Supplemental Materials

All supplemental figures, tables and data can be downloaded from the supplemental website (http://cmgm.stanford.edu/~kimlab/public_html/Liuetal/index.html)

Supplemental Experimental Procedures

Annotation of nuclei with highly variable locations or with ambiguous cell lineage:

The position of some nuclei is variable for different worms. For convenience, we represent the anterior nucleus of a pair by (lr), while the posterior one by (rl). For the pharyngeal muscles, the anterior nucleus is denoted (ap) and the posterior one is denoted (pa).

hyp3: anterior = ABp(lr)aapaaaa; posterior = ABp(rl)aapaaaa

hyp4: anterior = ABp(lr)aappaa; posterior = ABp(rl)aappaa

hyp6: anterior = ABp(lr)aappap; posterior = ABp(rl)aappap

hyp7: anterior = ABp(lr)aapppa; posterior = ABp(rl)aapppa

hyp7: anterior = ABp(lr)appppa; posterior = ABp(rl)appppa

hyp10: anterior = ABp(lr)ppppppp; posterior = ABp(rl)ppppppp

DB1/DB3: anterior = ABp(lr)paaaapp; posterior = ABp(rl)paaaapp

Pharyngeal muscle 2, dorsal pair: anterior = ABaraap(ap)apa; posterior = ABaraap(pa)apa

Pharyngeal muscle 2, left pair: anterior = ABalpaaa(ap)a(pa); posterior = ABalpaaa(pa)a(ap)

Pharyngeal muscle 2, right pair: anterior = ABarapaa(ap)a(pa); posterior = ABarapaa(pa)a(ap)

Commitment Algorithm: Comparison of scoring methods

Our commitment algorithm predicts the level of commitment to gene expression of each cell during the development of the worm. We make these predictions by combining the gene expression profiles of the annotated cells in the L1 worm, and the cell lineage. Our approach finds the commitment to gene expression of internal nodes (common ancestors of the cells annotated in the L1) that minimize the overall change in expression throughout development.

Minimizing the overall gene expression change can be scored in multiple ways. In the main text, we chose to minimize the linear sum of all changes along the tree. There are many ways one could score expression values in order to compute commitment to gene expression. We tried two methods (linear scoring and sum of squared changes) in order to determine whether the gene expression commitment algorithm was sensitive to the scoring method. As described below, both methods yielded similar results, and so chose to use linear scoring. This is a more intuitive solution, and also offers the flexibility of different penalties for an increase or decrease in gene commitment in the future.

However, we wanted to confirm the method would yield similar results using different scoring functions. We therefore tested the sum of squared changes (SSC) method. We are presented with 363 annotated gene expression values for each of the 363 terminal cells in the L1 lineage that were scored. As before, we want to minimize the overall changes in gene expression used during embryogenesis to generate the final pattern of gene expression. Rather than use linear scoring, we alter the scoring function so that we are minimizing the sum of the squared changes (rather than the linear sum of changes) between the parent, u, and daughter cells, v.

$$\sum_{(u,v)}(x_u - x_v)^2$$

As before, at any leaf node, we set

$$x_l = o_l$$

where o_l is the observed expression at the cell represented by leaf node, l.

To compare the results from the method described in the main text to the SSC approach described above, we analyzed the activity map generated for each. First, we compared the two scoring methods for similarity in gene expression commitment for 9 individual genes, and found that both methods were similar by visual inspection. Supplemental Figure 6A, second row, shows an example for *C08B11.3*. Second, we combined the results from all 93 genes to look at the overall change in molecular signature in the embryonic cell lineage. Through visual inspection, we again find that the molecular divergence maps for the SSC method (Supplementary Figure 6B, second row) and linear scoring strongly resemble each other. As a result, we are confident in our predictions, and choose the most intuitive description of the problem (linear scoring) as the reported results.

Commitment Algorithm: Modification of the cell lineage to accommodate un-annotated cells

We chose to disregard all un-annotated cells by removing the corresponding leaf nodes in the lineage. In addition, we excluded all internal nodes from the graph that contained only un-annotated leaves in their subtree. Finally, the tree was transformed into a bifurcating tree. All internal nodes that do not have two included children are also removed, and a direct path is created from the most recent included ancestor, and the first included descendent. The latter would meet one of two conditions. (1) It has two included daughters or (2) it is an annotated leaf node. Figure 5 shows the complete cell lineage, where the solid lines represent the modified lineage, and dotted lines show the excluded portions of the original lineage.

Commitment Algorithm: Sensitivity to Un-annotated Cells

The commitment algorithm predicts the changes in gene expression commitment based on a subset of annotated cells in the L1, and as a result we use a modified cell lineage in the method (described fully in the methods section). As in evolutionary biology, we make our predictions based only on observed data. Therefore, our commitment and subsequently the molecular divergence map examine differences between common ancestors of observed cells only.

However, unlike evolutionary biology where the total number of species present in Earth's history is unknown, we acknowledge that the complete cell lineage of *C.*

elegans is known. This gives us the opportunity to examine what effect these additional cells might have on the predictions.

We considered assigning all possible gene expression values for each of the cells, but this is not computationally feasible. We also considered random assignments of gene expression values to the set of un-annotated cells based on a distribution derived from the observed expression values in the known cells. This is biologically unsound because such a distribution would assume that gene expression values between cells with similar fates are not correlated. That is, any statistical analysis would assume that the expression levels of all cells are independent and identically distributed.

Instead, we analyze the system at two of the highest perturbation levels. We allow the un-annotated cells to receive either the highest or the lowest gene expression value. While there are other possible gene expression patterns that cause a higher level of perturbation, such information would be difficult to interpret since many would need to be analyzed, and as already stated the resulting distribution has little meaning.

For each gene, we solved the gene expression commitment map using either the maximum values or the minimum expression values for the un-annotated cells. In comparing these commitment maps, we found that the shared branches showed similar predictions. Supplemental Figure 6A, third and fourth rows, shows an example for *C08B11.3*.

We then combined the results from all 93 genes to form the gene molecular divergence map. We compared the original map using data only from annotated cells, to the two maps modeled using the maximum and minimum expression data for the un-annotated cells. We find that the vast majority of the shared branches between the molecular divergence maps remain consistent, indicating that the tree shown in Figure 5 in the text is robust to very large changes in expression in the un-annotated cells (Supplemental Figure 6B, third and fourth rows). This is largely due to the fact that the un-annotated cells are often segregated to entire subtrees. However, the un-annotated cells may have effects on certain nodes in the tree, particularly those cells that are ancestral to the un-annotated sublineage.

Supplemental Figure Legends

Supplemental Fig. 1. Consistency of gene expression measurements. (A) Correlations of gene expression between different individual worms from the same transgenic line. For every transgenic line, the Pearson correlation coefficient (R) for gene expression between all individual worms was calculated and averaged. For most strains, different individual worms have correlation coefficients for mCherry expression of R > 0.80 (Supplemental Figure 1A), indicating both that the annotation of cell nuclei is reliable and that the mCherry expression is reproducible. However, some transgenic lines showed variable expression of the mCherry reporter between individual worms, indicating that expression is affected by heterogeneity in growth, development or culture condition between different individuals within a population. The clearest case of variable expression is a strain expressing a *sod-3::mCherry* reporter. Expression of *sod-3* (which encodes an iron/manganese superoxide dismutase(Giglio et al., 1994)) starts to be expressed at hatching and is regulated by stress, so that small differences in developmental age (+/- 1.5 hours) or levels in stress could account for variability in *sod-3* expression between individuals. (B) Comparing expression correlations between worms from different transgenic lines derived from the same mCherry reporter construct with correlation among worms from the same transgenic line. For 12 mCherry reporters, multiple integrated transgenic lines were generated and used for single-cell gene expression analysis. The Pearson correlation coefficient for single cell gene expression between different worms was calculated. The y-axis shows the mean of the Pearson correlation coefficients between worms of same transgenic line, and the x-axis shows the mean of the Pearson correlation coefficients between worms of different transgenic lines. Error bars represent 95% confidential interval.

Supplemental Figure 2. Heterogeneous gene expression patterns among body wall muscle cells. A. Different expression patterns among body wall muscle cells based on their lineage. Columns represent different muscle cells, arranged by their lineage ancestry. Rows represent genes that are differentially expressed between muscles derived from different MS and D blastomeres ($p< 10^{-5}$, t-test). Gene expression levels were normalized for each gene so that the minimal and maximal expression values are 0 and 1 for each gene. Expression levels in these nuclei and p-values are in Supplemental Table 5B. B. An anterior-posterior gradient of gene expression in the body wall muscles. Rows represent different muscle cells, arranged from anterior to posterior. Rows represent genes that are differentially expressed in the anterior-posterior axis ($p< 10^{-5}$, linear regression), excluding lineage-specific genes shown in Supplemental Figure 5A. Color indicates level of expression. Expression levels in these nuclei and p-values are in Supplemental Table 5C.

Supplemental Figure 3. Expression of C08B11.3 in hypodermis 7 nuclei. A. C08B11.3 maintains expression in AB-derived hyp 7 nuclei at least until the end of L1 stage. Specific data are shown in Supplemental Table 6. B. Expression of C08B11.3 re-appears following photobleaching. The left figure shows expression *C08B11.3:mCherry* in hyp7 in a newly hatched L1 worm. The worms were photobleached to remove fluorescence from pre-existing mCherry protein (middle figure) and mCherry expression re-appeared 15 hours later (right figure) showing that mCherry fluorescence in the L1 involves new protein synthesis and is not solely due to residual protein from the embryonic AB lineage.

Supplemental Figure 4. Gene expression commitment in the embryo. For each gene, the gene commitment algorithm is used to predict commitment to express the gene in the embryonic lineage based on observed expression in the L1 larvae. Shown is the embryonic lineage. Dotted lines show the portions of the complete cell lineage that were unscored. The solid lines represent the modified lineage used for the analysis. Red indicates commitment to express the gene. O indicates cells in which expression has been previously observed and X indicates cells in which expression has previously been shown to shut off (Ardizzi and Epstein, 1987; Hallam et al., 2000; Horner et al., 1998; Kalb et al., 1998; Murray et al., 2008).

Supplemental Figure 5. Gene expression commitment in the embryo for each of the 93 genes.

Supplemental Figure 6. Commitment Algorithm: Comparison of scoring methods.

Ardizzi, J.P., and Epstein, H.F. (1987). Immunochemical localization of myosin heavy chain isoforms and paramyosin in developmentally and structurally diverse muscle cell types of the nematode Caenorhabditis elegans. The Journal of cell biology *105*, 2763-2770.

Giglio, M.P., Hunter, T., Bannister, J.V., Bannister, W.H., and Hunter, G.J. (1994). The manganese superoxide dismutase gene of Caenorhabditis elegans. Biochemistry and molecular biology international *33*, 37-40.

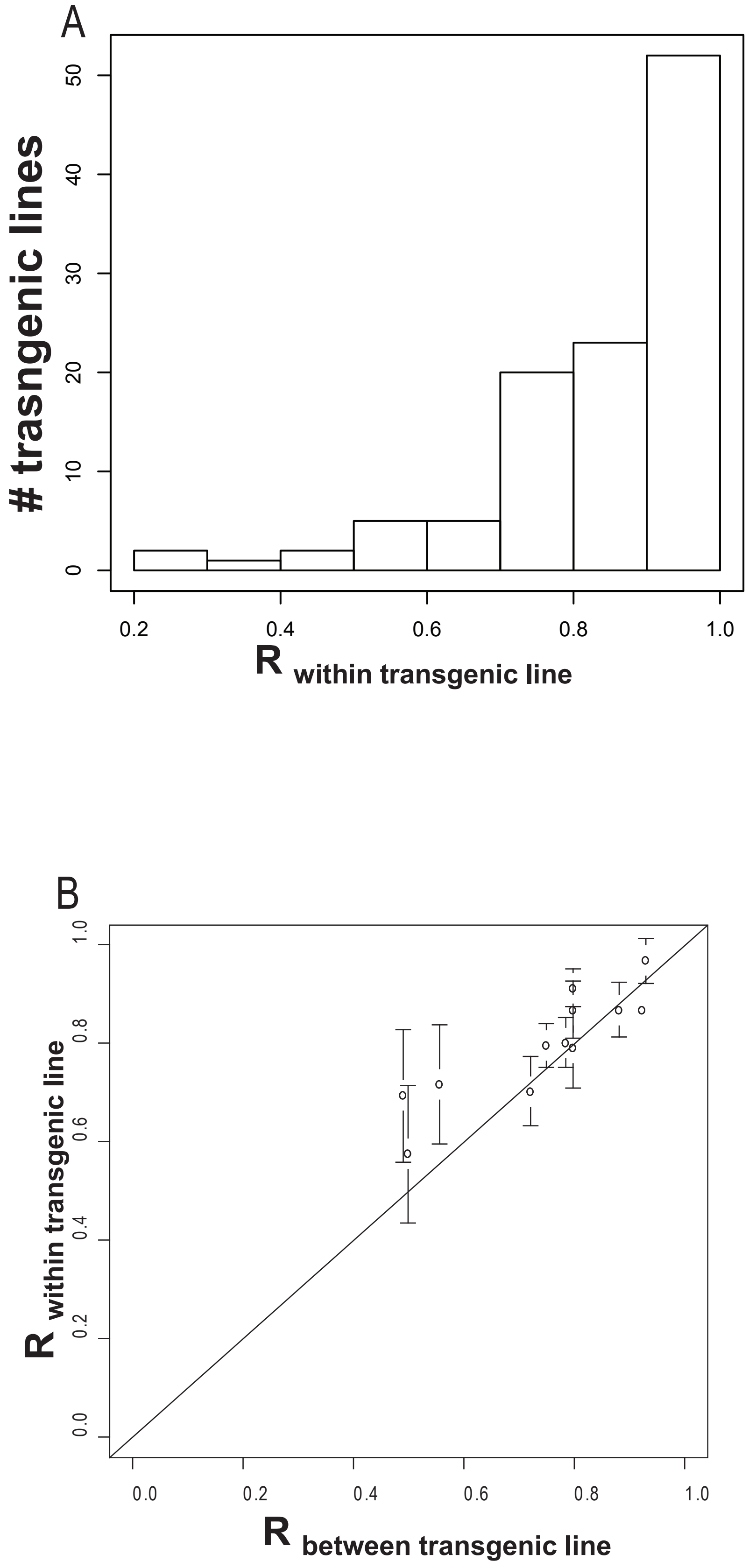
Hallam, S., Singer, E., Waring, D., and Jin, Y. (2000). The C. elegans NeuroD homolog cnd-1 functions in multiple aspects of motor neuron fate specification. Development (Cambridge, England) *127*, 4239-4252.

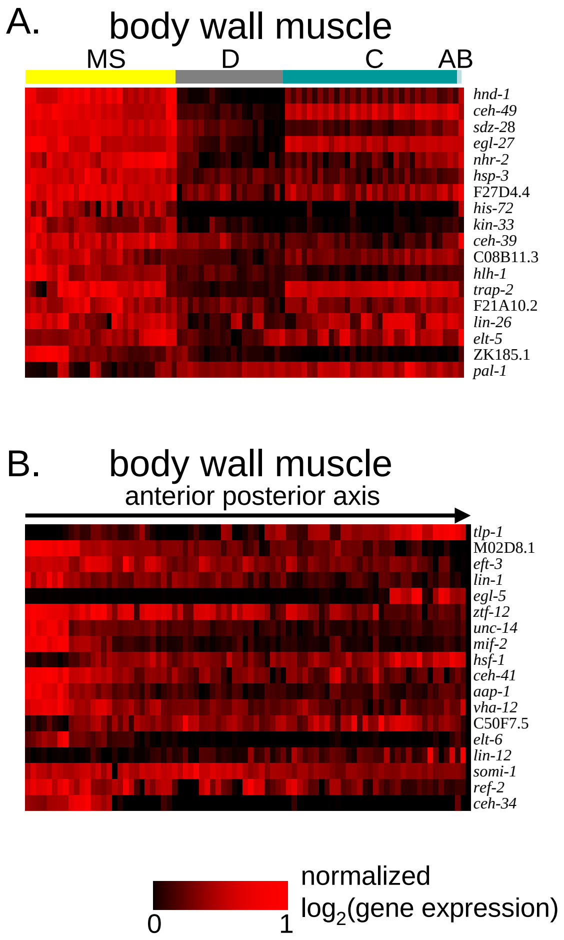
Horner, M.A., Quintin, S., Domeier, M.E., Kimble, J., Labouesse, M., and Mango, S.E. (1998). pha-4, an HNF-3 homolog, specifies pharyngeal organ identity in Caenorhabditis elegans. Genes & development *12*, 1947-1952.

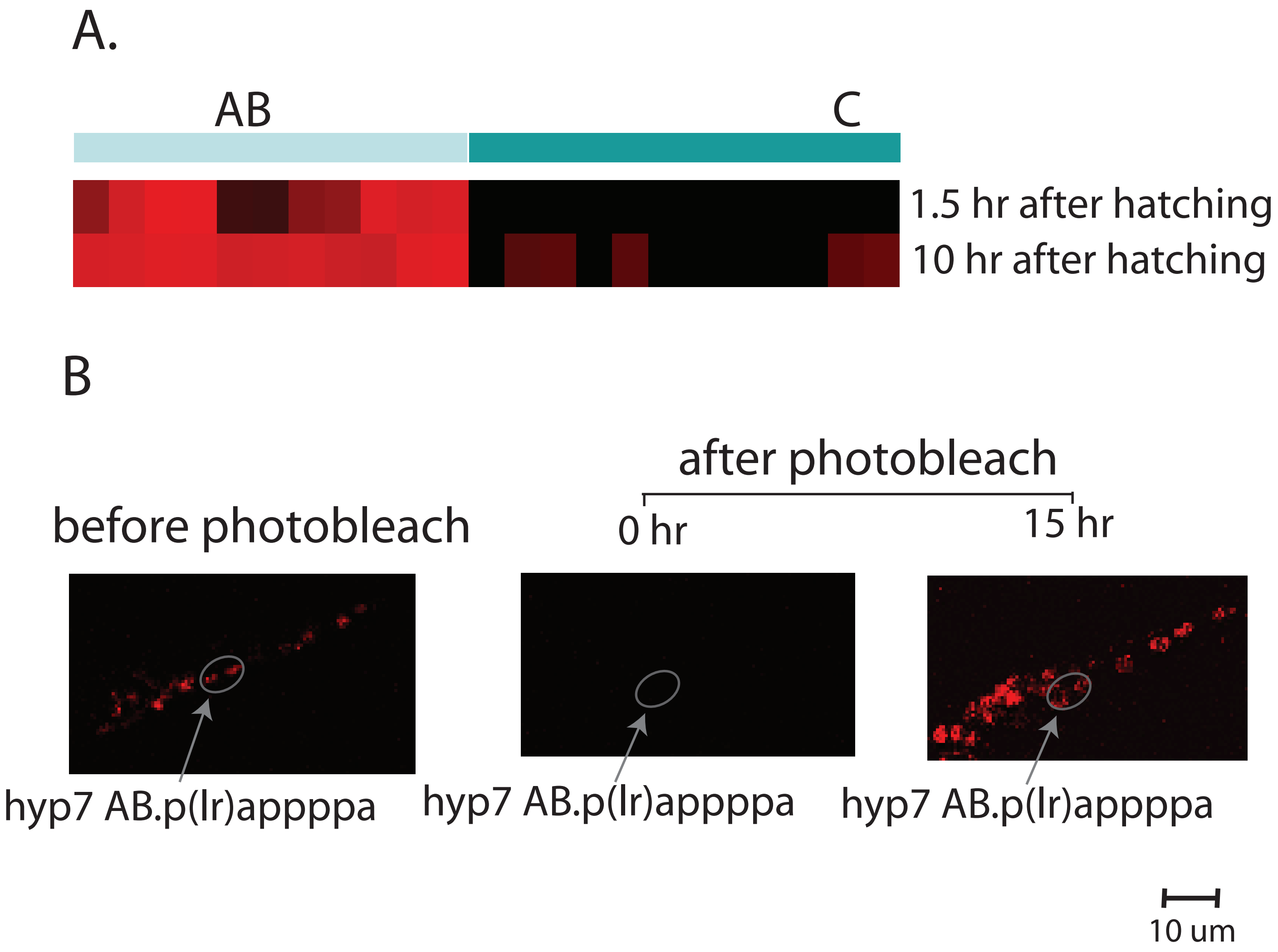
Kalb, J.M., Lau, K.K., Goszczynski, B., Fukushige, T., Moons, D., Okkema, P.G., and McGhee, J.D. (1998). pha-4 is Ce-fkh-1, a fork head/HNF-3alpha,beta,gamma homolog that functions in organogenesis of the C. elegans pharynx. Development (Cambridge, England) *125*, 2171-2180.

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

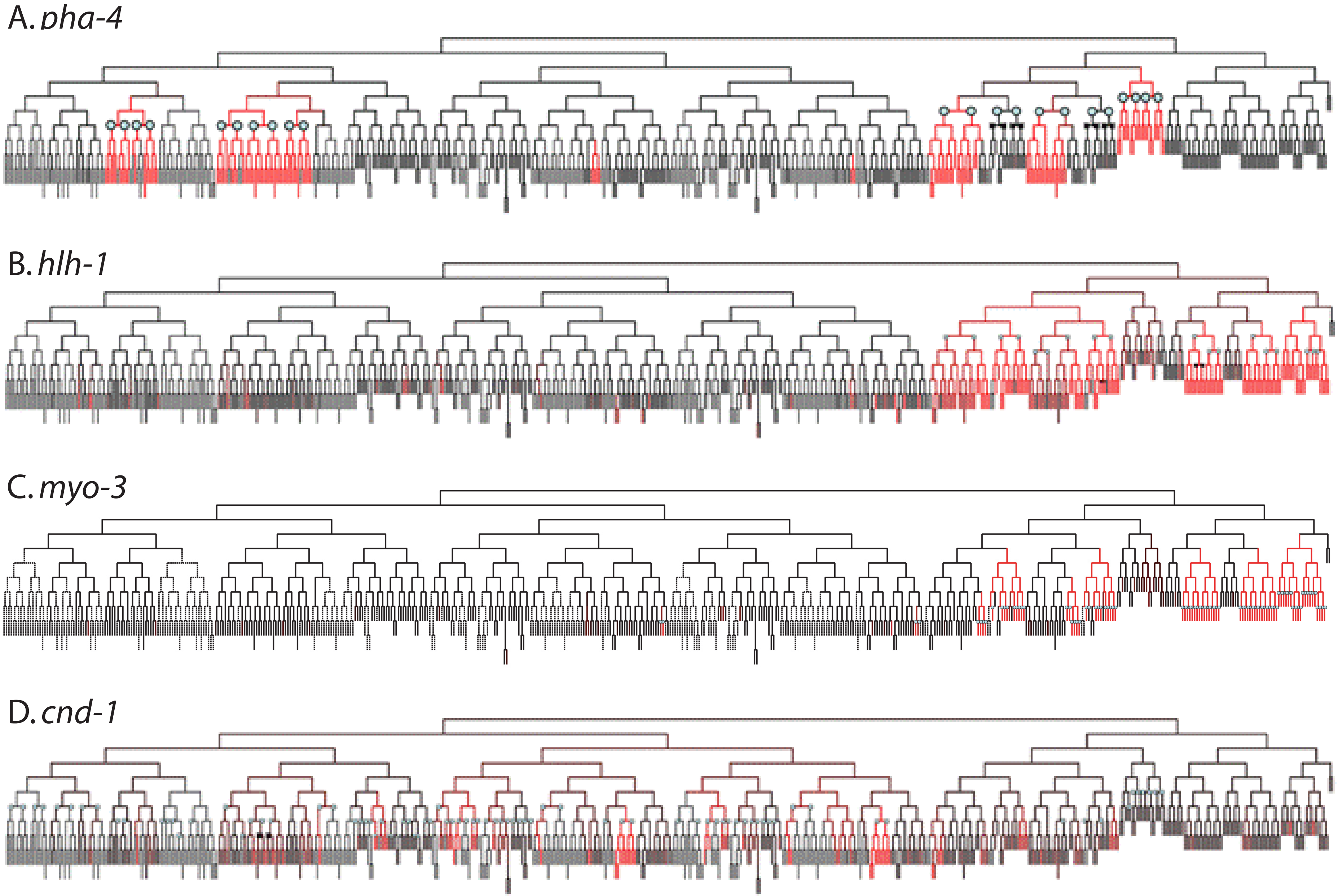
[G] Supplemental Figures




A.
body wall muscle
MS
D
C
AB
hnd-1
ceh-49
sdz-28
egl-27
nhr-2
hsp-3
F27D4.4
his-72
kin-33
ceh-39
C08B11.3
hlh-1
trap-2
F21A10.2
lin-26
elt-5
ZK185.1
pal-1
B.
body wall muscle
anterior posterior axis
tlp-1
M02D8.1
eft-3
lin-1
egl-5
ztf-12
unc-14
mif-2
hsf-1
ceh-41
aap-1
vha-12
C50F7.5
elt-6
lin-12
somi-1
ref-2
ceh-34
normalized
log2(gene expression)
0
1


A.
AB
C
1.5 hr after hatching
10 hr after hatching
B
after photobleach
before photobleach
0 hr
15 hr
hyp7 AB.p(lr)appppa
hyp7 AB.p(lr)appppa
hyp7 AB.p(lr)appppa
10 um

[G] Supplemental Figures

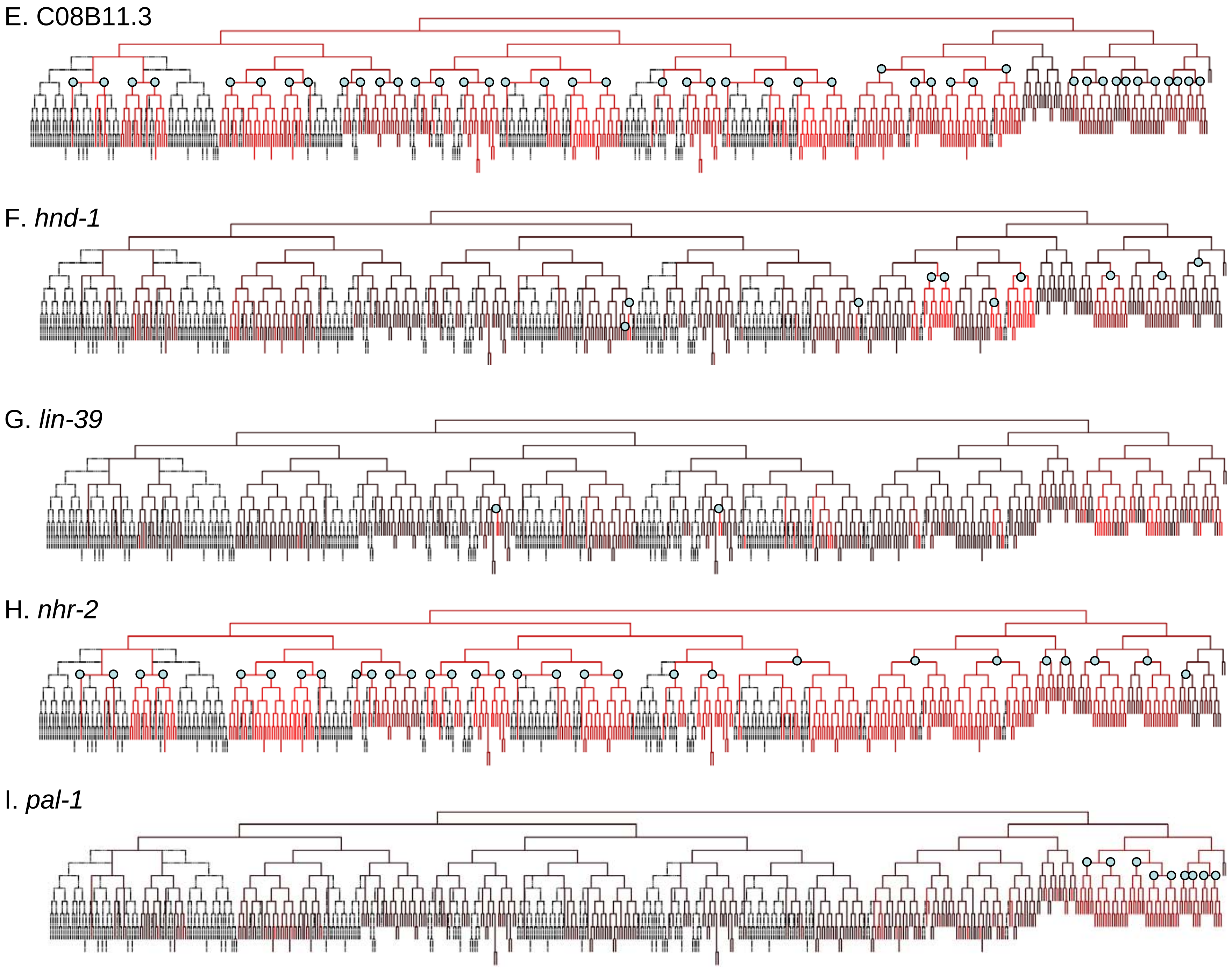


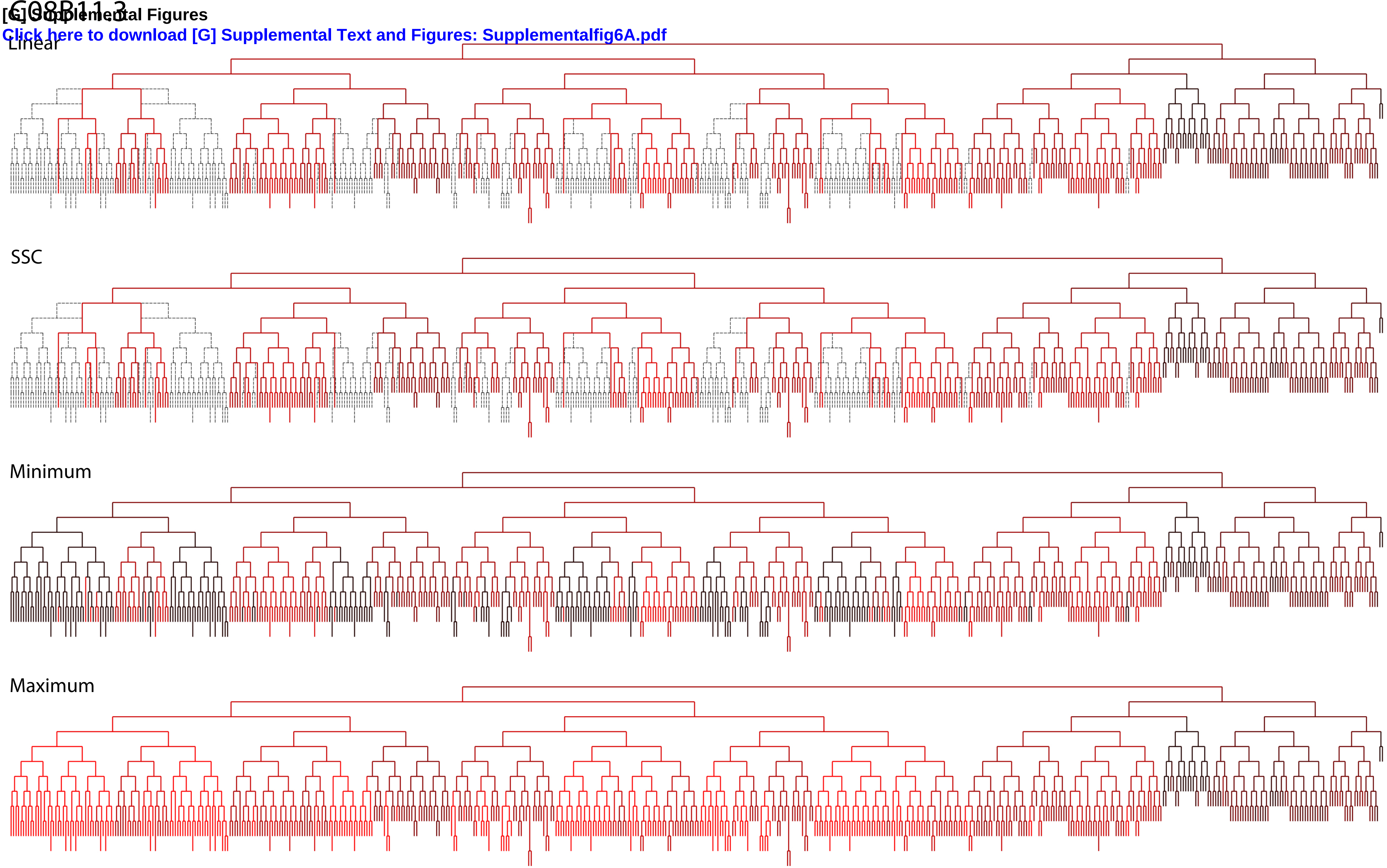
A. *pha-4*

B. *hlh-1*

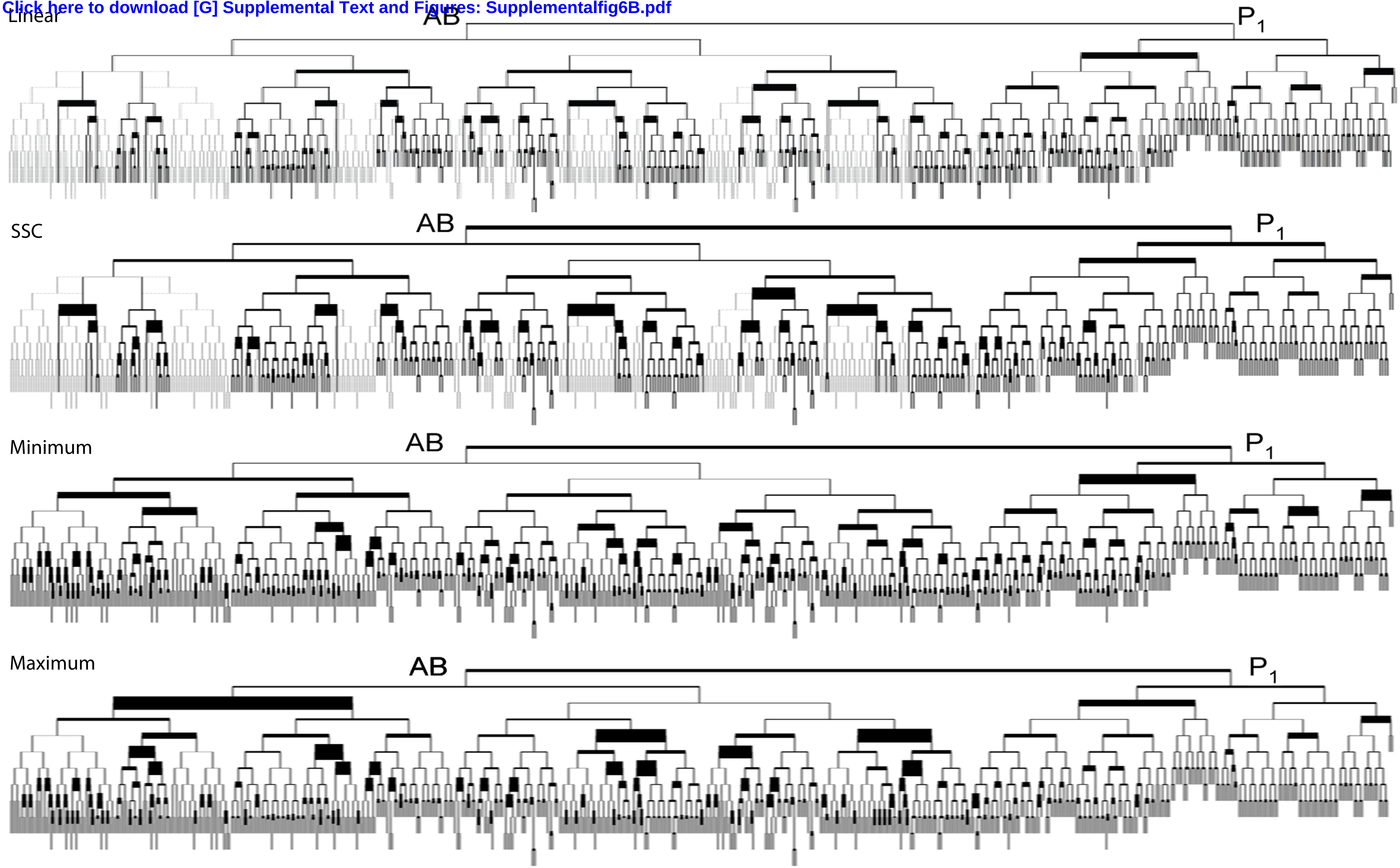
C. *myo-3*

D. *cnd-1*


E. C08B11.3
F. *hnd-1*
G. *lin-39*
H. *nhr-2*
I. *pal-1*


C08B11.3
Linear
SSC
Minimum
Maximum

Molecular Divergence Map



Linear

SSC

Minimum

Maximum