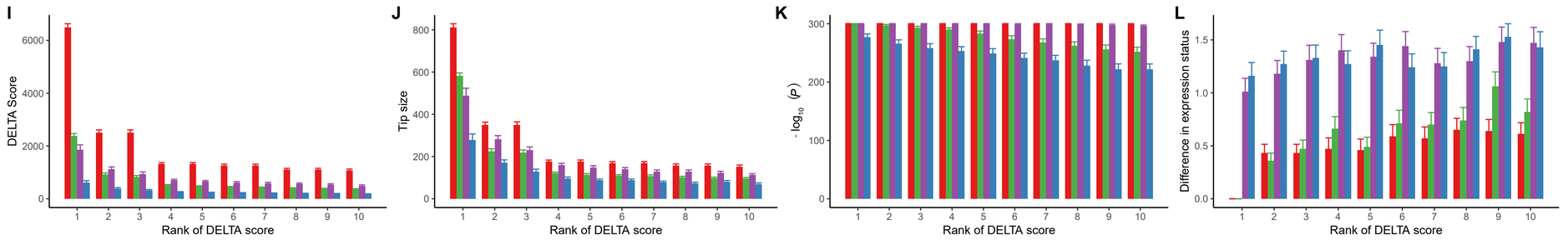
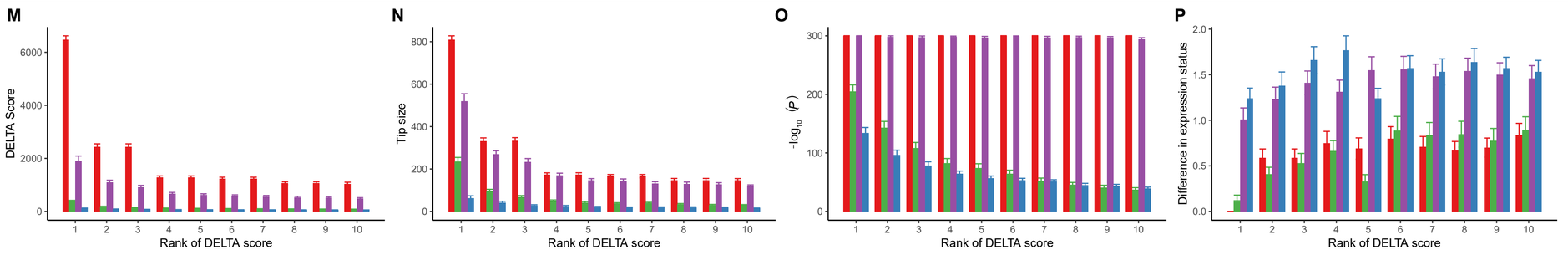
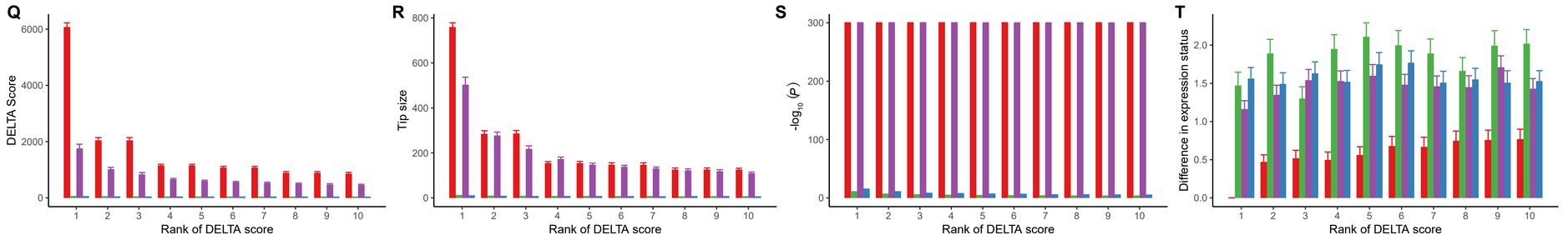
Cell Loss Rate	Performance	Legend of Tree Simulation Model
5%		Full autonomous tree 5% terminal cells loss 5% non−autonomous expression 5% terminal cell loss + 5% non−autonomous expression
10%		Full autonomous tree 10% terminal cells loss 5% non−autonomous expression 10% terminal cell loss + 5% non−autonomous expression
20%		Full autonomous tree 20% terminal cells loss 5% non−autonomous expression 20% terminal cell loss + 5% non−autonomous expression
50%		Full autonomous tree 50% terminal cells loss 5% non−autonomous expression 50% terminal cell loss + 5% non−autonomous expression
90%		Full autonomous tree 90% terminal cells loss 5% non−autonomous expression 90% terminal cell loss + 5% non−autonomous expression

Figure S3

Figure S3. DELTA performance under various cell loss rate, related to Fig. 2.

(A-D) These panels are similar to Fig. 2C-F, except that a terminal cell loss rate of 5% instead of 35% was used in the simulation. Note that only the green and blue results were affected.
(E-H) These panels are similar to Fig. 2C-F, except that a terminal cell loss rate of 10% instead of 35% was used in the simulation.
(I-L) These panels are similar to Fig. 2C-F, except that a terminal cell loss rate of 20% instead of 35% was used in the simulation.
(M-P) These panels are similar to Fig. 2C-F, except that a terminal cell loss rate of 50% instead of 35% was used in the simulation.
(Q-T) These panels are similar to Fig. 2C-F, except that a terminal cell loss rate of 90% instead of 35% was used in the simulation.

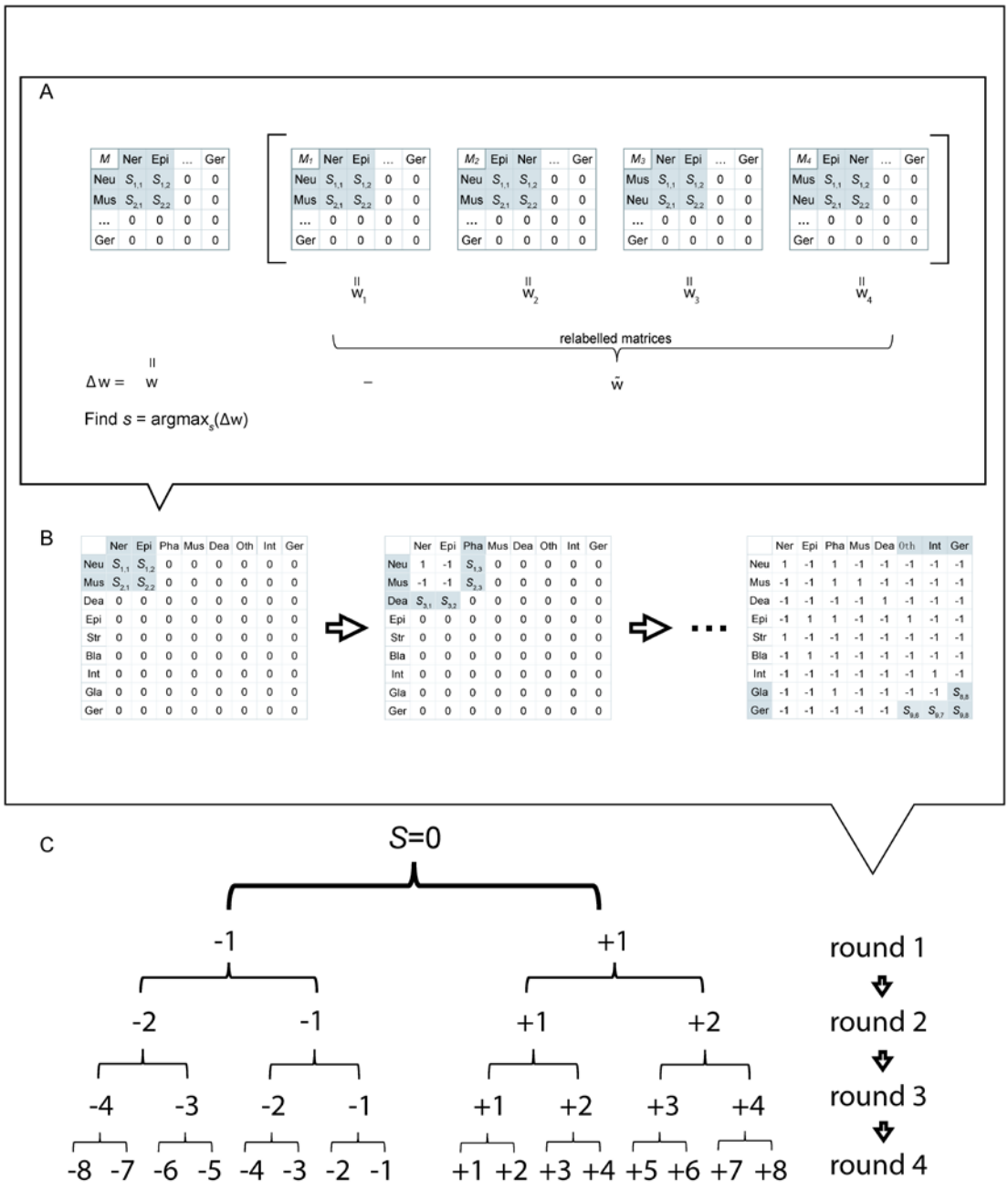

A
M
Ner
Epi
...
Ger
Neu
Mus
Ger
M1
M2
M3
M4
W1
W2
W3
W4
relabelled matrices
Δw =
w
–
w̃
Find s = argmax_s(Δw)
B
Ner
Epi
Pha
Mus
Dea
Oth
Int
Ger
Neu
Mus
Dea
Epi
Str
Bla
Int
Gla
Ger
C
S=0
-1
+1
-2
-1
+1
+2
-4
-3
-2
-1
+1
+2
+3
+4
-8 -7 -6 -5 -4 -3 -2 -1
+1 +2 +3 +4 +5 +6 +7 +8
round 1
round 2
round 3
round 4

Figure S4. Greedy strategy used in optimization of the scoring matrix for *C. elegans* vs *P. marina*, related to Fig. 7.

(A) Cell types containing more cells were arranged at the top left corner of the scoring matrix. Then elements in the matrix were optimized as sub-groups of n elements (we used n = 4 in this study). Only row and column labels related to the elements currently being optimized were permutated to generated control relabeled matrices, considerably reducing the computational time for $\widetilde{w}$. (B) Elements were optimized in groups as in (A) starting from the top-left (associated with more cells) to the bottom-right (associated with fewer cells), prioritizing those on the top-left to bottom-right diagonal line. (C) The scoring matrix was optimized following the scheme of (B) in four rounds, where each element of the matrix was chosen from two diverging options conditioned on the result of the previous round, such that Δw is maximized.

Transparent Methods

The DELTA algorithm

Given an alignment (not necessarily optimal) of two CLTs, two parameters are required to quantify the similarity of these two CLTs. First, for any pair of cell types containing one cell type from each of the two CLTs being compared, a "matching score" is required to describe their similarity in terms of developmental state. All the matching scores between the cell types from one CLT and the cell types from the other CLT can be summarized in a scoring matrix, analogous to the substitution matrix (e.g., PAM and BLOSUM) used in the sequence alignment, except that cell types from the two CLTs could be different. Second, similar to the gap penalty in sequence alignment, a pruning cost is used when some cells or sub-CLTs from one CLT are "pruned", meaning no correspondence in the other CLT can be found for them. In our implementation, the pruning cost is multiplied by the size (number of leaves, regardless of the cell type) of the pruned sub-CLTs and then simply subtracted from the alignment score when sub-CLT pruning is required. With the scoring matrix and pruning cost defined, the goodness of correspondence between terminal cell types and topology can be quantified as a score for the alignment between a pair of CLTs. The task of DELTA is then to find the alignment with the maximal possible score (the "DELTA score"), allowing necessary mismatches of types in terminal cells and pruning of sub-CLTs.

Finding the optimal alignment between CLTs is computationally intensive because a CLT remains unchanged or isomorphic by swapping any pair of sister sub-CLTs. Thus, a query CLT with 1,000 internal nodes, approximately the size of the full *C. elegans* CLT, could be aligned to another subject CLT in $2^{1,000}$ possible ways, not to mention the isomorphic transformation of the subject CLT and pruning of both CLTs. In DELTA, this issue is resolved by dynamic programming (Camacho et al., 2009; Needleman and Wunsch, 1970; Smith and Waterman, 1981) (Figure S1A), where the smallest sub-CLTs (those containing only one terminal cell) are aligned first and larger sub-CLTs are aligned by the best combination of the alignments of their two daughter sub-CLTs, be it a match, a mismatch, or pruning. The final alignment is extracted from the dynamic programming matrix by backtracing the subalignments from the matrix cell with the top DELTA score.

To gauge the statistical significance of an alignment, for each of the aligned CLTs, we generated 1,000 pseudo-CLTs by randomly coalescing the leaves of the real CLT and calculated the DELTA score between the 1,000 pairs of pseudo-CLTs. (details given in Supplemental Text). The distribution of the DELTA scores assessed by these randomized CLTs controls both the sizes of the CLT and the composition of the terminal cells. The DELTA scores of the 1,000 pairs of pseudo-CLTs were used to estimate a *P* value for the actual DELTA score by a *Z*-test.

We implemented the DELTA algorithm in C++, the source of which is available on GitHub. Three files are required as input for the DELTA algorithm, namely, two files with the CLTs to be aligned and a file defining the matching scores of different terminal cell types.

Simulated CLTs

We simulated CLTs using a previously published model (Lohaus et al., 2007). Briefly, the expression profile of a cell at time t is represented by vector $S(t)$, for which the elements $s_i(t)$ ($-1 \leq s_i(t) \leq 1$) indicate the expression of genes $i = 1, 2, \ldots,$ and N ($N > 2$), and the gene expression state of

the gene is considered "on" if $s_i(t) > 0$, or "off" otherwise. A regulatory network composed of these N genes is constructed as a $N \times N$ matrix R, for which the elements r_{ij} indicate the regulatory effect of gene i on the expression of gene j (Fig. 2A). We defined each gene to regulate an average of K other genes; thus, random values following a standard normal distribution were assigned to $N \times K$ random elements in R, the remaining elements of which were set to 0, indicating no regulation (Figure S2A). Therefore, the expression profile of the cell at the next discrete time point $S(t+1)$ can be calculated algebraically using $S(t)$ and R as $\mathrm{S(t+1) = f(R \times S(t))}$, where $\mathrm{f(x) = \frac{1-e^{-ax}}{1+e^{-ax}}}$ is a sigmoid activation function that determines how the expression of each gene is influenced by the total regulatory input from the interaction network (Azevedo et al., 2006; Siegal and Bergman, 2002) and a is the activation constant that determines the transitional shape of the sigmoid curve. Among the N genes, two have special roles as a cell cycle regulator or an asymmetric division regulator. For the cell cycle regulator (gene 1), the cell divides instantaneously into two daughter cells once $s_1(t) > 0$, and the $s_1(t)$ for both daughter cells are reset to -1. For the asymmetric division regulator (gene 2), if the cell divides when $s_2(t) \leq 0$, both daughter cells retain the original expression of $s_2(t)$; otherwise, one of the daughter cells will be assigned $s_2(t) = -1$, while the other daughter cell will retain the original $s_2(t)$. The developmental process is then simulated by initializing a single cell with a randomly generated $S(0)$, with which $S(1)$, $S(2)$, and so on, are calculated. Multiple rounds of cell division as dictated by the regulatory network will be recorded until $t = t_{max}$ or the "depth" of any terminal cells reaches d_{max}, where "depth" refers to the number of cell divisions a terminal cell undergoes since the zygote. This procedure gives rise to a CLT, where the terminal cell types are defined by $S(t_{max})$ (Fig. 2A). Note that t_{max} and d_{max} do not necessarily indicate the end but rather a "cross-section" of the full developmental process.

We performed 1,000 tree simulations for each condition and performed self-alignment using a scoring matrix defined by the number of genes with identical gene expression on/off states between a pair of cell types and a pruning cost of $1.25N$. In the main text, we showed results for $a = 100$, $N = 16$, $K = 2$, $t_{max} = 50$ and $d_{max} = 6$. We tried various settings (Figure S2B) and found our observation of DELTA's capability of associating CLTs with gene expression similarity to be robust for different settings. For example, most of the aligned sub-CLTs have very small P values (Figure S2C), which means that this DELTA score is much higher than that between two random trees. Additionally, the $\log_{10}(P \text{ value})$ is highly correlated with the match length (Figure S2C) and alignment scores (Figure S2D) of the subtrees, indicating that the more complex two subtrees are, the lower the chance that the alignments can be generated by chance.

The CLTs of other species might not be fully autonomous and deterministic, in contrast to those of *C. elegans*, and experimentally determined CLTs might not capture all terminal cells. We modeled these issues as two types of perturbations in the CLT simulation. On the one hand, the expression of each gene has a 5% probability of being negated at every time point in every cell. On the other hand, 35% of terminal cells were randomly removed, and an apparent CLT was reconstructed following the topology of the actual underlying CLT. In other words, any internal cells that lost all of their descendant leaves were also removed, and those that lost one of their daughter cells were replaced by the remaining daughter. We also tried various other loss probabilities (5%, 10%, 20%, 50% and 90%) of terminal cells in the simulation (Figure S3).

Experimentally determined cell lineage trees

The developmental CLT of wild-type *C. elegans* as determined by Sulston et al. (Sulston et al., 1983) and that of *P. marina* as determined by Houthoofd et al. (Houthoofd et al., 2003) were retrieved from previous publications (Azevedo et al., 2005; Yang et al., 2014). Briefly, the 671 terminal cells in *C. elegans* cell lineage (up to hermaphrodite embryogenesis) were categorized by standard anatomical descriptions (Sulston et al., 1983) as follows: 39 blast, 113 death, 93 epithelial (arcade, hypodermal, pharyngeal structural, rectal, and valves), 2 germ, 13 gland (coelomocytes, excretory system, and pharyngeal glands), 20 intestinal, 123 muscle (including the head mesodermal cell), 46 neural structural, and 222 neural cells. A DELTA comparison between *C. elegans* CLTs with anatomically defined terminal cell types was carried out using a scoring matrix where the cell pairs of identical types were scored as 10 and other pairs as -2, and the pruning cost was 1. For *P. marina*, the cell lineage with 638 terminal cells (up to muscle contraction) was classified as follows: 81 body muscle, 67 death, 2 germ, 131 hypodermal, 20 intestinal, 195 nervous system, 112 pharynx, and 30 other-fate cells (Houthoofd et al., 2003). Bilaterally symmetric sub-CLTs in *C. elegans* were extracted from previous reports (Sulston et al., 1983).

For gene expression along the *C. elegans* lineage tree, we downloaded the EPIC data (Murray et al., 2012). For genes with more than one biological replicate, only the one used as examples on the website was used. We further averaged the expression of each gene across the whole lifespan of each cell to generate its expression level for the individual cell. Each cell was then represented by the expression of all 130 genes recorded in the EPIC dataset, and one Pearson's correlation coefficient was calculated between concatenated expressions from all aligned cells in one CLT and that from the other CLT in the alignment. Since different CLT alignments involve different numbers of aligned cells, Pearson's correlation coefficients (R) were standardized by $z = R\sqrt{(n-3)/1.06}$ (Fisher's r-to-z transformation) before being compared (Fig. 4).

To evaluate the capability of DELTA to associate phenotypic changes in a CLT with their underlying genetic changes, we downloaded the CLT of *C. elegans* where 204 conserved genes were individually knocked down from the Digital Development database (Santella et al., 2016). The downloaded data contain the gene expression state ("ON" or "OFF") of three tissue markers, namely, cnd-1 (a subset of neurons), pha-4 (pharynx and gut), and nhr-25 (HYP). Assuming cells with the same lineal name in different experiment are cells with the same identity, we further combined the experimental replicates for the same mutant CLT by a simple majority rule. That is, the gene expression state of this marker in a specific cell is considered "ON" if it is supported by the majority (>50%) of the experimental replicates of the specific mutant strain. With the gene expression status of the three markers, each terminal cell was categorized into one of eight (2^3) types. To construct a scoring matrix as a DELTA parameter, the matching score between two cell types was defined as $10x - y$, where x and y are the number of markers with the same and opposite gene expression status, respectively, in the two cell types (Fig. 5A). The pruning parameter was set to 1. During DELTA comparison between two CLTs, it is possible for some terminal cells from one CLT to become internal in the other CLT. To ensure the comparability of CLTs, we removed cells that were recorded in only one of the two CLTs based on their lineal names and used the remaining ancestral internal cells (mother of the removed cells) as terminal. The previously discovered homeotic transformations were manually retrieved from the original report (Du et al., 2015). Note that four knockdown strains in the dataset did not contain information for all three tissue markers. The 18 homeotic transformation events derived from these four knockdown strains were excluded

from our analysis (Table S3).

Genomic and comparative genomic data

The expression level of *C. elegans* protein-coding genes and the number of synonymous (*dS*) and nonsynonymous (*dN*) substitutions between one-to-one orthologs in *C. elegans* and *Caenorhabditis briggsae* were obtained as previously described (Zhang and Yang, 2015). The confidence scores for experimentally determined protein-protein interactions and gene coexpression among *C. elegans* genes were extracted from the STRING database v10.5 (Szklarczyk et al., 2017).

Optimizing the scoring matrix between cell types from two CLTs

To find the proper scoring matrix for alignment between CLTs of two species, such that correspondence between subjectively defined cell types can be inferred from DELTA, we employed an expectation maximization algorithm to optimize the scoring matrix between cell types from two CLTs. The basic logic behind this algorithm is that a biologically meaningful scoring matrix should maximize $\Delta w = w - \widetilde{w}$, where *w* is the DELTA score from global alignment between the two CLTs being compared using a specific scoring matrix, and $\widetilde{w}$ is the expected DELTA score when the same pair of CLTs is being compared using a scoring matrix where cell types (labels of rows and columns in the scoring matrix) are randomly shuffled (Fig. 7A). For alignment between two CLTs with different cell types, e.g., *x* and *y*, there are *x* * *y* matching scores to be optimized. We employed a greedy grid search strategy to reduce the computational time of this optimization (Figure S4). Briefly, the scoring matrix was initialized by assigning 0 to all elements, and the scoring matrix was then optimized by four rounds of grid searches with increasing precision. Each round of grid searching was finished by progressively optimizing multiple groups of 4 elements, with elements associated with more cells optimized first (Figure S4).

The details of the optimization are as follows. In the first grid searching round, for the first four elements in the scoring matrix, we assigned -1 or 1 to each of the four elements, giving rise to 2^4 different scoring matrices, and calculated $\Delta w = w - \widetilde{w}$ for each matrix, where $\widetilde{w}$ is the averaged DELTA score between the two CLTs based on all the scoring matrices generated by permutating the cell types (column and row labels of the matrix) associated with the four elements being optimized (i.e., the maximum number of permutations is 4! * 4! - 1= 575) (Figure S4A). The one out of 2^4 scoring matrices with the largest Δw was chosen. The next four elements were then optimized similarly. The whole scoring matrix was optimized by progressively optimizing groups of 4 elements until all elements were scanned once (Figure S4B). In the second round of optimization, the elements assigned -1 and 1 were further optimized for a choice between -2 or -1 and 1 or 2, respectively, using a method similar to that used in the first round. The third round then continued, resulting in a scoring matrix with elements representing were one of (-4,-3,-2,-1,1,2,3,4), where -4 and -3 were from elements valued -2 in the previous round, -2 and -1 were from those valued -1, and so on. The final round of optimization gave rise to a scoring matrix with elements representing one of (-8,-7, -6,-5,-4,-3,-2,-1,1,2,3,4,5,6,7,8) (Figure S4C). The pruning cost was fixed at 10 times the maximal possible matching score during this process. In other words, the pruning cost was 10, 20,

40 and 80 for round 1, 2, 3 and 4, respectively. A high matching score between two cell types in this final scoring matrix indicates that the two cell types are closely related, as suggested by the DELTA alignment between CLTs.

Supplemental Text

Algorithmic details of DELTA

A cell lineage tree (CLT), as used in the current study, can be described as a directed, acyclic, connected graph $T = (V, E)$, where *V* and *E* are collections of all nodes (vectors) and edges in the tree, respectively. Each node $v \in V$ represents one single cell, and each edge $e \in E$ represents a descendant relationship from a mother cell to one of its daughter cells. The number of edges attached to a node (regardless the direction) is called the degree of the node, denoted by d_v. The root node of the CLT has $d_v = 2$, indicating the common ancestor of all the cells in the CLT, e.g. the zygote that gives rise to the whole *C. elegans* CLT. An internal node of the CLT has $d_v = 3$, indicating non-root cells that undergofurther division recorded by the CLT. Note that we consider CLT as an unordered tree, which means that swapping the two daughters of an internal node, along with their descendant subtrees, does not change the tree. A terminal node of the CLT has $d_v = 1$, which represents a terminal cell recorded by the CLT, but not necessarily the terminal of development (i.e., further divisions of the "terminal" cells are just not recorded as part of the CLT). All terminal nodes are labeled by their cell types, which could be anatomically defined as, for example, muscle, neural cells, or epigenetically defined such as CD4+ cells.

DELTA finds alignments between two trees with necessary pruning of some subtrees. For a tree $T = (V, E)$, a subtree T_v^u is a connected subgraph of *T*. T_v^u contains all nodes $v' \in V$ such that the path between *u* and *v'* starts with the edge $(u \to v)$, as well as all the edges attached to these nodes except $(u \to v)$. A pruning of subtree T_v^u includes three steps. First, remove all the nodes and edges in T_v^u; second, if *u* has two remaining neighbors attached to it by $(u' \to u)$ and $(u \to u'')$, connect them by a new edge $(u' \to u'')$; third, remove *u* and all edges attached to it.

Given a pair of trees $Q = (V, E)$ and $S = (V', E')$, an isomorphic alignment is a bijection A : $V \leftrightarrow V'$, such that for every pair of nodes with $v, u \in V$, we have $(v, u) \in E \Leftrightarrow \big(A(v), A(\mathrm{u})\big) \in E'$. A homeomorphic subtree alignment *A* between *Q* and *S* is defined as an isomorphic alignment between Q' and S'', where Q' is the result of zero or more pruning of subtrees in *Q*, and S' is the result of zero or more pruning of subtrees in *S*. Here all the subtree pruning in *Q* and *S* are collectively denoted as $\pi(A)$. If we further denote the alignment score between two nodes $v \in V$ and $v' \in V'$ as $a(v, v')$, and the cost for pruning a subtree $\hat{T}$ as $p(\hat{T})$. The score of a homeomorphic subtree alignment *A* between Q and *S* can then be expressed as

$$w(Q, S, A) = \sum_{(v,v') \in A} a(v, v') - \sum_{\hat{T} \in \pi(A)} p(\hat{T})$$

Given two CLTs, a scoring matrix *M* and a pruning coefficient *q* (>0), the DELTA algorithm finds the optimal *A* (with optimal/highest possible *w*) by dynamic programming (detailed below). Here, $p(\hat{T})$ equals to *q* times the number of terminal nodes in $\hat{T}$. If v and v' are internal nodes, $a(v, v')$ equals 0, or if v and v' are terminal nodes, $a(v, v')$ equals to $M(f_v, f_{v'})$, with f_v and $f_{v'}$ representing the cell type of v and v'. To find the optimal *w*, DELTA employs a modified implementation of previously described HSA algorithm (Milo et al., 2013) that was used to aligned

RNA structure trees, with simplifications tailored for the alignment of CLT.

The dynamic programming (DP) procedure by which DELTA finds the optimal alignment between two CLTs (*Q* and *S*) by recursively finding the optimal alignment between their subtrees. It starts from constructing a DP matrix with N_Q row and N_S column, where N_Q and N_S are the total number of nodes in *Q* and *S*, respectively. Each cell of the matrix will store the optimal *w* between the nodes/subtrees represented by the row and the column. For simplicity, we will hereinafter refer to optimal *w* between CLT *Q* and *S* as *w*(*Q*,*S*), with the optimal homeomorphic subtree alignment *A* implicitly indicated. To fill up the matrix, the *w* between terminal nodes are first directly determined by the scoring matrix of the terminal cell types. To calculate *w* between (i.e. to align) one leaf *v* and one internal node *u* with two daughter cells *l* and *r* (which could be internal nodes or leaves), one of the two subtrees (T_l^u or T_r^u) need to be pruned. Thus, we have $w(v,u) = \max\big(w(v,l) - p(T_r^u), w(v,r) - p(T_l^u)\big)$, where the optimal choice (indicating how should *v* and *u* be aligned) was stored for later traceback (Fig. 1D, top panel). For the DELTA score between two internal nodes *v* and *u*, with children that are respectively *l*/*r* and *l'*/*r'*, we have

$$w(v,u) = \max\begin{cases} w(l,l') + w(r,r') \\ w(l,r') + w(r,l') \\ w(v,l') - p(T_{r'}^u) \\ w(v,r') - p(T_{l'}^u) \\ w(l,u) - p(T_r^v) \\ w(r,u) - p(T_l^v) \end{cases}$$

Again, the optimal choice was stored for later traceback (Fig. 1D, middle panel). This process went recursively until the whole DP matrix was filled. For the global alignment between *Q* and *S*, the traceback procedure starts from *w*(*Q*,*S*), i.e., the one DP matrix cell representing the alignment between the roots of *Q* and that of *S*. The optimal alignment for each node was then determined by the optimal choice recorded in each DP matrix cell along the (branched) route of recursive traceback. For the local alignment, the one DP matrix cell with the highest *w* (not necessarily involving the root of CLT) was located, and the local alignment was extracted by recursive traceback starting from this very DP matrix cell. All the DP matrix cells on the route of back tracing were marked as "used", and the highest *w* in the unused DP matrix cells will similarly be used to extract the second local alignment result. More local alignments can be found by repetitively locating the highest *w* in the progressively smaller sets of unused DP matrix cells.

In assessing the statistical significance of a global or local alignment, the two aligned CLTs (or subtrees, in case of local alignment) were individually randomized and realigned by DELTA for 1,000 times. The 1,000 resulting DELTA scores were used to estimate a *P* value for the actual DELTA score by Z-test. The randomization of an individual CLT is conducted as follows. First, all the terminal cells of the original CLT were extracted from the tree to form a list of cells. Second, two cells were randomly chosen from the list and paired up as sister cells, creating a subtree represented by their (arbitrarily constructed) mother cell. Third, these two chosen cells were then removed from the list and replaced by their mother cell. Fourth, the second and their steps were repeated until only one cell is left in the list, thus creating a randomized CLT with the same terminal cells as the original CLT. The distribution of DELTA score assessed by these randomized CLTs thus controls both the sizes of the CLT, as well as the composition of the terminal cells.